MAX PLANCK INSTITUTE
FOR RESEARCH ON
COLLECTIVE GOODS

CHRISTOPH ENGEL
JOHANNES KRUSE

Discussion Paper
2025/14

LLM AS A LAW PROFESSOR: HAVING A LARGE LANGUAGE MODEL WRITE A COMMENTARY ON FREEDOM OF ASSEMBLY

LLM as a Law Professor: Having a Large Language Model Write a Commentary on Freedom of Assembly

Christoph Engel* **/ Johannes Kruse**

Abstract

In many jurisdictions, academia is at the service of legal practice. Law professors write commentaries that summarize the state of the art of doctrine, chiefly of jurisprudence. In the spirit of a proof of concept, using the guarantee of freedom of assembly in the European Convention on Human Rights, we show that this task can be completely outsourced to large language models. Using standard NLP metrics and an LLM as a judge approach, we develop an evaluation pipeline that works without costly human annotation. The commentaries fully written by GPT 4o, Gemini 2.5 flash or Kimi K2 Instruct are on par with their best human written competitor, the Guide provided by the Court itself.

* corresponding author: Prof. Dr. Christoph Engel, Max Planck Institute for Research on Collective Goods, Kurt-Schumacher-Straße 10, D 53113 Bonn, Germany, email: engel@coll.mpg.de

1 Introduction

In computer science, the "LLM-as-a-judge" approach is popular (Zheng, Chiang et al. 2023, Gu, Jiang et al. 2024). From a legal perspective it is a misnomer. The term does not stand for robo judges (on the involvement of LLMs in judicial decision making see e.g. Liu and Li 2024). It means the automation of benchmarking. Rather than having a set of human raters evaluate the performance of an LLM, evaluation is performed by an LLM as well. The term makes sense as LLMs possess the ability to pass judgement. They can come up with meaningful assessments even if the evaluation cannot be derived from first principles by mere operations of logic. But the standard "LLM-as-a-judge" application only calls on the generative ability of LLMs in a very limited way: the LLM generates a score. In this paper, we introduce a pipeline that capitalizes more profoundly on the generative abilities of LLMs. To mark the difference, we call it "LLM-as-a-law-professor".

We hasten to stress that we are not proposing a "robo law professor". For the foreseeable future, the evolution of the law as an academic discipline will remain in the hands of (human) legal scholars. Experienced academics may see connections to analogous literatures that, at least for the time being, would be harder for a language model to detect. Academics may also amalgamate summarization, analysis and categorization with suggestions for legal evolution. These additional functions of commentaries invite a division of labor. The large language model does the tedious part, i.e. structured review of case law. The professor focuses on the finish.

Three LLM-written commentaries on freedom of assembly. In this paper we show that a fairly involved task that hitherto had typically been performed by law professors can be delegated to machines: writing a commentary on a statutory provision. This task is regarded as highly valuable by the legal community: the "Guide" prepared by the Registry of the Court of Human Rights (European Court of Human Rights 2024) is also confined to the structured summary and analysis of the court's jurisprudence, i.e. to the task we show LLMs can perform. In a computer science perspective, the task can at first glance be characterized as summarization (see below Section 4). Yet in this application, LLMs are not just summarizing a single text. They are writing a structured report of the complete jurisprudence of the competent court(s) on one statutory provision, and organize it along the elements of doctrine that have emerged from its application.

For a proof of concept, we have chosen a provision of the European Convention of Human Rights. Art. 11 ECHR protects freedom of assembly. The provision is appealing for a set of reasons. While the Court does not publish all decisions, a fairly large amount of its rulings is digitized. On the database of the European Court of Human Rights, 1198 cases are posted that discuss Art. 11 ECHR.[1] The database is open source, so that we do not face copyright issues.[2] Originally, the format of a commentary has been developed in the German language jurisdictions. This paper addresses an international audience. As the European Convention on Human Rights also applies to Germany, the format has travelled. There is a number of

[1] We have scraped "judgements" and "decisions".

[2] For detail see https://www.echr.coe.int/copyright-and-disclaimer#:~:text=Persons%20wishing%20to%20use%20information,form%20available%20on%20the%20website.

commentaries on the provision that have been written by academics. We could compare the LLM commentary to these human written competitors. if there were no copyright barriers. Precisely for this reason, it is particularly fortunate that there is even a quasi-authoritative, commentary-like document, the "Guide" on the provision, prepared by the Registry of the Court (European Court of Human Rights 2024), which is open source. This gives us a benchmark for assessing the quality of our machine-written commentary. We use this application for a proof of concept: the structured summary of jurisprudence, in the format of a commentary, can be delegated to LLMs. The practical relevance of this proof will be highest where no such commentary exists yet.

Large language models develop very rapidly. We have had GPT-4o write the original version of the commentary. From the vantage point of today, that model suffers from a number of limitations, the most important being the inability to simultaneously handle larger amounts of text. In this second version of the project, we have added two up to date models, namely Gemini 2.5 flash and Kimi K2. The latter model is open source, so that we can also check the capability of models that are not proprietary. The open-source nature is particularly valuable for replicability. The results of our exercise are available on three separate websites:

http://professor-gpt.coll.mpg.de/html/overview.html
http://professor-gemini.coll.mpg.de/html/overview.html
http://professor-kimi.coll.mpg.de/html/overview.html

While professor-authored commentaries often offer ancillary services beyond our scope (e.g., practice guidance, legislative history, or comparative perspectives), our focus is the commentary's core function: the structured synthesis of legal reasoning. This consists in evaluating, condensing, and systematizing the case law within an established doctrinal framework - i.e., presenting jurisprudence in a form that makes the underlying doctrine explicit and navigable. This is not only the most laborious part. It is also why commentaries are so widely used in legal practice. Typically practitioners do not have time to, themselves, read through a rich and multifaceted body of jurisprudence. Commentaries are so popular precisely because they give practitioners easy and reliable access to the state of the art. It is this function that a commentary written by a large language model is able to fulfil.

Automating the evaluation of (LLM-written) commentaries. The commentary written by either LLM looks professional. The commentaries do not only inform readers about key elements of the doctrine of Art. 11 ECHR. All statements come with references to the paragraphs of rulings from which the commentary takes the information. These references are links, so that a skeptical reader can check herself. A second link takes the reader to the complete ruling posted on the Court's own website. This functionality is also valuable if a legal practitioner searches for cases that are closely related to her own, or that can serve as a source of inspiration for developing her own legal argument. To the eyes of the authors, who are both trained lawyers, the commentaries look convincing, and no different from human-written commentaries. With sufficient prompting, LLMs are not only able to summarize and analyze a larger body of jurisprudence. They do so in exactly the format to which practicing lawyers are used. But we do not want to stop at subjective impressions, and have tried to validate the quality of the commentaries objectively.

Now to the best of our knowledge, there is no established benchmark for the validation of software used to summarize the doctrine of a statutory provision in the light of a body of jurisprudence. This is why we develop our own. We combine an LLM as a judge approach (Zheng, Chiang et al. 2023, Gu, Jiang et al. 2024) with established NLP metrics, like ROUGE (Lin 2004) and BLEU (Reiter 2018). To avoid circularity, we feed another frontier model, Claude 4 Sonnet, each of the three LLM written commentaries as a RAG (retrieval augmented generation). We compare outcomes with the ones we get when, instead, not giving the LLM any information beyond its general training, or when feeding it the Guide prepared by the Court itself as a RAG. We have two sets of benchmarks. In the first set, we ask our LLM as a judge to predict the decision of cases that hinge on Art. 11 ECHR. We compare these predictions with groundtruth. In the second set, we assess the quality of legal argument, and compare it with the same measures for the human-written Guide.

For the prediction task, we use three sets of cases. A first set of test cases has been decided by the European Court of Human Rights. This set of cases has the highest external validity, but we cannot exclude that Claude has seen these cases at general training. This is why we add two more sets of cases. The second set of test cases exploits the fact that the German Constitution features a closely related fundamental freedom. We ask Gemini to predict the disposition of these cases, had they been decided under Art. 11 ECHR. A third set of test cases is fictitious, so that memorization is impossible.

Claude is fairly good at predicting (postdicting) decisions that the European Court of Human Rights has actually taken (accuracy is 88%). But Claude achieves this high performance even before given access to either the Guide or to one of the LLM-written commentaries. The respective RAG never improves performance. This suggests that memorization is indeed a serious concern. For the cases that have actually been decided by the German Constitutional Court, the LLM outcome predictor would have to think two steps ahead: from one language to another, and from one constitutional guarantee to another. Results suggest that this is not happening. Now zero-shot performance is much poorer. The machine-written LLMs are even better than the Guide. But all of them are far from perfect. It seems that predicting the outcomes of unseen cases is a fairly hard task, even for a frontier LLM like Claude 4 Sonnet.

The main purpose of writing a commentary is not improving the performance of another LLM; the main purpose is helping human jurists. Human jurists regularly use the available human-written commentaries. In our evaluation benchmark we therefore also compare the predictions made with the help of the Guide to the one made with the help of one of the machine-written commentaries. In this comparison, all three LLM-written commentaries shine, with convergence rates of 80% and higher across the board.

Courts do not only decide. They also justify their decisions. Commentaries help the court with the justification, and the parties with preparing their argument. In a second suite of benchmarks, we comparatively assess the performance of the Guide and our three LLM-written commentaries in these respects. Specifically, we compare their content (with criteria like legal precision, doctrinal relevance, or perceived utility for the task of an attorney), their structure (with criteria like logic of the presentation and its granularity, or perceived ease of navigation) and references to the jurisprudence of the European Court of Human Rights (with criteria like accuracy and completeness, but also redundancy). For the evaluation of citations,

we also use metrics that do not require (LLM) judgement (like ROUGE, Jaccard and BLEU). In the dimensions content and structure, the LLM-written commentaries are on par, sometimes even better than the human-written Guide. The LLM-written commentaries outperform the Guide when it comes to the quality of case-law citations. The Guide gets better marks for citation (non-)redundancy. But this result is the flipside of what one might consider an advantage of an LLM-written commentary: rather than using professional experience to select prominent rulings, it gives human users structured access to the complete evidence.

Summing up, LLM-written commentaries are not perfect. There remains room for improvement. Given the very rapid development of technology, one has reason to be optimistic that such improvements will be possible. Yet as a proof of concept our prototype is promising. It is conceivable that one more academic task is delegated to machines.

Organization of the paper. The remainder of the paper is organized as follows: in Section 2, we introduce the topic of our commentary, i.e. the guarantee of freedom of assembly in the European Convention on Human Rights. Section 3 is chiefly meant for legal readers without extensive exposure to the computer science literature on large language models. It explains the capabilities of large language models that make them a promising tool for the task. Section 4 relates our paper to the literature, and defines our contribution. Section 5 explains in detail how we have proceeded. Section 6 assesses the performance of the commentaries that the three LLMs have written. Session 7 discusses limitations. Section 8 concludes.

2 Freedom of Assembly, as Protected by the European Convention on Human Rights

We have selected an area of law where (a) there is sufficiently rich jurisprudence to make summarization meaningful and (b) summarizations, in the form of a commentary, are available that have been written by professional lawyers. Now commentaries are not a standard tool in legal practice in either the US or the UK. On the other hand, the lingua franca of the international academic community is English. This has led us to an international instrument that produces its output (at least predominantly) in English, while being applicable and of practical relevance in the German law speaking jurisdictions. The latter feature is responsible for the availability of human written commentaries in English language. The Germanic tradition has been picked up by this international jurisdiction.

Specifically, we would, in principle, be able to compare the commentary written by GPT with the following two types of competitors: The first type consists of technical-functional competitors that do not address Article 11 of the European Convention on Human Rights (ECHR), but offer tools for summarizing court decisions. In this context, the offerings of major commercial providers are particularly noteworthy. Thomson Reuters (with Westlaw AI-Assisted Research and Ask Practical Law AI) and LexisNexis (with Lexis+AI) both provide platforms featuring LLM-supported tools capable of summarizing court decisions, among other functionalities (Nexis 2023, Reuters 2023). In Germany, for example, Wolters Kluwer offers GPT-based summaries of court decisions (Kluwer 2024). To the best of our knowledge and based on publicly available documentation, these offerings do not appear to include the generation of structured case reports; moreover, they remain largely behind paywalls

On the other hand, content-related competitors that also offer a structured summary of the case law on Article 11 ECHR provide an ideal starting point for a comparative evaluation. In addition to the official guide to Article 11 ECHR, we considered several commentaries and a handbook. With regard to the latter, we have followed the official list promulgated by the Court itself.[3] Of the commentaries listed there, we comparatively examined, to the extent permitted by copyright law, the two English commentaries, one handbook (Grabenwarter 2014, Schabas 2017, and Villiger 2022) and two of the four German-language commentaries (Karpenstein and Mayer 2022, Mayer-Ladewig, Nettesheim et al. 2023).

The Guide on Article 11 ECHR (European Court of Human Rights 2024) serves as a gold standard, representing an exceptionally expert human summary of case law. The Guide is prepared by the Registry of the European Court of Human Rights (ECHR), not by the judges themselves. However the Registry should be equally, if not more, informed about the Court's case law. This is because the Registry is responsible for providing the legal and administrative services required by the Court (see Rule 18 of the Rules of Court, 28 March 2024). Additionally, a unit within the Registry, Jurisconsult, is tasked with ensuring the quality and consistency of the Court's case law (see Rule 18B).

A further reason for selecting freedom of assembly as protected by the European Convention on Human Rights is a parallel exercise by the two authors of this paper. In a companion project, we have programmed GPT to write a commentary on freedom of assembly as protected by article 8 Basic Law, i.e. by the German constitution (see the companion paper Engel and Kruse 2024). This makes it possible to compare the performance of GPT across both jurisdictions. We in particular are in a position to identify additional challenges present in the jurisprudence of the European Court of Human Rights. In the original version of our LLM commentary, this limitation was pronounced, as the jurisprudence of the Court is so rich that we had to chunk the input. Happily, the more recent LLMs that we have added in this new version of the paper no longer suffer from this limitation.[4]

While the main reason for selecting the application is thus pragmatic, freedom of assembly is an academically interesting and practically relevant topic in its own right. The guarantee reads:

> 1. Everyone has the right to freedom of peaceful assembly and to freedom of association with others, including the right to form and to join trade unions for the protection of his interests.
> 2. No restrictions shall be placed on the exercise of these rights other than such as are prescribed by law and are necessary in a democratic society in the interests of national security or public safety, for the prevention of disorder or crime, for the protection of health or morals or for the protection of the rights and freedoms of others. This Article shall not prevent the imposition of lawful restrictions on the exercise of these rights by members of the armed forces, of the police or of the administration of the State.

Freedom of assembly is a fundamental right in a democratic society and, like the right to freedom of expression, one of the foundations of such a society (Salát 2015, Butler 2016,

[3] https://www.echr.coe.int/convention-collections.

[4] For detail see below Section 5.

Rights 2024). This right has been instrumental in nearly every major social movement throughout history (Inazu 2010) and remains vital in the information and internet age (Lewis 2006). Recently, the right to assemble has been central to discussions on significant social and global political conflicts, including the COVID-19 pandemic, the Black Lives Matter movement, and pro-Palestine protests. The pandemic, in particular, highlighted the challenging balance between the right to assemble and other protected interests, such as public health (Kruse and Langner 2021).

