



CHRISTOPH ENGEL
YOAN HERMSTRÜWER
JOHANNES KRUSE

Discussion Paper
2026/6

CAN DOCTRINE HARNESS SILICON LAW CLERKS?

Can Doctrine Harness Silicon Law Clerks?*

Christoph Engel[†] Yoan Hermstrüwer[‡] Johannes Kruse[§]

Abstract

Large language models have arrived in the court room: not as decision-makers, but as decision-aids. Realistically, the human judge in the loop can only do so much. What can be done to prevent the AI clerk from surreptitiously introducing ideological bias, and from being manipulated by the parties? We investigate experimentally whether an institutional constraint that has a guiding effect on human judges also helps to discipline silicon law clerks. Specifically we study how strongly the obligation to reason in terms of established legal doctrine constrains AI outputs, on two channels: by directing attention to relevant precedents, and by imposing structure on the reasoning process. Six treatments are designed to isolate and reinforce these two channels through which legal doctrine operates. We emulate similarity-based reasoning with a Vector RAG and structural reasoning with a Graph RAG, and combine both interventions. We test seven LLMs, comparing both outcomes and a rich set of metrics for the quality of legal reasoning, using decisions by the European Court of Human Rights as ground truth. Even without RAG, frontier models produce impressive outputs. But RAG interventions align outcomes and explicit reasons significantly and substantially more with the human benchmarks.

Keywords: large language models as legal decision aids, legal reasoning, judicial decision-making, retrieval-augmented generation, algorithmic bias, human rights.

JEL codes: K40, K41, O33, C90, D83.

*We are grateful to Jonathan Choi, Michael Livermore, Renana Keydar, and participants of the OWCAL at the University of Virginia, and of the Law Tech Talks seminar at the University of Hong Kong for valuable suggestions. This project was funded by the Max Planck Institute for Research on Collective Goods and the University of Zurich.

[†]E-mail: engel@coll.mpg.de Address: Max Planck Institute for Research on Collective Goods, Kurt-Schumacher-Str. 10, 53115 Bonn, Germany.

[‡]E-mail: yoan.hermstruewer@ius.uzh.ch Address: University of Zurich, Rämistrasse 74, 8001 Zurich, Switzerland.

[§]E-mail: kruse@coll.mpg.de Address: Max Planck Institute for Research on Collective Goods, Kurt-Schumacher-Str. 10, 53115 Bonn, Germany.

1 Introduction

Legal realists have alerted legal scholars to the person of the judge and how the judicial mind shapes legal interpretation and judicial decision-making. On the legal realist view, adjudication is considered an exercise of sovereign power that the judge is expected to wield as she individually deems fit – not necessarily to increase her personal income (Posner 1993) but, in the perspective of the prominent attitudinal model, to advance the political goals she believes in (Segal and Spaeth 1993, 2002) or that the political party supports by which she has been appointed (Pinello 1999).

With the advent of large language models (LLMs), technically, the decision of legal cases and the application of legal doctrine could be delegated to LLMs. For the time being, outright delegation is illegal in the European Union under the EU AI Act, and only tentative elsewhere: there are traces in Brazil (Araújo et al. 2025), China (Balasubramanian et al. 2025; Wong 2024), India (Srinivasan, Tyagi, and Jha 2025), and even the US (Alaska) (Prescott 2024) and the EU (Dutch Digital Courts) Podder (2025). But LLMs as judicial decision aids are already implemented, for example in the Chinese province of Shenzhen (Liu and Li 2024) and increasingly investigated with a serious implementation focus (see, e.g., for international human rights adjudication Fikfak and Helfer 2025).

These developments make it both timely and meaningful to examine LLMs in judicial decision-making through a legal realist lens. When a general-purpose LLM is deployed in adjudication, there is little reason to suspect that it has been deliberately programmed to favor specific plaintiffs or defendants. The same is likely true for LLMs specifically developed for judicial use, such as the tool currently being rolled out to the Israeli judiciary. Yet deliberate design is not the only source of bias. LLMs can absorb systematic biases from their training data or acquire them during fine-tuning. This has been documented for gender (Bolukbasi et al. 2016), race (Bai et al. 2025), and ideology (Hartmann, Schwenzow, and Witte 2023; Feng et al. 2023). Most worryingly, LLM outputs can be easily manipulated to follow a defined ideological orientation (Chen et al. 2024). The judiciary therefore has every reason to anticipate that parties and their attorneys will attempt to draft submissions in ways that trigger a favorable assessment.

Now, the legal realist model has been criticized as overly simplistic in both the legal and the political science literature on judicial politics. We report this debate in greater detail below. Suffice it to note here that the literature converges towards an intermediate position. Even assuming that individual judges are motivated to advance a personal political cause, they can only do so within a fairly rich institutional arrangement. Using rational choice language, they can at most maximize their personal political convictions within the constraints resulting from these institutional safeguards (Epstein and Knight 1998; Epstein and Jacobi 2010). Many constraints that matter for human judges are immaterial for machine judges: they do not hope for promotions; they do not have to maintain their reputation in a professional community;

and they have little reason to fear that judicial powers could be taken away from them. But arguably one constraint matters for human and for machine judges alike: they know that they will have to justify their decision in a way that is consistent with applicable legal doctrine. This is the constraint on which we focus. We investigate empirically in which ways LLMs entrusted with judicial functions are guided by doctrine. Specifically, we ask: How does legal doctrine shape and influence the reason giving process and outcomes generated by LLMs tasked to predict the solution of legal cases?

Earlier research suggests that human judges are aware of the constraining effect of doctrine. This is why rules have a stronger disciplining force than standards (Smith and Todd 2015). If district courts dislike the effect, they try to conceal their own position that differs from the Supreme Court by strategically increasing the complexity of the reasons (Boston 2020). In anticipation, the Supreme Court designs its ruling such that they signal which arguments are most likely to attract review (Masood, Kassow, and Songer 2019).

To test the impact of legal doctrine empirically, we devise a novel approach, combining elements of computational and experimental methods. This approach is intended to directly examine whether doctrinal knowledge affects legal prediction and reasoning. We compile a dataset of cases pertaining to freedom of assembly granted by Article 11 of the European Convention on Human Rights (ECHR). The role of legal doctrine is particularly complex in supranational courts like the European Court of Human Rights (ECtHR), where judges must navigate not only legal interpretation but also the delicate relationship between international human rights law and national sovereignty. The Court has developed its own distinctive doctrinal apparatus to manage these challenges. Central to the Court’s doctrinal approach is the margin of appreciation doctrine, which allows national authorities discretion in implementing Convention rights (Greer 2000). Similarly, the Court’s commitment to treat the Convention as a living instrument allows rights to develop over time (Dzehtsiarou 2011). Through doctrinal concepts like proportionality analysis, the living instrument approach, and consensus-based interpretation, the Court has constructed a sophisticated framework that legitimizes its judicial authority while providing flexibility in addressing diverse legal and social contexts across member states. The Convention system thus offers an ideal context to test the effect of highly developed doctrinal concepts in LLM-generated case prediction and judicial reasoning.

To identify the causal impact of legal doctrine, we implement an experimental design with six treatments that progressively increase the sophistication of doctrinal access. In our *Zero Context* treatment, which serves as our baseline, we use a zero-shot prompt, meaning we simply ask the LLM to predict how the ECtHR will decide a real case based solely on the model’s parametric knowledge acquired during training. In our *Doctrinal Context* treatment, we supplement the case description with a structured explanation of Article 11 doctrine, including the four-stage analytical framework (Applicability, Interference, Justification, Consequences) and the specific legal questions associated with each doctrinal element.

The remaining four treatments employ retrieval-augmented generation (RAG) to give the

LLM access to alternative computational representations of legal doctrine. Specifically we have the RAG identify additional doctrinal material that it considers most relevant for the case at hand. This additional text is supplemented to the sketch of the respective case (that we retrieve from the court’s own website). In the *Vector RAG* treatment, we implement hybrid retrieval combining dense embedding similarity, sparse keyword matching, and cross-encoder reranking to identify relevant passages from the Court’s Reference Guide on Article 11 ECHR. In the *Graph RAG* treatment, context representation does not come in the form of semantically similar text chunks. Rather, the LLM retrieves legal doctrine through a structured knowledge graph encoding the hierarchical relationships between doctrinal concepts – stages of analysis, doctrinal elements, diagnostic questions, and illustrative precedents. Entry point detection guides the model to relevant portions of this doctrinal structure based on case characteristics.

Finally, we implement two treatments that combine vector- and graph-based context retrieval. In the *Vector-Graph RAG-additive* treatment, the LLM receives both vector-retrieved passages and graph-traversed content as separate, clearly demarcated context sections. Hence in this treatment, we leave it to the LLM to synthesize the input from both sources. In the *Vector-Graph RAG-fusion* treatment, we implement a three-step LLM-mediated synthesis process: a Legal Architect constructs an examination schema from graph context, an Evidence Curator populates this schema with evidence from vector context, and a Case Strategist produces a weighted analytical roadmap, identifying decisive questions and applicable precedents. This fusion process provides analytical structure without predetermining legal conclusions – all substantive assessments remain with the judgment-generating model.

We have four sets of dependent variables. We first exploit the fact that we have access to the ground truth – the actual decision of the ECtHR – and compare the models’ outcome predictions with that benchmark. Our second dependent variable comprises the reasons that the LLM predicts the Court will give for its decision. Using an LLM-as-a-judge approach, we compare those reasons with the Court’s actual reasoning across eight domain-specific criteria including legal accuracy, analytical depth, precedent quality, and methodological rigor. We also use standard natural language processing (NLP) metrics for the comparison. The third dependent variable is provided by the context that the respective RAG constructs based on the Court’s guidelines. Using the established RAGAS metrics (Es et al. 2024) tailored to RAG architectures, we gauge how faithfully and effectively the LLM uses the context provided by the respective RAG architecture. Our fourth dependent variable is the trajectory that the LLM takes to retrieve the doctrinal structure, as enshrined in the respective RAG architecture.

For our primary analysis, we use two models with knowledge cutoffs predating our 13 test cases. For these two LLMs we do not have to worry about memorization (Carlini et al. 2021; Carlini et al. 2022; C. Zhang et al. 2023). This selection includes both a standard model (GPT-4o) and a model with extended reasoning capabilities (DeepSeek-R1). However, the rapid development of LLMs means that restricting analysis to older models would miss the full potential of state-of-the-art systems. We therefore also test more recent models – Claude Sonnet

4.5, GPT-5 (in both low and high reasoning configurations), Gemini 2.5 Flash, and Gemini 3 Pro Preview – while acknowledging that their superior performance may partly reflect memorization of training data that included our test cases. We note however that, if memorization were near perfect, these models should perform nearly perfectly even when queried zero-shot. If we find treatment effects, we can still interpret them as marginal improvements resulting from our interventions (that are motivated by features of legal doctrine).

The remainder of our article proceeds as follows. In Section 2, we relate our investigation to the literature. In Section 3, we describe our research design. Section 4 presents our results. In Section 5 we offer a discussion. Section 6 concludes.

2 Literature

The contested power of doctrine. Legal doctrine is a central yet contested force of the judicial mind. As the dominant approach to legal analysis in the US, legal realism embodies the belief that legal doctrine is more malleable, less determinate, and less causal of judicial outcomes than the formalist approach assumes. Legal realists are skeptic about the constraining effect of legal doctrine on judicial decisions and argue that judicial reasoning is rarely ever the product of legal doctrine alone (Eskridge and Frickey 1987; Schauer 2013). Yet despite this skepticism, judicial decisions continue to speak the language of legal doctrine, and legal doctrine remains the currency of legal analysis writ large (Tiller and Cross 2006). While legal scholars, political scientists, and economists have explored whether and how legal doctrine can explain and predict the reasoning and outcomes underlying judicial decisions, no definitive answer has emerged from the academic debate.

The oldest theory of judicial decision-making is formalism. On this view, judicial decisions can be explained and predicted based on the facts, the applicable law, precedent, and judicial deliberation – and nothing more. Legal formalists have traditionally argued that doctrine operates as a genuine constraint, with judges mechanically applying predetermined rules to reach legally mandated outcomes (Pound 1908). The idea of legal doctrine as a mechanistic constraint rests on the belief that the law provides the single source of deduction in adjudication: any given legal decision is fully determinable by an applicable rule (Leiter 2010). This approach assumes that judges apply the law in a mechanical and straightforward manner, acting in Montesquieu’s words as “la bouche de la loi.”

By contrast, legal realists and political scientists working in the attitudinal tradition have characterized legal doctrine as largely epiphenomenal (Fuller 1934; Kronman 1988; McDougal 1947; Segal and Spaeth 2002; Schauer 2015, 2013). The attitudinal model holds that judges reason and decide based on their individual ideological and political predisposition and preferences, unconstrained by the law or institutional context. On this account, legal doctrine is either irrelevant to judicial decision-making or serves as formal post-hoc rationalization for decisions driven by ideological or political preferences. To many legal realists the process of

articulating a judicial opinion in the language of legal doctrine when the actual reasoning originated independently is mostly considered camouflage of the real reasons for the decision.

The strategic model combines aspects of the formalist and attitudinal models: it suggests that judges make decisions by not only taking into consideration their individual preferences, but also the broader legal and institutional context in which they operate. Specifically, while the model assumes that judges want to implement their preferred policies, they reason and decide in light of the long-term consequences of their choices and of the likely reactions of other actors. The strategic model generalizes the attitudinal model by recognizing that there is a range of policy over which justices may express their preferences without fear of reversal, but it also provides room for other forces shaping the judicial decision-making process, including professional norms and legal doctrine.

Recent scholarship has moved beyond this dichotomy (Lax 2011). On this view, one important way in which legal doctrine might constrain judicial reasoning is by structuring the question so as to manipulate the salience of certain aspects of the dispute and shape the reasoning and decision process through what behavioral economists call a framing effect. Rather than imposing a direct and hard constraint, legal doctrine constrains by drawing attention towards certain aspects of the case at hand, and rendering certain other aspects of the dispute irrelevant or secondary (Balganesh 2014). On this view, legal doctrine matters not because it eliminates judicial discretion; rather, it matters because it provides a map and instruments through which policy-oriented judges pursue their goals (see Dickinson 1931). Judicial preferences are fundamentally preferences over legal rules and elements of legal doctrine that govern the disposition of the case at hand. In essence, judges may be political actors, but they pursue their political goals through distinctly legal means: they shape, interpret, apply, and select legal doctrine to satisfy their political preferences.

Our experimental design extends this debate about the degree by which human judges are guided by doctrine to LLMs that are operating as judicial agents. We operationalize the competing theoretical perspectives through six treatments that vary in how and whether LLMs have access to legal doctrine. Each treatment can be understood as testing different mechanisms through which doctrine might constrain decision-making, corresponding to different theoretical assumptions about how judicial reasoning operates.

Our *Zero Context* treatment provides a computational analogue to the attitudinal model. In this treatment, we simply ask the LLM to predict how the ECtHR will decide a case without providing any access to legal doctrine or interpretive guidelines. Just as the attitudinal model posits that judges decide based on their preferences unconstrained by outside institutions, this treatment tests whether LLMs rely solely on patterns learned during training – potentially reflecting biases or heuristics embedded in their training data – without the constraining influence of explicit doctrinal knowledge. If it is indeed the case that doctrine is merely a post-hoc rationalization for decisions driven by other factors, then providing access to doctrine should make little difference to prediction accuracy or reasoning quality. This treatment thus oper-

ationalizes the strongest form of the realist critique: that legal doctrine is epiphenomenal to judicial decision-making.

In the *Doctrinal Context* treatment, we explicitly instruct the LLM to justify its decision, using the structure that the ECtHR has developed for the task, and make this structure explicit. In this treatment we do, however, not provide the LLM with any further guidance how to do this. Hence with this treatment we make the constraint salient. But we still leave the LLM on its own with implementation.

Our *Vector RAG* treatment provides the LLM with access to semantically relevant passages from the Court’s interpretive guidelines. Vector-based retrieval augmented generation uses semantic similarity search within a vector database to retrieve relevant text chunks based on their proximity in embedding space to the query. This approach tests whether doctrine constrains by structuring the question and framing the inquiry, drawing attention toward certain aspects of the dispute while rendering others less salient. Traditional Vector-RAG focuses on enhancing natural language processing tasks by retrieving relevant textual information to support the text-generating process. The LLM receives the bits and pieces from which it can infer a more general rule, much as a judge constructs doctrinal understanding from accumulated precedent. However, the distinction between genuine structural constraint and mere rationalization remains analytically important: in genuine constraint, doctrine compels the court to frame their analysis in particular ways, whereas in rationalization, doctrine merely provides post-hoc justification for conclusions reached through other means. Vector-RAG allows us to test whether access to doctrinal fragments genuinely shapes reasoning or merely provides vocabulary for rationalizing predetermined outcomes.

Our *Graph RAG* treatment offers access to knowledge graphs that structure data into entities and relationships, enabling deeper and more contextual retrieval by capturing structured relationships between doctrinal concepts. In this knowledge graph legal doctrine is represented not as a collection of semantically similar text fragments, but as a structured network of interconnected legal concepts, tests, and interpretive principles. Graph-RAG creates a knowledge graph that extracts entities, relationships, and key claims from text, allowing for more structured reasoning. This architecture yields a more sophisticated understanding of legal doctrine that goes beyond simple exposure to relevant materials to encompass highly structured knowledge of doctrinal relationships. The strategic model recognizes that judges understand not merely isolated doctrinal elements but the systematic relationships among legal concepts and the institutional constraints these relationships create. By representing legal doctrine as a structured graph, this treatment tests whether understanding the architecture of legal doctrine – how different principles relate to one another, which tests govern which situations, and how interpretive principles interact – provides additional constraining power beyond mere semantic relevance.

Our two *Vector-Graph RAG* treatments provide the LLM with access to both semantic similarity-based retrieval and structured doctrinal relationships. This hybrid approach combines vector-

based and graph-based retrieval to exploit the power of both techniques, motivated by research showing that hybrid systems outperform either approach individually (Sarmah et al. 2024). This hybrid architecture operationalizes what we might call doctrinal instrumentalism. Legal doctrine is today seen as important not for its own sake, but because of its connection to normative criteria that are deemed independently worthwhile for the law to adopt, producing an equilibrium where doctrinal and non-doctrinal considerations are treated as equally important for a fuller understanding of the law. Just as judges strategically deploy both specific doctrinal materials (found through their relevance to the case at hand) and systematic doctrinal structures (understood through their training and experience) to advance their legal reasoning, this treatment allows the LLM to leverage both forms of doctrinal knowledge.

