

SCALE: Structured Case-law-guided Anticipatory Legal Reasoning

Mann Khatri^{1*}, Santosh T.Y.S.S.², Oana Ichim³

Rajiv Ratn Shah⁴, Ponnurangam Kumaraguru⁵

¹Indraprastha Institute of Information Technology, Delhi ²Amazon, Seattle, U.S.A

³Graduate Institute of International and Development Studies, Geneva, Switzerland

⁴Indian Institute of Technology, Kanpur

⁵International Institute of Information Technology, Hyderabad

mannk@iiitd.ac.in, santoshtyss@gmail.com, oana.ichim@graduateinstitute.ch

rajivrtn@iitk.ac.in, pk.guru@iiit.ac.in

Abstract

LLMs have shown promise in legal NLP, but generating legal reasoning that is factually grounded, faithful to precedent, and consistent with judicial analysis remains difficult. Retrieval-augmented generation is often used to improve grounding, yet standard legal RAG faces two limitations: retrieved materials are hard to synthesize, and retrieval is driven by surface similarity rather than the latent legal issues and factual considerations that govern the analysis. We address these limitations with **SCALE** (Structured Case-law-guided Anticipatory Legal Reasoning), a framework for precedent-grounded legal reasoning. SCALE converts article-specific Case Law Guides into an *Issue-Factor Hierarchy*, identifies the issues and factors implicated by a case, retrieves the linked guide passages, and generates the final legal analysis. Across Llama-3.1 and Qwen3 models, expert evaluation shows that SCALE consistently outperforms prompt-only and retrieval-based baselines, with the largest gains in legal issue identification and precedent use. These gains are especially strong for smaller models, suggesting that explicit legal structure can compensate for weaker implicit reasoning capacity. We also compare expert assessment with an LLM-as-a-Judge framework and find only weak-to-moderate agreement, underscoring the limits of automated evaluation for high-stakes legal reasoning.

1 Introduction

Large language models have shown promise across legal NLP tasks, including legal judgment prediction (Wang et al., 2024; Shui et al., 2023; Jiang and Yang, 2023a; Deng et al., 2023; Santosh et al., 2024b), legal summarization (Bauer et al., 2023; Agarwal et al., 2022; Ash et al., 2024; Santosh et al., 2025b, 2024c,a), and legal case retrieval (Ma et al., 2024; Zhou et al., 2023; Mahari et al., 2024; Tyss

et al., 2024; Upadhyaya and Santosh, 2025), but generating legal reasoning that is factually grounded, faithful to precedent, and consistent with judicial analysis remains a substantial challenge (T.Y.S.S et al., 2024; Santosh et al., 2025a). In particular, when asked to explain whether a legal provision has been violated in light of a case’s facts, a model must identify the relevant legal questions, determine which aspects of the factual record are legally salient, and connect those facts to prior authority (Wang et al., 2024; Ye et al., 2018). Performing all of these steps implicitly within a single generation often leads to shallow comparisons, unsupported abstractions, or hallucinated precedent.

A common mitigation strategy is retrieval-augmented generation (RAG) (Lewis et al., 2020), in which prior cases or legal materials are retrieved and supplied as additional context (Cui et al., 2023; Wu et al., 2023; Ryu et al., 2023; Chouhan and Gertz, 2024). However, current RAG pipelines face two important limitations in legal reasoning settings. First, retrieved legal materials are typically long, heterogeneous, and not organized around the specific reasoning task, forcing the model to filter, synthesize, and align multiple sources during generation (Hou et al., 2025; Weidinger et al., 2025). This places a substantial burden on the model, particularly when the retrieved context mixes relevant discussion with irrelevant detail. Second, retrieval itself is usually driven by similarity between the case facts and candidate documents, even though legal relevance often depends on latent legal issues and factual considerations rather than surface resemblance (Deng et al., 2024; Yao et al., 2025; Santosh et al., 2025c). What must be retrieved is not merely text similar to the facts, but authority addressing the issues and factors that will govern the downstream analysis.

Recent work has begun to move beyond purely reactive, similarity-based retrieval by showing that retrieval can be improved when it is guided by legal

concepts, anticipated content, or other intermediate representations that better reflect the reasoning needs of the task (T.y.s.s et al., 2025a, 2022; T.y.s.s and Hernandez, 2025). This suggests that effective legal retrieval should be *anticipatory*: instead of retrieving only from the surface form of the input, the system should first identify the issues or factual dimensions likely to matter for the final analysis, and then retrieve materials aligned with those needs. However, in jurisdiction-specific legal reasoning, this remains a strong assumption. The relevant concepts are often tied to article-specific standards, fact-sensitive tests, and evolving case law, and may not be consistently recoverable from the model alone, especially in smaller models.

In this work, we address this limitation by externalizing that legal structure into an explicit representation that guides both retrieval and generation. We operationalize this idea in the domain of the European Court of Human Rights (ECHR) by leveraging *Case Law Guides*, which summarize how Convention articles have been interpreted across the Court’s jurisprudence, identify recurring legal questions, and highlight the factual considerations relevant to resolving them, together with references to supporting paragraphs from prior judgments. These guides provide a structured source of precedent that is more closely aligned with legal reasoning than raw case fragments alone.

Building on this source of structured precedent, we introduce **SCALE** (Structured Case-law-guided Anticipatory Legal Reasoning), a framework for precedent-grounded legal reasoning generation. SCALE converts article-specific Case Law Guides into an **Issue–Factor Hierarchy**, where *issues* represent the core legal questions arising under a Convention article and *factors* capture recurring factual considerations relevant to resolving those questions. At inference time, given the facts of a new case and an alleged article, SCALE identifies the relevant issues and factors from the IFH, then retrieves the corresponding guide passages, and finally generates a legal analysis grounded in this structured context. By organizing precedent around legal issues and recurring factual factors, SCALE makes retrieval both more targeted and easier to synthesize than standard similarity-based RAG.