3 The Power of Large Language Models

The human mind is a black box, and so are large language models. Precisely what makes large language models so powerful also makes them opaque. Large language models no longer require fully determined if-then relations. They can handle the characteristic open texture of legal decision-making (Bix 1991, Schauer 2013). They do not shy away from ambiguity (Ellsberg 1961, Edelman 1992, Etner, Jeleva et al. 2012). They strive at making sense of the available input as best they can (Weick 1995, Turner, Allen et al. 2023). Still at a rather high level, it can be described how large language models work.

Language models make predictions. More precisely: they predict the next token, which typically is the next word. Given the text they have received so far: what is the most likely continuation? The user therefore controls language models with the input they provide. These are referred to as prompts. A prompt need not consist of a single sentence. The most advanced language models can process very long texts, even an entire book. Experience has shown that it is not only important to tell the language model as precisely as possible what it should do. A whole art of particularly skillful ways of asking the computer questions has developed, prompt engineering (Sahoo, Singh et al. 2024), including applications to law (Choi 2023).

Language models use machine learning. Machine learning organizes large amounts of data. New observations are either classified (top down) in decision trees or they are assigned (bottom up) to other data points that are as closely related as possible (for an excellent introduction see James, Witten et al. 2022). Neural networks are particularly sophisticated instruments for this task. Not only can they process a large number of dimensions, they can also place these dimensions in complex relationships to one another. They can have an architecture that allows preliminary assignments to be checked and gradually refined (for background see Goodfellow, Bengio et al. 2016). Transformers do not just translate inputs (e.g. natural language) into long chains of probabilities; computers can deal with such chains much more effectively. Rather, they provide the neural network with an attention mechanism (Vaswani, Shazeer et al. 2017). They use rich training data sets in this translation process. In this way, the local classification task is embedded in the “knowledge” that the architecture has previously acquired (Lin, Wang et al. 2022). Language models build on all these elements and add a generative component. The output no longer merely consists of an assignment of a data point to a class. Rather, the model can write texts (or generate images, or output sounds) (Chang, Wang et al. 2024).

The size of a language model refers to both the number of its parameters and the size of its training corpus. Large language models (LLMs) are models that contain billions of parameters and are trained with huge corpora that can be as large as the complete (freely accessible) Internet.[5] The training data also includes legal information. However, legal texts regularly make up only a fraction of the data (Colombo, Pires et al. 2024). Before we present a prototype of the commentary without an author and report on our experience with GPT as an annotator, we first outline the current state of research.

4 Related Work, and our Contribution

To the best of our knowledge, at the time of writing this paper, language models in general and GPT in particular had not yet been used to write a legal commentary – except for our own companion project on the parallel provision in the German constitution (Engel and Kruse 2024). After we had posted the original version of this paper,[6] a related paper has come out (Santosh, Aly et al. 2025). In this section, we position our paper in the broader literature on legal NLP, and on the summarization of legal text in particular.

Legal NLP. Language models have already been used for a variety of legal tasks (Kapoor, Henderson et al. 2024): from legal education (Choi, Hickman et al. 2021, Choi and Schwarcz 2023), the explanation of ambiguous legal concepts (Savelka, Ashley et al. 2023), empirical legal research (Drápal, Westermann et al. 2023, Livermore, Herron et al. 2024) to legal practice (Rodgers, Armour et al. 2023, Bilgin and Licato 2024, Trozze, Davies et al. 2024). The use of LLMs has already achieved considerable success. For example, GPT-4 was able to answer questions on the US Bar Exam with an average accuracy rate that would have been sufficient to pass in all states (Katz, Bommarito et al. 2024) or showed high accuracy in extracting legal information from Employment Tribunal judgments (de Faria, Xie et al. 2024). The subsumption skills (statutory reasoning) of LLMs have also been examined (Trozze, Davies et al. 2024, Zou, Zhang et al. 2024). Even simple subsumption tasks caused difficulties though for the (now outdated) GPT-3 model: answers were only correct in around 4 out of 5 cases (Blair-Stanek, Holzenberger et al. 2023).

Hallucinations have attracted particular attention (Mik 2023). If the language model hallucinates, it generates results that are not at all based on the input given to them (Dahl, Magesh et al. 2024a). In the present context, this could mean that the model "references" a non-existent ruling, or one that discusses a different human right. A study has found that ChatGPT-4 answered legal questions incorrectly in 58% of cases (Dahl, Magesh et al. 2024a). Deroy et al. investigated the extent to which GPT-3.5 Turbo is suitable for summarizing court decisions. They identified multiple hallucinations and came to the conclusion that language models are not yet capable of providing fully automated summaries of legal texts (Deroy, Gosh et al. 2023). Mindful of this risk, we introduce two precautions, one generic and one specific. As a generic safeguard, we do not only design a pipeline that automates the writing of a commentary. We also provide an evaluation suite that checks accuracy. As a specific safeguard, the LLM-written commentaries do not only summarize the rich jurisprudence of

[5] Current frontier models are said to use 1.8 trillion (GPT-4o), 405 billion (Llama 3.1 405B) and 176 billion parameters (Mixtral 8x22B); Claude 3.5 Sonnet Opus has not disclosed the number of parameters.
[6] On Oct 21, 2024, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4994131.

the European Court of Human Rights on freedom of assembly. Each statement comes with references to supporting paragraphs from rulings of the court, and each reference comes with a link to the raw text. Hence whenever a legal user is skeptical, she may immediately, and swiftly, check back.

Legal Summarization. Judicial opinions shape doctrine and guide future decisions, requiring practitioners to review extensive case corpora to identify relevant precedents. Yet the growing volume and pace of decisions exceed human reading capacity, rendering manual synthesis impractical. Recent work addresses this challenge through automatic single-case summarization, which condenses individual opinions for faster and more accurate uptake (Jain, Borah et al. 2021, Akter, Çano et al. 2025). Most approaches are extractive, prioritizing faithfulness to the source (Akter, Çano et al. 2025), and include both unsupervised and supervised variants (Bhattacharya, Hiware et al. 2019). However, extractive methods are only of limited use for legal purposes, as they lack expert guided synthesis.

In response, recent research has shifted toward abstractive and hybrid summarization approaches (Ragazzi, Moro et al. 2024, Gao, Yu et al. 2025). Models such as BART, BERT-Summ, and BigBird have been adapted to legal texts (Shukla, Bhattacharya et al. 2022), often using length-aware pipelines that segment long decisions into semantically coherent units before generation (Moro and Ragazzi 2022). To improve factual accuracy, entailment-based re-ranking selects faithful summaries (Feijo and Moreira 2023), while summary re-ranking (Elaraby, Zhong et al. 2023), argument mining (Elaraby and Litman 2022) and dynamic retrieval-augmented generation (Ajay Mukund and Easwarakumar 2025) enhance content selection and organization.

With the rise of large language models (LLMs), foundational models have further advanced summarization capabilities. Initially applied in other domains, such as long-book summarization with GPT-4 and GPT-3.5 Turbo (Chang, Lo et al. 2023), LLM-based pipelines have more recently been tailored to legal texts (Cui, Ning et al. 2023, Pont, Galli et al. 2023, Deroy, Ghosh et al. 2024, Benedetto, Cagliero et al. 2025, Santosh, Aly et al. 2025).

A legal commentary reduces and filters information. This is why it can be brought under the rubric of text summarization. Yet a legal commentary is more ambitious: it (i) integrates multiple cases under coherent doctrinal principles, (ii) identifies patterns and developments in a body of jurisprudence, and (iii) contextualizes decisions within broader legal frameworks. In the words of the Guide, a commentary intends “to inform legal practitioners about the fundamental judgments and decisions delivered by the Strasbourg Court” (p. 5). If they have access to a reliable commentary, practitioners do not need to evaluate the wealth of potentially relevant material themselves. They can limit themselves to looking up the references that the commentary flags as directly relevant.

Our first contribution is an automated pipeline that analyzes, categorizes, and organizes large corpora of judicial decisions according to a predefined doctrinal framework, using large language models (LLMs). The key innovation lies in injecting domain-specific legal expertise directly into the pipeline: we do not let the doctrinal structure guiding the analysis emerge from the case law, but externally provide this expert knowledge to the LLM, with the help of elaborate prompts. This approach goes beyond “summarizing summaries.” Prior research has

shown that domain-specific knowledge improves the accuracy and relevance of summaries (Sharma and Singh 2025).

Jurisprudence cannot be fully understood by analyzing individual decisions in isolation. Only by embedding judicial reasoning within doctrinal structures do key principles and cross-references become visible. This explains the finding of Santosh, Aly et al. (2025), who report that their structured, LLM-generated reports on user-specified topics often exhibit a degree of fragmentation that "fails to capture the interconnected nature of legal issues" and "can hinder a comprehensive understanding of the broader legal landscape and reduce the utility of the generated reports for complex legal analyses". Our LLM-written commentaries address this concern by explicitly providing the LLM with the underlying doctrinal logic.

Evaluation metrics. The very point of a legal commentary is division of labor. The legal practitioner does not have to spend days or weeks reading potentially relevant court rulings. Rather she trusts that the commentary informs her about the current interpretation of the statutory provisions that matter for her case. If the commentary has been written by human experts, the main source of trust is the reputation of the authors. For LLM-written commentaries, the authority of the author cannot serve as a signal of trustworthiness. This is why explicit evaluation is critical.

One obvious option is the evaluation by human experts. Yet human legal experts are a scarce and expensive resource. Requiring human validation for every LLM output would fundamentally undermine the potential of LLM-assisted workflows. There would be little point in generating an entire commentary for less than $50 if its validation costs hundreds or even thousands of dollars. More importantly even: LLM-written commentaries are most appealing in legal domains that are important for smaller, more specialized legal audiences. In such specialized domains, a sufficient population of validation experts may be very hard to find. For both reasons, automated evaluation is desirable. In that spirit, Xu and Ashley (2023) have developed an evaluation framework based on questions and answers. Santosh, Aly et al. (2025) have proposed their G-Eval approach for assessing legal summaries. We introduce a pipeline that it specifically geared towards the evaluation of legal commentaries (and legal scholarship more broadly). It consists of two components: predicting case outcomes, and comparatively assessing legal argument.

Outcome predictions. Legal practitioners chiefly consult a commentary if they want to find out whether a case is worth litigating. Courts and administrators want to understand which decision of a case would be most in line with the existing body of jurisprudence. For both use cases, there is a straightforward validation metric. Adopting an LLM-as-a-judge approach (Zheng, Chiang et al. 2023, Gu, Jiang et al. 2024) one provides an evaluator LLM with the facts of cases, and asks them to predict outcomes. One compares responses when just asking (zero shot) with responses when the evaluator LLM has access to the commentary in question through a RAG (for background see Lewis, Perez et al. 2020, Gao, Xiong et al. 2023, Gupta, Ranjan et al. 2024). This approach also lends itself to compare the performance of different commentaries including, in our case, the Guide written by the registry of the court.

Content evaluation. The second arm of our evaluation pipeline uses text analysis to assess the quality of the commentaries written by the LLMs. The earliest and most widely adopted

evaluation paradigm is based on lexical overlap: a machine-generated summary is considered high quality if it shares a sufficient number of words or phrases with one or more human-written reference summaries. Within this paradigm, two prominent metrics dominate (Akter, Çano et al. 2025):

- ROUGE (Recall-Oriented Understudy for Gisting Evaluation) - Measures the proportion of words or n-grams from the reference that also appear in the generated summary (recall-focused) (Lin 2004).
- BLEU (Bilingual Evaluation Understudy) - Measures the proportion of words in the generated summary that are also present in the reference (precision-focused) (Reiter 2018).

However, these metrics suffer from semantic blindness: they cannot distinguish a valid paraphrase from a nonsensical sentence sharing the right keywords (Reiter 2018). To address this, newer metrics incorporate semantic similarity and factual consistency:

- BERTScore - Uses contextual embeddings from language models such as BERT to compute cosine similarity between token representations, enabling recognition of conceptual equivalence even when different words are used (Deutsch and Roth 2020).
- NLI-based metrics (Natural Language Inference, e.g., SummaC) - Assess whether the source text (premise) logically supports statements in the summary (hypothesis) (Laban, Schnabel et al. 2022).
- QA-based metrics (Question Answering) - Generate questions from the summary and verify that a QA system produces consistent answers based on both the source text and the summary (Fabbri, Wu et al. 2021).

We offer such quantitative metrics as part of our evaluation pipeline, but chiefly rely on an LLM-as-a-judge approach (Liu, Iter et al. 2023, Wu, Gong et al. 2023, Gu, Jiang et al. 2024). To capture the specific nature of a legal commentary, multiple elements must be evaluated, which together determine its overall quality. Following Santosh, Aly et al. (2025), we assess both the content of the commentary and its structure (hierarchical organization and headings). In addition, we analyze case references and citations, which represent a significant quality dimension in commentary creation. Beyond these objective metrics, our evaluation pipeline also includes a subjective dimension, following Liu, Iter et al. (2023), estimating the user experience of three distinct user groups: lawyers, judges, and affected parties.

5 Programming GPT to Write a Commentary

Jurisprudence of the European Court of Human Rights. The first step of the process on which we report in this paper does not require a large language model. Happily the European Court of Human Rights is very transparent. On its website, it posts the complete text of 68,956 decisions. This is a large number, but still only a fraction of the cases that have been submitted: in the 10 years from 2014 to 2023, 467,300 cases have been allocated to one of the decision-making bodies set up by the Convention.[7] Were this a paper intending to causally analyze the jurisprudence of the court, we would have to worry about selection. Yet for our purposes, this limitation is mild. It would only matter if the court had kept rulings confidential that are of

[7] https://www.echr.coe.int/documents/d/echr/stats-analysis-2023-eng?download=true, p. 6. For the early years of the Convention bodies, see https://www.echr.coe.int/documents/d/echr/survey_19591998_bil.

high importance for predicting the future decision of analogous cases. Theoretically, we cannot exclude this possibility. But it is a very unlikely concern. Informing the general public, and governments for that matter, about the development of its jurisprudence is the most important policy lever of the court. This is reflected in a very elaborate and outspoken policy of promulgating such developments.[8]

In principle, as the first step of the process, we might have downloaded all rulings that the court has posted, and would then have filtered them for freedom of assembly. We did not have to do that though as the court maintains a very well-organized database.[9] This has allowed us to directly filter cases that the database highlights for discussing Art. 11 ECHR.[10] This gave us a wider set of 1198 cases.

Downloading these cases was a bit of a challenge, as the database of the court is constructed as a dynamic website.[11] We could therefore not directly target the .html code with the `beautiful soup` package in `python`, and had to mimic browsing to the dynamic site for each case, with the help of the `selenium` package. On each site, we had the program click the .pdf button, and download the file linked to it.[12] Next we have extracted the raw text from each file, using the `PyPDF2` package. After a series of data cleaning steps[13] we had our set of 1198 raw data files.