Recent work on Graph RAGs explicitly motivates such hybridization by arguing that many complex questions require both textual and relational information – referred to as “hybrid questions” – which complicates the retrieval process for non-hybrid systems (Lee et al. 2024). Such architectures address complementary failure modes: vector retrieval supplies high-resolution semantic matching for granular textual grounding, whereas graph-based retrieval exploits relational structure to support systematic coverage and multi-hop integration (Lee et al. 2024). Comparative evaluations further suggest that hybrid variants improve both coverage and factual grounding by automatically refining the output through agentic feedback and adaptive retrieval (Lee et al. 2024). In our implementation, we therefore distinguish two hybrid regimes. In the *Vector-Graph RAG-additive* treatment, the LLM receives both vector-retrieved passages and graph-traversed content as separate, clearly demarcated context sections, and is left to synthesize these sources independently. By contrast, in the *Vector-Graph RAG-fusion* treatment we implement a three-step LLM-mediated synthesis process: a Legal Architect constructs an examination schema from graph context, an Evidence Curator populates this schema with evidence from vector context, and a Case Strategist produces a weighted analytical roadmap identifying decisive questions and applicable precedents. This fusion process provides analytical structure without predetermining legal conclusions – all substantive assessments remain with the judgment-generating model.

AI and law. Our paper also contributes to the fast-growing literature that investigates the power of artificial intelligence, and of LLMs in particular, for legal tasks (for an overview see Wilf-Townsend and Tobia 2025). We task the LLM with predicting how the ECtHR would decide the respective case. This is how our experiment is related to the broad literature on prediction engines (summarized by Cui, Shen, and Wen 2023; Feng, Li, and Ng 2022), which has produced mixed results (Stillwell and Harrington 2025; Schweitzer and Salerno 2025). Outcome prediction predates LLMs, and has started with more traditional machine learning methods (Medvedeva, Vols, and Wieling 2020; Bagherian-Marandi, Ravanshadnia, and Akbarzadeh-T 2021; Medvedeva, Wieling, and Vols 2023; Bertalan and Ruiz 2024; McConnell et al. 2021). More recent attempts use generic (Salaün, Langlais, and Benyekhlef 2021; Wei et al. 2025) or task-specific transformer architectures (Min and Noh 2025) or, most recently, GPT-5 (Levy

2025). A task specific benchmark has also been developed (Sesodia et al. 2025). Our project differs from this literature in that for us outcome prediction is a performance measure, not the ultimate goal.

Papers come closer that try to improve prediction accuracy by adding an “explainer” step (Benedetto et al. 2025), instruct the LLM to explicitly match facts of the case to legal rules (Jiang and Yang 2023; H. Zhang et al. 2023) or to justify the prediction (Bex and Prakken 2021; Yamada et al. 2025). In our experiments, we also ask for these subtasks. But the main point of our exercise is the manipulation of alternative methods for generating a legal justification, mapped after key features of legal doctrine.

There is also a growing literature that investigates the ability of LLMs to engage in legal reasoning. At a high level, legal reasoning consists of two interrelated components: the interpretation of the applicable legal rules (Waldon et al. 2025), and the application of these rules to the facts of the case. Both elements have attracted interest. LLMs are promising decision aids if the interpretative canon is “ordinary meaning” as, arguably, this is what a generic LLM has picked up at training (Engel and McAdams 2024; Kruse 2025). The task gets more involved if the applicable rule exhibits “open texture”, i.e. if its interpretation may differ across contexts, and is open to the evolution of normative convictions. Attempts are made to capture this with LLMs as well (Canavotto and Horty 2023; Guitton, Gubelmann, et al. 2025; Guitton, Tamò-Larrieux, et al. 2025). It has also been tested experimentally whether LLMs are as open to the balancing of explicit normative goals with equity as human decision-makers (Katz and Zamir 2025). Related investigations aim at extracting principles for the interpretation of openly phrased statutory rules from jurisprudence (Mainali et al. 2020; Lüders and Stohlmann 2025). In this paper, we test LLMs on cases decided by the ECtHR. The human rights protected by the European Convention essentially define normative concerns, and require the Court to balance the interests of citizens engaging in some protected activity (an assembly, in the application that we test) with competing governmental goals (like the smooth functioning of traffic).

Multiple papers have tried to engage LLMs with the task of delineating the normative landscape with the help of precedent (Odekerken, Bex, and Prakken 2024), explicitly capturing the scope for interpretation inherent in drawing analogies across cases (Blass and Forbus 2023; Rigoni 2025). While the ECHR follows the continental law model, in interpreting the broadly defined human rights, the Court heavily relies on (its own) jurisprudence. In our *Vector RAG* treatment, we give the LLM access to a structured summary of this jurisprudence.

We do not only ask the LLM to predict the decision of the ECtHR, but also ask for an explanation. This is how our paper is related to the literature that tries to customize the concept of explainability to legal tasks (Kesari et al. 2024; Mumford, Atkinson, and Bench-Capon 2023; Steging, Renooij, and Verheij 2021; Park and Park 2024).

Yet another strand of literature explores the capability of LLMs to match facts with legal rules, and hence engage in what the legal literature discusses under the heading of syllogism (K. Zhang et al. 2025), exploring the ability of LLMs to properly capture aggravating factors

in criminal law (Rodríguez Rodríguez et al. 2024), or to discriminate between alternative legal rules in the light of the facts of the case (Fungwacharakorn, Tsushima, and Satoh 2021). Our task implies syllogism. But our focus is not on the general feasibility of algorithmic syllogism. Rather, we are interested in the guiding force of legal doctrine in legal reasoning.

Papers are closest that investigate the relationship between legal argument and (predicted) outcomes. Santin et al. (2023) use a corpus of decisions by the Court of Justice of the European Union (CJEU) on fiscal state aid to engage in what they call “argumentation structure prediction”: by which set of legal arguments can the known decision of the court best be explained? Savelka et al. (2023) investigate how giving an LLM access to a (customized) “legal information retrieval module”, using sentences from pertinent case law, improves accuracy. Wardas and Matthes (2025) test the ability of LLMs to classify contract clauses as “valid”, “unfair” or “void” when giving the LLM access to alternative legal materials. These three papers share elements of our interventions. But they do not systematically evaluate which features of legal doctrine are most effective guardrails for LLMs tasked with making legal decisions.

3 Research Design

To empirically test the impact of legal doctrine on judicial reasoning delegated to LLMs, we develop a comprehensive experimental pipeline.¹ The pipeline consists of the following integrated components: six experimental treatments, a vector-based retrieval architecture, a graph-based retrieval architecture, a combined retrieval architecture, and a comprehensive evaluation framework.

Our pipeline prioritizes reproducibility and scalability. All components are implemented as modular, asynchronous Python functions enabling parallel processing across multiple models and cases. The system supports diverse LLM families, including DeepSeek-R1, three of OpenAI’s GPT models, Anthropic’s Claude Sonnet 4.5, and two of Google’s Gemini models, facilitating robustness testing across different model architectures. Most models operate with temperature set to 0 to ensure deterministic output; reasoning models (DeepSeek-R1, GPT-5 variants) require non-zero temperatures as specified by their APIs.

Independent of treatment, the LLM tasked with deciding the case receives a sketch of the facts that we take from the website of the Court (the “case description”). For every case, the Registry of the Court prepares such a summary. It is circulated among the parties (and potentially additional stakeholders) in preparation of the decision by the Court. An example can be found in the Appendix.

1. The study was pre-registered with the Open Science Framework (<https://osf.io/ksnfr2q7b/overview>). The pre-registration covered four treatments; the *Doctrinal Context* and *Vector-Graph RAG-fusion* treatments were added as the project evolved.

3.1 Treatments

The pipeline implements six treatments that progressively increase the sophistication of doctrinal access:

1. **Zero Context:** Models receive only the case description, relying solely on parametric knowledge acquired during training. This establishes a lower bound for performance and reveals what legal knowledge LLMs have internalized.
2. **Doctrinal Context:** Models receive the case description supplemented by a structured explanation of Article 11 ECHR doctrine, including the four-step pronged doctrinal test framework (Scope, Interference, Justification, Consequences) and the specific legal questions associated with each doctrinal element.
3. **Vector RAG:** Retrieval-augmented generation using hybrid search that combines dense embedding similarity, BM25 sparse retrieval, and cross-encoder reranking to identify relevant passages from the ECtHR Reference Guide (European Court of Human Rights 2025).
4. **Graph RAG:** Graph-based retrieval where LLM-selected entry points guide traversal through a structured knowledge graph encoding the hierarchical relationships of Article 11 doctrine.
5. **Vector-Graph RAG-additive:** Both vector and graph contexts are retrieved independently and presented as separate, clearly distinct sections in the prompt, without integration.
6. **Vector-Graph RAG-fusion:** A three-step LLM-mediated fusion process synthesizes vector and graph contexts into an integrated analytical framework before judgment generation.

To evaluate the accuracy of the decision and, more importantly, of legal reasoning, we need a credible benchmark. We choose decisions of the ECtHR on freedom of assembly, as protected by Art. 11 ECHR. Hundreds of decisions on the provision have been published (for background see Engel and Kruse 2026). There is no need to infer doctrine from this body of jurisprudence, as the Registry of the Court regularly publishes a comprehensive document (European Court of Human Rights 2025). The Court routinely circulates a short summary of the facts (see Appendix for a sample). We use these summaries as semi-official stimulus materials. We ask the LLM to decide whether, given these facts, the defendant country has violated Art. 11 ECHR. We require not only the (prediction of the) decision, but also an explicit justification. We compare the decision and the justification with the actual decision by the ECtHR.

This design helps with evaluation, as we can consider the actual decision of the court as the gold standard. But we face an obvious problem: we must worry that the LLM has already seen the actual decision at training, and just repeats what it has seen before (C. Zhang et al. 2023). We prevent memorization by using two LLMs with knowledge cutoffs predating 13 recent ECtHR decisions, so that we can ensure genuine out-of-sample prediction while maintaining access to ground truth for evaluation. This holds for DeepSeek-R1 (July 2023) and GPT-4o (October 2023). This selection of LLMs includes a reasoning model (DeepSeek-R1), which arguably matters for a task that hinges on the degree of elaboration. Still, the development of LLMs is so rapid that we would not see the full potential of state of the art models, were we to exclude the latest models. This is why, in addition, we also test GPT-5 with low and high degree of reasoning (September 2024), Claude Sonnet 4.5, Gemini 2.5 Flash, and Gemini 3 Pro Review (January 2025, respectively). We acknowledge that the higher performance of these models might result from memorization. This is why, when interpreting results, we focus on treatment effects rather than on the specific LLM used.

3.2 Vector RAG: Construction and Implementation

The vector database employs ChromaDB with a hybrid retrieval architecture. Document chunking follows a paragraph-based strategy: the system identifies explicitly numbered paragraphs in the ECtHR Reference Guide – matching the regular expression pattern $\text{\textasciitilde{s}*(\text{\textasciitilde{d}}+)\text{\textasciitilde{s}}+}$ – and groups them into non-overlapping blocks of three consecutive paragraphs. For example, paragraphs 1-3 form one chunk, paragraphs 4-6 form the next, and so forth. This approach preserves semantic coherence within the court’s own organizational structure while maintaining manageable chunk sizes. Each chunk receives metadata recording its constituent paragraph numbers (e.g., “1-3”) and source document.

The embedding pipeline supports multiple backends. By default, we use OpenAI’s text-embedding model (1,536 dimensions); alternatively, the sentence-transformers all-MiniLM-L6-v2 model (384 dimensions) provides a local option. All embeddings are stored in ChromaDB with cosine distance metrics. Retrieval implements Reciprocal Rank Fusion (RRF) combining three signals:

- **Dense retrieval:** Embedding similarity search returning the top 60 candidates
- **Sparse retrieval:** BM25 keyword matching returning the top 60 candidates
- **Reranking:** Cross-encoder (ms-marco-MiniLM-L-6-v2) rescoring the combined candidates

The RRF score for each document d is computed as:

$$\text{RRF}(d) = \sum_{r \in \{\text{dense}, \text{sparse}\}} \frac{1}{k + \text{rank}_r(d)} \quad (1)$$

where $k = 60$ is a smoothing parameter. The top 30 documents by RRF score are reranked by the cross-encoder, with the final top 20 forming the retrieval context, subject to a 12k-token budget.

The *Vector RAG* treatment queries this database using the case description as the retrieval query. The prompt presents retrieved chunks as context and instructs the model to integrate the retrieved legal principles into its judgment while maintaining judicial voice.

3.3 Graph RAG: Construction and Implementation

While vector embeddings capture semantic similarity, legal doctrine requires a structured representation to encode the hierarchical relationships that characterize legal reasoning. The knowledge graph explicitly encodes these structural relationships, transforming the ECtHR Reference Guide into a computationally traversable network of interconnected legal concepts, principles, and precedents stored in Neo4j.

The graph foundation derives from the ECtHR’s own structuring of Article 11 doctrine. The hierarchy comprises three node types:

- **Stage nodes** represent the four fundamental phases in Article 11 analysis: Applicability of Article 11, Interference, Justification, and Consequences.
- **Doctrinal Element nodes** decompose each stage into legal tests:
 - Applicability: Substantive Scope, Personal Scope
 - Interference: Existence of Interference, Positive Obligations
 - Justification: Prescribed by Law (with sub-elements Accessibility, Foreseeability, Safeguards against Abuse), Legitimate Aim (seven enumerated aims: National Security, Public Safety, Prevention of Disorder, Prevention of Crime, Protection of Health, Protection of Morals, Rights of Others), Necessary in a Democratic Society
 - Consequences: Just Satisfaction
- **Diagnostic Question nodes** capture the specific legal questions associated with each doctrinal element, formulated as interrogatives that courts must answer.

The hierarchical structure is established through typed relationships: `CONTAINS_ELEMENT` links stages to their constituent doctrinal elements, `ASSESSED_BY` connects elements to their diagnostic questions, and `NEXT_STAGE` establishes the sequential flow of legal analysis between stages.

Beyond the static scaffold, graph enrichment leverages LLM-guided extraction from the 27 source documents of the Reference Guide. Three extraction passes operate across all documents:

1. **Court decision extraction** identifies ECtHR cases referenced in doctrinal texts, creating `CaseRuling` nodes linked to relevant doctrinal elements via `REFERENCES_CASE` relationships.
2. **Legal question extraction** identifies specific legal issues addressed by each section, creating `LegalQuestion` nodes connected via `ASKS` relationships.
3. **Factual scenario extraction** captures illustrative examples demonstrating how abstract principles apply to concrete situations, stored as `FactualScenario` nodes linked via `HAS_EXAMPLE` relationships.

A second enrichment phase constructs `LegalFinding` nodes that encode complete legal arguments, linking these to supporting case law via `CITES_CASE` relationships with relevance annotations. Document provenance is preserved through `APPEARS_IN` relationships connecting all extracted entities to their `SourceFile` nodes.

Each of the Reference Guide's 27 documents maps to specific doctrinal nodes based on topical alignment, with documents potentially connecting to multiple nodes when addressing intersecting legal issues. This multi-dimensional mapping enables context-appropriate retrieval regardless of the analytical entry point.

The *Graph RAG* treatment operationalizes legal doctrine as a navigable network. Entry point detection employs LLM analysis of the case description to select 3-7 relevant nodes from a validated set of 22 doctrinal elements (the Stage and Doctrinal Element nodes). This constraint to pre-validated nodes prevents hallucination of non-existent legal concepts while ensuring flexibility to adapt to diverse case characteristics.

Graph traversal retrieves connected content up to depth 5 and is subject to a 20k-token budget. The traversal query matches paths from entry points to content nodes (`LegalQuestion`, `CaseRuling`, `FactualScenario`, `LegalFinding`), returning node properties, relationship types, and path depth. Retrieved content is organized into structured categories:

- **Primary questions:** Core legal issues directly connected to entry points
- **Primary cases:** Court decisions at depth 1-2, considered authoritative precedent
- **Supporting cases:** Decisions at depth 3-5, providing secondary authority
- **Factual patterns:** Scenarios illustrating principle application
- **Procedural guidance:** Questions containing procedural or safeguard terminology
- **Legal principles:** Extracted from questions addressing doctrinal tests

The retrieved content then undergoes LLM-guided relevance filtering, which retains only case-relevant items and enriches referenced court decisions with model-knowledge summaries of their holdings and factual backgrounds.

3.4 Vector-Graph RAG: Construction and Implementation

Having established both vector and graph representations of legal doctrine, we test two treatments that leverage these complementary approaches.

The *Vector-Graph RAG-additive* treatment combines both retrieval methods without LLM-mediated integration. The system retrieves vector context (allocated 50% of the token budget) and graph context (allocated 50%) independently, presenting them as separate, clearly demarcated sections:

```
=== VECTOR CONTEXT (Document Retrieval) ===  
[Top 20 chunks from hybrid retrieval]  
=== GRAPH CONTEXT (Knowledge Graph) ===  
[Structured content from graph traversal]
```

The model receives explicit instructions to synthesize insights from both sources but must perform this integration independently. This treatment tests whether simply providing both structured and unstructured context – without pre-processing – enables effective reasoning.

The *Vector-Graph RAG-fusion* treatment implements a three-step synthesis process that leverages the complementary strengths of both retrieval methods. The fusion process uses GPT-5 with high reasoning effort.

Step 1: Legal Architect. The first step constructs an examination schema from graph context, defining the analytical structure for the specific case. The output specifies: (i) the examination stages applicable to the case, (ii) the specific legal questions arising from the case facts, (iii) mappings between questions and relevant precedents, and (iv) definitions of applicable legal standards. This step produces *structure only* – no legal assessments or predictions about outcomes.

Step 2: Evidence Curator. The second step populates the examination schema with evidence from vector context. For each element in the schema, the system extracts: (i) relevant text passages with source attribution, (ii) case citations with specific paragraph references (e.g., “§91”), (iii) canonical formulations – exact phrases from case law carrying doctrinal weight, and (iv) gaps where no relevant evidence was found. This step organizes evidence without evaluating its implications.

Step 3: Case Strategist. The third step creates a weighted solution outline identifying analytical priorities. Each doctrinal element receives a weight: critical, important, or routine. The output identifies decisive questions – those likely to determine the outcome – with their applicable precedents. Again, this step identifies *what* is important without assessing *how* issues should be decided.

The resulting “solution outline” serves as an analytical roadmap for the generating model. It specifies the structure of analysis, the questions requiring answers, and the precedents to consider, but leaves all legal determinations to final judgment generation. The generating model receives explicit instructions:

“The outline contains NO legal assessments or predictions. It provides the STRUCTURE for your analysis, the QUESTIONS you must answer, the PRECEDENTS you must consider, and the EVIDENCE available. YOU must make ALL legal determinations. The outline is your roadmap; you are the judge who decides the destination.”

This separation ensures that the fusion process enhances analytical structure without pre-determining conclusions, enabling clean comparison between treatments that differ in context organization rather than substantive legal guidance.

3.5 Evaluation Framework

The final element of our experimental pipeline assesses (i) the accuracy of the outcome and (ii) the quality of the legal reasoning applied.