We evaluate SCALE using models from the Llama-3.1 and Qwen3 families, spanning both smaller and larger parameter scales. A legal expert evaluates generated reasoning traces along six dimensions: identification of the appropriate

legal issues, factual faithfulness, coherence and logical consistency, relevance of the precedent invoked, correctness of precedent application, and alignment with the ground-truth judicial reasoning. Across these dimensions, SCALE improves over baselines based on implicit model reasoning, standard RAG, retrieval from Case Law Guides without explicit synthesis into an issue–factor hierarchy, and anticipation-only retrieval variants. Notably, the gains are especially pronounced for smaller models, which benefit more from structured legal scaffolding and in some cases become comparable to substantially larger models. To complement expert assessment, we also employ an LLM-as-a-Judge evaluation framework. While automated evaluation provides a useful scalable signal, we find only weak positive correlation with expert judgments, indicating that LLM-based evaluators capture some aspects of reasoning quality but remain limited for assessing fine-grained legal analysis.

Our contributions are threefold:

- We identify two core limitations of current legal RAG for reasoning tasks: the difficulty of synthesizing dense, unstructured retrieved materials, and the mismatch between similarity-based retrieval and the latent issue- and factor-level structure required for legal analysis.
- We propose **SCALE**, a structured legal reasoning framework that externalizes article-specific legal structure from ECHR Case Law Guides into an Issue–Factor Hierarchy and uses it to support anticipatory, issue-aligned, and factor-specific retrieval.
- We provide a multi-model evaluation combining expert assessment and LLM-as-a-Judge analysis, showing that structured retrieval improves the quality of legal reasoning traces—especially for smaller models—while also highlighting the limits of automated evaluation in this domain.

2 Related Work

Legal judgment prediction has a long history, from early rule-based and quantitative approaches (Lawlor, 1963; Nagel, 1960; Ulmer, 1963; Keown, 1980) to recent machine learning methods, including traditional classifiers (Aletras et al., 2016; Medvedeva et al., 2018), deep neural models (Luo et al., 2017; Zhong et al., 2018; Khatri et al., 2023; Lyu et al., 2023), and transformer-based systems

with legal-domain pretraining (Liu et al., 2022; Chalkidis et al., 2019, 2020; Masala et al., 2024). While these methods improved predictive performance, much of this work remained focused on outcomes rather than faithful legal reasoning, motivating efforts on rationale extraction and explanation generation (Malik et al., 2021; Chalkidis et al., 2021; Feng et al., 2022; Zhong et al., 2020; Wu et al., 2022). However, subsequent work showed that legal models can exploit spurious correlations and shortcuts, raising doubts about whether such explanations reflect genuine legal reasoning (T.y.s.s et al., 2022; Xu et al., 2023).

The rise of LLMs shifted attention toward richer legal analysis and prompted work on structured prompting, including Chain-of-Thought, IRAC-style prompting, syllogistic reasoning, and prompts enriched with legal definitions or rhetorical roles (Shui et al., 2023; Shu et al., 2024; Wei et al., 2022; Yu et al., 2023; Kang et al., 2023; Jiang and Yang, 2023b; Deng et al., 2023; Khatri et al., 2025). Yet LLM-generated reasoning remains vulnerable to hallucination and unsupported analysis (Dahl et al., 2024). To improve grounding, recent work has increasingly adopted retrieval-augmented generation in legal settings, including precedent-aware pipelines, unified legal retrievers, and faithfulness-oriented decoding strategies (Lewis et al., 2020; T.y.s.s et al., 2024; Wu et al., 2023; Magesh et al., 2025; Li et al., 2025; T.y.s.s et al., 2025b). More recent work has also moved beyond raw similarity-based retrieval toward concept-guided or anticipatory retrieval, such as LeCoPCR and LexKeyPlan (T.y.s.s et al., 2025c; T.y.s.s and Hernandez, 2025). A parallel line of work structures legal reasoning around explicit *factors*: Wang et al. (2026) learn sparse weights over LLM-generated reasoning questions for issue relevance, while Zhang et al. (2026) show that LLMs struggle to reliably infer factor–concern–issue hierarchies without explicit guidance. For generation, Shi et al. (2026a) decompose statutes into premise–situation–conclusion triplets and combine them with case facts to generate judicial opinions. Our work builds on this direction, but differs in the source and form of structure, using an explicit *Issue–Factor Hierarchy* derived from Case Law Guides rather than a flat set of statutory triplets, and in how that structure is applied: we use it at inference time to drive issue-aligned, factor-specific retrieval over a large doctrinal corpus, without any model training.

Evaluation of legal reasoning remains less devel-

oped than prediction and retrieval. Prior work has shown disagreement among both legal experts and judges (Xu et al., 2023, 2024), and recent work argues for evaluating reasoning traces directly rather than relying on text similarity or outcome agreement alone (Hou et al., 2024; Han et al., 2025). Closest to our use of structured legal issues, Lee et al. (2026) extract issue trees from decided judgments to evaluate LLM reasoning, whereas we construct issue–factor hierarchies from pre-outcome case law guides to guide issue-aligned retrieval and reasoning. Similarly, Shi et al. (2026b) advocate fine-grained, expert-designed rubrics for evaluating legal reasoning. We adopt this fine-grained perspective in our downstream evaluation, assessing generated reasoning across multiple legal dimensions through expert review and LLM-as-a-Judge comparison. More detailed discussion is provided in Appendix A.