Quite a number of the raw data files are actually not discussing freedom of assembly. The main reason is the construction of Art. 11 ECHR. The provision also covers freedom of association (406 cases). Other files on closer inspection do not discuss either human right (101 cases). This leaves us with an actual set of 691 cases.

Classifying individual rulings. For the analysis we have separately used three different LLMs: `GPT-4o-2024-08-06`, `Gemini-2.5-flash`, and `Kimi-K2-Instruct`[14], always setting `temperature` to 0. For GPT we have additionally set `top_p` to 1, `presence_penalty` to 0, and `frequency_penalty` to 0. For Gemini and Kimi-K2 we have achieved the same specification by keeping the default parameters.

Gemini-2.5-flash has a context window of 1M tokens, which we have never reached. GPT-4o and Kimi-K2 Instruct have an impressive context window of 128,000 tokens. Still the totality of the 691 cases uses 28 MB, too much even for a frontier model.[15] Additionally, performance may decrease when processing inputs that approach the token limit. A recent study found that large language models (LLMs) exhibit a "lost in the middle" effect when handling extensive amounts of text, similar to the serial position effect observed in humans (Feigenbaum and Simon 1962, Murdock Jr 1962). When large volumes of information are

[8] See most notably the definition of „key cases“, ECHR 2025, p. 8, and the establishment of a separate knowledge-sharing institution, https://www.echr.coe.int/knowledge-sharing.
[9] https://tinyurl.com/3d8a8v5k.
[10] The exact filtering steps are documented in the ReadMe document.
[11] For background see Koppanati 2021, Chen, Chen et al. 2021.
[12] Some old cases where on static websites, so that we had to split the process, after discriminating between static and dynamic websites. For detail see the ReadMe document, and the associated code.
[13] Documented in the ReadMe document.
[14] As this access is known to be more stable than Kimi's own website, we have used together.ai as the conduit.
[15] A rough estimate is 6 characters per token, which would result in a total of 4,666,667 tokens.

present, LLM performance significantly drops if relevant data is situated in the middle of a document rather than at the beginning or end (Liu, Lin et al. 2024, Zhang, Zhang et al. 2025). This poses particular challenges when processing a large set of court decisions. We therefore proceed iteratively. We first analyze each ruling individually, and only in a second step aggregate over the summaries that GPT has written for each individual ruling.

The ultimate goal of this first step is the multidimensional classification of each individual case. As our process starts with individual rulings, we cannot have the large language model infer doctrine from the complete body of jurisprudence. Rather we inform it with the help of the prompt. One may of course consider this a limitation of the approach. But this limitation is in no way different from the approach of human commentators. They would also not start from scratch, but would build on the doctrine established in the earlier jurisprudence of the court. Moreover, in the case of Art. 11 ECHR the mapping between the wording of the provision and the structure of the established doctrine is very close. Hence merely by reading the relevant provision, one would already come close to an understanding of the relevant doctrine. This implies that our upfront intervention is indeed very mild.

The prompt additionally adds structure to the process. We inform the large language model about the typical content of a ruling, and ask it to focus on the opinion of the court, using the statements of the parties only to the extent that they help better understand the court's decision. In the spirit of chain of thought prompting (Wei, Wang et al. 2022), we ask the language model to first characterize in natural language whether, and if so how, the ruling addresses each individual element of the established doctrine. If, for the element of doctrine in question, the response is positive, we further ask the language model to note the paragraph or paragraphs within which the court discusses this element of doctrine. Finally, to facilitate the next steps of the process, we ask GPT to respond in JSON format whether the ruling discusses each individual element of doctrine, "Yes" or "No".

This intermediate step already yields an interesting observation: the discussion of the elements of doctrine is very unevenly distributed (Figure 1). The large majority of the rulings spell out whether the governmental act or omission against which the complaint is directed falls into the substantive and personal scope of freedom of assembly, and whether the act interferes with this human right. A not much smaller number of rulings also applies the "necessary in a democratic society" test. By contrast, discussions of the various aims that the provision considers legitimate are rarer. Interestingly, the most frequently invoked justification is prevention of disorder, not the protection of the rights of others, let alone the remaining legitimate aims. Finally it is remarkable how frequently the court explicitly discusses whether the applicant deserves some form of just satisfaction.

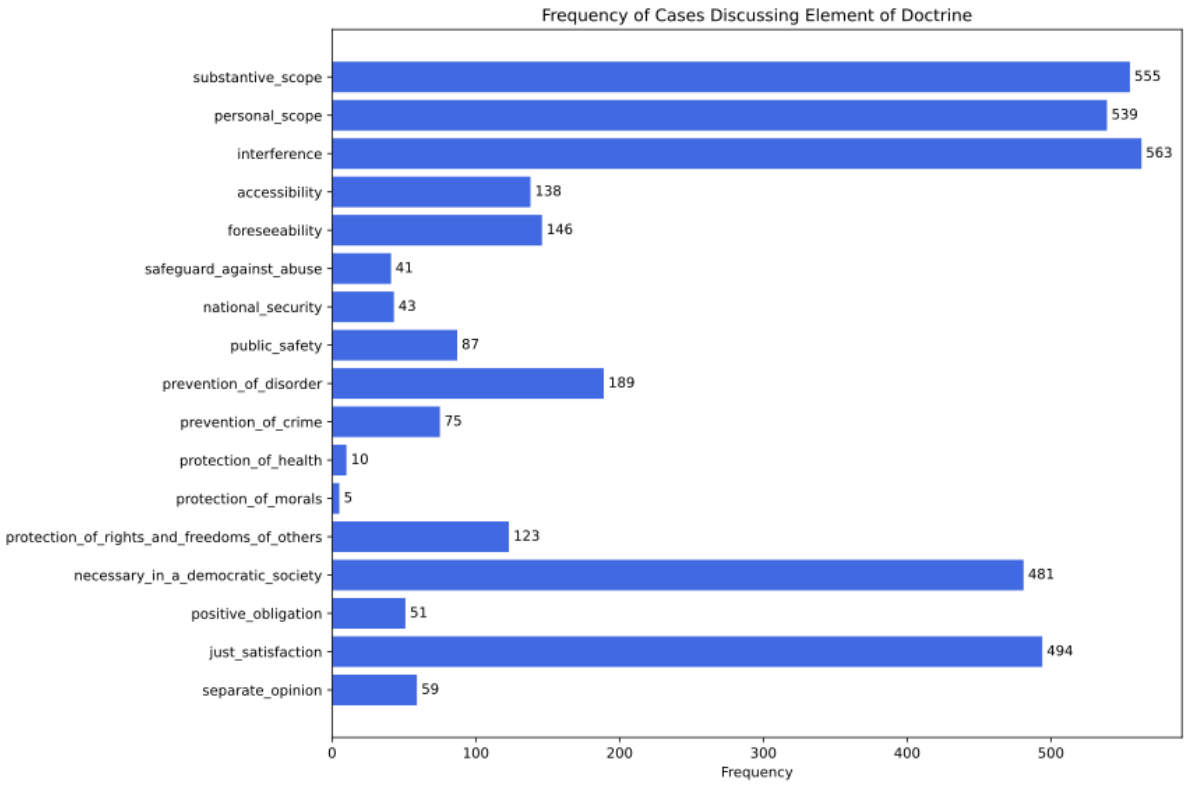

Figure 1 Intensity of doctrinal discussion (based on GPT)

Summarization across rulings. The summarization across rulings starts with creating a folder, separately for each element of doctrine, with paragraphs of all rulings that the respective LLM has identified as discussing the element in question. These are therefore not generic, but aspect-based summaries (Santosh, Aly et al. 2024).

At this point of our pipeline, we see a clear difference between the slightly older model GPT-4o and the two more modern LLMs. For GPT, just combining all these relevant paragraphs into one single file and asking GPT to summarize this file is not feasible. There is just too much information. Even if, technically, the file fits within the token window, results do not look convincing. GPT is overwhelmed with the sheer amount of text. This is why we have to handle the input data in batches of 100 rulings.[16]

While this split makes the process manageable, it creates a new challenge. When summarizing the next batch, GPT does not remember how it had summarized earlier batches. This is why there is pronounced heterogeneity both in style, but also in substance, across per batch summaries. We react the following way: we first ask GPT to extract the subtopics it finds in each per batch summary, and to bring the resulting list of subtopics into a coherent order. We then have GPT revisit each batch of input data, adding the more coherent list of subtopics to the system prompt. For the rest of the process, we work with the result of this second round

[16] Specifically, we split the input file at case numbers 100, 200 …, 600. Effectively therefore each batch contains less than extracts from 100 rulings, as no element of doctrine is discussed in each and every ruling.

of extraction. Happily then the resulting summaries are considerably more coherent across batches.

We concatenate these revised per batch summaries, separately for each element of doctrine, and for each subtopic within the element of doctrine that GPT has identified. To these raw summaries, we apply a number of data cleaning steps. In earlier steps of our process, we have singled out paragraphs that truly discuss the element of doctrine in question. Yet this step does not filter out text that conjointly discusses more than one element of doctrine. Despite the fact that we always have reminded GPT of the structure of doctrine, it repeatedly has summarized everything that has been said, not just what has been said about the element of doctrine in question. In a first cleaning step, we have asked GPT to remove text from the summaries that belongs to other elements of doctrine. A further challenge results from the tendency of GPT to be redundant. In a second cleaning step, we have asked it to also remove repetitive statements. Finally in a third pass, we have asked GPT to double check whether summaries of one element of doctrine still cover material that belongs to other elements of doctrine, and to remove such text as well.

Only months later, and following the request by one of our referees, we have repeated the generation of the complete commentary with Gemini 2.5 flash, and with Kimi K2. As it turns out, both these newer models no longer face the same problem. We can right away feed the complete list of paragraphs from individual rulings that either model has identified as discussing the respective element of the doctrine of freedom of assembly. Technical progress not only simplifies the pipeline. We also have reason to expect that information extraction is more coherent, as all classifications are directly performed on the raw input data.

The complete code for all three LLMs is available for scrutiny.[17] In the Appendix, we illustrate the approach with the master for writing the system prompts for the final summarization step.

Presentation. We present the result on a set of websites.[18] The commentary is structured hierarchically. At the top level, it is organized by the elements of doctrine that result from the wording of Art. 11 ECHR. The pipeline that we need for GPT is not only more involved. GPT also organizes the commentary by subtopics. As the output is so rich, we also use these subtopics in the presentation of the commentary on the website. Neither Gemini nor Kimi K2 are so verbose. For these two LLMs we therefore only structure the output by the established elements of the doctrine of Art. 11 ECHR. On the page for each (sub)topic, there is a general definition, and a list of applications. Each statement comes with references to the paragraphs from which the respective LLM has taken the statement in question. Each reference is clickable and leads to the wording of the respective paragraph. We also provide a list of cases, complete with the case number assigned by the court, and a link to the full text of the ruling.

[17] The complete code is available at the following GitHub repository: https://github.com/JohKrus/LLM_as_a_Law_Professor.
[18] See the links in the introduction.

6 Evaluation

This paper does not stop with the proof of concept: surveying, analyzing, summarizing and systematizing a rich body of jurisprudence in accordance with established doctrine is a task for which large language models can be of considerable help. We also provide a method for testing the performance of (human- or computer-generated) structured reports of case law. On this basis, we investigate the performance of our commentary.

We proceed in three steps. Copyright strongly limits the first step. Technically, we could scrape - or, alternatively, digitize and extract - commercial commentaries, and then proceed exactly as we do with the Guide, i.e. feed them into a RAG, and compare their performance with the performance of our LLM-written commentaries. But very likely this process would no longer be covered by fair use rights. This is why, in this first step, we only provide a few fairly brushing metrics that position the LLM-written commentaries in comparison with commercial competitors. We can go much deeper in the remaining two steps of our evaluation pipeline. For these steps we only use the Guide prepared by the Registry of the European Court of Human Rights and our three LLM-written commentaries as input. We use these inputs for predicting the outcomes of a set of test cases (second evaluation step), and for a set of metrics assessing the content, the structure and the referencing of the legal argument (third evaluation step).

Comparison with commercially provided human written commentaries. Table 2 summarizes the comparison with commercially provided commentaries. We cannot say anything about quality. All the table clearly shows is coverage with respect to case citations: the commercial commentaries and the Guide only cover a small portion of jurisprudence, while the LLM-written commentaries provide structured access to the complete evidence that the European Court of Human Rights has made publicly available. Hence consulting our commentary, researchers, or practicing jurists for that matter, can check the complete jurisprudence on a debated element of doctrine. No other commentary provides this functionality (on the question of redundancy, see below).

Metric	GPT 4o	Gemini 2.5	Kimi K2	Guide	Karpenstein/ Mayer	Schabas	Villiger	Meyer-Ladewig/ Nettesheim/v. Raumer
Total Number of Case Citations (Unweighted)	15.306	11.383	9.240	267	88	430	42	94
Total Number of Case Citations (Normalized by Text Length, per 1,000 characters)	30,5	57,3	123,2	4,3	3,6	3,2	3,5	4,1
Number of Unique Cases Cited (Unweighted)	624	579	518	118	43	134	21	56
Number of Unique Cases Cited (Normalized by Text Length)	1,23	2,91	6,91	1,9	1,7	1	1,75	2,4

Table 1 Comparative Comprehensiveness

Prediction of case outcomes. Existing benchmarking tools in legal NLP chiefly sometimes on human alignment, comparing the output generated by the LLM with labels assigned by human legal experts. This procedure is straightforward if such labels exist anyways as, for instance,

with the US Supreme Court (Spaeth, Epstein et al. 2023) or with the US Federal Courts of Appeal (Songer 2011). But human legal experts are a scarce and expensive resource. It would seriously limit the assistance by language models if their output could only be used once validated by a set of human annotators. There would be little point in writing the entire commentary for less than $50, if validation costs many hundreds if not thousands of $. Yet actually there is no need for such investments. The purpose of legal doctrine is guiding legal decision-making. The most important use case in legal practice is a party assessing whether a case is worth bringing, and how to argue in court. Upon a moment's reflection, one sees that there is an abundance of labels created by human legal experts: the actual rulings by the competent court. Actually, these labels are not only authored by humans with domain specific expertise. These humans even have authority to decide on the matter. Making judicial decisions predictable and amenable to forecasting is an important purpose of writing a commentary. Comparing the predictions made when having access to either commentary is our first validation method.

One can of course not know with certainty how the competent court will decide a case it has not yet seen. On the other hand, if one asks the LLM to decide a case that the competent court has published, one cannot exclude that the LLM merely memorizes the decision it has seen at training, training-data contamination (Kruse 2025). Our validation tool addresses these opposing challenges with triangulation. It combines the postdiction of cases that the European Court of Human Rights has decided with the prediction of constructed cases. For the former task, the most interesting benchmark is alignment with the actual outcomes.