Measuring the accuracy of the outcome is straightforward: we compare predicted outcomes (Violation/Non-Violation) with actual Court decisions. Moreover, when a model reaches a different conclusion, we also evaluate whether that conclusion is defensible given the limited facts provided – a determination made by the first available judge model. This defensibility assessment prevents penalizing models for reasonable interpretations that differ from outcomes reached with fuller information.

Measuring the quality of legal reasoning is much harder. Tackling this key challenge in legal AI research, we design an evaluation framework that combines computational metrics with domain-specific legal criteria. The rationale of our approach is to assess the quality of LLM-generated reasoning beyond surface-level textual similarity while accounting for the information asymmetry between the AI system and the ECtHR.

This information asymmetry represents a key challenge: the AI systems under evaluation receive only simplified case descriptions, not the full factual record available to the ECtHR. Our design explicitly accounts for this information asymmetry by instructing evaluators to assess reasoning quality given limited information rather than penalizing models for reaching different conclusions than the Court. This “fair evaluation” principle ensures that models are judged on methodological rigor and analytical depth rather than merely on outcome alignment.

Our evaluation framework rests on two evaluation methods: one based on the LLM-as-a-judge paradigm and one tailored to the evaluation of outputs generated using RAG architectures (RAGAS). We use both evaluation methods to produce two distinct scores measuring the overall quality of legal reasoning: an answer score and a composite score.

Our **Answer Score** measures the quality of the predicted decision itself, independent of how it was produced. It aggregates LLM-as-a-judge evaluations of legal reasoning quality (85%), RAGAS answer-level metrics (12%), and textual similarity via BLEU (3%) for all six treatments (non-RAG treatments and RAG treatments). This score captures whether the predicted decision reflects sound legal reasoning and coherent arguments, regardless of whether the LLM relied on parametric knowledge or retrieved context. The answer score thus enables direct comparison of output quality across all six treatments on a common scale.

Our **Composite Score** incorporates an additional dimension for RAG treatments: context quality. For non-RAG treatments (*Zero Context* and *Doctrinal Context*), the composite score equals the answer score, as these methods do not retrieve external information. For our four RAG treatments (*Vector RAG*, *Graph RAG*, *Vector-Graph RAG-additive*, *Vector-Graph RAG-fusion*), the composite score weights answer quality at 75% and context quality at 25%:

$$S_{\text{composite}} = \begin{cases} S_{\text{answer}} & \text{if non-RAG treatment} \\ 0.75 \cdot S_{\text{answer}} + 0.25 \cdot S_{\text{context}} & \text{if RAG treatment} \end{cases} \quad (2)$$

The context score evaluates retrieval effectiveness through six RAGAS metrics: faithfulness (whether the answer is grounded in retrieved context), context relevance, precision, recall, utilization, and entity recall. This component penalizes RAG systems that retrieve irrelevant material or fail to leverage retrieved information effectively.

The distinction matters for interpretation. A RAG treatment might achieve a high answer score but a lower composite score if context retrieval is superficial or misguided. This might indicate that strong model capabilities compensate for weak retrieval. Conversely, a treatment with moderate answer quality but excellent context quality suggests effective retrieval of context that is not sufficiently considered at the generation step. By reporting both metrics, we can assess whether performance differences result from reasoning capabilities, retrieval effectiveness, or their interaction.

LLM-as-a-judge: Legal Reasoning. The available judge models include Gemini 2.5 Flash, Gemini 3 Pro Preview, GPT-4o, GPT-5 (in both low and high reasoning configurations), and Claude Sonnet 4.5. Scores are averaged across all selected evaluators. All judges operate at temperature 0 (or temperature 1 for reasoning models requiring it) to ensure consistent and reproducible evaluations. Each judge scores the generated reasoning across eight domain-specific criteria:

1. Legal Accuracy (10% weight): correct identification and application of Article 11 principles and the three-part test;
2. Analytical Depth (16% weight): thoroughness of legal analysis, including systematic treatment of all relevant issues and consideration of counter-arguments;

3. Precedent Quality (16% weight): effective use of ECtHR case law, with specific paragraph citations rewarded over vague references;
4. Factual Grounding (7% weight): proper anchoring of analysis in provided facts while acknowledging limitations;
5. Methodological Rigor (13% weight): adherence to ECtHR’s established analytical methodology and proper sequencing;
6. Source Transparency (9% weight): explicit referencing of source materials and traceability of legal claims;
7. Argumentative Coherence (9% weight): internal consistency and logical flow of reasoning; and
8. Professional Presentation (5% weight): appropriate judicial tone and precise legal terminology.

Judge evaluations are normalized from a 1–5 scale to 0–1 and contribute 85% of the total answer quality score.

RAGAS: Contextual Reasoning. The RAGAS (Retrieval-Augmented Generation Assessment) metrics provide a standardized computational assessment tailored to outputs generated using RAG architectures. Answer-level metrics include answer relevancy (5% weight), answer similarity (5% weight), and answer correctness (2% weight). Context-level metrics, applied only to RAG treatments, include faithfulness (25% of context score), context relevance (15%), context precision (15%), context recall (15%), context utilization (15%), and context entity recall (15%). A BLEU score contributes an additional 3% to the answer quality component.

To ensure robust statistical inference, we conduct three fully independent runs of the entire evaluation pipeline, generating distinct observations for each case-model-treatment-evaluator combination. With multiple judges and three runs, each case-method pair yields numerous independent observations (e.g., with three judges: 9 observations per case-method). These observations are preserved as raw data points rather than being aggregated prematurely, enabling mixed-effects modeling, variance-component decomposition, and bootstrap resampling for confidence intervals.

4 Results

We begin by evaluating the binary alignment of model outputs with the gold standard decisions of the ECtHR, followed by an assessment of the aggregated legal reasoning scores (Section 4.1). We then analyze whether the quality of outcome predictions can be explained through the quality of the legal reasoning (Section 4.2). Finally, we turn to our LLM-as-a-judge

metrics as a measure of the quality of legal reasoning (Section 4.3) and to RAGAS metrics assessing the quality of contextual reasoning (Section 4.4).

4.1 Conclusion Accuracy and Overall Legal Reasoning Quality

The baseline level of conclusion accuracy in the *Zero Context* treatment (Figures 1 and 2) suggests that current LLMs are highly capable at predicting judicial decisions. Even when deprived of specific legal context, models correctly predict the ECtHR outcome in nearly 75% of cases. This baseline performance likely reflects the prevalence of landmark ECtHR principles and jurisprudence within the models’ training data. However, this high accuracy at the lower bound suggests that the models may be relying on memorized patterns rather than reasoned legal deduction.

To estimate treatment effects on conclusion accuracy, we specify linear probability models with a binary indicator for correct conclusion as the dependent variable. Table 2 presents the results across four specifications. Columns (1) and (3) estimate unconditional treatment effects relative to the *Zero Context* baseline, with data pooled across LLMs; columns (2) and (4) add LLM fixed effects to absorb systematic differences in model capability. Columns (1)–(2) report robust standard errors, while columns (3)–(4) cluster standard errors at the case level to account for the repeated observation structure.

Notably, the *Doctrinal Context* treatment does not significantly improve conclusion accuracy relative to the *Zero Context* baseline. By contrast, our RAG-based treatments yield a substantial increase in predictive accuracy, with the *Vector-Graph RAG-fusion* treatment achieving near-perfect alignment with the ECtHR’s actual decisions.

Table 1: Conclusion Accuracy by Treatment and LLM

Treatment	DeepSeek-R1	GPT-4o	GPT-5 low	GPT-5 high	Claude 4.5 Sonnet	Gemini 2.5 Flash	Gemini 3 Pro	Total
Zero Context	0.641 (0.482)	0.769 (0.423)	0.769 (0.423)	0.847 (0.363)	1.000 (0.002)	0.692 (0.464)	0.385 (0.489)	n=819
Doctrinal Context	0.770 (0.423)	0.769 (0.423)	0.615 (0.489)	0.924 (0.268)	0.923 (0.268)	0.847 (0.363)	0.393 (0.491)	n=819
Vector RAG	1.000 (0.002)	1.000 (0.002)	0.615 (0.489)	0.539 (0.501)	1.000 (0.002)	1.000 (0.002)	0.461 (0.500)	n=819
Graph RAG	0.872 (0.336)	0.923 (0.268)	0.846 (0.362)	0.846 (0.362)	1.000 (0.001)	0.923 (0.268)	0.692 (0.464)	n=819
Vector-Graph RAG-additive	1.000 (0.003)	1.000 (0.001)	0.846 (0.362)	0.769 (0.423)	1.000 (0.001)	1.000 (0.002)	0.692 (0.464)	n=819
Vector-Graph RAG-fusion	0.847 (0.363)	1.000 (0.002)	1.000 (0.002)	1.000 (0.001)	1.000 (0.001)	1.000 (0.002)	0.846 (0.362)	n=819
Total	n=702	n=702	n=702	n=702	n=702	n=702	n=702	N=4914

Note: Standard deviations in parentheses.

The composite score (Figure 3), which weights answer quality at 75% and context quality at 25% for RAG treatments (see Section 3), shows the true impact of our treatments. Here, we observe a staircase effect in performance. Interestingly, the provision of the four-pronged legal test in the *Doctrinal Context* treatment fails to improve the quality of legal reasoning relative to the *Zero Context* baseline. A significant boost in legal reasoning quality only emerges with the introduction of dynamic retrieval in the *Vector RAG* and *Graph RAG* treatments.

To estimate treatment effects on overall legal reasoning quality, we regress the composite score on treatment indicators using OLS. Table 4 presents the results. The specification

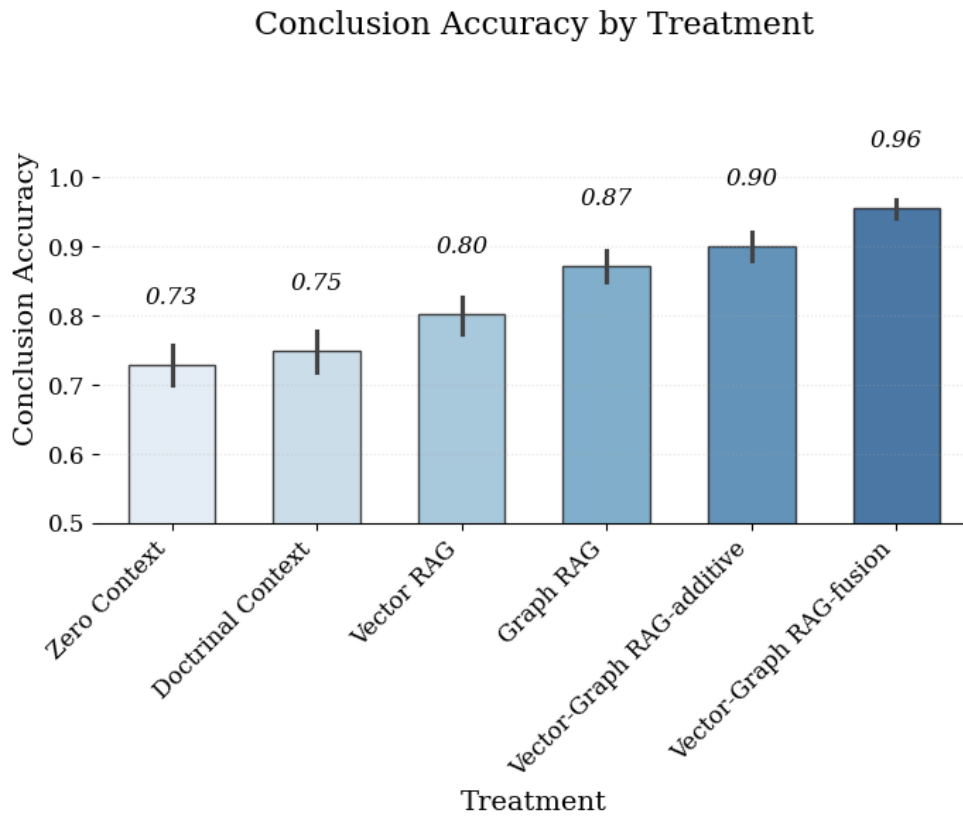


Figure 1: Conclusion Accuracy by Treatment

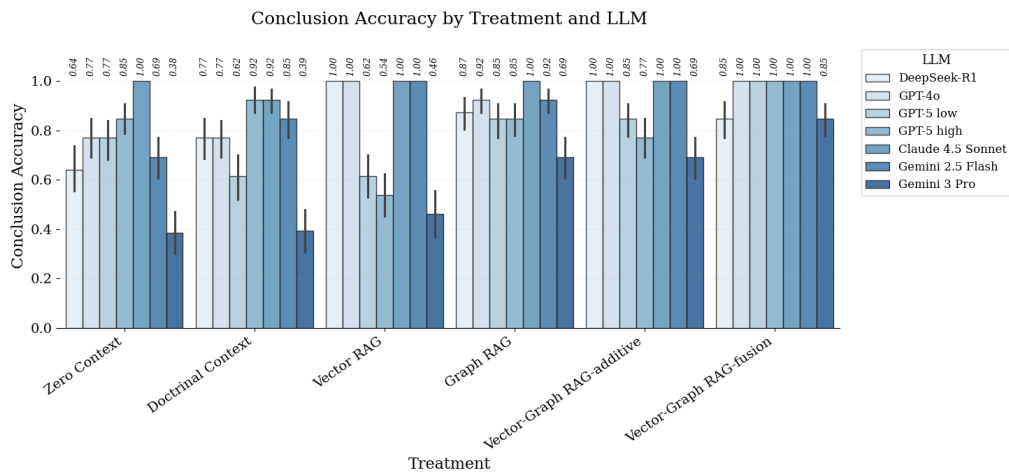


Figure 2: Conclusion Accuracy by Treatment and LLM

Table 2: Treatment Effects on Conclusion Accuracy

	(1)	(2)	(3)	(4)
<i>Panel A: Treatment (ref: Zero Context)</i>				
Doctrinal Context	0.020 (0.022)	0.020 (0.020)	0.020 (0.033)	0.020 (0.033)
Vector RAG	0.073*** (0.021)	0.073*** (0.019)	0.073** (0.030)	0.073** (0.030)
Graph RAG	0.143*** (0.019)	0.143*** (0.019)	0.143*** (0.054)	0.143*** (0.054)
Vector-Graph RAG-additive	0.172*** (0.019)	0.172*** (0.018)	0.172*** (0.056)	0.172*** (0.056)
Vector-Graph RAG-fusion	0.227*** (0.017)	0.227*** (0.017)	0.227*** (0.063)	0.227*** (0.063)
<i>Panel B: LLM (ref: DeepSeek-R1)</i>				
GPT-4o		0.055*** (0.017)		0.055 (0.067)
GPT-5 low		-0.073*** (0.020)		-0.073 (0.068)
GPT-5 high		-0.034* (0.020)		-0.034 (0.046)
Claude 4.5 Sonnet		0.132*** (0.014)		0.132** (0.054)
Gemini 2.5 Flash		0.055*** (0.017)		0.055 (0.045)
Gemini 3 Pro		-0.277*** (0.022)		-0.277*** (0.084)
Intercept	0.729*** (0.016)	0.749*** (0.020)	0.729*** (0.078)	0.749*** (0.091)
LLM fixed effects	No	Yes	No	Yes
Clustered SEs	No	No	Yes	Yes
N	4914	4914	4914	4914
R^2	0.049	0.156	0.049	0.156
Adj. R^2	0.048	0.154	0.048	0.154
Clusters (cases)	–	–	13	13

Notes: This table reports the results of a linear probability model. The dependent variable is a binary indicator for correct conclusion. Columns (1)–(2) report robust standard errors; columns (3)–(4) report standard errors clustered at the case level. Each case–LLM combination is observed three times. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

structure mirrors Table 2: columns (1) and (3) estimate treatment effects pooled across LLMs, columns (2) and (4) include LLM fixed effects, and standard errors are robust in columns (1)–(2) and clustered at the case level in columns (3)–(4).

The most striking result is the dominance of the *Vector-Graph RAG-fusion* treatment, which achieves a composite score that is significantly higher than that of the *Vector-Graph RAG-additive* approach. This finding suggests that simply providing an LLM with context from a vector database and a knowledge graph is insufficient for complex judicial tasks. Instead, the retrieved information must be pre-synthesized into an analytical roadmap to be fully operationalized in our judicial context.

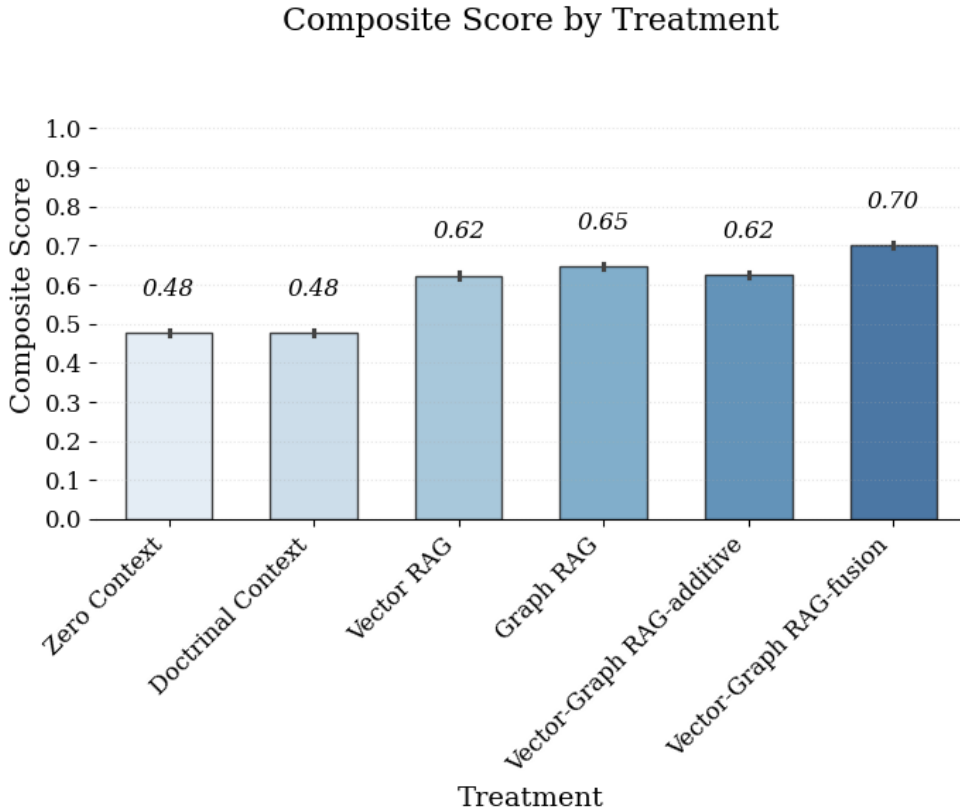


Figure 3: Composite Score by Treatment

Table 3: Composite Score by Treatment and LLM

Treatment	DeepSeek-R1	GPT-4o	GPT-5 low	GPT-5 high	Claude 4.5 Sonnet	Gemini 2.5 Flash	Gemini 3 Pro	Total
Zero Context	0.466 (0.114)	0.418 (0.129)	0.517 (0.086)	0.561 (0.065)	0.455 (0.121)	0.463 (0.111)	0.460 (0.102)	n=819
Doctrinal Context	0.478 (0.103)	0.390 (0.138)	0.528 (0.079)	0.536 (0.072)	0.421 (0.124)	0.487 (0.095)	0.504 (0.090)	n=819
Vector RAG	0.629 (0.109)	0.601 (0.107)	0.650 (0.109)	0.640 (0.113)	0.588 (0.109)	0.595 (0.145)	0.662 (0.099)	n=819
Graph RAG	0.637 (0.101)	0.570 (0.116)	0.702 (0.059)	0.706 (0.057)	0.570 (0.129)	0.673 (0.119)	0.674 (0.074)	n=819
Vector-Graph RAG-additive	0.615 (0.093)	0.547 (0.120)	0.675 (0.075)	0.670 (0.099)	0.554 (0.128)	0.653 (0.071)	0.660 (0.081)	n=819
Vector-Graph RAG-fusion	0.717 (0.086)	0.631 (0.093)	0.747 (0.032)	0.761 (0.040)	0.615 (0.106)	0.730 (0.052)	0.710 (0.052)	n=819
Total	n=702	n=702	n=702	n=702	n=702	n=702	n=702	N=4914

Note: Standard deviations in parentheses.