3 Task Definition

We formulate the task as generating precedent-grounded legal reasoning: Given the factual description of a case and a specified alleged article, the system must produce a natural language justification explaining how the facts relate to the legal standards under that article and whether those standards have been violated. We study this task in the domain of the European Court of Human Rights (ECHR), which adjudicates complaints by individuals against states concerning alleged violations of rights protected under the European Convention on Human Rights. We condition reasoning task on a specified alleged article rather than requiring the model to infer the applicable provision from the facts alone. This design choice is motivated by both ecological validity and evaluation rigour. In ECHR litigation, applicants advance complaints under particular Convention articles, and the court evaluates the merits of those specific allegations. Providing the alleged article therefore reflects the real decision-making setting. It also avoids a major confound in ECHR datasets: the *Facts* section is typically drafted after judgment and often contains lexical cues that implicitly reveal the relevant article. If the model were required to jointly infer the article and reason about its application, performance could partly reflect topic classification rather than legal analysis (T.y.s.s et al., 2022; Medvedeva and McBride, 2023). By fixing the alleged article as input, we isolate the core task of interest: apply-

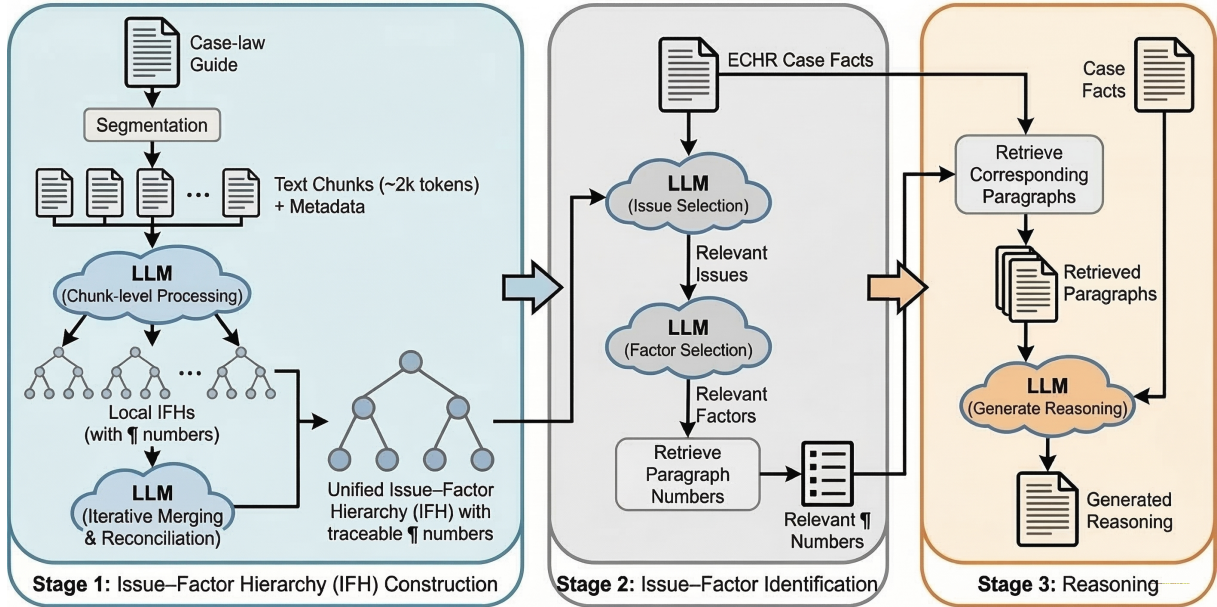


Figure 1: Overview of the **SCALE** legal reasoning generation pipeline. **Stage 1 (IFH Construction)**: A case-law guide is segmented into text chunks. An LLM generates local issue-factor hierarchies (IFHs) with paragraph references for each chunk, which are iteratively merged into a unified, traceable IFH. **Stage 2 (Issue-Factor Identification)**: Using ECHR case facts, an LLM identifies relevant issues and factors within the unified IFH to retrieve specific associated paragraph numbers. **Stage 3 (Reasoning)**: The full text of the selected paragraphs is retrieved from the source guide and combined with the case facts to enable an LLM to generate the final legal reasoning.

ing a given legal provision to a factual record in a precedent-grounded manner.

3.1 Methodology

To address the limitations of similarity-based retrieval for legal reasoning, we propose **SCALE**, a framework that identifies the legal issues and factual considerations implicated by a case before retrieving supporting precedent. **SCALE** uses ECHR *Case Law Guides*, which summarize article-specific jurisprudence, recurring legal questions, relevant factual considerations, and supporting paragraph references, enabling issue-aligned, factor-specific retrieval instead of unstructured case-fragment retrieval. Although we evaluate **SCALE** using the official ECHR Case Law Guides, analogous resources can be automatically constructed from judgment corpora using LLMs, as shown in prior work such as LexGenie (T.y.s.s et al., 2025a), making the framework applicable beyond the ECHR to other jurisdictions and legal systems where such guides are not available a priori.

SCALE consists of three components (See Figure 1): (a) **Issue-Factor Hierarchy (IFH) construction**: We transform each article-specific Case Law Guide into a hierarchy of legal issues and

associated factual factors. (b) **Issue and factor identification**: Given a new case and an alleged article, we identify which issues and factors in the hierarchy are implicated by the facts. (c) **Reasoning generation**: We retrieve the guide passages linked to the selected factors and generate the final legal analysis.

3.1.1 Issue-Factor Hierarchy Construction

The goal of this component is to convert each Case Law Guide into a structured representation of the legal framework associated with an article. The resulting **Issue-Factor Hierarchy (IFH)** contains two types of elements. **Issues** are the core legal questions that arise under a Convention article and structure the Court’s analysis (e.g., *Does the interference pursue a legitimate aim?*). **Factors** are recurring factual considerations relevant to resolving a given issue (e.g., *gravity of harm, vulnerability of the applicant, nature of state interference*).

To construct the hierarchy, each guide is segmented into chunks of approximately 2,000 tokens. Each chunk is then processed independently by an LLM to extract the legal issues, their associated factors, and the paragraph numbers supporting them, producing a set of local chunk-level hierarchies. To create the local IFH, we use the prompt shown in

Figure 2. Figures 17 and 18 show a single input chunk and the IFH derived from it.

In the next step, these local hierarchies are iteratively merged by an LLM into a single global IFH representing the full guide. The merging prompt instructs the model to reconcile duplicates, align conceptually similar issues, and preserve consistent factor granularity (see Figure 3). During this step, paragraph references from the local hierarchies are consolidated so that each issue and factor remains linked to the relevant parts of the original guide. We selected the chunk size reported above by comparing multiple alternatives and retaining the configuration that produced the most coherent and legally meaningful hierarchies in a single-blind expert review.