For the constructed cases, the benchmark is comparative: how well are predictions aligned when either giving the LLM-as-a-judge access to the LLM written commentary, or giving it access to a competing human written commentary? For our specific use case we add a third layer. The German Constitution enshrines a fundamental freedom (Art. 8 I GG) that is very similar to Art. 11 ECHR.[19] We therefore also test the LLM on a set of cases that it has not seen (they have not subsequently been filed with the ECtHR), but that the German Constitutional Court has decided, which gives us quasi-labels. We acknowledge that these test cases are less clean. The cases are taken from the official website of the German Constitutional Court. We cannot exclude that these cases have been part of general training. Yet the decisions are in German, and discuss freedom of assembly as protected by Art. 8 of the German Constitution, not by Art. 11 ECHR. Both features make it less likely that the respective LLM reports the decision by the German Constitutional Court that it memorizes. We have 18 cases decided by the European Court of Human Rights, 27 cases decided by the German Constitutional Court, and 10 fictitious cases. [20]

Our LLM-as-a-judge is `claude-sonnet-4-20250514`. As it is a frontier model, we can be confident about quality. On the other hand, none of the LLM-written comments has been written by one of the models provided by Anthropic, so that we do not have to worry about circularity. We set temperature to 0.7 and generate 5 independent outcome predictions for each case and commentary (or no commentary, if we elicit zero shot responses). This

[19] The quickest way to see the similarity is comparing one of the LLM-written commentaries on Art. 11 ECHR with our LLM-written commentary on Art. 8 I GG, https://kommentar-ohne-autor.coll.mpg.de/code_update/darstellung/html/uebersicht.html.

[20] Sketches are available in Appendix, Table 2: Test cases examples.

repetition allows us to check how confident Claude is in its prediction about the case outcome.[21] We ask for an explicit justification of the prediction in natural language, in the spirit of a chain-of-thought instruction. In one version, we merely ask for a prediction, without giving the model access to any additional information. In the remaining four versions, we give the LLM access to a RAG (for background see Lewis, Perez et al. 2020, Gao, Xiong et al. 2023, Gupta, Ranjan et al. 2024). The RAG consists of the content of the respective commentary. Specifically, we split the content at the main headings of the commentary. Hence there is, for instance, a chunk with the content of the commentary on the "substantive scope" of Art. 11 ECHR, and another chunk on the "legitimate aim: prevention of disorder". For encoding we use Anthropic's embedding model, i.e. `voyage-3-large`.[22] We use the same model to embed the facts of the test case. We ask Claude to select the 5 most related chunks from the RAG, and to add them to the prompt.

Figure 2 compares the predictions with the actual decisions made by either the European Court of Human Rights, or the German Constitutional Court. Accuracy is high with the cases originally decided by the European Court of Human Rights, close to 80% or higher. But performance is best in the zero-shot condition and with Gemini (88%). Accuracy is much lower with the cases originally decided by the German Constitutional Court (at most 54%). Both findings suggest that memorization is indeed a concern. The database of the European Court of Human Rights has been widely used in legal NLP, as it is rich, multi-lingual, and freely available. Quite likely the cases from the German Constitutional Court draw a more realistic picture. The low accuracy suggests that the prediction of case outcomes remains a hard task.[23] But in relative terms, we see the expected effect of a commentary: access to any commentary improves accuracy. Interestingly the effect of the LLM-written commentaries is even larger than the effect of the human-written Guide.

[21] Variance turns out to be rather low: in 6.5% of all cases, one response differs from the majority, in 5.8% two responses differ. Variance is not substantially different between courts, or between models. Detailed analyses of variance are available from the authors upon request.

[22] In the commentary written by GPT-4o, the sections of the commentary are particularly long, see http://professor-gpt.coll.mpg.de/html/overview.html, which is why at encoding we hit the limit of 32 K tokens. This is why, for this commentary, at encoding we employ Langchain's RecursiveCharacterTextSplitter, setting chunks to 1000 tokens, with 200 tokens overlap.

[23] One should not be misled to think that the performance is even below chance level. The response variable is of course binary. But in 2/3 of the 27 cases, the German Constitutional Court has held that freedom of assembly is violated, which is why guessing would not be a helpful strategy. We do of course not disclose this base rate to Claude.

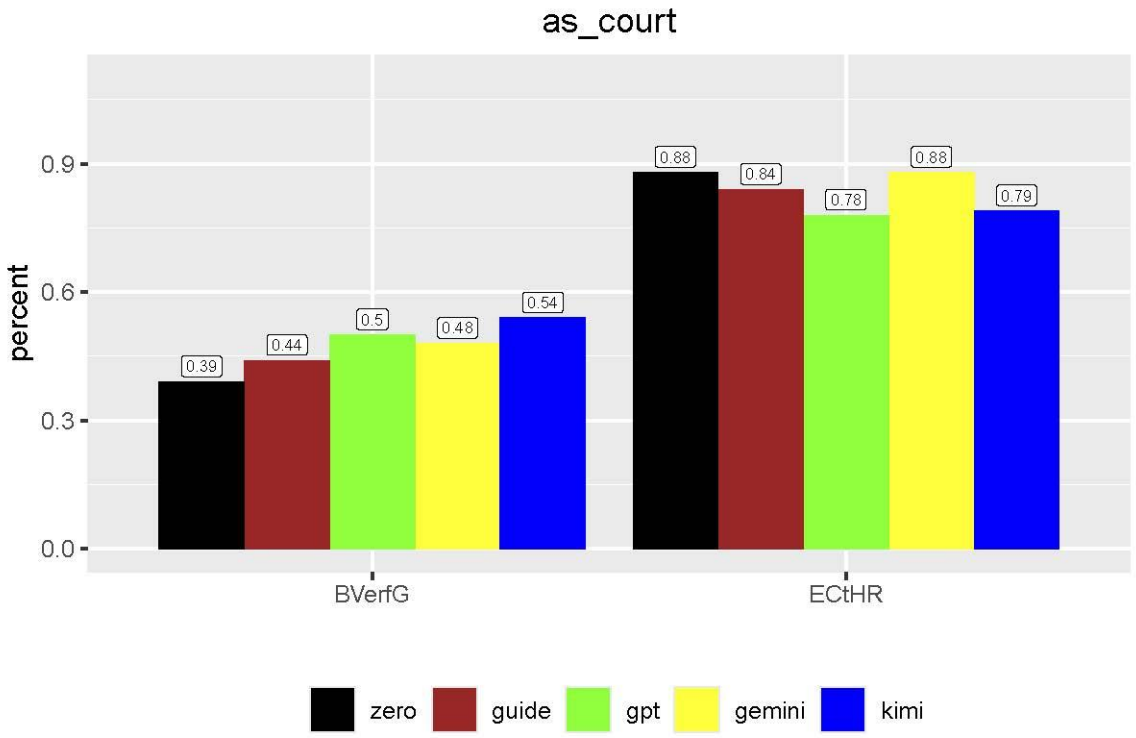

Figure 2
Comparison of Prediction with Actual Case Outcomes

Figure 3 adds a complementary metric. It compares outcomes with access to one of the LLM-written commentaries to the predictions Claude makes when given access to the Guide. This metric shows how well the LLM-written commentaries fare in comparison with their human-written competitor. Results show that convergence is always high (between 78% and 94%). In the realistic comparison with another commentary, the LLM-written commentaries match the performance of a commentary written by human domain experts. Essentially, what can be achieved with the help of a commentary is achieved. Interestingly, though, the marginal effect of having access to a commentary over just general training is not huge (and in the case of the 10 fictious cases, all LLM-written commentaries are even less close to the predictions based on the Guide, compared with the zero-shot response).

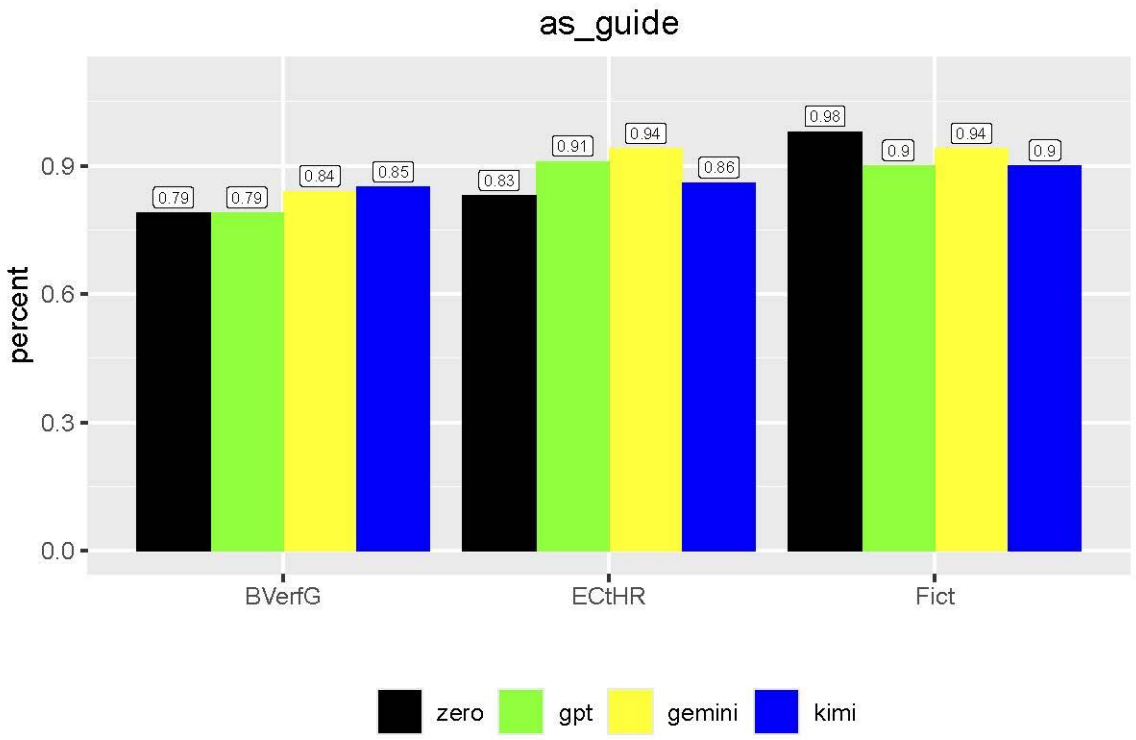

Figure 3
Comparison of Predictions with Prediction Based on Guide

Content. In the next step, we analyze the content of the Guide, compared with the content of the three LLM-written commentaries. We continue to use Claude Sonnet 4 as our LLM-as-a-judge. We ask each question 3 times, and ask Claude to score the respective commentary on a 5 point Likert scale, with higher values defining higher quality. For the first 5 measures we ask Claude to adopt the perspective of an objective observer. For the last 3 measures, we ask Claude to adopt the role of either a lawyer preparing an argument, a judge preparing a ruling, or an affected party assessing whether her case is worth bringing. All prompts are in the Appendix. Figure 4 summarizes these metrics.

The main message is: in all eight dimensions, scores for all four commentaries are close to each other. The LLM-written commentaries are comparable in quality with the human-written Guide. Sometimes the Guide is a tad better (legal precision; citation quality), sometimes one of the LLM-written commentaries is better. On all measures, the commentary written by GPT-4o is poorest. Interestingly, the commentary written with the help of Kimi-K2 Instruct is always the best among the LLM-written commentaries. This is remarkable given Kimi-K2 is not even a reasoning model.

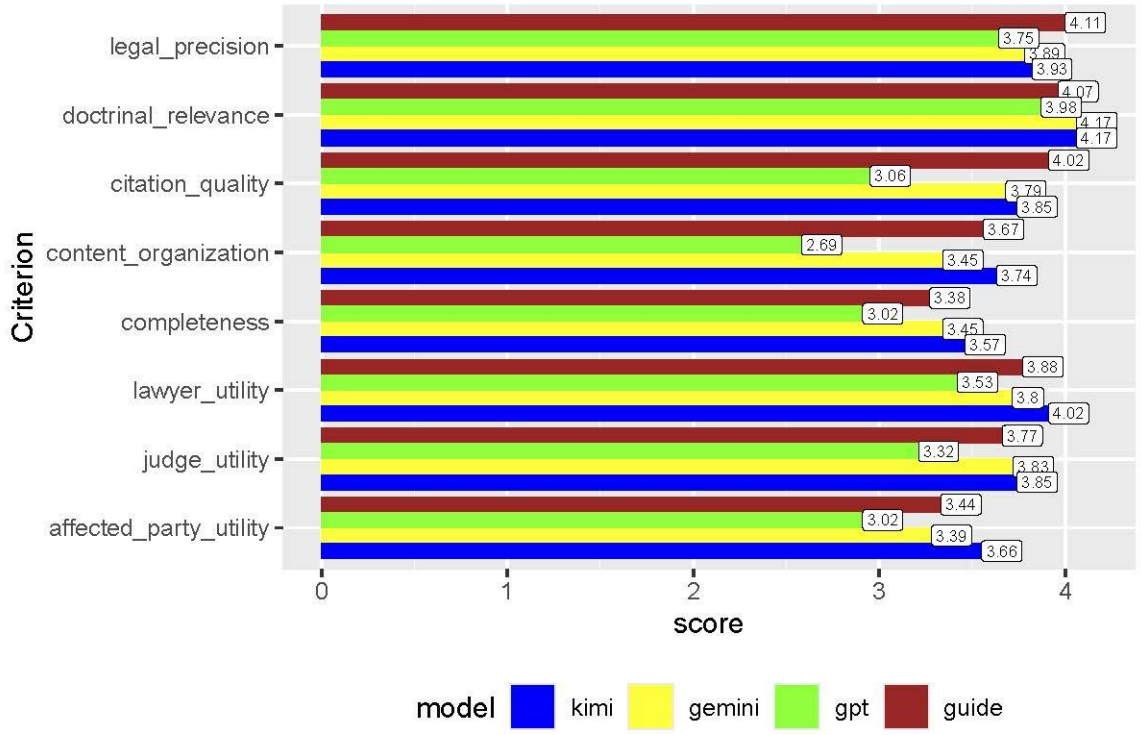

Figure 4
Content Analysis

Structure. The main difference between plain vanilla summarization and a commentary is a presentation of the rich material that is structured in a way that is congenial to the legal tasks for which the commentary is consulted. Using the same approach as with content analysis, we ask Claude Sonnet 4 to classify the performance of either commentary in this regard. We again ask 3 times each, and request a score on a 5 point Likert scale. For the first 4 measures we ask Claude to adopt the perspective of an objective observer. For the last 3 measures, we ask Claude to adopt the role of either a lawyer preparing an argument, a judge preparing a ruling, or an affected party assessing whether her case is worth bringing. All prompts are in the Appendix.

Figure 5 summarizes these metrics. Results are very similar to the ones for content analysis. Overall, the three LLM-written commentaries are on par with the Guide. Actually, in this respect, the Guide never outperforms all LLMs. Gemini is (slightly) better for hierarchical logic, coherence and the expected experience of an affected party.