A comparison of Tables 3 and 5 reveals a consistent pattern: composite scores for RAG treatments are systematically lower than their corresponding answer scores. Our best-performing

Table 4: Treatment Effects on Composite Score

	(1)	(2)	(3)	(4)
<i>Panel A: Treatment (ref: Zero Context)</i>				
Doctrinal Context	0.001 (0.006)	0.001 (0.005)	0.001 (0.004)	0.001 (0.004)
Vector RAG	0.146*** (0.006)	0.146*** (0.006)	0.146*** (0.011)	0.146*** (0.011)
Graph RAG	0.170*** (0.006)	0.170*** (0.005)	0.170*** (0.010)	0.170*** (0.010)
Vector-Graph RAG-additive	0.148*** (0.006)	0.148*** (0.005)	0.148*** (0.009)	0.148*** (0.009)
Vector-Graph RAG-fusion	0.225*** (0.005)	0.225*** (0.005)	0.225*** (0.012)	0.225*** (0.012)
<i>Panel B: LLM (ref: DeepSeek-R1)</i>				
GPT-4o		-0.064*** (0.006)		-0.064*** (0.012)
GPT-5 low		0.046*** (0.005)		0.046*** (0.011)
GPT-5 high		0.055*** (0.005)		0.055*** (0.013)
Claude 4.5 Sonnet		-0.057*** (0.006)		-0.057*** (0.011)
Gemini 2.5 Flash		0.010* (0.006)		0.010 (0.008)
Gemini 3 Pro		0.021*** (0.005)		0.021** (0.010)
Intercept	0.477*** (0.004)	0.475*** (0.005)	0.477*** (0.008)	0.475*** (0.013)
LLM fixed effects	No	Yes	No	Yes
Clustered SEs	No	No	Yes	Yes
N	4914	4914	4914	4914
R^2	0.378	0.475	0.378	0.475
Adj. R^2	0.377	0.474	0.377	0.474
Clusters (cases)	–	–	13	13

Notes: This table reports the results of an OLS regression model. The dependent variable is the composite score $\in [0, 1]$, combining conclusion accuracy and reasoning quality. Columns (1)–(2) report robust standard errors; columns (3)–(4) report standard errors clustered at the case level. Each case–LLM combination is observed three times. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

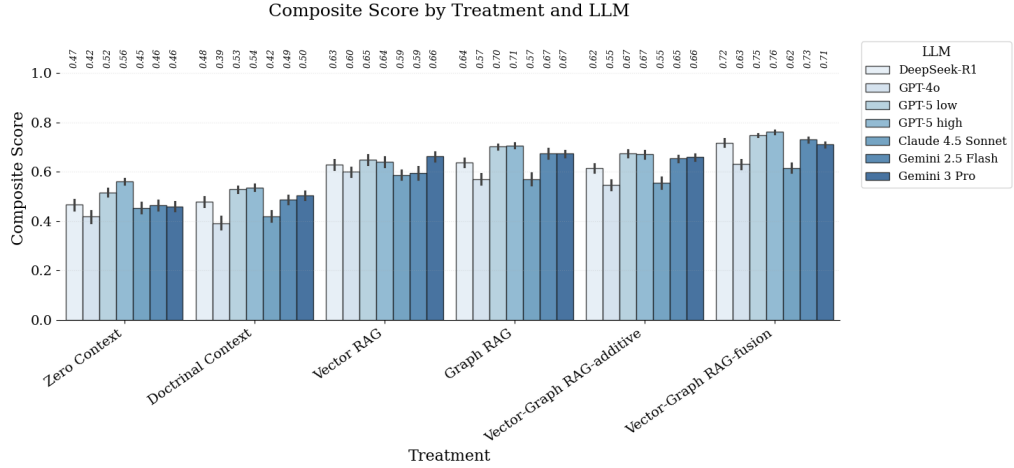


Figure 4: Composite Score by Treatment and LLM

treatment, *Vector-Graph RAG-fusion*, achieves a mean answer score of 0.780 but a composite score of 0.702. This gap indicates that even the best RAG architecture is imperfect at context retrieval: LLMs generate high-quality legal reasoning despite, rather than purely because of, the retrieved context. The context metrics penalize retrieved elements of legal doctrine that are not fully utilized or relevant to the legal case at hand.

Overall, however, the aforementioned staircase effect remains relatively stable across both metrics. *Vector-Graph RAG-fusion* dominates on answer score (0.780) and composite score (0.702), followed by the *Graph RAG* and *Vector RAG* treatments. It is worth noting that *Vector-Graph RAG-additive* performs no better than *Vector RAG* on the composite score and no better than *Graph RAG* on the answer score. The mere addition of vector-based retrieval and a knowledge graph does not seem to improve predictions by much. Rather, our results suggest that a densely structured retrieval pipeline and pre-synthesis is required for genuine improvements in legal reasoning quality.

Table 5: Answer Score by Treatment and LLM

Treatment	DeepSeek-R1	GPT-4o	GPT-5 low	GPT-5 high	Claude 4.5 Sonnet	Gemini 2.5 Flash	Gemini 3 Pro	Total
Zero Context	0.622 (0.153)	0.558 (0.172)	0.689 (0.115)	0.747 (0.086)	0.606 (0.162)	0.618 (0.148)	0.614 (0.136)	n=819
Doctrinal Context	0.637 (0.137)	0.521 (0.184)	0.704 (0.105)	0.715 (0.096)	0.561 (0.166)	0.649 (0.127)	0.673 (0.119)	n=819
Vector RAG	0.679 (0.127)	0.662 (0.140)	0.709 (0.128)	0.701 (0.139)	0.641 (0.134)	0.633 (0.194)	0.729 (0.120)	n=819
Graph RAG	0.703 (0.128)	0.619 (0.150)	0.779 (0.069)	0.788 (0.057)	0.605 (0.161)	0.739 (0.140)	0.760 (0.088)	n=819
Vector-Graph RAG-additive	0.703 (0.125)	0.616 (0.159)	0.784 (0.089)	0.767 (0.111)	0.605 (0.163)	0.748 (0.095)	0.766 (0.102)	n=819
Vector-Graph RAG-fusion	0.802 (0.084)	0.699 (0.117)	0.838 (0.029)	0.855 (0.020)	0.660 (0.132)	0.812 (0.054)	0.794 (0.058)	n=819
Total	n=702	n=702	n=702	n=702	n=702	n=702	n=702	N=4914

Note: Standard deviations in parentheses.

4.2 Does Legal Reasoning Explain Accurate Legal Conclusions?

A natural question emerging from our analysis is whether the accuracy of legal conclusions reached by LLMs can be explained by the quality of their legal reasoning. To examine the relationship between composite score and conclusion accuracy, we estimate logistic regression models, with conclusion accuracy as the dependent variable. Table 6 presents the results.

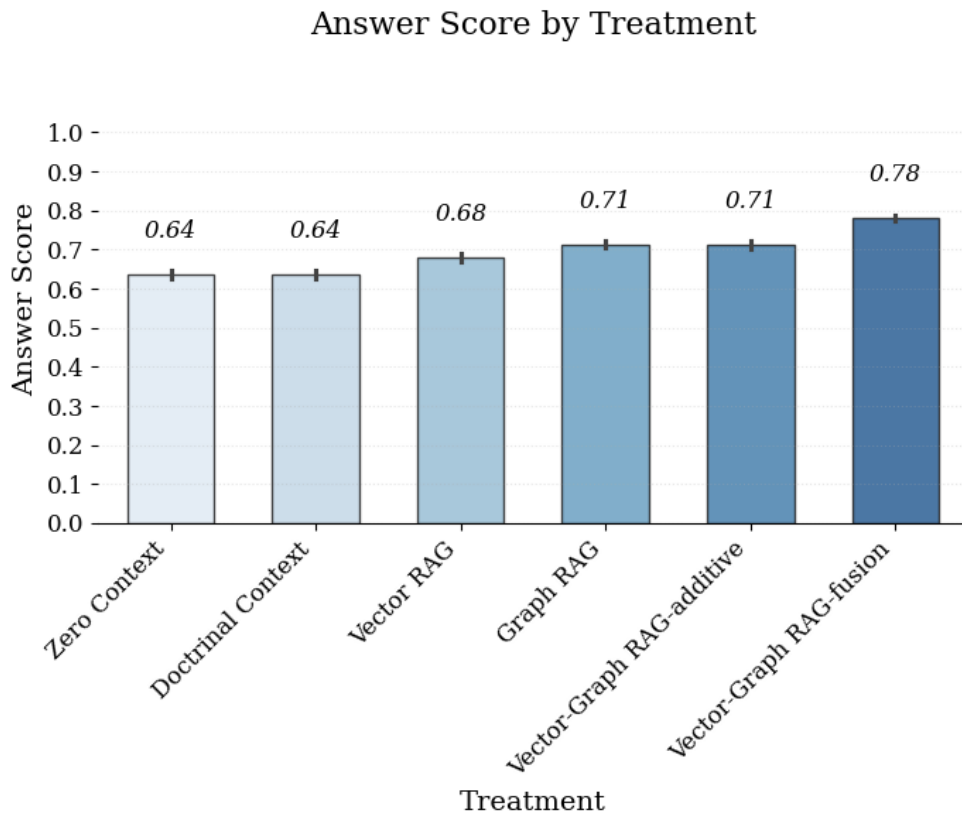


Figure 5: Answer Score by Treatment

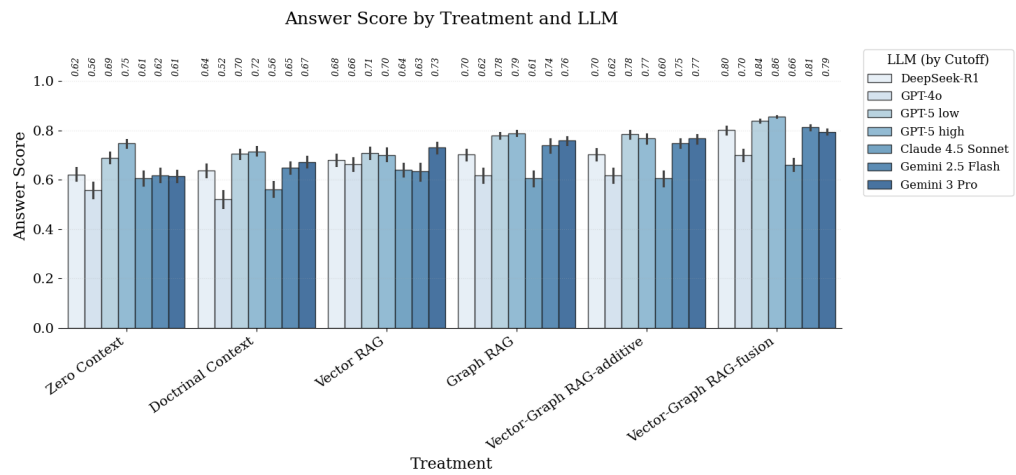


Figure 6: Answer Score by Treatment and LLM

Model 1 estimates the unconditional effect of composite score, with data pooled across treatments and LLMs; Model 2 introduces treatment dummies and their interactions with composite score to examine heterogeneity across RAG treatments, with data pooled across LLMs.

Table 6: Composite Score and Treatment Interactions

	Model 1		Model 2	
	Log-Odds	ME	Log-Odds	ME
Composite Score	2.733*** (0.254) [2.235, 3.230]	0.431*** (0.039) [0.354, 0.507]	2.972*** (0.675) [1.649, 4.294]	0.453*** (0.102) [0.253, 0.654]
Doctrinal Context			1.023* (0.461) [0.119, 1.926]	0.156* (0.070) [0.018, 0.294]
Vector RAG			2.736*** (0.571) [1.617, 3.854]	0.417*** (0.086) [0.248, 0.587]
Graph RAG			1.115 (0.609) [−0.079, 2.308]	0.170 (0.093) [−0.012, 0.352]
Vector-Graph RAG-additive			3.610*** (0.719) [2.201, 5.019]	0.551*** (0.109) [0.337, 0.764]
Vector-Graph RAG-fusion			2.231* (1.039) [0.195, 4.267]	0.340* (0.158) [0.030, 0.650]
Score × Doctrinal Context			−1.769 (0.958) [−3.648, 0.109]	−0.270 (0.146) [−0.556, 0.016]
Score × Vector RAG			−4.311*** (0.996) [−6.263, −2.359]	−0.658*** (0.151) [−0.953, −0.362]
Score × Graph RAG			−1.067 (1.050) [−3.125, 0.990]	−0.163 (0.160) [−0.476, 0.151]
Score × Vector-Graph RAG-additive			−4.655*** (1.201) [−7.008, −2.301]	−0.710*** (0.182) [−1.067, −0.352]
Score × Vector-Graph RAG-fusion			−1.602 (1.567) [−4.673, 1.468]	−0.244 (0.239) [−0.712, 0.224]
N	4914		4914	
Pseudo R^2	0.024		0.057	
AUC-ROC	0.604		0.666	

Notes: This table reports the results of a logistic regression model. Standard errors in parentheses. 95% confidence intervals in brackets. ME = average marginal effect. Reference category: Zero Context. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

In Model 1, composite score is a strong predictor of conclusion accuracy ($\beta = 2.733$, $p < 0.001$, Figure 7). A 0.1-unit increase in composite score is associated with a 4.31 percentage point increase in the probability of reaching a correct conclusion. The odds ratio of 15.4 indicates that each unit increase in composite score multiplies the odds of a correct prediction

by a factor of fifteen. In short, higher-quality legal reasoning – as captured by the composite score – translates into more accurate predictions of ECtHR decisions. Despite this highly significant relationship, the model’s explanatory power remains modest (Pseudo $R^2 = 0.024$, AUC = 0.604). This indicates that factors other than legal reasoning quality also substantially influence the accuracy of predicted legal conclusions.

Model 2 reveals treatment heterogeneity (Figure 8). The main effects indicate that, at mean composite score, several RAG treatments significantly outperform the *Zero Context* baseline: *Vector-Graph RAG-additive* (ME = 0.551, $p < 0.001$), *Vector RAG* (ME = 0.417, $p < 0.001$), *Vector-Graph RAG-fusion* (ME = 0.340, $p < 0.05$), and *Doctrinal Context* (ME = 0.156, $p < 0.05$). *Graph RAG* alone shows no significant improvement over the baseline. The interaction terms reveal a striking pattern: the relationship between reasoning quality and conclusion accuracy is particularly weak in the most effective RAG treatments. For *Vector RAG*, the interaction coefficient is strongly negative ($\beta = -4.311$, $p < 0.001$), as is the interaction for *Vector-Graph RAG-additive* ($\beta = -4.655$, $p < 0.001$). These negative interactions indicate that while composite score strongly predicts accuracy in the *Zero Context* baseline, this relationship flattens considerably when LLMs have access to effective retrieval of the doctrinal context.

One potential explanation is that LLMs may operate based on a compensation mechanism: RAG systems appear to be substitutes for legal reasoning quality inherited from the training data. When an LLM produces lower-quality reasoning, it benefits substantially from the contextual knowledge provided through RAG – the retrieved legal materials compensate for deficiencies in the model’s autonomous reasoning. Conversely, when LLMs engage in high-quality reasoning, they gain less from RAG – their baseline performance is already high – and, partly, even seem to become less accurate in their conclusions. Put differently, RAG narrows the performance gap between instances of strong and weak legal reasoning: it lifts the floor, but it doesn’t raise the ceiling and, sometimes, even lowers it.

4.3 LLM-as-a-judge: Legal Reasoning

Our LLM-as-a-judge evaluation decomposes reasoning quality into eight domain-specific criteria (Figure 9). While several metrics show a general upward trend, the most significant performance shifts across treatments are observed for source transparency and precedent quality.

Our results reveals a stark contrast in how models use legal authority across different levels of doctrinal access. In the *Zero Context* and *Doctrinal Context* treatments, scores for precedent quality remain at a functional minimum. This suggests that while LLMs can identify general legal principles from memory, they are unable to provide the specific, localized citations required for judicial reasoning in our context.

A sharp upward shift occurs with the introduction of the RAG treatments. The *Vector-Graph RAG-fusion* treatment, in particular, achieves the highest scores in source transparency and precedent quality. This peak suggests that the Evidence Curator and Case Strategist steps

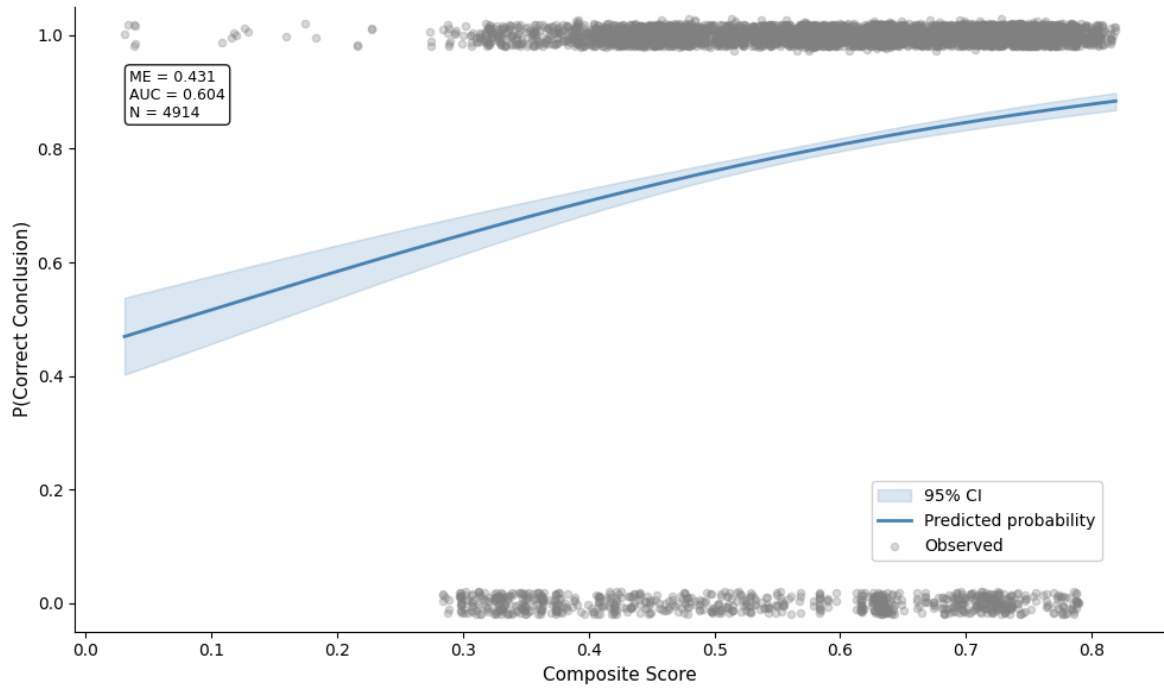


Figure 7: Predicted probability of correct conclusion as a function of composite score (Model 1). The solid curve shows fitted values from Model 1; the shaded region represents 95% confidence intervals computed via the delta method. Gray points are observed outcomes (0 = incorrect, 1 = correct), jittered vertically for visibility. ME = average marginal effect; AUC = area under the ROC curve; N = 4,914.