3.1.2 Issue and Factor Identification

At inference time, the model is given the facts of a new ECHR case together with the issues present in the constructed IFH corresponding to the alleged article. In the first stage, the model selects the provided issues that are relevant to the case (Figure 4). In the second stage, the LLM is again given case facts along with each selected issue and its corresponding list of factors from the IFH. The model evaluates the associated factors and identifies those supported by the facts (Figure 5). For every selected factor, we then collect the paragraph references attached to that factor in the IFH (see Figure 18) and retrieve the corresponding passages from the original Case Law Guide. This yields issue-aligned and factor-specific precedent retrieval.

3.1.3 Reasoning Generation

In the final stage, the model receives the case facts together with the retrieved guide passages associated with the selected issues and factors. Using this structured context, it generates a legal justification for the alleged article, explaining how the relevant precedent bears on the facts of the case and whether the article has been violated.

4 Experiments

Dataset: We construct our evaluation set from ECHR judgments available through the HUDOC database. Following prior work on ECHR judgment prediction (Chalkidis et al., 2019), we focus on the ten most frequently occurring Convention articles. To assess reasoning performance, we randomly sample 25 cases decided in 2025, covering all ten articles. Because a single case may involve

multiple alleged articles, we convert each case into article-specific instances, yielding 37 evaluation instances. Further, we collect approximately 17,000 ECHR cases decided before 2025 to serve as the retrieval corpus for retrieval-based baselines.

Baselines: We compare SCALE against baselines that vary along two dimensions: the *source of external knowledge* (historical cases vs. Case Law Guides) and the *form of the retrieval query* (full case facts vs. anticipatory abstractions of the facts). We use *anticipatory* (A-*) to denote retrieval in which the system first predicts the legal principles (See Figure 10 for prompt) that the downstream reasoning will need, and queries with those predictions, rather than retrieving directly on raw factual similarity. This design allows us to isolate the contributions of the retrieval source, anticipatory querying, and the explicit issue-factor structure introduced by SCALE. **(a) Parametric Reasoning (PR).** The model is prompted with the case facts and the alleged article, without access to external legal material (See Figure 8 and 9). **(b) CaseRAG.** We retrieve supporting context from historical ECHR judgments, using the full *Facts* section of the target case as the retrieval query. **(c) Anticipatory CaseRAG (A-CaseRAG).** Instead of querying with the raw facts, this baseline first prompts an LLM to abstract the case facts into legal principles, following the intuition of recent work on concept-guided and anticipatory retrieval (T.y.s.s and Hernandez, 2025; T.y.s.s et al., 2025c). These abstractions are then used to retrieve relevant passages from the historical case index. **(d) GuideRAG.** This baseline replaces historical judgments with the official ECHR Case Law Guides. For each alleged article, we build an article-specific index over guide paragraphs and use the full case facts to retrieve the top-k passages. **(e) Anticipatory GuideRAG (A-GuideRAG).** This baseline combines article-specific Case Law Guides with anticipatory querying. As in A-CaseRAG, an LLM first abstracts the case facts into legal principles, which are then used to retrieve guide passages from the index of the alleged article. Once we retrieve all the passages in each baseline, we use the reasoning prompt (Figure 6 and 7).

In contrast, SCALE also retrieves from article-specific Case Law Guides, but introduces an explicit **Issue-Factor Hierarchy** that identifies the legal issues and factual factors implicated by the case before retrieving the linked guide passages.

Experimental Setup: We evaluate SCALE using

Criteria	Models	PR	CaseRAG	A-CaseRAG	GuideRAG	A-GuideRAG	SCALE
Legal Issue Identification	<i>Llama3.1 8B</i>	3.54	3.46	3.72	3.88	4.05	4.43
	<i>Llama3.1 70B</i>	3.70	3.61	3.84	4.01	4.15	4.39
	<i>Qwen3 8B</i>	3.67	3.60	3.90	4.07	4.22	4.56
	<i>Qwen3 14B</i>	3.80	3.72	4.00	4.15	4.28	4.59
Factual Faithfulness	<i>Llama3.1 8B</i>	3.19	3.28	3.42	3.39	3.57	3.91
	<i>Llama3.1 70B</i>	3.33	3.29	3.48	3.56	3.61	3.98
	<i>Qwen3 8B</i>	3.41	3.47	3.60	3.55	3.74	4.15
	<i>Qwen3 14B</i>	3.53	3.58	3.69	3.74	3.82	4.24
Logical Coherence	<i>Llama3.1 8B</i>	3.14	3.23	3.34	3.31	3.43	3.69
	<i>Llama3.1 70B</i>	3.29	3.38	3.42	3.49	3.54	3.82
	<i>Qwen3 8B</i>	3.26	3.35	3.33	3.43	3.48	3.90
	<i>Qwen3 14B</i>	3.41	3.47	3.56	3.54	3.65	4.00
Precedent Relevance	<i>Llama3.1 8B</i>	3.00	3.24	3.50	3.66	3.87	4.25
	<i>Llama3.1 70B</i>	3.16	3.36	3.62	3.79	3.99	4.29
	<i>Qwen3 8B</i>	3.14	3.34	3.69	3.84	4.06	4.41
	<i>Qwen3 14B</i>	3.26	3.45	3.78	3.93	4.13	4.47
Precedent Application	<i>Llama3.1 8B</i>	2.96	3.19	3.43	3.58	3.81	4.19
	<i>Llama3.1 70B</i>	3.14	3.31	3.55	3.71	3.93	4.23
	<i>Qwen3 8B</i>	3.08	3.29	3.61	3.77	3.99	4.36
	<i>Qwen3 14B</i>	3.21	3.42	3.74	3.88	4.08	4.42
Judicial Alignment	<i>Llama3.1 8B</i>	3.12	3.26	3.39	3.47	3.64	3.93
	<i>Llama3.1 70B</i>	3.26	3.34	3.49	3.63	3.76	4.03
	<i>Qwen3 8B</i>	3.28	3.41	3.55	3.69	3.80	4.18
	<i>Qwen3 14B</i>	3.40	3.48	3.66	3.78	3.91	4.27

Table 1: Mean expert evaluation scores (1–5 Likert scale) for generated legal reasoning across models and baselines.

two open-weight model families across multiple parameter scales: Llama-3.1 (Grattafiori et al., 2024) (8B and 70B) and Qwen3 (Yang et al., 2025) (8B and 14B). For dense retrieval, we use the Qwen3 4B embedding model to encode both documents and queries, with FAISS (Johnson et al., 2019). More details are provided in Appendix D and GitHub¹.