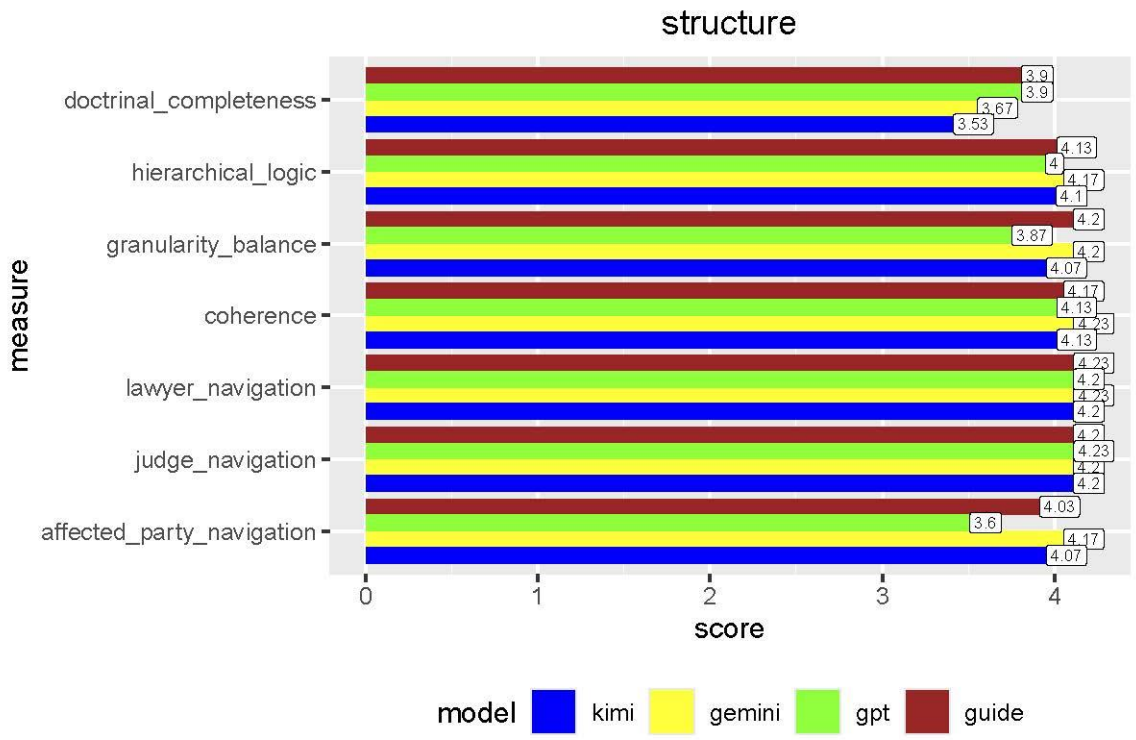

Figure 5
Structural Analysis

Citations. Commentaries are not the least so popular since they help practising lawyers with the laborious search for pertinent court decisions. Our last set of measures gauges how well the competing commentaries perform in this respect. We again ask Claude Sonnet 4 to classify the performance of either commentary in this regard. We ask 3 times each, and request a score on a 5 point Likert scale. All prompts are in the Appendix. As Figure 6 shows, in this dimension the LLM-written commentaries also perform well. Actually in all dimensions, the commentary written by Gemini even outperforms the Guide.

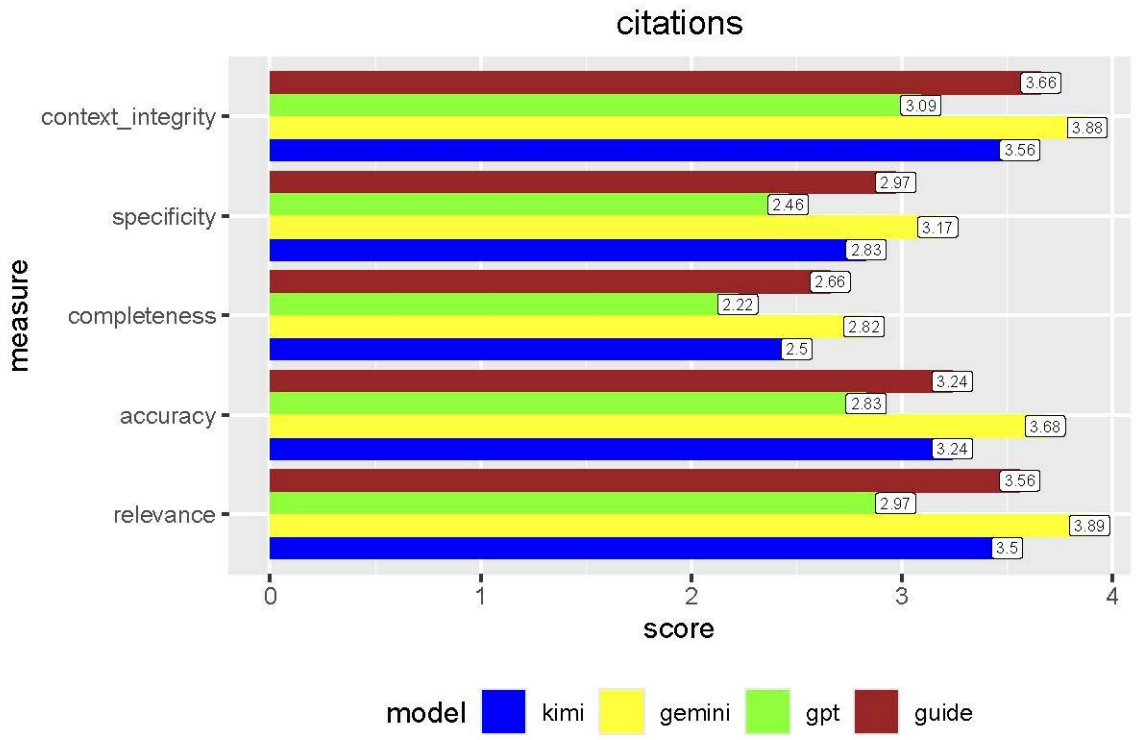

Figure 6
Citation Quality

Yet if we supplement the responses of our LLM as a judge with standard automated metrics (Figure 7) or with measures of non-redundancy (Figure 8), the Guide is clearly ahead, by quite a margin. Upon a moment's reflection, this result is expected. As the coverage of jurisprudence in all human-written commentaries is highly selective, in our pipeline for writing the three commentaries we have privileged coverage over fast access to landmark rulings. If one wanted that, one would need to modify the pipeline and add a step in which the LLM selects what it believes to be the most informative rulings. For this, one could rely on the ECtHR's official selection of judgments, decisions, and advisory opinions that significantly contribute to the development, clarification, or modification of the Court's case-law.[24] We have considered it more important to inform interested readers about the richness of the jurisprudence of the European Court of Human Rights. Also, because, as a factual matter, no two cases are alike, thereby giving users the best possible chance of locating the decision that most closely matches their own case.

[24] https://ks.echr.coe.int/documents/d/echr-ks/key-cases-2025-eng.

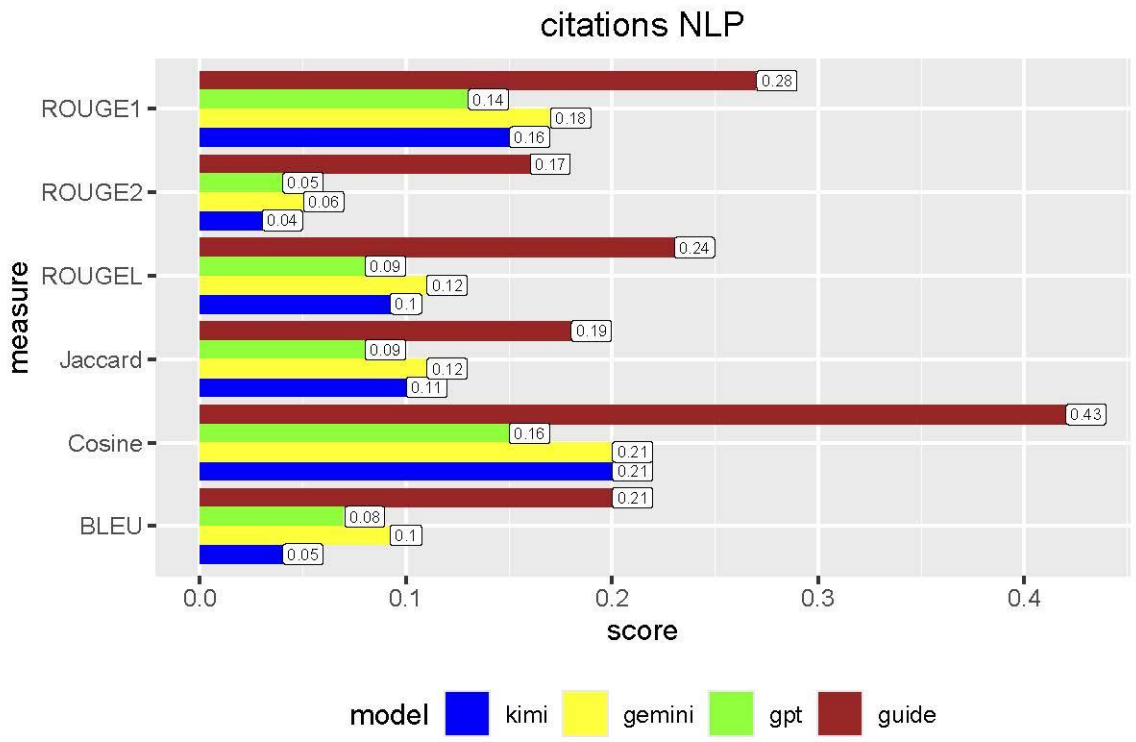

Figure 7
Mechanical NLP Metrics for the Informativeness of Citations

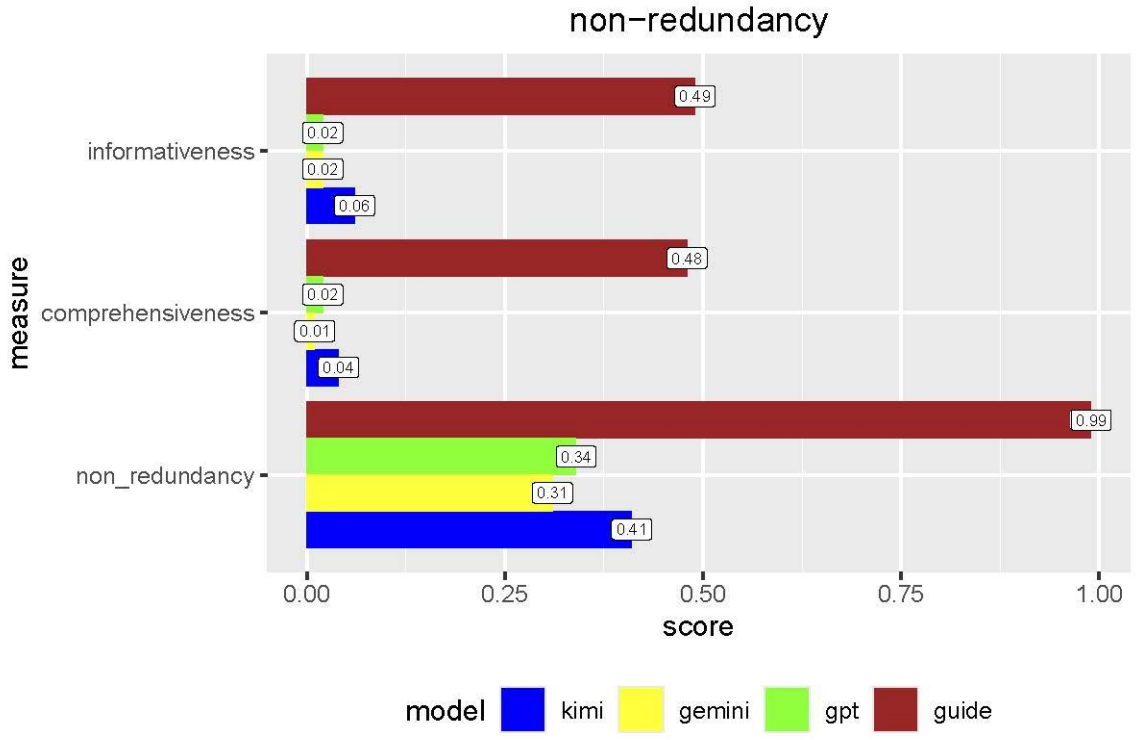

Figure 8
Non-Redundancy Metrics

7 Limitations

Limitations arise, first, from the use of LLMs. LLMs hallucinate, that is, they may generate results that are not grounded in the input given to them, be it at training or at inference (Dahl, Magesh et al. 2024b). In the present context, this could mean that the model "references" a non-existent ruling, or one that discusses a different human right. A study has found that ChatGPT-4 answered legal questions incorrectly in 58% of cases (Dahl, Magesh et al. 2024b). Hallucinations have also been observed when asking GPT-3.5 turbo to summarize court decisions (Deroy, Ghosh et al. 2023). Mindful of this risk, we introduce two precautions, one generic and one specific. As a generic safeguard, we do not only design a pipeline that automates the writing of a commentary. We also provide an evaluation suite that checks accuracy. As a specific safeguard, the LLM-written commentaries do not only summarize the rich jurisprudence of the European Court of Human Rights on freedom of assembly. Each statement comes with references to supporting paragraphs from rulings of the court, and each reference comes with a link to the raw text. Hence whenever a legal user is skeptical, she may immediately, and swiftly, check back.

LLMs are only able to write a usable commentary after fairly heavy prompting. At the prompting stage, the researcher must "teach" the LLM the doctrine of the provision she wants the LLM to summarize. Consequently for each legal provision professional legal input is required. One cannot scale the process by a generic prompt like "summarize the doctrine of provision X". But human commentators would also not start from scratch. In particular if the jurisprudence on a certain provision is rich, employing an LLM would dramatically speed up the process, and would also make it less likely that relevant rulings are overlooked. Nonetheless, the pipeline we have developed can be used only by experts, and only in areas of law in which they possess at least foundational knowledge.

Limitations must also be considered with respect to our evaluation pipeline. As a newcomer, one would naturally want to compare the LLM-generated commentaries with the existing offerings. While this would be technically feasible with our pipeline, it fails due to legal hurdles. The services of online commercial providers (e.g. Wolters Kluwer), and the commentaries published as a hardcopy by commercial publishers, are copyright protected and behind paywalls, which precludes a comparison.

If one evaluates the commentaries written by LLMs by their ability to postdict decisions taken by the competent or related courts, one may be disappointed by the fact that for unseen cases (like the ones originally decided by the German Constitutional Court on the equivalent guarantee in the German constitution) accuracy is very low. Yet this also holds for the commentary written by the court itself. This suggests that predicting the outcome of legal cases is still a difficult task, even for high end LLMs. However the intended reader of a commentary is not an LLM, but professionally trained lawyers. For them it is critical that they are precisely and succinctly informed about the position of the competent court, and that they can check back the interpretation that matters for their specific case with closely related rulings of the court. A rich set of metrics shows that the LLM-written commentaries are at least as helpful for these classic use cases as their human written competitor, the Guide prepared by the registry of the Court.

Technological advances have eliminated two further limitations: GPT-4o cannot handle the complete raw text on individual elements of the doctrine of Art. 11 ECHR. When using this LLM, we must split up the process into manageable chunks. This entails the need for a more involved pipeline that harmonizes the separate summaries written for each chunk. Gemini 2.5 flash has a context window of 1 M tokens, which is much wider than needed for the task. Interestingly, despite the fact that nominally the context window of Kimi K2 Instruct is the same as for GPT 4o (128 K tokens), for all elements of the doctrine of Art. 11 ECHR it gracefully handles the complete input at once.