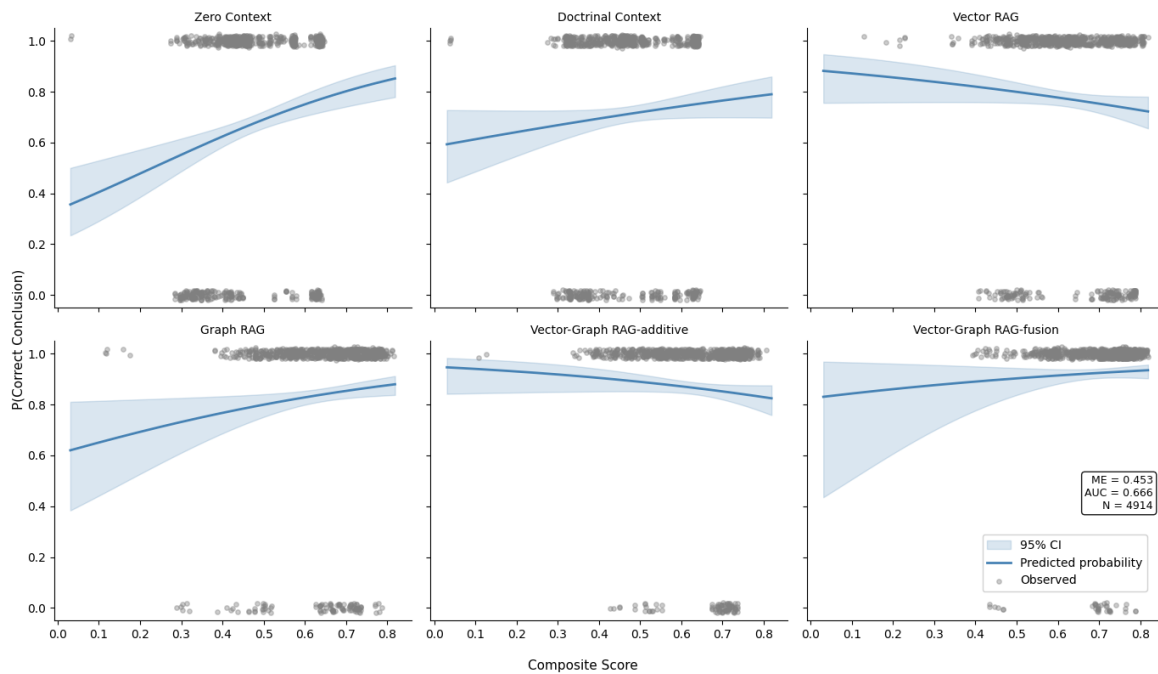


Figure 8: Predicted probability of correct conclusion as a function of composite score by treatment (Model 2). Each panel displays the fitted values from Model 2 for one treatment, with 95% confidence intervals (delta method) and observed outcomes. Gray points are observed outcomes (0 = incorrect, 1 = correct), jittered vertically for visibility. ME = average marginal effect; AUC = area under the ROC curve; N = 4,914.

in the fusion process are critical. By pre-populating the examination schema with specific paragraph references and weighted priorities before generation, the architecture effectively eliminates the vague referencing common in simpler retrieval models. This RAG architecture results in highly traceable and authoritative legal justifications.

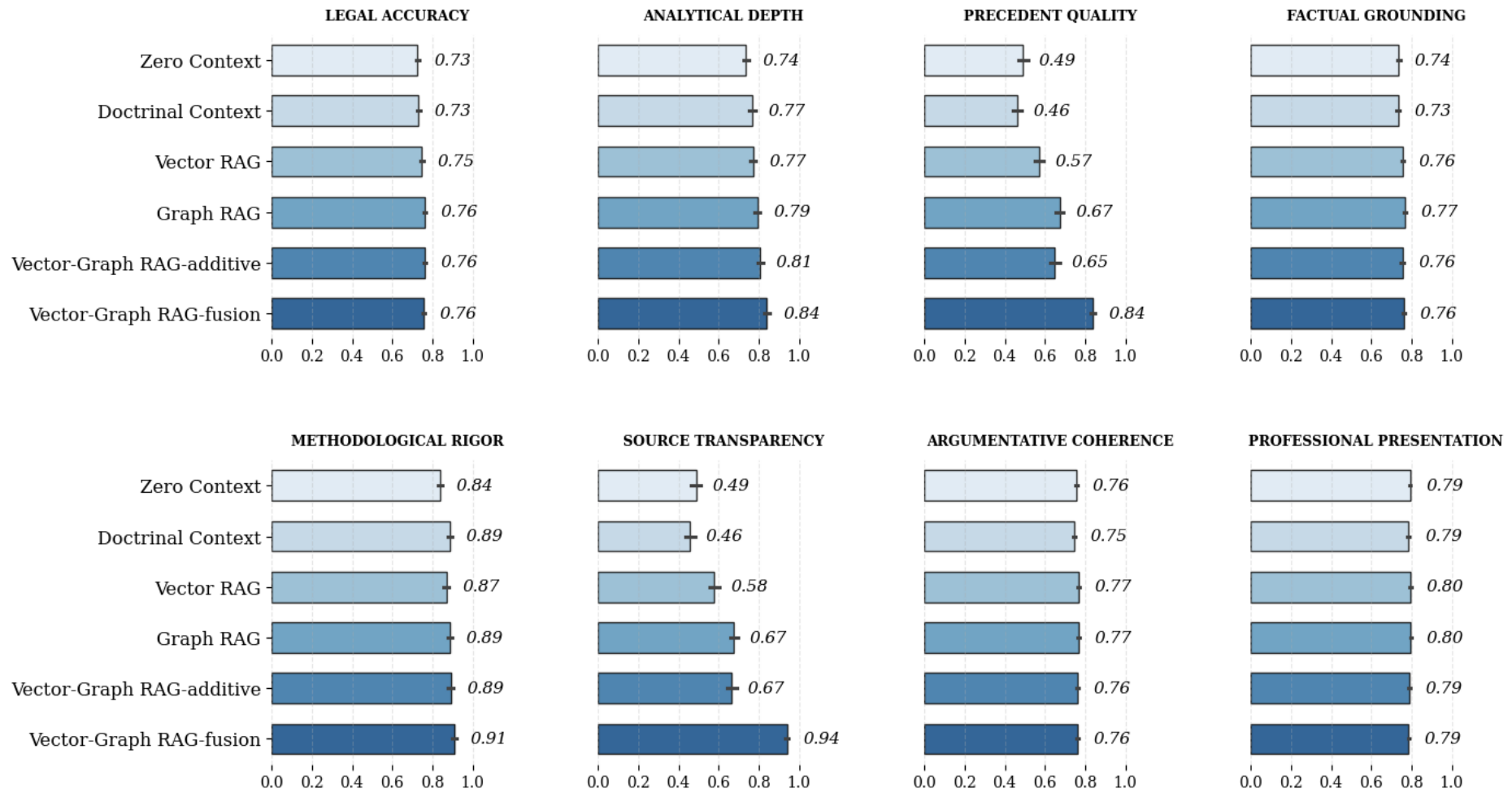


Figure 9: Detailed Performance Breakdown: Comparison of eight normalized legal metrics across all RAG treatments. Error bars represent 95% confidence intervals.

The global performance ranking (Figure 10) synthesizes these granular shifts. The monotonic increase in scores, culminating in the *Vector-Graph RAG-fusion* treatment, reflects the additive benefit of precise precedent integration and high source transparency.

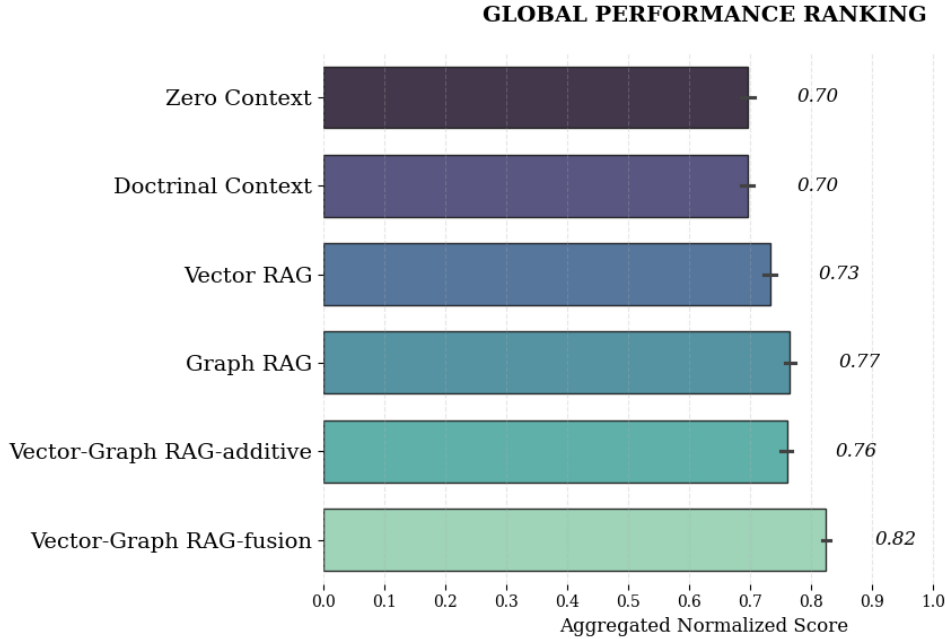


Figure 10: Global Performance Ranking: Aggregated mean scores across all evaluation dimensions. Higher scores indicate superior alignment with doctrinal accuracy and methodological rigor.

4.4 RAGAS: Contextual Reasoning

The RAGAS automated metrics provide a computational assessment of the reasoning process. As shown in Figure 11, we observe significant performance shifts in the six experimental treatments for our three answer-level metrics: answer relevance, answer similarity, and answer correctness.

Because these three RAGAS metrics do not require a retrieved context for calculation, they provide a direct comparison between the *Zero Context* and *Doctrinal Context* treatments on the one hand, and the RAG treatments on the other hand.

We observe a small but significant upward shift in answer relevance, answer correctness and answer similarity as we move from *Zero Context* to *Vector-Graph RAG-fusion*. The shift in correctness and similarity indicates that our RAG architectures incrementally improve the factual accuracy of the verdict and the underlying reasoning process. The shift in relevance suggests that, by filtering out doctrinal noise, our RAG architectures ensure that the final justification is precisely tailored to the interference at hand, and does not simply produce generic reasons. Overall, even when the LLMs have a solid intuition regarding legal reasoning, our RAG architectures significantly improve the quality of the resulting text.

The remaining six RAGAS metrics center on how effectively the LLMs retrieve and operationalize doctrinal information to construct a justification.

Faithfulness essentially provides a hallucination check by measuring how strictly the generated justification is anchored in the retrieved context. We observe that the *Vector RAG* treatment outperforms all other RAG architectures. The additive architecture performs worst. This suggests that vector-based retrieval is most effective at ensuring that legal claims are directly supported by the ECtHR Reference Guide.

Context precision evaluates the signal-to-noise ratio of the retrieved items. The shift toward higher precision in the *Graph RAG* and the *Vector-Graph RAG-fusion* treatments indicates that the pure knowledge graph and the fusion architecture work best at filtering out irrelevant elements of legal doctrine, as defined embedded in graph paths or vector chunks. Again, the additive architecture performs worst.

Context utilization measures the extent to which the model actually incorporates the retrieved context into its final reasoning. A similar pattern as for context precision emerges. Notably, the additive architecture performs worst.

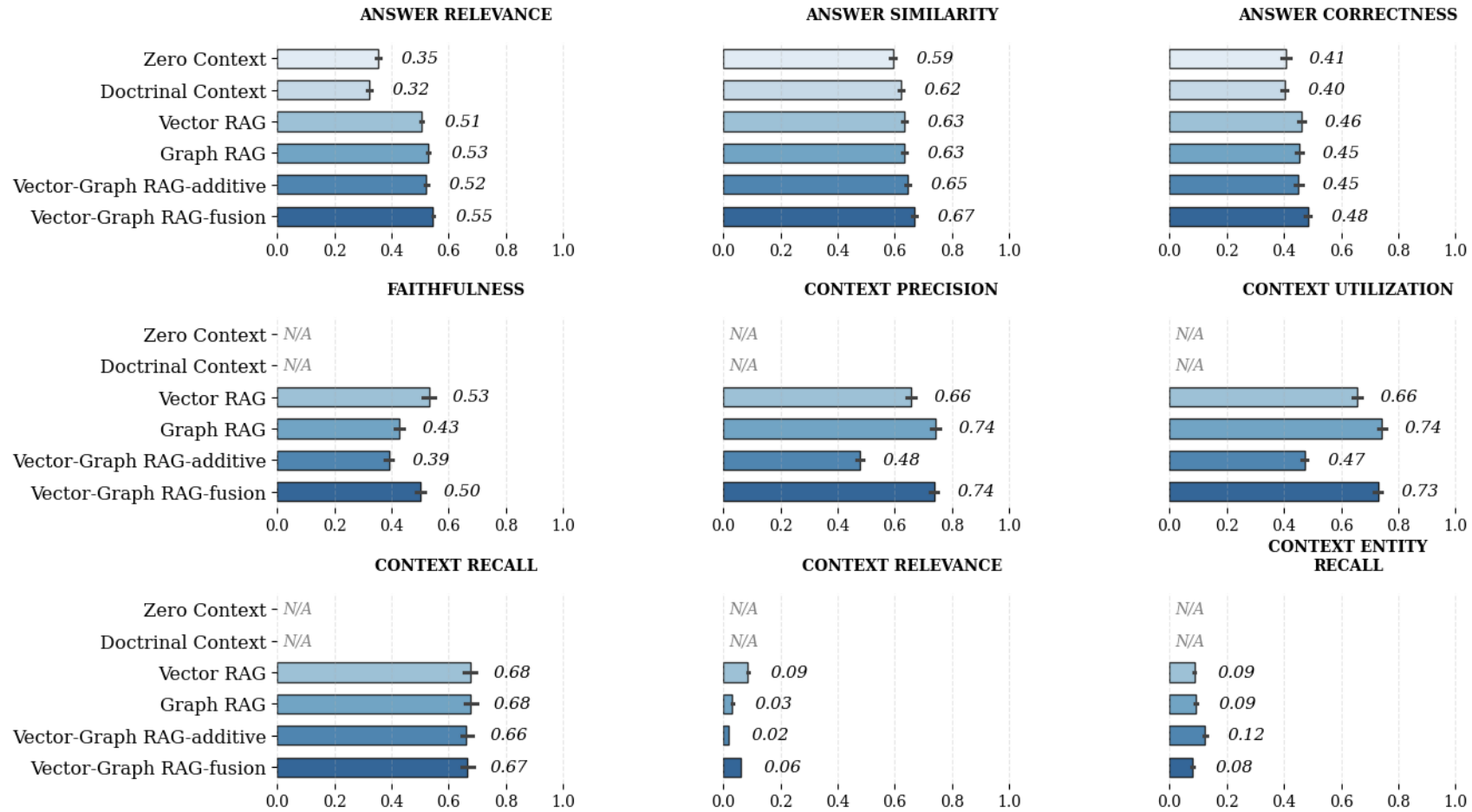


Figure 11: RAGAS automated evaluation results across nine dimensions. Error bars indicate 95% confidence intervals.

The global RAGAS ranking (Figure 12) synthesizes all RAGAS dimensions into a single index. The dominance of the *Vector-Graph RAG-fusion* treatment across both answer-based and context-based metrics confirms that our RAG architecture is not merely finding more information; rather, it is actively translating that information into higher-quality legal reasoning. These results provide evidence that the integration of structured and unstructured context does not merely assist the retrieval process; it qualitatively and quantitatively improve the quality and correctness of the judicial output itself.

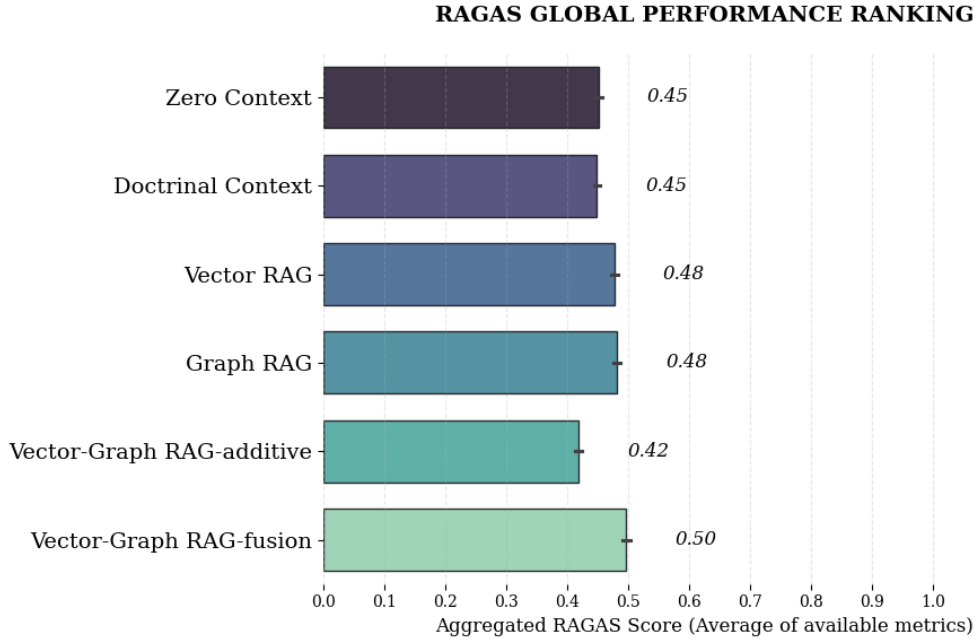


Figure 12: Global Performance Ranking across all treatments. The RAGAS mean is calculated based on available metrics for each treatment to ensure a fair baseline comparison.