5 Results and Discussion

5.1 Expert Evaluation

To assess the quality of generated legal reasoning, we conduct an evaluation study with a legal expert experienced in the jurisprudence of the ECHR, including prior roles as a case lawyer and Programme Advisor at the Council of Europe. The expert evaluates reasoning traces for 37 case–article pairs across 4 models and 6 methods, scoring each output on six dimensions on a 1–5 scale: **Legal Issue Identification** (whether the reasoning identifies the legally relevant questions), **Factual Faithfulness** (whether it accurately represents the facts of the case), **Logical Coherence** (whether it is internally consistent and follows a clear argumentative progression), **Precedent Relevance** (whether the reasoning invokes precedent that is relevant and useful for the case), **Precedent Application** (whether the

cited or referenced precedent is applied correctly and accurately in context), and **Judicial Alignment** (how closely the generated reasoning matches the reasoning adopted by the Court). The prompts can be found in Appendix D.

Parametric reasoning vs. retrieval. *PR* often produces plausible legal analysis, but without external support it is more prone to generic reasoning, weak precedent support, and divergence from the Court’s actual line of analysis. Relative to *PR*, retrieval-based baselines generally improve Precedent Relevance, Precedent Application, and Judicial Alignment, confirming that access to external legal material is necessary for generating faithful, precedent-aware explanations.

Raw-fact retrieval vs. anticipatory retrieval. Comparing *CaseRAG* with *A-CaseRAG*, and *GuideRAG* with *A-GuideRAG*, replacing raw-fact queries with abstracted legal principles generally leads to more relevant retrieval and better downstream reasoning. These improvements are most visible in Legal Issue Identification and Precedent Relevance, and in many settings also improve Factual Faithfulness. Anticipatory querying helps in two ways. First, it removes redundant and legally peripheral details from the full case facts, producing more focused retrieval queries. Second, it encourages the model to anticipate the legal principles

¹<https://github.com/khatrimann/SCALE>

needed to construct the eventual argument, rather than retrieving only what is superficially similar to the input. This supports our claim that legal retrieval should be guided by downstream reasoning needs rather than by raw factual similarity alone.

Historical cases vs. Case Law Guides. Comparing *CaseRAG* with *GuideRAG*, guide-based retrieval generally performs better than retrieval from historical judgments. This difference is most apparent on Legal Issue Identification, Precedent Relevance, and Judicial Alignment. Historical cases distribute relevant principles across many documents and often require substantial synthesis after retrieval, whereas the Case Law Guides provide denser, article-specific summaries of the Court’s reasoning. This directly addresses the first limitation of standard legal RAG: retrieved legal materials are often long, heterogeneous, and not organized around the reasoning task. By consolidating recurring legal questions and factual considerations into a more compact resource, the guides reduce the burden on the model to filter and synthesize dispersed precedent.

Free-form anticipation vs. explicit issue-factor structure. The most important comparison is between *A-GuideRAG* and SCALE. Both methods retrieve from article-specific Case Law Guides and both move beyond raw-fact querying. However, *A-GuideRAG* still relies on free-form abstractions of the case generated by the model, whereas SCALE constrains retrieval through an explicit **Issue-Factor Hierarchy**. The additional gains achieved by SCALE therefore isolate the value of explicit legal structure beyond both retrieval source and anticipatory querying. These gains are especially strong on Precedent Relevance and Precedent Application, with consistent improvements in Legal Issue Identification, Factual Faithfulness, and Judicial Alignment. This comparison speaks directly to the second limitation of standard legal RAG: legal relevance is not determined by similarity alone, but by latent issues and factual considerations that may not be consistently recoverable from the model in free form. The IFH helps address this by making article-specific legal structure explicit, allowing retrieval to be guided by the issues and factors that govern the downstream analysis rather than by unconstrained abstractions alone.

Fluency vs. legal quality. Differences in Logical Coherence are generally smaller than those in issue identification or precedent use. This is consis-

tent with the intuition that even weaker baselines can produce logically valid legal-sounding prose, while the harder challenge lies in selecting the right issues and grounding them in the right authority. SCALE still improves coherence, but its largest benefit is the reduction of retrieval mismatch and the provision of a clearer reasoning structure.

Effect across model scales. The benefits of SCALE are consistent across model sizes, but the relative improvements are especially strong for smaller models. In several settings, 8B models equipped with SCALE become comparable to, or approach, the performance of substantially larger models. This supports our central claim that externalizing legal structure can compensate for weaker implicit reasoning and retrieval-planning capacity. Detailed qualitative discussion is provided in Appendix B.

5.2 LLM-as-a-Judge Evaluation

To complement expert assessment, we also evaluate the generated reasoning using GPT-4.1 as an LLM-as-a-Judge. The evaluation prompt is adapted to the criterion being scored: for Logical Coherence, the judge sees only the generated reasoning; for Factual Faithfulness and Legal Issue Identification, it sees the reasoning together with the case facts; for Precedent Relevance and Precedent Application, it sees the reasoning together with the retrieved precedent text; and for Judicial Alignment, it sees the reasoning together with the ground-truth reasoning.