If the legal community wanted to move from our prototype to production, it might worry about the proprietary nature of frontier models. OpenAI has just taken down older models, including GPT-4o,[25] so that the process could not even be replicated. This makes it particularly appealing that the performance of the commentary written by the open source model Kimi K2 Instruct is close, although it is not even a thinking model.

8 Conclusion

The main purpose for writing this paper has been providing a proof of concept. As we show, a task of direct benefit for the legal community can be delegated to large language models. LLMs can write a structured summary of a large body of jurisprudence, and can provide the result in the format of a commentary. For each element of the doctrine of a statutory provision, the interpretation given by the competent court, as well as its application to concrete cases, are presented. The quality of LLM-written commentaries is on par with human-written commentaries.

As our application, we have chosen freedom of assembly, as protected by Art. 11 ECHR, as there is competition. Since we have started this project in the spirit of a proof of concept, we wanted the possibility to comparatively assess quality. Yet knowing that a large language model can write a very usable commentary is even more important for areas of law that have not been structured in this way. There are multiple reasons for the absence of a commentary. Probably the practically most important reason is what in data analysis would be called the long tail. Provisions that are of high practical importance have long been covered by commentaries, often by multiple competing commentaries. But many provisions are only relevant for a limited class of cases. For such matters, often no commentary is available.

A second use case is motivated by different professional traditions. Commentaries are standard in the German speaking countries. There are some examples in other language families. But in many jurisdictions, including the UK and the US, the classic German commentary is not available for any statutory or jurisprudential rule. Given our proof of concept, it would be easy and relatively cheap to try the format of a commentary in these jurisdictions as well. In addition, the prototype presented here can also be used (with minor modifications) for legal literature formats that play a significant role in common law jurisdictions, like “annotations” to constitutions and statutes (such as the United States Code

[25] But has brought them back in response to user protest, https://platform.openai.com/docs/models, visited on August 17, 2025.

Annotated and the United States Code Service). They provide not only the actual texts of the statute but also a summary of the cases that interpret the statute (Reimann 2020). A commentary written with the help of our prototype might also facilitate the preparation of a Restatement.

The most attractive feature of a commentary written by a large language model is timing. In areas of law that are both very active fields of jurisprudence and of high practical importance, commentaries might perhaps be updated every year. But especially if the field is active, a lot may happen during a year. And in many other fields, practitioners must wait much longer than a year before the next edition of the commentary becomes available. By contrast a large language model can be programmed such that the commentary is updated at very short intervals.

The primary interest in commentaries certainly stems from legal practice. But the structured analysis of jurisprudence is also useful for academic research. It is a straightforward next step to translate the classification that is necessary for writing the commentary into code. This provides academic research of judicial policy making with high-quality, fine grained sets of features that did not exist before.

Actually, this paper does not stop at providing the proof of concept. We also provide a method for testing the performance of (human-written or computer-generated) structured reports of legal doctrine. This method can easily travel beyond our specific use case. The necessary ingredients are generic: a set of test cases, and a machine-readable version of the summary. With the help of the equivalent machinery, researchers might, for instance, test how deeply the alternative decision of a leading case, or the expected decision of a contested case that is still undecided, would likely have changed the outcome of past, and would change the outcome of future cases. Our suite of quality metrics, for content, structure and the informativeness of references, could also be used to gauge the quality of scholarly contributions to law. Once more, with the advent of large language models the frontier of legal practice and legal research has moved out, by quite a margin.

References

Ajay Mukund, S and KS Easwarakumar (2025). "Optimizing Legal Text Summarization through Dynamic Retrieval-Augmented Generation and Domain-Specific Adaptation." Symmetry **17**(5): 633.

Akter, Mousumi, Erion Çano, Erik Weber, Dennis Dobler and Ivan Habernal (2025). A Comprehensive Survey on Legal Summarization: Challenges and Future Directions. arXiv preprint arXiv:2501.17830.

Benedetto, Irene, Luca Cagliero, Michele Ferro, Francesco Tarasconi, Claudia Bernini and Giuseppe Giacalone (2025). "Leveraging Large Language Models for Abstractive Summarization of Italian Legal News." Artificial Intelligence and Law: 1-21.

Bhattacharya, Paheli, Kaustubh Hiware, Subham Rajgaria, Nilay Pochhi, Kripabandhu Ghosh and Saptarshi Ghosh (2019). "A Comparative Study of Summarization Algorithms Applied to Legal Case Judgments." European conference on information retrieval: 413-428.

Bilgin, Onur and John Licato (2024). Determining Legal Relevance with Llms Using Relevance Chain Prompting. The International FLAIRS Conference Proceedings. **37**.

Bix, Brian (1991). "Hla Hart and the" Open Texture" of Language." Law and Philosophy **10**: 51-72.

Blair-Stanek, Andrew, Nils Holzenberger and Benjamin Van Durme (2023). Can Gpt-3 Perform Statutory Reasoning? Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law.

Butler, Judith (2016). “We, the People”: Thoughts on Freedom of Assembly. What Is a People? G. Didi-Huberman, S. Khiari, J. Rancière et al., Columbia University Press**:** 49-64.

Chang, Yapei, Kyle Lo, Tanya Goyal and Mohit Iyyer (2023). Booookscore: A Systematic Exploration of Book-Length Summarization in the Era of Llms. arXiv preprint arXiv:2310.00785.

Chang, Yupeng, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang and Yidong Wang (2024). "A Survey on Evaluation of Large Language Models." ACM Transactions on Intelligent Systems and Technology **15**(3): 1-45.

Choi, Jonathan H (2023). "How to Use Large Language Models for Empirical Legal Research." Journal of Institutional and Theoretical Economics **180**: 214-233.

Choi, Jonathan H, Kristin E Hickman, Amy B Monahan and Daniel Schwarcz (2021). "Chatgpt Goes to Law School." Journal of Legal Education **71**: 387-400.

Choi, Jonathan H and Daniel Schwarcz (2023). "Ai Assistance in Legal Analysis. An Empirical Study." Journal of Legal Education **73**: ***.

Colombo, Pierre, Telmo Pessoa Pires, Malik Boudiaf, Dominic Culver, Rui Melo, Caio Corro, Andre FT Martins, Fabrizio Esposito, Vera Lúcia Raposo and Sofia Morgado (2024). Saullm-7b: A Pioneering Large Language Model for Law. arXiv preprint arXiv:2403.03883.

Cui, Jiaxi, Munan Ning, Zongjian Li, Bohua Chen, Yang Yan, Hao Li, Bin Ling, Yonghong Tian and Li Yuan (2023). Chatlaw: A Multi-Agent Collaborative Legal Assistant with Knowledge Graph Enhanced Mixture-of-Experts Large Language Model. arXiv preprint arXiv:2306.16092.

Dahl, Matthew, Varun Magesh, Mirac Suzgun and Daniel E Ho (2024a). "Large Legal Fictions. Profiling Legal Hallucinations in Large Language Models." Journal of Legal Analysis **16**: 64-93.

Dahl, Matthew, Varun Magesh, Mirac Suzgun and Daniel E. Ho (2024b). "Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models." arXiv preprint arXiv:2401.01301.

de Faria, Joana Ribeiro, Huiyuan Xie and Felix Steffek (2024). Automatic Information Extraction from Employment Tribunal Judgements Using Large Language Models. arXiv preprint arXiv:2403.12936.

Deroy, Aniket, Kripabandhu Ghosh and Saptarshi Ghosh (2023). "How Ready Are Pre-Trained Abstractive Models and Llms for Legal Case Judgement Summarization?" arXiv preprint arXiv:2306.01248.

Deroy, Aniket, Kripabandhu Ghosh and Saptarshi Ghosh (2024). "Applicability of Large Language Models and Generative Models for Legal Case Judgement Summarization." Artificial Intelligence and Law: 1-44.

Deroy, Aniket, Kripabandhu Gosh and Saptarshi Gosh (2023). How Ready Are Pre-Trained Abstractive Models and Llms for Legal Case

Judgement Summarization?

Deutsch, Daniel and Dan Roth (2020). Understanding the Extent to Which Summarization Evaluation Metrics Measure the Information Quality of Summaries. arXiv preprint arXiv:2010.12495.

Drápal, Jakub, Hannes Westermann and Jaromir Savelka (2023). Using Large Language Models to Support Thematic Analysis in Empirical Legal Studies. JURIX.

Edelman, Lauren B (1992). "Legal Ambiguity and Symbolic Structures. Organizational Mediation of Civil Rights Law." American Journal of Sociology **97**(6): 1531-1576.

Elaraby, Mohamed and Diane Litman (2022). Arglegalsumm: Improving Abstractive Summarization of Legal Documents with Argument Mining. arXiv preprint arXiv:2209.01650.

Elaraby, Mohamed, Yang Zhong and Diane Litman (2023). Towards Argument-Aware Abstractive Summarization of Long Legal Opinions with Summary Reranking. arXiv preprint arXiv:2306.00672.

Ellsberg, Daniel (1961). "Risk, Ambiguity, and the Savage Axioms." Quarterly Journal of Economics **75**: 643-669.

Engel, Christoph and Johannes Kruse (2024). "Kommentar Ohne Autor. Können Sprachmodelle Das Kommentieren Übernehmen?" Juristenzeitung ***: ***.

Etner, Johanna, Meglena Jeleva and Jean-Marc Tallon (2012). "Decision Theory under Ambiguity." Journal of Economic Surveys **26**(2): 234-270.

European Court of Human Rights (2024). Guide on Article 11 of the European Convention on Human Rights.

Fabbri, Alexander R, Chien-Sheng Wu, Wenhao Liu and Caiming Xiong (2021). Qafacteval: Improved Qa-Based Factual Consistency Evaluation for Summarization. arXiv preprint arXiv:2112.08542.

Feigenbaum, Edward A and Herbert A Simon (1962). "A Theory of the Serial Position Effect." British Journal of Psychology **53**(3): 307-320.

Feijo, Diego de Vargas and Viviane P Moreira (2023). "Improving Abstractive Summarization of Legal Rulings through Textual Entailment." Artificial intelligence and law **31**(1): 91-113.

Gao, Wei, Shuai Yu, Yongbin Qin, Caiwei Yang, Ruizhang Huang, Yanping Chen and Chuan Lin (2025). "Lsdk-Legalsum: Improving Legal Judgment Summarization Using Logical Structure and Domain Knowledge." Journal of King Saud University Computer and Information Sciences **37**(1): 1-15.

Gao, Yunfan, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun and Haofen Wang (2023). Retrieval-Augmented Generation for Large Language Models: A Survey. arXiv preprint arXiv:2312.10997.

Goodfellow, Ian, Yoshua Bengio and Aaron Courville (2016). Deep Learning, MIT press.

Gu, Jiawei, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma and Honghao Liu (2024). A Survey on Llm-as-a-Judge. arXiv preprint arXiv:2411.15594.

Gupta, Shailja, Rajesh Ranjan and Surya Narayan Singh (2024). A Comprehensive Survey of Retrieval-Augmented Generation (Rag): Evolution, Current Landscape and Future Directions. arXiv preprint arXiv:2410.12837.

Inazu, John D (2010). "The Forgotten Freedom of Assembly." Tulane Law Review **84**: 565-612.

Jain, Deepali, Malaya Dutta Borah and Anupam Biswas (2021). "Summarization of Legal Documents: Where Are We Now and the Way Forward." Computer Science Review **40**: 100388.

James, Gareth, Daniela Witten, Trevor Hastie and Robert Tibshirani (2022). An Introduction to Statistical Learning, Springer.

Kapoor, Sayash, Peter Henderson and Arvind Narayanan (2024). Promises and Pitfalls of Artificial Intelligence for Legal Applications**:** 5.

Karpenstein, Ulrich and Franz C. Mayer (2022). Emrk. Konvention Zum Schutz Der Menschenrechte Und Grundfreiheiten. Kommentar. München, Beck.

Katz, Daniel Martin, Michael James Bommarito, Shang Gao and Pablo Arredondo (2024). "Gpt-4 Passes the Bar Exam." Philosophical Transactions of the Royal Society A **382**(2270): 20230254.

Kluwer, Wolters (2024). Effizientere Juristische Recherche Durch Generative Ki.

Kruse, Johannes (2025). The Ordinary Meaning Bot. Simulating Human Surveys with Llms. MPI Collective Goods Discussion Paper.

Kruse, Johannes and Christian Langner (2021). "Covid-19 Vor Gericht: Eine Quantitative Auswertung Der Verwaltungsgerichtlichen Judikatur." Neue Juristische Wochenschrift **74**: 3707-3712.

Laban, Philippe, Tobias Schnabel, Paul N Bennett and Marti A Hearst (2022). "Summac: Re-Visiting Nli-Based Models for Inconsistency Detection in Summarization." Transactions of the Association for Computational Linguistics **10**: 163-177.

Lewis, Charlie (2006). The Right of Assembly and Freedom of Association in the Information Age. Human Rights in the Global Information Society. R. F. Jorgensen**:** 151-184.

Lewis, Patrick, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih and Tim Rocktäschel (2020). "Retrieval-Augmented Generation for Knowledge-Intensive Nlp Tasks." Advances in Neural Information Processing Systems **33**: 9459-9474.

Lin, Chin-Yew (2004). Rouge: A Package for Automatic Evaluation of Summaries. Text Summarization Branches Out**:** 74-81.

Lin, Tianyang, Yuxin Wang, Xiangyang Liu and Xipeng Qiu (2022). "A Survey of Transformers." AI open **3**: 111-132.

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.

Liu, Nelson F, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni and Percy Liang (2024). "Lost in the Middle: How Language Models Use Long Contexts." Transactions of the Association for Computational Linguistics **12**: 157-173.

Liu, Yang, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu and Chenguang Zhu (2023). G-Eval: Nlg Evaluation Using Gpt-4 with Better Human Alignment. arXiv preprint arXiv:2303.16634.

Livermore, Michael A, Felix Herron and Daniel Rockmore (2024). "Language Model Interpretability and Empirical Legal Studies." Journal of Institutional and Theoretical Economics **180**: 244-276.

Mayer-Ladewig, Jens, Martin Nettesheim and Stefan von Raumer (2023). Emrk. Europäische Menschenrechtskonvention. Handkommentar. Baden-Baden, Nomos.

Mik, Eliza (2023). "Caveat Lector: Large Language Models in Legal Practice." Rutgers Business Law Journal **19**: 70-128.

Moro, Gianluca and Luca Ragazzi (2022). "Semantic Self-Segmentation for Abstractive Summarization of Long Documents in Low-Resource Regimes." Proceedings of the AAAI conference on artificial intelligence **36**(10): 11085-11093.

Murdock Jr, Bennet B (1962). "The Serial Position Effect of Free Recall." Journal of Experimental Psychology **64**(5): 482-488.

Nexis, Lexis (2023). Lexisnexis Launches Lexis+ Ai, a Generative Ai Solution with Linked Hallucination-Free Legal Citations.

Pont, Thiago Dal, Federico Galli, Andrea Loreggia, Giuseppe Pisano, Riccardo Rovatti and Giovanni Sartor (2023). Legal Summarisation through Llms: The Prodigit Project. arXiv preprint arXiv:2308.04416.