5 Discussion

Endowing LLMs with judicial authority is no longer futuristic. One may feed in the documents submitted by the parties and evidence that has been properly introduced into the procedure, and may ask one of the frontier models to draft the disposition. In the Chinese province of Shenzhen, this is already daily practice (Liu and Li 2024), and Israel has started out its *Legal-Chat* that offers the same service. In Germany, the systems FRAUKE and OLGA do the same for two sets of mass litigation: the rights of airline passengers, and damage claims resulting from *Dieselgate*, a fraudulent scheme put in place by German car manufacturers. Admittedly, in all these jurisdictions, the output is treated as a draft that the competent judge is supposed to check and authorize. But it would technically clearly be feasible to take the human judge out of the loop. Even if the jurisdiction keeps requiring some kind of human involvement, it will be tempting for judges who are aching under a high caseload to essentially trust the LLM. Yet LLMs are not foolproof. They suffer from biases and, more worryingly for the judiciary, can be

manipulated by the parties. From a law and policy perspective, it is therefore crucial how well the LLM-based drafting or decision-making process is contained, in the sense required by the legal rules that it is meant to apply.

Now, it cannot be taken for granted either that humans entrusted with judicial authority act as unconditionally faithful agents; they may be suspected to follow their own political agenda. To contain this risk, the law has embedded judicial decision-making into a rich institutional arrangement. Judges need legal training; they are selected for professional expertise. Whether they get a chance to be promoted hinges on the perception that they are trustworthy. The parties are allowed to plead their case, so that alternative views about the facts and the law are presented to the judge. Most court decisions may be appealed, or may entail a complaint to a constitutional or human rights court. None of these safeguards have a direct analogy with LLMs. But another safeguard applies to human judges and LLM judges alike: they have to justify their decision using the language of the applicable legal rules.

It is this constraint that we investigate experimentally. In two ways, our data validates the constraining force of doctrine for LLM decision-making. Even if we just give the LLM the sketch of a case that has (later) been decided by the ECtHR (*Zero Context*), in 75% of all cases, the LLM comes to the same conclusion, with the content of the justification being very similar. This suggests that state of the art LLMs have picked up a fairly good grasp of the law (in our case, of the rights guaranteed by the ECHR) at training or instruction fine-tuning. Moreover, we find clear treatment effects, both on outcomes, and on a host of metrics for the quality of legal reasoning. Our treatments are designed to capture progressively more features of legal doctrine. With a simple prompt used in our *Doctrinal Context* treatment, we tell the LLM to reason in terms of established doctrine of the applicable human right. The *Vector RAG* also gives the LLM access to a summary account of the established doctrine that it can scan for similarity. The *Graph RAG* provides the LLM with the conceptual structure of the applicable legal provision. Our two *Vector-Graph RAG* architectures combine both inputs. The fusion version does not give both sets of input disjunctly. It rather integrates structure and illustrations seamlessly. All of these interventions matter. The more elaborate the representation of legal doctrine, the stronger its guiding effect.

The most worrying technical limitation of our experiment results from the risk of memorization. The decisions of the ECtHR are in the public domain. The HUDOC database of the Court has very likely been scraped by the available foundation models. We find the biggest effects with LLMs that have finished training after the stimulus cases have been published. We can therefore not exclude that the accuracy of both outcomes and legal reasoning is so high because the LLM has recognized the case, and has checked the decision by the Strasbourg Court. Note, however, that we also find smaller, yet significant effects with two LLMs that have a knowledge cutoff date prior to the decision of all stimulus cases. Moreover, if the LLMs had simply checked the Court's decision, we should not be able to observe any treatment effects.

As we were chiefly concerned about memorization, we have limited the set of cases to the

13 cases that have been decided after the knowledge cutoff date of these two LLMs. 13 cases is not a huge number. We cannot exclude that features of these cases have made the effect of doctrine particularly visible. We also only test cases that require the application of freedom of assembly. We cannot exclude that the effect of doctrine is different for other human rights, or for other legal provisions for that matter. Yet we have given priority to depth over breadth. For our research question, we have considered it more important to capture multiple facets of doctrine with the help of a rich set of metrics. The resulting computational load was already quite substantial. Still it would of course be desirable, in future work, to replicate the effects of doctrine with different legal provisions.

Except for DeepSeek-R1, all LLMs that we have used are proprietary. DeepSeek-R1 is open weights, but not open source. Hence we do not have full control over the process, and in particular not over the training data. We also do not know which attempts the providers have made in instruction fine-tuning to align law-related outputs with existing doctrine. It certainly would be an interesting (but challenging) companion project to use an open source model to replicate the experiment. But available open source models are much smaller, and hence also less powerful. More importantly, frontier models are the most likely base model for any tool to which legal decision-making is delegated (the Israeli model *Legal-Chat* is based on Gemini). Therefore, we are testing the most likely candidates for legal decision-making by LLMs.

6 Conclusion

Could the decision of (some) legal cases responsibly be delegated to LLMs? How much attention would legislators have to pay to safeguards if they permit the use of LLMs as auxiliary tools in the court room? The responses to these questions depend on the degree of control: against inadvertent bias, but also against intentional manipulation through the parties. Our experiment tests one guardrail that is implementable: the legislator requests the decisions to be justified in the categories of existing legal doctrine. The experiment shows that doctrine has indeed a strong guiding effect, and that the effect can even be strengthened by the way how doctrine is integrated with the facts of the case. Building legal doctrine into LLMs is a promising technique for containing the exercise of sovereign powers by machines.

References

- Araújo, Gabriele S, Ewaldo EC Santana, Antonio FL Jacob Junior, and Fábio MF Lobato. 2025. “The Artificial Intelligence Integration in the Brazilian Legal Sector: A Systematic Review.” *Simpósio Brasileiro de Sistemas de Informação (SBSI)*, 575–584.

- Bagherian-Marandi, Navid, Mehdi Ravanshadnia, and Mohammad-R Akbarzadeh-T. 2021. "Two-layered fuzzy logic-based model for predicting court decisions in construction contract disputes." *Artificial Intelligence and Law* 29 (4): 453–484.
- Bai, Xuechunzi, Angelina Wang, Ilia Sucholutsky, and Thomas L Griffiths. 2025. "Explicitly unbiased large language models still form biased associations." *Proceedings of the National Academy of Sciences* 122 (8): e2416228122.
- Balasubramanian, Natarajan, Chenyang Pan, Jian Xie, and Wenjian Xu. 2025. "AI in the Court." *Available at SSRN* 5845764.
- Balganesh, Shyamkrishna. 2014. "The Constraint of Legal Doctrine." *University of Pennsylvania Law Review* 163:1843.
- Benedetto, Irene, Alkis Koudounas, Lorenzo Vaiani, Eliana Pastor, Luca Cagliero, Francesco Tarasconi, and Elena Baralis. 2025. "Boosting court judgment prediction and explanation using legal entities." *Artificial Intelligence and Law* 33 (3).
- Bertalan, Vithor Gomes Ferreira, and Evandro Eduardo Seron Ruiz. 2024. "Using attention methods to predict judicial outcomes." *Artificial Intelligence and Law* 32 (1): 87–115.
- Bex, Floris, and Henry Prakken. 2021. "On the relevance of algorithmic decision predictors for judicial decision making." In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, 175–179.
- Blass, Joseph, and Kenneth D Forbus. 2023. "Analogical Reasoning, Generalization, and Rule Learning for Common Law Reasoning." In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, 32–41.
- Bolukbasi, Tolga, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. "Man is to computer programmer as woman is to homemaker? Debiasing word embeddings." *Advances in Neural Information Processing Systems* 29.
- Boston, Joshua. 2020. "Strategic Opinion Language on the US Courts of Appeals" [in en]. *Journal of Law and Courts* 8, no. 1 (April): 1–26. ISSN: 2164-6570, 2164-6589, accessed July 23, 2026. <https://doi.org/10.1086/704633>. <https://www.cambridge.org/core/journals/journal-of-law-and-courts/article/strategic-opinion-language-on-the-us-courts-of-appeals/E0ED0DC2D15086E4846CBA946E72CD63>.
- Canavotto, Ilaria, and John Horty. 2023. "Reasoning with hierarchies of open-textured predicates." In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, 52–61.

- Carlini, Nicholas, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2022. "Quantifying memorization across neural language models." In *The Eleventh International Conference on Learning Representations*.
- Carlini, Nicholas, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. "Extracting training data from large language models." In *30th USENIX security symposium (USENIX Security 21)*, 2633–2650.
- Chen, Kai, Zihao He, Jun Yan, Taiwei Shi, and Kristina Lerman. 2024. "How Susceptible are Large Language Models to Ideological Manipulation?" *arXiv preprint arXiv:2402.11725*.
- Cui, Junyun, Xiaoyu Shen, and Shaochun Wen. 2023. "A survey on legal judgment prediction: Datasets, metrics, models and challenges." *IEEE Access* 11:102050–102071.
- Dickinson, John. 1931. "Legal Rules: Their Function in the Process of Decision." *University of Pennsylvania Law Review* 79:833–868.
- Dzehtsiarou, Kanstantsin. 2011. "European consensus and the evolutive interpretation of the European Convention on Human Rights." *German Law Journal* 12 (10): 1730–1745.
- Engel, Christoph, and Johannes Kruse. 2026. "LLM as a law professor: Having a Large Language Model Write a Commentary on Freedom of Assembly." *Artificial Intelligence and Law*, ***.
- Engel, Christoph, and Richard H McAdams. 2024. "Asking GPT for the ordinary meaning of statutory terms." *University of Illinois Journal of Law, Technology & Policy*, 235.
- Epstein, Lee, and T. Jacobi. 2010. "The Strategic Analysis of Judicial Decisions." *Annual Review of Law and Social Science* 6:1–18.
- Epstein, Lee, and Jack Knight. 1998. *The Choices Justices Make*. Washington, DC: C.Q. Press.
- Es, Shahul, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. "Ragas: Automated evaluation of retrieval augmented generation." In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, 150–158.
- Eskridge, William N., and Philip P. Frickey. 1987. "Legislation Scholarship and Pedagogy in the Post-Legal Process Era." *University of Pittsburgh Law Review* 48:691–731.
- European Court of Human Rights. 2025. *Guide on Article 11 of the European Convention on Human Rights: Freedom of Assembly and Association*. Strasburg: European Council.
- Feng, Shangbin, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. "From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models." *arXiv preprint arXiv:2305.08283*.

- Feng, Yi, Chuanyi Li, and Vincent Ng. 2022. "Legal Judgment Prediction: A Survey of the State of the Art." In *IJCAI*, 5461–5469.
- Fikfak, Veronika, and Laurence R Helfer. 2025. "Automating International Human Rights Adjudication." *Michigan Journal of International Law* 46:69.
- Fuller, L. L. 1934. "American Legal Realism." *University of Pennsylvania Law Review* 82:429–462.
- Fungwacharakorn, Wachara, Kanae Tsushima, and Ken Satoh. 2021. "On semantics-based minimal revision for legal reasoning." In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, 50–59.
- Greer, Steven C. 2000. *The Margin of Appreciation: Interpretation and Discretion under the European Convention on Human Rights*. Vol. 17. Council of Europe.
- Guitton, Clement, Reto Gubelmann, Ghassen Karray, Simon Mayer, and Aurelia Tamò-Larrieux. 2025. "Identifying open-texture in regulations using LLMs." *Artificial Intelligence and Law*, 1–45.
- Guitton, Clement, Aurelia Tamò-Larrieux, Simon Mayer, and Gijs van Dijck. 2025. "The challenge of open-texture in law." *Artificial Intelligence and Law* 33 (2): 405–435.
- Hartmann, Jochen, Jasper Schwenzow, and Maximilian Witte. 2023. "The political ideology of conversational AI: Converging evidence on ChatGPT's pro-environmental, left-libertarian orientation." *arXiv preprint arXiv:2301.01768*.
- Jiang, Cong, and Xiaolei Yang. 2023. "Legal syllogism prompting: Teaching large language models for legal judgment prediction." In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, 417–421.
- Katz, Ori, and Eyal Zamir. 2025. "Law, Justice, and Artificial Intelligence."
- Kesari, Aniket, Daniela Sele, Elliott Ash, and Stefan Bechtold. 2024. "A legal framework for explainable artificial intelligence." *Center for Law & Economics Working Paper Series* 9.
- Kronman, Anthony. 1988. "Jurisprudential Responses to Legal Realism." *Cornell Law Review* 73:335–376.
- Kruse, Johannes. 2025. "The Ordinary Meaning Bot: Simulating Human Surveys with LLMs." *MPI Collective Goods Discussion Paper*, nos. 2025/12.
- Lax, Jeffrey R. 2011. "The New Judicial Politics of Legal Doctrine." Contemporary synthesis of research on how legal doctrine shapes judicial behavior, *Annual Review of Political Science* 14:131–157.

- Lee, Meng-Chieh, Qi Zhu, Costas Mavromatis, Zhen Han, Soji Adeshina, Vassilis N. Ioannidis, Huzefa Rangwala, and Christos Faloutsos. 2024. "HybGRAG: Hybrid Retrieval-Augmented Generation on Textual and Relational Knowledge Bases." Accepted to ACL 2025, *arXiv preprint arXiv:2412.16311*.
- Leiter, Brian. 2010. "Legal Formalism and Legal Realism: What is the Issue?" *Legal Theory* 16:111–133.
- Levy, Haim V. 2025. "Simulating Judgment with GPT-5: Large Language Models as Judicial Reasoners." *Available at SSRN* 5613031.
- Liu, John Zhuang, and Xueyao Li. 2024. "How do judges use large language models? Evidence from Shenzhen." *Journal of Legal Analysis* 16 (1): 235–262.
- Lüders, Kilian, and Bent Stohlmann. 2025. "Classifying proportionality-identification of a legal argument." *Artificial Intelligence and Law* 33 (4): 1051–1078.
- Mainali, Nischal, Liam Meier, Elliott Ash, and Daniel L Chen. 2020. "Automated classification of modes of moral reasoning in judicial decisions." In *Computational Legal Studies*, 77–94. Edward Elgar Publishing.
- Masood, Ali S., Benjamin J. Kassow, and Donald R. Songer. 2019. "The Aggregate Dynamics of Lower Court Responses to the US Supreme Court" [in en]. *Journal of Law and Courts* 7, no. 2 (October): 159–186. ISSN: 2164-6570, 2164-6589, accessed July 23, 2026. <https://doi.org/10.1086/703067>. <https://www.cambridge.org/core/journals/journal-of-law-and-courts/article/aggregate-dynamics-of-lower-court-responses-to-the-us-supreme-court/1A67714120E06C89A0A23A50306C8B83>.
- McConnell, Devin J, James Zhu, Sachin Pandya, and Derek Aguiar. 2021. "Case-level prediction of motion outcomes in civil litigation." In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, 99–108.
- McDougal, Myres S. 1947. "Law School of the Future: From Legal Realism to Policy Science in the World Community." *Yale Law Journal* 56:1345–1355.
- Medvedeva, Masha, Michel Vols, and Martijn Wieling. 2020. "Using machine learning to predict decisions of the European Court of Human Rights." *Artificial Intelligence and Law* 28 (2): 237–266.
- Medvedeva, Masha, Martijn Wieling, and Michel Vols. 2023. "Rethinking the field of automatic prediction of court decisions." *Artificial Intelligence and Law* 31 (1): 195–212.
- Min, Hyunsik, and Byeongjoon Noh. 2025. "TRACS-LLM: LLM-based traffic accident criminal sentencing prediction focusing on imprisonment, probation, and fines." *Artificial Intelligence and Law*, 1–22.

- Mumford, Jack, Katie Atkinson, and Trevor Bench-Capon. 2023. "Combining a legal knowledge model with machine learning for reasoning with legal cases." In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, 167–176.
- Odekerken, Daphne, Floris Bex, and Henry Prakken. 2024. "Precedent-based reasoning with incomplete information for human-in-the-loop decision support." *Artificial Intelligence and Law*, 1–46.
- Park, Jee Won, and Sunmi Park. 2024. "When Correct Isn't Enough: Deconstructing Legal Causal Reasoning Capability in Large Language Models." *Available at SSRN 5774579*.
- Pinello, Daniel R. 1999. "Linking party to judicial ideology in American courts: A meta-analysis." *The Justice System Journal*, 219–254.
- Podder, Sumanta Narayan. 2025. "Interim Relief in Dutch Digital Courts: Integrating Compliance Frameworks from EU Digital Laws with Article 47 ECHR Fundamental Rights Protection." *Available at SSRN 5575416*.
- Posner, Richard A. 1993. "What do judges and justices maximize? (The same thing everybody else does)." *Supreme Court Economic Review* 3:1–41.
- Pound, Roscoe. 1908. "Mechanical Jurisprudence." *Columbia Law Review* 8:605–623.
- Prescott, JJ. 2024. "Next Steps in Online Courts: Accelerating Access to Justice Through Court Technology." *Alaska Law Review* 41:93.
- Rigoni, Adam. 2025. "Toward representing interpretation in factor-based models of precedent." *Artificial Intelligence and Law* 33 (1): 199–226.
- Rodríguez Rodríguez, Carlos Rafael, Yarina Amoroso Fernández, Denis Sergeevich Zuev, Marieta Peña Abreu, and Yeleny Zulueta Veliz. 2024. "M-LAMAC: a model for linguistic assessment of mitigating and aggravating circumstances of criminal responsibility using computing with words." *Artificial Intelligence and Law* 32 (3): 697–739.
- Salaün, Olivier, Philippe Langlais, and Karim Benyekhlef. 2021. "Labels distribution matters in performance achieved in legal judgment prediction tasks." In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, 268–269.
- Santin, Piera, Giulia Grundler, Andrea Galassi, Federico Galli, Francesca Lagioia, Elena Palmieri, Federico Ruggeri, Giovanni Sartor, and Paolo Torroni. 2023. "Argumentation structure prediction in CJEU decisions on fiscal state aid." In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, 247–256.