Before interpreting these scores directly, we first examine their correspondence with expert judgment. Table 3 reports Spearman correlations between expert scores and LLM-as-a-Judge scores across the six dimensions. The correlations are uniformly positive, but remain weak to moderate overall, indicating that the automated evaluator captures broad differences in reasoning quality without reliably reproducing expert assessment at a fine-grained legal level. Agreement is strongest for Logical Coherence and Judicial Alignment, both of which are comparatively accessible from the context provided to the judge: coherence can be assessed from the internal structure of the generated reasoning, while judicial alignment becomes a direct comparison task once the ground-truth reasoning is supplied. Correlation is lower for Factual Faithfulness and Legal Issue Identification, where the judge must distinguish legally salient facts and issues from merely plausible analysis. Agreement

Criteria	Models	PR	CaseRAG	A-CaseRAG	GuideRAG	A-GuideRAG	SCALE
Legal Issue Identification	<i>Llama3.1 8B</i>	3.63	3.71	3.77	3.75	3.88	4.00
	<i>Llama3.1 70B</i>	3.77	3.80	3.87	3.91	3.94	4.05
	<i>Qwen3 8B</i>	3.70	3.78	3.84	3.82	3.92	4.06
	<i>Qwen3 14B</i>	3.82	3.87	3.92	3.96	3.98	4.10
Factual Faithfulness	<i>Llama3.1 8B</i>	3.51	3.60	3.57	3.63	3.61	3.74
	<i>Llama3.1 70B</i>	3.59	3.62	3.65	3.68	3.66	3.79
	<i>Qwen3 8B</i>	3.56	3.64	3.62	3.59	3.68	3.81
	<i>Qwen3 14B</i>	3.67	3.73	3.71	3.76	3.75	3.88
Logical Coherence	<i>Llama3.1 8B</i>	3.74	3.80	3.82	3.77	3.84	3.90
	<i>Llama3.1 70B</i>	3.85	3.88	3.91	3.93	3.92	3.98
	<i>Qwen3 8B</i>	3.71	3.76	3.73	3.79	3.81	3.88
	<i>Qwen3 14B</i>	3.82	3.86	3.90	3.87	3.93	3.97
Precedent Relevance	<i>Llama3.1 8B</i>	3.40	3.59	3.62	3.68	3.70	3.82
	<i>Llama3.1 70B</i>	3.53	3.67	3.70	3.77	3.79	3.87
	<i>Qwen3 8B</i>	3.45	3.60	3.66	3.71	3.75	3.85
	<i>Qwen3 14B</i>	3.57	3.70	3.75	3.80	3.83	3.91
Precedent Application	<i>Llama3.1 8B</i>	3.38	3.55	3.58	3.63	3.68	3.79
	<i>Llama3.1 70B</i>	3.51	3.64	3.68	3.73	3.77	3.85
	<i>Qwen3 8B</i>	3.42	3.58	3.63	3.68	3.72	3.83
	<i>Qwen3 14B</i>	3.55	3.68	3.73	3.78	3.81	3.89
Judicial Alignment	<i>Llama3.1 8B</i>	3.46	3.54	3.57	3.60	3.61	3.71
	<i>Llama3.1 70B</i>	3.59	3.66	3.68	3.72	3.71	3.80
	<i>Qwen3 8B</i>	3.49	3.57	3.61	3.64	3.66	3.75
	<i>Qwen3 14B</i>	3.62	3.69	3.72	3.75	3.77	3.84

Table 2: Mean LLM-as-a-Judge scores (1–5 Likert scale) for generated legal reasoning across models and baselines.

is weakest for Precedent Relevance and Precedent Application. Even when precedent text is provided, deciding whether the authority is the most appropriate for the case and whether it is applied correctly remains a legally demanding judgment.

Table 2 presents the LLM-as-a-Judge results across models and baselines. Broadly, the automated evaluator preserves the overall ranking observed in expert evaluation: *SCALE* remains the strongest method overall, and guide-based, anticipatory retrieval methods generally outperform prompt-only reasoning and weaker retrieval baselines. However, the margins between methods are smaller than in expert evaluation, and intermediate baselines exhibit more frequent reversals. In particular, the automated evaluator shows less consistent separation between raw-fact and anticipatory retrieval, and between case-based and guide-based retrieval, especially on Factual Faithfulness, Logical Coherence, and Judicial Alignment. Relative to the expert, GPT-4.1 appears more sensitive to fluency, structural consistency, and broad similarity to judicial style, and less sensitive to subtle failures of legal grounding. As a result, rhetorically polished outputs may receive favorable automated scores even when they miss a legally decisive issue, rely on only loosely relevant authority, or apply precedent imprecisely.

Criterion	Spearman ρ
Legal Issue Identification	0.18
Factual Faithfulness	0.21
Logical Coherence	0.31
Precedent Relevance	0.11
Precedent Application	0.09
Judicial Alignment	0.34

Table 3: Spearman correlation between expert and LLM-as-a-Judge scores across the six evaluation dimensions.

These findings highlight both the usefulness and the limitations of LLM-as-a-Judge in this setting. It provides a scalable way to compare many model and retrieval configurations and broadly identifies the same top-performing method as expert review. At the same time, its limited agreement with expert scores suggests that it should not be treated as a substitute for domain-specific legal evaluation, particularly when the goal is to assess whether reasoning is grounded in the right precedent and applies that precedent correctly. In high-stakes legal reasoning tasks, automated evaluation is therefore best understood as a complementary signal rather than a definitive measure of quality.

5.3 Quality of the Issue–Factor Hierarchy

We also assess the quality of the automatically constructed Issue–Factor Hierarchies through expert re-

view. A legal expert evaluates representative IFHs for four Convention article case law guides (Articles 5, 6, 8, and 10) on a 1–5 Likert scale, considering issue- and factor-level *completeness*, *precision/correctness*, and *redundancy*. The resulting scores are consistently high, indicating that the constructed IFHs faithfully capture the legal issues and factual factors summarized in the Case Law Guides while remaining largely non-redundant. Full evaluation criteria, article-level scores, and qualitative discussion are provided in Appendix C.

6 Conclusion

We presented **SCALE**, a structured framework for precedent-grounded legal reasoning in the European Court of Human Rights. By converting article-specific Case Law Guides into an explicit *Issue–Factor Hierarchy* and using that hierarchy to guide retrieval, SCALE enables reasoning that is more issue-aligned, factor-specific, and grounded in relevant authority. Across multiple open-weight model families and parameter scales, SCALE consistently outperforms prompt-only and retrieval-based baselines. The largest gains appear in legal issue identification and precedent use, showing that explicit legal structure helps models recover the salient questions in a case and connect them to more appropriate authority. These improvements are especially pronounced for smaller models, suggesting that structured retrieval can compensate for weaker implicit reasoning and retrieval-planning capacity. More broadly, our results show that structured precedent summaries are more effective than retrieval over dispersed case fragments, and that decomposing reasoning into issue identification, factor selection, and final generation yields explanations that are more faithful and interpretable than one-step generation over unstructured context. We also find only partial agreement between expert and LLM-based evaluation, indicating that automated evaluation is useful as a scalable signal but remains less reliable for fine-grained judgments about precedent selection and application.