Ragazzi, Luca, Gianluca Moro, Stefano Guidi and Giacomo Frisoni (2024). "Lawsuit: A Large Expert-Written Summarization Dataset of Italian Constitutional Court Verdicts." Artificial Intelligence and Law: 1-37.

Reimann, Mathias (2020). Legal “Commentaries” in the United States. Division of Labor. Juristische Kommentare. Ein Internationaler Vergleich. D. Kästle-Lamparter, N. Jansen and R. Zimmermann. Tübingen, Mohr Siebeck**:** 277-294.

Reiter, Ehud (2018). "A Structured Review of the Validity of Bleu." Computational Linguistics **44**(3): 393-401.

Reuters, Thomson (2023). Introducing Ai-Assisted Research. Legal Research Meets Generative Ai.

Rights, European Court of Human (2024). Guide on Article 11 of the European Convention on Human Rights.

Rodgers, Ian, John Armour and Mari Sako (2023). "How Technology Is (or Is Not) Transforming Law Firms." Annual Review of Law and Social Science **19**(1): 299-317.

Sahoo, Pranab, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal and Aman Chadha (2024). A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications. arXiv preprint arXiv:2402.07927.

Salát, Orsolya (2015). The Right to Freedom of Assembly.

Santosh, TYS, Mahmoud Aly and Matthias Grabmair (2024). Lexabsumm: Aspect-Based Summarization of Legal Decisions. arXiv preprint arXiv:2404.00594.

Santosh, TYS, Mahmoud Aly, Oana Ichim and Matthias Grabmair (2025). Lexgenie: Automated Generation of Structured Reports for European Court of Human Rights Case Law. arXiv preprint arXiv:2503.03266.

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). "On the Open Texture of Law." Grazer Philosophische Studien **87**: 197-215.

Sharma, Saloni and Piyush Pratap Singh (2025). "Advancements in Legal Text Summarization: Integrating Inlegalbert for Effective Extractive Summarization." International Journal of System Assurance Engineering and Management **16**(4): 1382-1397.

Shukla, Abhay, Paheli Bhattacharya, Soham Poddar, Rajdeep Mukherjee, Kripabandhu Ghosh, Pawan Goyal and Saptarshi Ghosh (2022). Legal Case Document Summarization: Extractive and Abstractive Methods and Their Evaluation, arXiv.

Songer, Donald R (2011). U.S. Courts of Appeals Databases 1925-2011.

Spaeth, Harold, Lee Epstein, Ted Ruger, Sarah C. Benesh, Jeffrey A Segal and Andrew D Martin (2023). "Supreme Court Database Code Book."

Trozze, Arianna, Toby Davies and Bennett Kleinberg (2024). "Large Language Models in Cryptocurrency Securities Cases: Can a Gpt Model Meaningfully Assist Lawyers?" Artificial Intelligence and Law: 1-47.

Turner, John R, Jeff Allen, Suliman Hawamdeh and Gujjula Mastanamma (2023). "The Multifaceted Sensemaking Theory: A Systematic Literature Review and Content Analysis on Sensemaking." Systems **11**(3): 145.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser and Illia Polosukhin (2017). Attention Is All You Need. Advances in Neural Information Processing Systems.

Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le and Denny Zhou (2022). "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models." Advances in neural information processing systems **35**: 24824-24837.

Weick, Karl E. (1995). Sensemaking in Organizations. Thousand Oaks, Sage Publications.

Wu, Ning, Ming Gong, Linjun Shou, Shining Liang and Daxin Jiang (2023). "Large Language Models Are Diverse Role-Players for Summarization Evaluation." CCF international conference on natural language processing and Chinese computing: 695-707.

Xu, Huihui and Kevin Ashley (2023). Question-Answering Approach to Evaluating Legal Summaries. arXiv preprint arXiv:2309.15016.

Zhang, Junhao, Richong Zhang, Fanshuang Kong, Ziyang Miao, Yanhan Ye and Yaowei Zheng (2025). Lost-in-the-Middle in Long-Text Generation: Synthetic Dataset, Evaluation Framework, and Mitigation. arXiv preprint arXiv:2503.06868.

Zheng, Lianmin, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li and Eric Xing (2023). Judging Llm-as-a-Judge with Mt-Bench and Chatbot Arena. Advances in neural information processing systems. **36:** 46595-46623.

Zou, Xinrui, Ming Zhang, Nathaniel Weir, Benjamin Van Durme and Nils Holzenberger (2024). Reframing Tax Law Entailment as Analogical Reasoning. arXiv preprint arXiv:2401.06715.

Appendix

Prompts
Prompt for summarizing individual ruling

This is case <number>

INTRODUCTION
Task

I would like to help lawyers who have to decide on the legality of assemblies. These lawyers typically do not have the time to read all the relevant decisions of the European Court of Human Rights (and the former Human Rights Committee) themselves before making their decision. It is therefore very important that my readers can rely on my summary of the case law.

(In what follows, I will simplify the exposition by only speaking of the "court". Please read this as a shorthand that also encompasses the former Human Rights Committee).

To prepare this summary, I would like to proceed in two steps. In the first step, I will ask you to work out, separately for each decision in which the court mentions freedom of assembly as protected by Article 11 of the European Convention on Human Rights, the court's statements on the interpretation of this fundamental freedom. In the second step, I will then combine these individual summaries into a single text. This question is limited to the first step.

Structure of the raw data

In the following I quote the full text of a decision by the European Court on Human Rights. I am only interested in how the court itself interprets freedom of assembly. The part of the decision in which the court expresses and justifies its own interpretation is directly relevant.

The court often also reports on the views of the parties (and any other third parties that have been heard). In addition, the court often gives reasons as to whether it considers the application to be admissible. If the complainant applies for an interim measure, the court regularly reports on whether the complainant's interest in provisional protection outweighs the sovereign's interest in the provisional effectiveness of its decision. These passages of the ruling can be important for understanding the court's own opinion. This is particularly the case when the court formulates its opinion in disagreement with the opinion of one of the parties. As such, however, I am not interested in the opinions of other parties to the proceedings.

Judges are authorized to formulate a separate or concurring opinion. If a judge makes use of this power, this opinion is stated at the end of the decision. Please consider individual opinions only to the extent that they indicate how the decision of the majority of the court is to be understood.

The decision of the European Court of Human Rights often depends on more than one fundamental freedom. I am only interested in the interpretation of Article 11 of the European Convention on Human Rights, and only on the interpretation of freedom of assembly (not of freedom of association, which is also protected by the same article). However, the interpretation of other fundamental freedoms can be relevant in order to understand the court's statements on freedom of assembly, as protected by Article 11 of European Convention on Human Rights.

Please proceed in two steps: in the first step, decide which parts of the text of the decision are at all relevant to how the court itself interprets freedom of assembly. In the second step, summarize how the court interprets freedom of assembly.

Please note: the decision is available on the court's website for download and is not protected by copyright.

Structure of the doctrine of fundamental freedoms

In the following, I explain the structure of the doctrine of fundamental freedoms as it has been developed in the jurisprudence of the European Court of Human Rights. These explanations should make it easier for you to understand the ruling. However, please only answer the numbered questions formulated below.

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, i.e. who can invoke the fundamental freedom against the state.

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. On the other hand, the applicant may not have suffered a sufficiently significant disadvantage, in the sense of Art. 35 III b of the Convention.

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. This test consists of the following elements
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, namely
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

Provided the state act serves a legitimate aim, it must, on balance, be necessary in a democratic society. In this regard, the state enjoys a certain margin of appreciation, but its scope depends, inter alia, on the severity of the intervention, and on the normative relevance of the legitimate aim.

IV. positive obligation

The European Convention on Human Rights not only requires that member states refrain from unjustified interferences. The members state may also be required to proactively intervene on behalf of those meant to enjoy their freedom of assembly.

V. just satisfaction

If the Court finds that the state has violated freedom of assembly, it may decide whether the applicant should be granted just satisfaction, in the sense of Art. 41 of the Convention.

TEXT OF INDIVIDUAL CASE

<full text>

END CASE TEXT

RESPONSE FORMAT
Summarization

Please answer all of the following questions. For each case document, please first determine whether the European Court of Human Rights discusses the respective question in this decision. If so,

a) please summarize how the court interprets the respective element of doctrine in general, and how it motivates this element,

b) please summarize which concrete features of the case are relevant for (not) bringing the case under the rubric of the respective element of doctrine.

Questions

Here is the list of elements of the doctrine to which the ruling might speak, and which I ask you to assess: does the court address this element of the doctrine?

1. substantive scope

Does the governmental act about which the applicant complains come under the substantive purview of freedom of assembly?

2. personal scope

Are the individuals or institutions who complain about the governmental act protected by freedom of assembly?

3. interference

Does the governmental act interfere with freedom of assembly?

4. prescribed by law: accessibility

Is the legal rule that empowered government to intervene sufficiently accessible?

5. prescribed by law: foreseeability

Has the intervention been sufficiently foreseeable, given the legal rule that empowered government?

6. prescribed by law: safeguards against abuse

Does the legal rule that empowered government come with sufficient safeguards against abuse?

7. legitimate aim: national security

Is the intervention legitimately motivated by the desire to protect national security?

8. legitimate aim: public safety

Is the intervention legitimately motivated by the desire to protect public safety?

9. legitimate aim: prevention of disorder

Is the intervention legitimately motivated by the desire to prevent disorder?

10. legitimate aim: prevention of crime

Is the intervention legitimately motivated by the desire to prevent crime?

11. legitimate aim: protection of health

Is the intervention legitimately motivated by the desire to protect health?

12. legitimate aim: protection of morals

Is the intervention legitimately motivated by the desire to protect morals?

13. legitimate aim: protection of the rights and freedoms of others

Is the intervention legitimately motivated by the desire to protect the rights and freedoms of others?

14. necessary in a democratic society

Is the intervention considered necessary in a democratic society? Does the court discuss the member state's margin of appreciation?

15. positive obligation

Has government been under the positive obligation to protect the applicant against interference with their freedom of assembly?

16. just satisfaction

Did the court discuss whether the applicant should receive just satisfaction?

17. separate opinion

Have one or several judges added a separate opinion?

Format and length of answer

Coverage

Please comment on each of the 17 points. Please also indicate if the court makes no statement on this issue in the ruling.

The length of your response should be proportionate to the level of detail with which the court discusses freedom of assembly.

Numbering of paragraphs to be cited

Segments of the rulings are numbered. Numbers are always in the format

<new paragraph>
<number>.
<text of the paragraph>

Many of the rulings posted on the website of the Court did not originally have paragraphs numbered, or had a more idiosyncratic way of numbering. I have harmonised numbering. If there had been original numbers that I no longer want to use, they now feature in the following format:

"<number>"

Hence numbers to be ignored are enclosed in "< >". Please ignore them in your summary, and only refer to the new numbers.

Please list the relevant number of the paragraph after your summary of the statement, in the following format:

<your statement> (number)

Hence enclose the number in round brackets.

If a statement covers more than one paragraph, please list them in the following format if paragraphs are consecutive

<your statement> (first_number, last_number)

and if paragraphs are not consecutive

<your statement> (number1, number2 ..., numberN)

Reasoning: Response in Natural Language

Please answer all questions in normal language first.

!! IMPORTANT !!

1. DO NOT REPEAT THE PROMPT. ONLY REPORT YOUR REASONING, AND YOUR ASSESSMENT
2. In a later step, I want to write a document that collects <case_number> and <paragraph_number>, separately for each element of doctrine. The code I am using to write this document works with a regular expression. The regular expression expects

(<paragraph_number1>, <paragraph_number2>, ... <paragraph_numberN>)

Hence the reference to paragraphs must be included in brackets, and it must ONLY contain paragraph numbers, separated by commas. Please do NOT add any further information to these references, like

- "para"
- "paragraph"
- "page"
- "§"
- "Appendix"
- "appended table"
- "operative part"

Within these references, do not add any explanations.
Do only cite paragraphs by their number, not any subheadings (like "a" or "i")

If you want to add any information to your explanation in natural language, and want to put some text into brackets, please consistently use SQUARE brackets, not ordinary round brackets.

JSON

Finally, please also summarize your answers in JSON format (but do not exclusively report JSON output; always also summarize the decision of the court, on each element of the doctrine, in natural language).

For the JSON output, please follow the pattern below:

"Yes" OR "No" always means: has the court explicitly discussed the respective element of the doctrine of freedom of assembly?

For the placeholder "<key>" at the top of the JSON output, please use the key that is used in the first line of this prompt. Hence if this first line reads "case_12", then please write as first line in the JSON output "case_12".

In a later step, I want to extract the JSON block into a separate file. Therefore the beginning of this section should be marked with "**JSON_DATA_START**", and the end should be marked with "**JSON_DATA_END**"

JSON_DATA_START

```
{
  "case_<key_from_first_line_of_prompt>": [
   {
    "substantive_scope": "Yes" OR "No",
    "personal_scope": "Yes" OR "No",
    "interference": "Yes" OR "No",
    "accessibility": "Yes" OR "No",
    "foreseeability": "Yes" OR "No",
    "safeguard_against_abuse": "Yes" OR "No",
    "national_security": "Yes" OR "No",
    "public_safety": "Yes" OR "No",
    "prevention_of_disorder": "Yes" OR "No",
    "prevention_of_crime": "Yes" OR "No",
    "protection_of_health": "Yes" OR "No",
    "protection_of_morals": "Yes" OR "No",
    "protection_of_rights_and_freedoms_of_others": "Yes" OR "No",
    "necessary_in_a_democratic_society": "Yes" OR "No",
    "positive_obligation": "Yes" OR "No",
    "just_satisfaction": "Yes" OR "No",
    "separate_opinion": "Yes" OR "No"
  }
  ]
}
```

JSON_DATA_END

Prompt for summarizing element of doctrine

This is the raw text for <element of doctrine>

INTRODUCTION
Task

I would like to help lawyers who have to decide on the legality of assemblies. These lawyers typically do not have the time to read all the relevant decisions of the European Court of Human Rights (and the former Human Rights Committee) themselves before making their decision. It is therefore very important that my readers can rely on my summary of the case law.

(In what follows, I will simplify the exposition by only speaking of the "court". Please read this as a shorthand that also encompasses the former Human Rights Committee).

To prepare this summary, I am proceeding in three steps. In the first step, I have asked you to work out, separately for each decision in which the court mentions freedom of assembly as protected by Article 11 of the European Convention on Human Rights, the court's statements on the interpretation of this fundamental freedom. These per case summaries reference each statement with the number of the paragraphs on which your summary draws. In the second step, separately for each element of the established doctrine of freedom of assembly, I have compiled a document with the paragraphs from all rulings that you had singled out to discuss the respective element of doctrine. In this third step, I am asking you to examine this document, for the element of doctrine defined at the beginning of this prompt.