- Sarmah, Bhaskarjit, Dhagash Mehta, Benika Hall, Rohan Rao, Sunil Patel, and Stefano Pasquali. 2024. "Hybridrag: Integrating knowledge graphs and vector retrieval augmented generation for efficient information extraction." In *Proceedings of the 5th ACM International Conference on AI in Finance*, 608–616.
- Savelka, Jaromir, Kevin D Ashley, Morgan A Gray, Hannes Westermann, and Huihui Xu. 2023. "Explaining legal concepts with augmented large language models (GPT-4)." *arXiv preprint arXiv:2306.09525*.
- Schauer, Frederick. 2013. "Legal Realism Untamed." *Texas Law Review* 91:749–780.
- . 2015. *The Force of Law*. Reassessment of the role of coercion in distinguishing law from other social rules. Cambridge, MA: Harvard University Press.
- Schweitzer, NJ, and Jessica M Salerno. 2025. "The (Dis) utility of Large Language Model AI in Predicting Juror Decisions." *Available at SSRN* 5706582.
- Segal, Jeffrey A., and Harold J. Spaeth. 1993. *The Supreme Court and the Attitudinal Model*. New York: Cambridge University Press.
- . 2002. *The Supreme Court and the Attitudinal Model Revisited*. New York: Cambridge University Press.
- Sesodia, Magnus, Alina Petrova, John Armour, Thomas Lukasiewicz, Oana-Maria Camburu, Puneet K Dokania, Philip Torr, and Christian Schroeder de Witt. 2025. "AnnoCaseLaw: A Richly-Annotated Dataset For Benchmarking Explainable Legal Judgment Prediction." *arXiv preprint arXiv:2503.00128*.
- Smith, Joseph L., and James A. Todd. 2015. "Rules, Standards, and Lower Court Decisions" [in en]. *Journal of Law and Courts* 3, no. 2 (October): 257–275. ISSN: 2164-6570, 2164-6589, accessed July 23, 2026. <https://doi.org/10.1086/681544>. <https://www.cambridge.org/core/journals/journal-of-law-and-courts/article/rules-standards-and-lower-court-decisions/DCC1ED46B4987FB7D9BF05C54540122E>.
- Srinivasan, S, Nidhi Tyagi, and Radhey Shyam Jha. 2025. "Employing Artificial Intelligence for Adjudication in Indian Courts." *Available at SSRN* 5277233.
- Steging, Cor, Silja Renooij, and Bart Verheij. 2021. "Discovering the rationale of decisions: towards a method for aligning learning and reasoning." In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, 235–239.
- Stillwell, Hayley, and Sean A Harrington. 2025. "Michael Scott Is Not a Juror: The Limits of AI in Simulating Human Judgment." *Available at SSRN* 5400737.
- Tiller, Emerson H., and Frank B. Cross. 2006. "What is Legal Doctrine?" *Northwestern University Law Review* 100 (1): 517–534.

- Waldon, Brandon, Nathan Schneider, Ethan Wilcox, Amir Zeldes, and Kevin Tobia. 2025. "Large Language Models for Legal Interpretation? Don't Take Their Word for It." *Don't Take Their Word for It* (February 03, 2025).
- Wardas, Oliver, and Florian Matthes. 2025. "LLMs for Legal Subsumption in German Employment Contracts." *arXiv preprint arXiv:2507.01734*.
- Wei, Bin, Yaoyao Yu, Leilei Gan, and Fei Wu. 2025. "An LLMs-based neuro-symbolic legal judgment prediction framework for civil cases." *Artificial Intelligence and Law*, 1–35.
- Wilf-Townsend, Daniel, and Kevin Tobia. 2025. "AI-Generated Legal Texts." *Available at SSRN* 5243382.
- Wong, Jeremy. 2024. "Digital Justice: Unleashing the Promises and Containing the Perils of AI in China's Courts." *Available at SSRN* 5199005.
- Yamada, Hiroaki, Takenobu Tokunaga, Ryutaro Ohara, Akira Tokutsu, Keisuke Takeshita, and Mihoko Sumida. 2025. "Japanese tort-case dataset for rationale-supported legal judgment prediction." *Artificial Intelligence and Law* 33 (3): 783–807.
- Zhang, Chiyuan, Daphne Ippolito, Katherine Lee, Matthew Jagielski, Florian Tramèr, and Nicholas Carlini. 2023. "Counterfactual memorization in neural language models." *Advances in Neural Information Processing Systems* 36:39321–39362.
- Zhang, Han, Zhicheng Dou, Yutao Zhu, and Ji-Rong Wen. 2023. "Contrastive learning for legal judgment prediction." *ACM Transactions on Information Systems* 41 (4): 1–25.
- Zhang, Kepu, Weijie Yu, Zhongxiang Sun, and Jun Xu. 2025. "Syler: A framework for explicit syllogistic legal reasoning in large language models." In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, 4117–4127.

Appendix

A Example Case Sketch

Communicated Case:[SUBJECT MATTER OF THE CASE: “The application concerns a gathering dispersed by the police on the grounds that the demonstration had not been declared to the Silopi (Şırnak) sub-prefecture and that the demonstrators, including the three applicants, members of the political party Demokratik Toplum Partisi, intended to make a public statement to the press on the highway leading to the Habur border gate, located on the Iraqi border. The police sprayed the demonstrators with water cannons and, during the altercations, the applicant Sevaahir Bayındır fractured her hip. Following the complaint lodged by the applicants, the domestic authorities refused to authorize the prosecution of the police officers involved. Relying on Article 3 of the Convention, the applicant Sevaahir Bayındır argued that the use of force by the police constituted inhuman and degrading treatment and complained about the lack of an effective investigation into the matter. The applicants complained of a violation of their right to freedom of expression and assembly under Articles 10 and 11 of the Convention. Finally, relying on Article 13, they complained of the lack of a remedy for submitting their complaints in this regard.” QUESTIONS: “1. In view of the injury sustained by the applicant Sevaahir Bayındır during the intervention by the law enforcement authorities, was he subjected to inhuman or degrading treatment in violation of Article 3 of the Convention? 2. In view of the procedural protection against inhuman or degrading treatment (see *Bouyid v. Belgium* [GC], no. 23380/09, §§ 114-123, ECHR 2015), did the refusal of the domestic authorities to authorize the opening of proceedings against the law enforcement officers concerned violate Article 3 of the Convention? 3. Was there interference with the applicants’ freedom of expression and peaceful assembly within the meaning of Articles 10 § 1 and 11 § 1 of the Convention? If so, was that interference prescribed by law and necessary within the meaning of Articles 10 § 2 and 11 § 2? 4. Did the applicants have at their disposal, as required by Article 13 of the Convention, an effective domestic remedy through which they could have raised their complaints under Articles 10 and 11 of the Convention?”]

B Technical Implementation Details

B.1 Treatment Summary

Table 7 summarizes the six experimental treatments and their context configurations.

B.2 Model Configuration

Table 8 details the configuration for each LLM used in the experiments.

Treatment	Context Provided	Retrieval	Integration
Zero Context	None	—	—
Doctrinal Context	Static doctrinal framework	—	—
Vector RAG	Retrieved chunks	Hybrid	—
Graph RAG	Graph traversal results	Entry-point guided	—
Vector-Graph RAG-additive	Both (separate sections)	Both	None
Vector-Graph RAG-fusion	Both (synthesized)	Both	3-step LLM

Table 7: Summary of experimental treatments

Model Key	Model Identifier	Temp.	Timeout (s)	Thinking
openai_gpt_4o	gpt-4o	0	300	No
openai_gpt_5_low	gpt-5 (effort: low)	—	450	Yes
openai_gpt_5_high	gpt-5 (effort: high)	—	600	Yes
gemini_2_5_flash	gemini-2.5-flash	0	—	Yes
gemini_3_pro_preview	gemini-3-pro-preview	0	—	Yes
anthropic_claude_4_5	claude-sonnet-4.5-20251101	0	600	No
deepseek_r1	deepseek-r1	0.6	900	Yes

Table 8: Model configurations. GPT-5 variants do not support explicit temperature settings. DeepSeek-R1 uses the recommended temperature of 0.6 to prevent repetitive outputs. Gemini models with thinking capabilities support standard temperature parameters.

B.3 Token Budget Configuration

Table 9 specifies the token budgets governing retrieval and context construction.

Parameter	Value	Description
PROMPT_TOKEN_BUDGET	100,000	Maximum total prompt tokens
VECTOR_TOP_K	20	Chunks retrieved from vector database
VECTOR_MAX_TOKENS	12,000	Token budget for vector context
GRAPH_ENTRYPOINTS_MAX	15	Maximum graph entry points
GRAPH_MAX_DEPTH	5	Maximum traversal depth in graph
GRAPH_MAX_TOKENS	20,000	Token budget for graph context
INITIAL_RETRIEVAL_K	60	Initial candidates per retrieval method
RERANK_TOP_K	30	Candidates passed to cross-encoder
ADDITIVE_VECTOR_SHARE	0.5	Vector share in additive treatment

Table 9: Token budget configuration parameters

B.4 Paragraph-Based Chunking Algorithm

Documents are chunked using paragraph-number detection rather than fixed token windows. The algorithm identifies explicitly numbered paragraphs matching the regular expression:

```
pattern = re.compile(r'^\s*(\d+)\.\s+', re.MULTILINE)
```

Consecutive numbered paragraphs are grouped into non-overlapping blocks of three:

```

GROUP_SIZE = 3

paragraphs = []
current_number = None
current_lines = []

for line in document:
    match = pattern.match(line)
    if match:
        if current_number is not None:
            paragraphs.append({
                "number": current_number,
                "text": "\n".join(current_lines)
            })
            current_number = int(match.group(1))
            current_lines = [line]
        else:
            current_lines.append(line)

# Group into blocks of 3
chunks = []
for i in range(0, len(paragraphs), GROUP_SIZE):
    group = paragraphs[i:i+GROUP_SIZE]
    chunk_text = "\n\n".join(p["text"] for p in group)
    chunk_label = f"{group[0]['number']}-{group[-1]['number']}"
    chunks.append({
        "content": chunk_text,
        "paragraph_number": chunk_label,
        "paragraph_numbers": [p["number"] for p in group]
    })

```

Each chunk receives metadata including `paragraph_numbers` (list of integers), `paragraph_number` (range label, e.g., “1-3”), and `chunk_type` (“paragraph_group”).

B.5 Hybrid Retrieval Pipeline

The *Vector RAG* treatment employs a four-stage retrieval pipeline implementing Reciprocal Rank Fusion (RRF):

Stage 1: Dense Retrieval. Query embedding (OpenAI text-embedding or all-MiniLM-L6-v2) compared against chunk embeddings using cosine similarity; top 60 candidates returned.

Stage 2: Sparse Retrieval. BM25 scoring over tokenized chunks using the LangChain BM25Retriever; top 60 candidates returned.

Stage 3: RRF Combination. For each document d appearing in either result set:

$$\text{RRF}(d) = \sum_{r \in \{\text{dense}, \text{sparse}\}} \frac{\mathbf{1}[d \in r]}{k + \text{rank}_r(d)} \quad (3)$$

where $k = 60$ is a smoothing constant that prevents excessive weight on top-ranked documents.

Stage 4: Cross-Encoder Reranking. Top 30 documents by RRF score are rescored using cross-encoder/ms-marco-MiniLM-L-6-v2; final top 20 selected subject to the 12,000-token budget. If the budget is exceeded, the final chunk is truncated.

B.6 Knowledge Graph Schema

The Neo4j knowledge graph implements the following schema:

B.6.1 Node Types

- Stage: Main analytical phases (4 nodes)
 - Applicability of Article 11
 - Interference
 - Justification
 - Consequences
- DoctrinalElement: Legal tests and criteria (18 nodes)
 - Substantive Scope, Personal Scope
 - Existence of Interference, Positive Obligations
 - Prescribed by Law, Legitimate Aim, Necessary in a Democratic Society
 - Accessibility, Foreseeability, Safeguards against Abuse
 - National Security, Public Safety, Prevention of Disorder, Prevention of Crime, Protection of Health, Protection of Morals, Rights of Others
 - Just Satisfaction

- `DiagnosticQuestion`: Specific legal questions for each element
- `CaseRuling`: ECtHR decisions extracted from Reference Guide
- `LegalQuestion`: Legal issues identified in documents
- `FactualScenario`: Illustrative examples from case law
- `LegalFinding`: Extracted legal arguments with citations
- `SourceFile`: Reference Guide source documents (27 files)

B.6.2 Relationship Types

- `CONTAINS_ELEMENT`: `Stage` → `DoctrinalElement`
- `ASSESSED_BY`: `DoctrinalElement` → `DiagnosticQuestion`
- `NEXT_STAGE`: `Stage` → `Stage` (sequential analytical flow)
- `HAS_FINDING`: `DoctrinalElement` → `LegalFinding`
- `CITES_CASE`: `LegalFinding` → `CaseRuling` (with relevance property)
- `REFERENCES_CASE`: `DoctrinalElement` → `CaseRuling`
- `ASKS`: `DoctrinalElement` → `LegalQuestion`
- `HAS_EXAMPLE`: `DoctrinalElement` → `FactualScenario`
- `APPEARS_IN`: `Entity` → `SourceFile`

B.7 Graph Traversal Query

Entry point traversal uses variable-depth path matching in Cypher:

```
UNWIND $start_nodes AS start_node_name
MATCH (start_node) WHERE start_node.name = start_node_name

CALL {
  WITH start_node
  OPTIONAL MATCH path = (start_node)-[*1..5]-(content)
  WHERE content:LegalQuestion
    OR content:CaseRuling
    OR content:FactualScenario
    OR content:LegalFinding
```

```

WITH DISTINCT content,
      length(path) as depth,
      [r IN relationships(path) | type(r)][-1] as rel_type
ORDER BY depth
LIMIT 600
RETURN content, rel_type, depth
}

```

```

RETURN start_node.name AS start_point,
      labels(content)[0] AS content_type,
      properties(content) AS properties,
      depth,
      rel_type
ORDER BY depth, content_type

```

Retrieved content is organized by depth: nodes at depth 1–2 are classified as primary authority; nodes at depth 3–5 as supporting authority.

B.8 Document-to-Node Mapping

Each of the 27 Reference Guide documents maps to specific doctrinal nodes:

```

DOCUMENT_NODE_MAPPING = {
  "A.Importance of the right to freedom of peaceful assembly.txt":
    ["Substantive Scope"],
  "B.Classification of complaints under Articles 9, 10 and or 11.txt":
    ["Substantive Scope"],
  "C.Scope of the right to freedom of assembly.txt":
    ["Substantive Scope"],
  "C.1.Form and type of assembly.txt":
    ["Substantive Scope"],
  "C.2.Freedom of forum.txt":
    ["Substantive Scope"],
  "C.3.Peaceful assembly.txt":
    ["Substantive Scope"],
  "D.Positive obligations.txt":
    ["Positive Obligations"],
  "D.1.Obligation to ensure the peaceful conduct of an assembly.txt":
    ["Positive Obligations"],
  "D.2.Counter-demonstrations.txt":
    ["Positive Obligations"],
}

```

```

"E.Restrictions on the right to freedom of assembly.txt":
    ["Existence of Interference", "Necessary in a Democratic Society"],
"E.1.Interference with the exercise of the right to freedom of assembly.txt":
    ["Existence of Interference"],
"E.2.Justification of restrictions.txt":
    ["Accessibility", "Foreseeability", "Safeguards against Abuse",
    "Legitimate Aim", "Necessary in a Democratic Society"],
"E.2.a.Prescribed by law.txt":
    ["Accessibility", "Foreseeability", "Safeguards against Abuse"],
"E.2.b.Legitimate aim.txt":
    ["Legitimate Aim"],
"E.2.c.Necessary in a democratic society.txt":
    ["Necessary in a Democratic Society"],
"E.2.c.i.Narrow margin for content-based interference.txt":
    ["Necessary in a Democratic Society"],
"E.2.c.ii.Narrow margin for general ban.txt":
    ["Necessary in a Democratic Society"],
"E.2.c.iii.Wider margin for disruption sanctions.txt":
    ["Necessary in a Democratic Society"],
"E.2.c.iv.Chilling effect.txt":
    ["Necessary in a Democratic Society"],
"E.2.c.v.Sanctions - nature and severity.txt":
    ["Necessary in a Democratic Society"],
"E.2.c.vi.Dispersal and the use of force.txt":
    ["Necessary in a Democratic Society"],
"F.Prior notification and authorisation procedures.txt":
    ["Existence of Interference"],
"F.1.Aim of notification and authorisation procedures.txt":
    ["Existence of Interference"],
"F.2.Unlawful assembly.txt":
    ["Existence of Interference"],
"F.3.Spontaneous assembly.txt":
    ["Existence of Interference"],
"G.Reprehensible conduct.txt":
    ["Necessary in a Democratic Society"]

```

```

}

```

C Prompt Templates

C.1 Common Prompt Components

All treatments share the following prompt components:

C.1.1 Role Definition

You ARE the European Court of Human Rights (ECtHR). Your task is to author a judgment based on your general legal knowledge and the following case description. Please decide whether freedom of assembly, as protected by Art. 11 ECHR, has been violated, and explain the reasoning that supports your decision.

C.1.2 Critical Instructions

****CRITICAL INSTRUCTIONS:****

- Adopt the formal, authoritative, and declarative voice of the Court.
- ****NEVER**** refer to "The Court" as an external entity. Instead, state findings directly.
- Treat the case description as the complete record.

C.1.3 Output Format

Provide your answer strictly in the following format:

Conclusion (regarding Art. 11 ECHR): [Violation / Non Violation]

Legal Reasoning (regarding Art. 11 ECHR): [Your formal judgment.]

C.2 Zero Context Treatment Prompt

You ARE the European Court of Human Rights (ECtHR). Your task is to author a judgment based on your general legal knowledge and the following case description.

Please decide whether freedom of assembly, as protected by Art. 11 ECHR, has been violated, and explain the reasoning that supports your decision.

****CRITICAL INSTRUCTIONS:****

- Adopt the formal, authoritative, and declarative voice of the Court.
- ****NEVER**** refer to "The Court" as an external entity. Instead, state findings directly.

- Treat the case description as the complete record.

Case Description: "{case_description}"

Provide your answer strictly in the following format:

Conclusion (regarding Art. 11 ECHR): [Violation / Non Violation]

Legal Reasoning (regarding Art. 11 ECHR): [Your formal judgment.]

C.3 Doctrinal Context Treatment Prompt

You ARE the European Court of Human Rights (ECtHR). Your task is to author a judgment based on your general legal knowledge, the following case description and the following legal background.

Please decide whether freedom of assembly, as protected by Art. 11 ECHR, has been violated, and explain the reasoning that supports your decision.

****CRITICAL INSTRUCTIONS:****

- Adopt the formal, authoritative, and declarative voice of the Court.
- ****NEVER**** refer to "The Court" as an external entity. Instead, state findings directly.
- Treat the case description as the complete record.

Case Description: "{case_description}"

Legal Background: To facilitate your task, in this section you are informed about a) the doctrinal structure of Art. 11 ECHR and b) the key content of each doctrinal element:

Structure of the doctrine of fundamental freedoms

The court proceeds in five steps:

- I. Scope
- II. Interference
- III. Restrictions
- IV. Positive Obligation
- V. Just Satisfaction

- I. Scope

The scope of the provision refers to the aspect of reality that is protected from state interference by Article 11 of the European Convention on Human Rights. The scope has a substantive and a personal dimension. The substantive scope determines which conduct is protected. The personal scope of protection defines who is the holder of the fundamental freedom.