Limitations

Our experiments are conducted only in the ECHR setting. This is a deliberate choice: ECHR judgments provide a comparatively clean environment for studying precedent-grounded reasoning because the *Facts* and *Law* sections are clearly separated. In many other jurisdictions, including sys-

tems such as India, the United Kingdom, and the United States (Hou et al., 2025; Malik et al., 2021), factual narration and legal analysis are often interwoven, making the reasoning task harder to isolate without additional segmentation and introducing further confounds (T.y.s.s. et al., 2024).

Second, SCALE depends on the availability and quality of article-level Case Law Guides. In the ECHR, these guides provide a strong source of structured precedent, enabling a controlled study of issue-aligned and factor-specific retrieval. Comparable resources, however, may be incomplete, unavailable, or uneven in quality in other legal systems. While such guide-like resources can in principle be constructed automatically from judgment corpora, as explored in prior work such as LexGenie (T.y.s.s. et al., 2025a), this introduces an additional source of error. Inaccuracies in automatically generated guides may propagate into the construction of the Issue–Factor Hierarchy and, in turn, affect retrieval and reasoning quality.

Third, expert evaluation in legal reasoning is inherently difficult to scale. Reliable assessment requires evaluators with deep knowledge of the relevant jurisdiction, substantially limiting the pool of suitable experts and making large-scale evaluation costly in both time and resources. This challenge is amplified by the length and complexity of the generated reasoning traces: to assess a single output, an expert may need to inspect the case facts, the reasoning itself, and the cited or implied precedents and references. Even our relatively small expert study—37 case–article pairs evaluated across 4 models and 6 methods—results in 888 reasoning traces to review. Moreover, disagreement is itself a structural feature of legal decision-making: prior work has documented disagreement among legal experts (Xu et al., 2023), and even judges may disagree in the same case (Xu et al., 2024). Our expert evaluation should therefore be understood as a strong but necessarily limited signal rather than a definitive measure of reasoning quality.

These limitations suggest several directions for future work: validating SCALE in jurisdictions with less structured judgments, studying its robustness when guide-like resources are generated automatically rather than curated manually, and developing evaluation protocols that combine expert review with more reliable scalable assessment for complex legal reasoning. SCALE can also be expanded to Long-Form Legal QA (Louis et al., 2024), where the same issue-aligned, factor-

specific retrieval could ground answers in relevant legal provisions rather than precedent alone, indicating that our approach generalizes beyond the reasoning-generation setting studied here. Such tasks would also broaden our evaluation, since long-form legal QA offers expert-annotated answers and gold provision-level annotations that complement the expert review protocol used in this work. We experimented with SCALE on smaller models; whether these gains persist at larger scales remains open, though recent scaling analyses of contextual entrainment (Kukreja et al., 2026).

Ethical considerations

This work studies structured retrieval for precedent-grounded legal reasoning in ECHR. The proposed system is intended as an assistive tool for analysis, not as a substitute for legal professionals or judicial decision-making. Even when models produce fluent and precedent-supported reasoning, they may still omit relevant issues, misapply authority, or generate persuasive but incorrect legal analysis. Human oversight therefore remains essential.

Our experiments use publicly available ECHR judgments and official Case Law Guides. However, both legal corpora and the underlying language models reflect institutional, historical, and social patterns that may be reproduced in generated outputs. In addition, the pretrained models used in this work inherit biases from their broader training data, which may affect what issues are emphasized, how facts are framed, or which forms of reasoning appear more plausible. Improved faithfulness to precedent therefore does not by itself guarantee fairness, neutrality, or normative desirability.

We also emphasize that automated evaluation is not a substitute for legal expertise. Although LLM-as-a-Judge provides a scalable signal, our results show only limited agreement with expert assessment, particularly on precedent-specific dimensions. For high-stakes legal applications, both system use and system evaluation require domain-aware human oversight.

Use of AI writing tools for paraphrasing the contents of the paper During the writing process, the authors used ChatGPT and Claude as tools to refine the language and enhance the overall flow of the manuscript. Prior to submission, a comprehensive review was carried out and appropriate revisions were made, with the authors assuming complete responsibility for the published content.

Acknowledgements: We acknowledge the support of the IHUB-ANUBHUTI-IIITD FOUNDATION set up under the NM-ICPS scheme of the Department of Science and Technology, India. Along with we also thank Infosys Center for AI at IIIT Delhi.

References

- Abhishek Agarwal, Shanshan Xu, and Matthias Grabmair. 2022. [Extractive summarization of legal decisions using multi-task learning and maximal marginal relevance](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1857–1872, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Nikolaos Aletras, Dimitrios Tsarapatsanis, Daniel Preoŧiuc-Pietro, and Vasileios Lamos. 2016. Predicting judicial decisions of the european court of human rights: A natural language processing perspective. *PeerJ computer science*, 2:e93.
- Elliott Ash, Aniket Kesari, Suresh Naidu, Lena Song, and Dominik Stambach. 2024. [Translating legalese: Enhancing public understanding of court opinions with legal summarizers](#). In *Proceedings of the 2024 Symposium on Computer Science and Law, CSLAW '24*, page 136–157, New York, NY, USA. Association for Computing Machinery.
- Emmanuel Bauer, Dominik Stambach, Nianlong Gu, and Elliott Ash. 2023. Legal extractive summarization of us court opinions. *arXiv preprint arXiv:2305.08428*.
- Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. 2019. [Neural legal judgment prediction in English](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4317–4323, Florence, Italy. Association for Computational Linguistics.
- Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. [LEGAL-BERT: The muppets straight out of law school](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online. Association for Computational Linguistics.
- Ilias Chalkidis, Manos Fergadiotis, Dimitrios Tsarapatsanis, Nikolaos Aletras, Ion Androutsopoulos, and Prodromos Malakasiotis. 2021. [Paragraph-level rationale extraction through regularization: A case study on European court of human rights cases](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 226–241, Online. Association for Computational Linguistics.
- Guhong Chen, Liyang Fan, Zihan Gong, Nan Xie, Zixuan Li, Ziqiang Liu, Chengming Li, Qiang Qu,