Specifically, I want you to formulate a new text that achieves the following:
a) how does the court define the element of doctrine?
Keep in mind: it is possible that the definition has changed over time, or that it has become more elaborate. Then please do not only report the latest version, but the development over time as well. Note that cases are reported in chronological order. Hence the higher the case_number, the later the decision has been taken.
b) in which ways has the court applied its definition? Which are the reported features of the case to which the court refers when deciding that the element of doctrine is (or is not) fulfilled? Such detail is important for my readers as it allows them to compare their own case with the features of decided cases, and to predict how the court would decide that new case.

Things to keep in mind

1. character of the quotes from the rulings
In the first step of this pipeline, you have (rightly) been inclusive. This has two implications:
a) read all quotes from a ruling in conjunction. They have often been split over multiple paragraphs, and can sometimes only be understood when reading them in the light of earlier or later paragraphs (from that same ruling).
b) Not so rarely, not all cited paragraphs are really important for extracting either the definition of the element of doctrine, or the specifics of its application to the case in question. Hence in the document you are now writing, do not feel pressed to include everything that is part of the raw data. Rather be concise.

2. coverage
It happens that the court discusses more than one element of doctrine, or even more than one human right. If this is the case, please ONLY report what helps with the interpretation of the element of doctrine (of freedom of assembly) that is mentioned on top of this prompt.

Let me explain why: at a later point, I will also ask you to write the analogous summary for the remaining elements of the doctrine of freedom of assembly. Yet for the users of the commentary that I want to write with your help, it is important that they can zero in on the element of doctrine that is critical for their individual case.

Structure of the doctrine of fundamental freedoms

As this may help you understand which (elements of) a quote truly address the element of doctrine mentioned on top of this prompt, let me remind you of the structure of the doctrine of fundamental freedoms as it has been developed in the jurisprudence of the European Court of Human Rights.

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, i.e. who can invoke the fundamental freedom against the state.

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. On the other hand, the applicant may not have suffered a sufficiently significant disadvantage, in the sense of Art. 35 III b of the Convention.

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. This test consists of the following elements
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, namely
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

Provided the state act serves a legitimate aim, it must, on balance, be necessary in a democratic society. In this regard, the state enjoys a certain margin of appreciation, but its scope depends, inter alia, on the severity of the intervention, and on the normative relevance of the legitimate aim.

IV. positive obligation

The European Convention on Human Rights not only requires that member states refrain from unjustified interferences. The members state may also be required to proactively intervene on behalf of those meant to enjoy their freedom of assembly.

V. just satisfaction

If the Court finds that the state has violated freedom of assembly, it may decide whether the applicant should be granted just satisfaction, in the sense of Art. 41 of the Convention.

RAW TEXT

END RAW TEXT

RESPONSE FORMAT
Summarization

a) Please summarize how the court interprets the respective element of doctrine in general, and how it motivates this element. Which purpose is it meant to serve, according to the jurisprudence of the court? Does the court discuss alternative interpretations? If the interpretation has changed over time, please report the different versions, and how they have evolved (recall that case numbers reflect temporal order: later cases have higher case numbers).

b) Please summarize how the court has applied its definition of the element of doctrine to individual cases. For which situations of life the element of doctrine has been fulfilled, and for which it has not? Please be as specific as possible. The more the ruling has been detailed, the more report. If possible, do not just write a long list of more or less relevant cases. Such a list is much less helpful for users than the structured report of detail. Please do order this summary by substance matter, not by case_numbers.

References to paragraphs of cases

For every statement in your summary, add a reference to one or more paragraphs of court cases. Please add these references to the end of the statement, in the format

(<case_number>_<paragraph_number>)

If applicable enclose multiple references in these brackets.

What I do not want

Let me use an example to explain what I do not want: when asked about "interference", in an earlier attempt, I have received the following response:

"The European Court of Human Rights (ECtHR) defines "interference" with the right to freedom of assembly as any action by public authorities that restricts or impedes the exercise of this right. This includes measures taken before, during, or after an assembly, such as bans, dispersals, arrests, and punitive measures (1_97, 2_103, 2_239, 3_93, 4_71, 5_56, 5_87, 6_81, 6_88, 6_89, 7_81, 7_82, 9_242, 9_243, 10_2, 10_38, 10_39, 10_40, 10_41, 11_94, 12_97, 12_98, 13_86, 14_88, 15_66, 16_88, 18_99, 19_73, 22_48, 22_49, 22_50, 22_125, 23_114, 24_108, 24_109, 25_106, ..."

This is not sufficiently helpful. The first sentence essentially repeats the doctrinal element. The second sentence lumps a multitude of potential interferences together. Users do not learn anything about the specifics of the cases that the court has considered to interfere with freedom of assembly. They do not learn about the conditions under which the court has considered the governmental act NOT to interfere with freedom of intervention.

Also the summary should be much more detailed. Several hundred snippets should not just be summarized by a single paragraph (even if the paragraph is a long one).

Prompts for evaluating structure

You are an expert legal evaluator specializing in human rights law and ECHR jurisprudence. Provide precise numerical scores following the given criteria.

Rate the doctrinal completeness of this Article 11 EMRK commentary/guide structure.

Check if it covers:
- Freedom of assembly (all aspects)
- Positive and negative obligations (detailed)
- All interference types
- Complete three-part justification test
- Margin of appreciation doctrine
- All procedural aspects
- Historical evolution
- Comparative analysis

Structure:
{structure}

Note: Academic legal commentary REQUIRES exhaustive coverage.
More sections and details = Better completeness.
Evaluate the completeness (1=very incomplete, 5=fully comprehensive)

Rate the hierarchical organization of this Article 11 EMRK commentary/guide structure.

Assess:
- Logical flow (applicability → interference → justification) even with extensive detail
- Clear structure despite comprehensive coverage
- Proper nesting of numerous subtopics
- Management of complex legal relationships
- Systematic organization of multiple levels

Structure:
{structure}

Note: Complex, multi-level hierarchies are EXPECTED and POSITIVE.
Do NOT penalize depth.
Evaluate the hierarchical organization (1=poorly organized, 5=excellently organized):

Rate the coherence and logical flow of this Article 11 EMRK commentary/guide structure.

Consider:
- Progressive building from basic to complex
- Smooth transitions between sections
- Internal consistency
- Comprehensive treatment maintains logical thread
- Detailed analysis remains coherent

Structure:
{structure}

Evaluate the coherence (1=incoherent, 5=highly coherent)

Rate the granularity balance of this Article 11 EMRK commentary/guide structure.

Evaluate:
- Appropriate level of detail
- Neither too superficial nor overly fragmented
- Practical usability

Structure:
{structure}

Evaluate the granularity balance (1=poorly balanced, 5=optimally balanced):

As a human rights lawyer, rate how easily you can navigate this structure to find:
- Admissibility requirements
- Relevant precedents
- Argumentation strategies

Structure:
{structure}

Evaluate navigability for lawyers (1=very difficult, 5=very easy)

As an ECHR judge, rate how easily you can navigate this structure to find:
- Established principles
- Doctrinal evolution
- Consistency requirements

Structure:
{structure}

Evaluate navigability for judges (1=very difficult, 5=very easy)

As someone affected by Article 11 issues, rate how easily you can understand:
- Your rights
- What authorities can/cannot do
- Available remedies

Structure:
{structure}

Evaluate accessibility for affected parties (1=very difficult, 5=very easy)

Prompts for evaluating content

You are an expert legal evaluator specialized in ECHR jurisprudence and G-Eval methodology. Follow the structured evaluation steps precisely and provide numerical scores only.

SCORING INSTRUCTIONS - Use the full 1-5 scale:

Score 5: Exceptional - Reference-quality content (rare)
Score 4: Good - Above average with minor gaps
Score 3: Adequate - Meets basic requirements (most common)
Score 2: Below average - Significant deficiencies
Score 1: Poor - Fundamental problems

Respond with ONLY a number between 1.0 and 5.0 (e.g., 3.2)

Evaluate Doctrinal Relevance

You will assess how well the content aligns with its section heading in an Article 11 EMRK commentary/guide.

Evaluation Criteria:
Doctrinal Relevance (1-5) - The degree to which the content addresses the doctrinal element indicated by the heading.

Evaluation Steps:
1. Identify the key doctrinal concept(s) in the heading: "{heading}"
2. Check if the content addresses these concepts comprehensively
3. Assess whether all content is relevant to the heading
4. Identify any missing elements that should be covered
5. Check for off-topic digressions
6. Assign a score from 1 to 5

Heading: {heading}

Content:
{content_truncated}

Score (1.0-5.0)

Evaluate Content Organization

Evaluation Criteria:
Content Organization (1-5) - The logical progression of ideas and systematic presentation.

Evaluation Steps:

1. Identify the organizational structure used
2. Check for logical progression of ideas
3. Assess transition quality between paragraphs
4. Verify systematic presentation of principles
5. Evaluate overall coherence
6. Assign a score from 1 to 5

Heading: {heading}

Content:
{content_truncated}

Score (1.0-5.0)

Evaluate Legal Precision

Evaluation Criteria:
Legal Precision (1-5) - Correct use of legal terminology and accurate statement of principles.

Evaluation Steps:
1. Identify all legal terms and principles stated
2. Verify correct usage of legal terminology
3. Check accuracy of legal principles
4. Assess precision of case law characterization
5. Identify any ambiguous statements
6. Assign a score from 1 to 5

Heading: {heading}

Content:
{content_truncated}

Score (1.0-5.0)

Evaluate Completeness

Evaluation Criteria:
Completeness (1-5) - The extent to which all relevant aspects of the topic are addressed.

Evaluation Steps:
1. Identify what should be covered under heading: "{heading}"
2. Check coverage of fundamental principles
3. Assess treatment of exceptions and special cases
4. Verify inclusion of procedural aspects if relevant
5. Check if landmark cases are included
6. Check for recent jurisprudential developments

7. Assign a score from 1 to 5

Heading: {heading}

Content:
{content_truncated}

Score (1.0-5.0)

Evaluate Citation Quality

Evaluation Criteria:
Citation Quality (1-5) - Appropriateness and accuracy of case law citations.

Evaluation Steps:
1. Identify all case citations in the content
2. Assess relevance of cited cases to the topic
3. Check if landmark cases are included
4. Verify accuracy of case characterization
5. Evaluate balance between old and recent cases
6. Assign a score from 1 to 5

Heading: {heading}

Content:
{content_truncated}

Score (1.0-5.0)

Evaluate Content Utility for {persona_data['description']}

You are evaluating from the perspective of: {persona_data['description']}
Your primary needs: {', '.join(persona_data['primary_needs'])}

Evaluation Criteria:
Practical Utility (1-5) - How well the content serves your professional needs.

Evaluation Steps:
1. Assess if content addresses your primary needs
2. Check for actionable insights relevant to your work
3. Evaluate clarity for your expertise level
4. Determine time-saving value
5. Consider practical applicability
6. Assign a score from 1 to 5

Heading: {heading}

Content:
{content_truncated}

Score (1.0-5.0)

```
PERSONAS = {
   'lawyer': {
      'description': 'Human rights lawyer preparing ECHR applications',
      'primary_needs': [
         'Identifying viable legal arguments',
         'Finding supporting precedents quickly',
         'Understanding procedural requirements',
         'Assessing success probability'
      ]
   },
   'judge': {
      'description': 'ECHR judge reviewing Article 11 cases',
      'primary_needs': [
         'Doctrinal consistency verification',
         'Evolution of jurisprudence',
         'Distinguishing factors between cases',
         'Systematic interpretation'
      ]
   },
   'affected_party': {
      'description': 'Assembly organizer/participant or public authority',
      'primary_needs': [
         'Understanding rights and obligations',
         'Assessing legality of measures',
         'Practical compliance guidance',
         'Risk assessment'
      ]
   }
}
```

Prompts for evaluating references

You are an expert legal evaluator specializing in ECHR jurisprudence. Always respond with valid JSON.

Evaluate how much NEW information each additional paragraph provides beyond what previous paragraphs already covered.

Rate the information value (0-1):
- Does it provide direct support for the statement?
- How specific and relevant is the information?

Return JSON: {{"score": <0-1>, "info": "brief description"}}"""

Rate the ADDITIONAL information value (0-1):
- What NEW support does this add?
- Consider factual diversity of underlying cases

Return JSON: {{"score": <0-1>, "info": "brief description"}}

Evaluate citation with multiple evaluations and averaging

Evaluate how well this court paragraph supports the statement.

STATEMENT: "{statement}"

CITATION: {citation}

COURT PARAGRAPH: "{court_paragraph[:1000]}"

Evaluate each criterion (1-5 scale):

1. **RELEVANCE**: Does the court paragraph directly address the legal point made in the statement? Consider if the legal principle, rule, or finding in the paragraph matches what the statement claims.

2. **ACCURACY**: Does the paragraph accurately support the specific claim? Check if the statement correctly represents what the court actually said without distortion or misrepresentation.

3. **COMPLETENESS**: Does the paragraph provide complete support for all aspects of the statement? Consider whether key elements of the statement are fully covered or if important parts lack support.

4. **SPECIFICITY**: How specific and precise is the support? Generic or vague connections score low; exact, detailed matches score high.

5. **CONTEXT INTEGRITY**: Is the paragraph's meaning preserved when used for this statement? Check if taking this paragraph out of its original context changes or distorts its legal meaning.

Return JSON:
{{
  "relevance": <1-5>,
  "accuracy": <1-5>,
  "completeness": <1-5>,

"specificity": <1-5>,
"context_integrity": <1-5>,
"overall_score": <weighted average>,
"justification": "Brief explanation"

Frequency and content of retrieved chunks when predicting case outcomes with the help of a RAG that accesses a commentary

1. **Guide**

chunk	freq	header
21	55	Peaceful assembly
14	54	Assembly as a form of expression and expression of opinion during assembly: Articles 10 and 11
24	40	Scope of the right to freedom of assembly
3	38	Dispersal and the use of force
12	34	Importance of the right to freedom of peaceful assembly and its link with the right to freedom of expression
20	25	Interference with the exercise of the right to freedom of assembly
5	10	Unlawful assembly
7	10	Narrow margin of appreciation for interference based on the content of views expressed during an assembly
15	4	Necessary in a democratic society
4	2	Spontaneous assembly
7	2	Narrow margin of appreciation for interference based on the content of views expressed during an assembly
17	2	Classification of complaints under Articles 9, 10 and/or 11: Religious meetings
16	1	Legitimate aim

2. **Gemini**

chunk	freq	header
7	54	prevention of crime
14	54	foreseeability
5	53	separate opinion
8	49	personal scope
12	45	positive obligation
6	13	substantive scope
1	3	protection of rights and freedoms of others
4	2	prevention of disorder
10	1	protection of morals
11	1	accessibility

3. **GPT**

chunk	freq	header
1	54	protection of rights and freedoms of others
14	54	foreseeability
16	45	protection of health
0	42	interference

7	35	prevention of crime
2	20	public safety
15	14	national security
11	5	accessibility
12	3	positive obligation
4	1	prevention of disorder
6	1	substantive scope
13	1	safeguard against abuse

4. Kimi

chunk	freq	header
0	54	interference
9	52	necessary in a democratic society
3	50	just satisfaction
2	40	public safety
5	24	separate opinion
8	24	personal scope
14	12	foreseeability
7	11	prevention of crime
1	4	protection of rights and freedoms of others
12	4	positive obligation