II. Interference

The challenged state act interferes with the fundamental freedom if it makes the conduct protected by the fundamental right impossible or significantly more difficult. This may include a chilling effect.

III. Restrictions

If the challenged state act interferes with the fundamental freedom, its justification depends on whether the formal and substantive limits of the fundamental freedom are observed.

1. prescribed by law

The formal limits require that the interference is prescribed by law:

- a) accessibility
- b) foreseeability
- c) sufficient safeguards against abuse

2. legitimate aim

The substantive limits require that the interference serves one of the aims listed in alinea 2 of Article 11:

- a) national security
- b) public safety
- c) prevention of disorder
- d) prevention of crime
- e) protection of health
- f) protection of morals
- g) protection of the rights and freedoms of others

3. necessary in a democratic society

The state act must, on balance, be necessary in a democratic society. The state enjoys a certain margin of appreciation.

IV. Positive obligation

The member state may be required to proactively intervene on behalf of those meant to enjoy their freedom of assembly.

V. Just satisfaction

If the Court finds a violation, it may grant just satisfaction under Art. 41 of the Convention.

In preparing your response, please consider each of the following questions. If you come to the conclusion that the respective question is not relevant, just skip it.

[16 diagnostic questions follow, numbered 1-16, covering:]

1. substantive scope
2. personal scope
3. interference
4. prescribed by law: accessibility
5. prescribed by law: foreseeability
6. prescribed by law: safeguards against abuse
- 7-13. legitimate aims (national security through rights of others)
14. necessary in a democratic society
15. positive obligation
16. just satisfaction

Provide your answer strictly in the following format:

Conclusion (regarding Art. 11 ECHR): [Violation / Non Violation]

Legal Reasoning (regarding Art. 11 ECHR): [Your formal judgment.]

C.4 Vector RAG Treatment Prompt

You ARE the European Court of Human Rights (ECtHR). Your task is to author a judgment based on your general legal knowledge, the following case description and the following retrieved context.

Please decide whether freedom of assembly, as protected by Art. 11 ECHR, has been violated, and explain the reasoning that supports your decision.

****CRITICAL INSTRUCTIONS:****

- Adopt the formal, authoritative, and declarative voice of the Court.
- ****NEVER**** refer to "The Court" as an external entity. Instead, state findings directly.
- Treat the case description as the complete record.

Case Description: "{case_description}"

****Retrieved Context** ({context_tokens} tokens, {num_chunks_retrieved} chunks via {retrieval_method} search):******

{retrieved_context}

Provide your answer strictly in the following format:

Conclusion (regarding Art. 11 ECHR): [Violation / Non Violation]

Legal Reasoning (regarding Art. 11 ECHR): [Your formal judgment.]

C.5 Graph RAG Treatment Prompts

C.5.1 Entry Point Detection Prompt

You are a senior legal analyst at the European Court of Human Rights (ECtHR).

Your task is to identify the 3-7 most relevant doctrinal starting points for analyzing a legal case under Article 11 ECHR (freedom of assembly).

****CASE DESCRIPTION TO ANALYZE:****

{case_description}

****ALLOWED NODE NAMES (YOU MUST ONLY USE THESE EXACT NAMES):****

[

"Applicability of Article 11",

"Interference",

"Justification",

"Consequences",

"Substantive Scope",

"Personal Scope",

"Existence of Interference",

"Positive Obligations",

"Prescribed by Law",

```
"Legitimate Aim",
"Necessary in a Democratic Society",
"Just Satisfaction",
"Accessibility",
"Foreseeability",
"Safeguards against Abuse",
"National Security",
"Public Safety",
"Prevention of Disorder",
"Prevention of Crime",
"Protection of Health",
"Protection of Morals",
"Rights of Others"
]
```

****GRAPH STRUCTURE INFORMATION:****

The knowledge graph follows this hierarchy:

- ****Stages**** (main analysis phases): Applicability of Article 11, Interference, Justification, Consequences
- ****Checkpoints**** (doctrinal tests): Substantive Scope, Personal Scope, Existence of Interference, Positive Obligations, Prescribed by Law, Legitimate Aim, Necessary in a Democratic Society, Just Satisfaction
- ****Subpoints**** (detailed criteria): Accessibility, Foreseeability, Safeguards against Abuse, National Security, Public Safety, Prevention of Disorder, Prevention of Crime, Protection of Health, Protection of Morals, Rights of Others

****RULES:****

1. You **MUST ONLY** select from the **ALLOWED NODE NAMES** list above
2. Do **NOT** create new names or variations
3. Select 3-7 most relevant nodes for this case
4. Consider nodes from different hierarchy levels for comprehensive analysis

Provide your answer as a JSON array of strings.

Example: ["Applicability of Article 11", "Existence of Interference", "Prescribed by Law", "Necessary in a Democratic Society"]

C.5.2 Final Judgment Prompt

You **ARE** the European Court of Human Rights (ECtHR). Your task is to

author a judgment based on your general legal knowledge, the following case description and the following retrieved context.

Please decide whether freedom of assembly, as protected by Art. 11 ECHR, has been violated, and explain the reasoning that supports your decision.

****CRITICAL INSTRUCTIONS:****

- Adopt the formal, authoritative, and declarative voice of the Court.
- ****NEVER**** refer to "The Court" as an external entity. Instead, state findings directly.
- Treat the case description as the complete record.

****CONTEXT INSTRUCTIONS:****

- Use **ONLY** the graph context provided below.
- Apply the legal framework and precedents from the graph to analyze the case.

****CASE DESCRIPTION:****

{case_description}

****GRAPH CONTEXT (filtered & enriched):****

{graph_context_json}

Provide your answer strictly in the following format:

Conclusion (regarding Art. 11 ECHR): [Violation / Non Violation]

Legal Reasoning (regarding Art. 11 ECHR): [Your formal judgment.]

C.6 Vector-Graph RAG-additive Treatment Prompt

You ARE the European Court of Human Rights (ECtHR). Your task is to author a judgment based on your general legal knowledge, the following case description and the following retrieved context.

Please decide whether freedom of assembly, as protected by Art. 11 ECHR, has been violated, and explain the reasoning that supports your decision.

****CRITICAL INSTRUCTIONS:****

- Adopt the formal, authoritative, and declarative voice of the Court.
- ****NEVER**** refer to "The Court" as an external entity. Instead,

state findings directly.
- Treat the case description as the complete record.

****CONTEXT INSTRUCTIONS:****

- You have access to TWO separate context sources below.
- Synthesize insights from BOTH sources to form your legal reasoning.
- The vector context provides textual evidence; the graph context provides legal structure.

****CASE DESCRIPTION:****

{case_description}

*****== VECTOR CONTEXT (Document Retrieval) ***==**

{vector_context_text}

*****== GRAPH CONTEXT (Knowledge Graph - filtered & enriched) ***==**

{graph_context_text}

Provide your answer strictly in the following format:

Conclusion (regarding Art. 11 ECHR): [Violation / Non Violation]

Legal Reasoning (regarding Art. 11 ECHR): [Your formal judgment.]

C.7 Vector-Graph RAG-fusion Treatment Prompts

The fusion treatment employs three specialized prompts executed sequentially using GPT-5 with high reasoning effort.

C.7.1 Step 1: Legal Architect Prompt

You are a LEGAL ARCHITECT specializing in ECtHR Article 11 jurisprudence. Your expertise lies in constructing systematic legal examination frameworks. You think in doctrinal structures, hierarchical analysis, and logical sequences.

IMPORTANT: You do NOT make legal assessments or predictions. You only BUILD STRUCTURE.

YOUR TASK:

Construct a comprehensive EXAMINATION SCHEMA for an Article 11 ECHR case. You are building the STRUCTURE for legal analysis - NOT making assessments.

CASE TO ANALYZE:

{case_description}

GRAPH CONTEXT (Knowledge Graph Data):

{graph_context_json}

CRITICAL RULES:

- You must NOT include any legal assessments, predictions, or conclusions
- You ONLY provide: structure, questions, definitions, precedent mappings
- For each element, state WHAT must be examined, not HOW it should be decided

REQUIREMENTS:

- Create the COMPLETE examination structure (all relevant stages and elements)
- For EACH element, formulate the SPECIFIC LEGAL QUESTIONS that arise in THIS case
- Assign RELEVANT PRECEDENTS from the graph context to each element
- Include DEFINITIONS for key doctrinal concepts
- Note SYSTEMATIC RELATIONSHIPS between elements

OUTPUT FORMAT (JSON):

```
{
  "case_type": "Brief characterization",
  "examination_stages": [
    {
      "stage": "Stage name",
      "stage_number": 1,
      "purpose": "What this stage examines",
      "elements": [
        {
          "element": "Element name",
          "definition": "Legal definition/standard from case-law",
          "questions_for_this_case": [
            {
              "question": "Specific legal question arising here",
              "why_arises": "What facts trigger this question"
            }
          ],
          "relevant_precedents": [
            {
              "case": "Case name",
```

```

        "paragraphs": ["§XX", "§YY"],
        "relevance": "What aspect of law it addresses"
    }
],
    "sub_elements": []
}
],
    "sequence_note": "When to examine / what must be established first"
}
],
    "key_legal_questions": [...],
    "examination_flow": "Stage1 → Stage2 → ..."
}

```

Respond with valid JSON only.

C.7.2 Step 2: Evidence Curator Prompt

You are an EVIDENCE CURATOR specializing in ECtHR case law. Your expertise lies in identifying, categorizing, and linking legal evidence to doctrinal points. You excel at extracting precise paragraph references and canonical formulations.

IMPORTANT: You do NOT make legal assessments. You only ORGANIZE EVIDENCE.

YOUR TASK:

Populate the examination schema with EVIDENCE from the vector context.

You are ORGANIZING EVIDENCE - NOT making legal assessments.

CASE:

{case_description}

EXAMINATION SCHEMA TO POPULATE:

{examination_schema_json}

VECTOR CONTEXT (Document Chunks - {num_chunks} chunks):

{vector_chunks_json}

CRITICAL RULES:

1. You must NOT include any legal assessments or conclusions
2. You ONLY: find evidence, extract citations, identify canonical language
3. For each piece of evidence, state WHAT it says, not WHAT it means

for the case

REQUIREMENTS:

1. For EACH element in the schema, search vector chunks for relevant passages
2. Extract PARAGRAPH REFERENCES (§§) where cases are cited
3. Identify CANONICAL FORMULATIONS (exact phrases from case-law)
4. Note the SOURCE of each piece of evidence
5. Flag elements where NO relevant evidence was found

OUTPUT FORMAT (JSON):

```
{
  "populated_elements": [
    {
      "element": "Element name",
      "stage": "Parent stage",
      "questions": ["Question 1", "Question 2"],
      "evidence_passages": [
        {
          "text": "Exact quote from vector context",
          "source": "Source file name",
          "chunk_id": 0
        }
      ],
      "precedents_found": [
        {
          "case": "Full case name",
          "citation": "no. XXXXX/XX, YEAR",
          "paragraphs": ["§91", "§155"],
          "holdings_quoted": [
            {"paragraph": "§91", "text": "Exact holding text"}
          ]
        }
      ],
      "canonical_formulations": [
        {
          "phrase": "Exact wording in quotes",
          "source": "Case §para"
        }
      ],
      "evidence_found": true
    }
  ]
}
```

```

    }
  ],
  "elements_without_evidence": ["Element names with gaps"],
  "all_cases_cited": [...]
}

```

Respond with valid JSON only.

C.7.3 Step 3: Case Strategist Prompt

You are a CASE STRATEGIST specializing in ECtHR litigation.
Your expertise lies in identifying decisive issues and weighting analytical priorities.

IMPORTANT: You identify WHAT is important, but you do NOT assess HOW it should be decided. You prepare the framework; the judge decides the outcome.

YOUR TASK:

Create a WEIGHTED SOLUTION OUTLINE that identifies priorities and decisive issues. You determine WHAT is important - the judge decides HOW to resolve it.

CASE:

{case_description}

EXAMINATION SCHEMA:

{examination_schema_json}

POPULATED EVIDENCE:

{populated_evidence_json}

CRITICAL RULES:

1. You must NOT include legal assessments, predictions, or conclusions
2. You must NOT use "Assessment:" sections
3. You ONLY: weight importance, identify decisive questions, map citations
4. Use (critical), (important), (routine) for weighting

REQUIREMENTS:

1. WEIGHT each element by importance for THIS case (without predicting outcome)
2. Identify the DECISIVE QUESTIONS that will determine the outcome

3. Map which PRECEDENTS apply to which questions
4. Create a TEXT OUTLINE that the judge can use as a roadmap

OUTPUT:

First, provide JSON:

```
{
  "weighted_elements": [
    {
      "element": "Element name",
      "weight": "critical/important/routine",
      "weight_reason": "Why this weight (based on facts, not prediction)",
      "key_questions": ["Question 1", "Question 2"],
      "applicable_precedents": ["Case1 §X", "Case2 §Y"]
    }
  ],
  "decisive_questions": [
    {
      "question": "The legal question to be answered",
      "element": "Where located",
      "why_decisive": "Why this question matters (without predicting answer)",
      "key_precedents": ["Case §para"]
    }
  ],
  "citations_by_element": {
    "Element name": ["Case1 §X", "Case2 §Y"]
  }
}
```

Then, provide a TEXT-BASED SOLUTION OUTLINE:

ARTICLE 11 ECHR - SOLUTION OUTLINE

I. CASE OVERVIEW

[Brief factual summary relevant to legal analysis]

II. EXAMINATION STRUCTURE

[The stages to examine: Applicability → Interference → Justification]

III. DETAILED ANALYSIS

A. [STAGE NAME] //

[Element Name]

- Legal Question: [Specific question for this case]
- Legal Standard: [Definition/test from case-law]
- Key Precedent: [Case name, §§]
"[Exact holding quoted]"
- Evidence: [What evidence is available - NOT what it proves]

IV. DECISIVE QUESTIONS

[List the questions that will determine the outcome - without answering them]

V. KEY CITATIONS

[Precedents with §§ that must appear in the judgment]

VI. CANONICAL FORMULATIONS

[Exact phrases from case-law to incorporate]

Provide JSON first, then the text outline.

C.7.4 Final Judgment Prompt (Fusion)

You ARE the European Court of Human Rights (ECtHR). Your task is to author a judgment based on your general legal knowledge, the following case description and the solution outline prepared for this case.

Please decide whether freedom of assembly, as protected by Art. 11 ECHR, has been violated, and explain the reasoning that supports your decision.

****CRITICAL INSTRUCTIONS:****

- Adopt the formal, authoritative, and declarative voice of the Court.
- ****NEVER**** refer to "The Court" as an external entity. Instead, state findings directly.
- Treat the case description as the complete record.

UNDERSTANDING YOUR CONTEXT: THE SOLUTION OUTLINE

You have been provided with a SOLUTION OUTLINE - a structured analytical framework specifically prepared for this case. This outline was created through a systematic three-step process by specialized legal analysts.

IMPORTANT: The outline contains NO legal assessments or predictions.

It provides:

- The STRUCTURE for your analysis
- The QUESTIONS you must answer
- The PRECEDENTS you must consider
- The EVIDENCE available

YOU must make ALL legal determinations. The outline is your roadmap; you are the judge who decides the destination.

HOW TO USE THIS OUTLINE:

1. FOLLOW THE STRUCTURE - Move through the examination stages in order:
Applicability → Interference → Justification
2. ANSWER EVERY QUESTION - Each legal question in the outline must receive an explicit answer in your judgment.
3. APPLY THE PRECEDENTS - For each question, consider the precedents listed. State their holdings and explain whether they apply.
Cite cases with paragraph references: "Kudrevičius v. Lithuania [GC], §157"
4. USE CANONICAL LANGUAGE - Incorporate the established formulations from case-law where they apply.
5. RESPECT THE WEIGHTS - Elements marked (critical) require detailed analysis. Elements marked (routine) can be addressed briefly.

{critical_elements_summary}

{decisive_questions_summary}
6. TRANSFORM INTO JUDGMENT - The outline is analytical skeleton; your judgment is living legal reasoning.

CASE DESCRIPTION:
{case_description}

SOLUTION OUTLINE

{solution_outline_text}

Provide your answer strictly in the following format:

Conclusion (regarding Art. 11 ECHR): [Violation / Non Violation]

Legal Reasoning (regarding Art. 11 ECHR): [Your formal judgment.]

D Evaluation Framework Details

D.1 Composite Score Calculation

For *Zero Context* and *Doctrinal Context* treatments:

$$S_{\text{composite}} = S_{\text{answer}} \quad (4)$$

For RAG treatments (*Vector RAG*, *Graph RAG*, *Vector-Graph RAG-additive*, *Vector-Graph RAG-fusion*):

$$S_{\text{composite}} = 0.75 \times S_{\text{answer}} + 0.25 \times S_{\text{context}} \quad (5)$$

D.2 Answer Score Components

$$S_{\text{answer}} = 0.85 \times S_{\text{judge}} + 0.12 \times S_{\text{RAGAS-answer}} + 0.03 \times S_{\text{BLEU}} \quad (6)$$

Where:

- S_{judge} = Weighted average of eight LLM-judge criteria (normalized 0–1)
- $S_{\text{RAGAS-answer}}$ = RAGAS answer metrics (relevance, similarity, correctness)
- S_{BLEU} = BLEU score against reference judgment

Criterion	Weight
Legal Accuracy	10%
Analytical Depth	16%
Precedent Quality	16%
Factual Grounding	7%
Methodological Rigor	13%
Source Transparency	9%
Argumentative Coherence	9%
Professional Presentation	5%
Total	85%

Table 10: Judge criteria weights (summing to 85% of answer score)

D.3 LLM-as-Judge Criteria Weights

D.4 RAGAS Metrics

D.4.1 Answer-Level Metrics (Applied to All Treatments)

- **Answer Relevance (5%)**: Measures how relevant the generated answer is to the question
- **Answer Similarity (5%)**: Semantic similarity between generated and reference answers
- **Answer Correctness (2%)**: Factual accuracy of the generated answer

D.4.2 Context-Level Metrics (Applied to RAG Treatments Only)

Metric	Weight (of context score)
Faithfulness	25%
Context Relevance	15%
Context Precision	15%
Context Recall	15%
Context Utilization	15%
Context Entity Recall	15%
Total	100%

Table 11: RAGAS context metrics and weights

D.5 Evaluation Pipeline Structure

Each case-model-treatment combination undergoes three independent evaluation runs. With six available judge models, this produces up to 18 observations per case-method pair (6 judges $\times$ 3 runs). The output preserves individual evaluator rows with the following structure:

```
case_id, method, evaluator, run_id,
judge_legal_accuracy, judge_analytical_depth, ...,
```

```
ragas_answer_relevancy, ragas_faithfulness, ...,  
bleu_score, composite_score
```