- Hamid Alinejad-Rokny, Shiwen Ni, and 1 others. 2025a. Agentcourt: Simulating court with adversarial evolvable lawyer agents. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5850–5865.
- Xi Chen, Mao Mao, Shuo Li, and Haotian Shanguan. 2025b. [Debate-feedback: A multi-agent framework for efficient legal judgment prediction](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 462–470, Albuquerque, New Mexico. Association for Computational Linguistics.
- Ashish Chouhan and Michael Gertz. 2024. [LexDrafter: Terminology drafting for legislative documents using retrieval augmented generation](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10448–10458, Torino, Italia. ELRA and ICCL.
- Jiaxi Cui, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. 2023. Chatlaw: Open-source legal large language model with integrated external knowledge bases. *CoRR*.
- Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E Ho. 2024. Large legal fictions: Profiling legal hallucinations in large language models. *Journal of Legal Analysis*, 16(1):64–93.
- Francisco de Arriba-Pérez, Silvia García-Méndez, Francisco J González-Castaño, and Jaime González-González. 2022. Explainable machine learning multi-label classification of spanish legal judgements. *Journal of King Saud University-Computer and Information Sciences*, 34(10):10180–10192.
- Chenlong Deng, Kelong Mao, Yuyao Zhang, and Zhicheng Dou. 2024. [Enabling discriminative reasoning in LLMs for legal judgment prediction](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 784–796, Miami, Florida, USA. Association for Computational Linguistics.
- Wentao Deng, Jiahuan Pei, Keyi Kong, Zhe Chen, Furu Wei, Yujun Li, Zhaochun Ren, Zhumin Chen, and Pengjie Ren. 2023. [Syllogistic reasoning for legal judgment analysis](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13997–14009, Singapore. Association for Computational Linguistics.
- Yi Feng, Chuanyi Li, and Vincent Ng. 2022. [Legal judgment prediction via event extraction with constraints](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 648–664, Dublin, Ireland. Association for Computational Linguistics.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Sophia Simeng Han, Yoshiki Takashima, Shannon Zejiang Shen, Chen Liu, Yixin Liu, Roque K Thuo, Sonia Knowlton, Ruzica Piskac, Scott J Shapiro, and Arman Cohan. 2025. Courtreasoner: Can llm agents reason like judges? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 35279–35294.
- Abe Hou, William Jurayj, Nils Holzenberger, Andrew Blair-Stanek, and Benjamin Van Durme. 2024. [Gaps or hallucinations? scrutinizing machine-generated legal analysis for fine-grained text evaluations](#). In *Proceedings of the Natural Legal Language Processing Workshop 2024*, pages 280–302, Miami, FL, USA. Association for Computational Linguistics.
- Abe Bohan Hou, Orion Weller, Guanghui Qin, Eugene Yang, Dawn Lawrie, Nils Holzenberger, Andrew Blair-Stanek, and Benjamin Van Durme. 2025. [CLERC: A dataset for U. S. legal case retrieval and retrieval-augmented analysis generation](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7913–7928, Albuquerque, New Mexico. Association for Computational Linguistics.
- Wanhong Huang, Yi Feng, Chuanyi Li, Honghan Wu, Jidong Ge, and Vincent Ng. 2024. [CMDL: A large-scale Chinese multi-defendant legal judgment prediction dataset](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5895–5906, Bangkok, Thailand. Association for Computational Linguistics.
- Cong Jiang and Xiaolei Yang. 2023a. [Legal syllogism prompting: Teaching large language models for legal judgment prediction](#). In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law, ICAIL ’23*, page 417–421, New York, NY, USA. Association for Computing Machinery.
- Cong Jiang and Xiaolei Yang. 2023b. Legal syllogism prompting: Teaching large language models for legal judgment prediction. In *Proceedings of the nineteenth international conference on artificial intelligence and law*, pages 417–421.
- Cong Jiang and Xiaolei Yang. 2024. Agents on the bench: Large language model based multi agent framework for trustworthy digital justice. *arXiv preprint arXiv:2412.18697*.
- Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.
- Xiaoxi Kang, Lizhen Qu, Lay-Ki Soon, Adnan Trakic, Terry Zhuo, Patrick Emerton, and Genevieve Grant. 2023. [Can ChatGPT perform reasoning using the IRAC method in analyzing legal scenarios like a lawyer?](#) In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13900–13923, Singapore. Association for Computational Linguistics.

- Robert Keown. 1980. Mathematical models for legal prediction. *Computer/lj*, 2:829.
- Mann Khatri, Mirza Yusuf, Yaman Kumar, Rajiv Ratn Shah, and Ponnuram Kumaraguru. 2023. Exploring graph neural networks for indian legal judgment prediction. *arXiv preprint arXiv:2310.12800*.
- Mann Khatri, Mirza Yusuf, Rajiv Ratn Shah, and Ponnuram Kumaraguru. 2025. Structured definitions and segmentations for legal reasoning in llms: a study on indian legal data. In *International Conference on Big Data Analytics*, pages 295–304. Springer.
- Dikshant Kukreja, Kshitij Sah, Gautam Gupta, Avinash Anand, Rajiv Ratn Shah, Zhengkui Wang, Aik Beng Ng, and Erik Cambria. 2026. [Better and worse with scale: How contextual entrainment diverges with model size](#). In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 30194–30209, San Diego, California, United States. Association for Computational Linguistics.
- Reed C Lawlor. 1963. What computers can do: analysis and prediction of judicial decisions. *American Bar Association Journal*, pages 337–344.
- Jinu Lee, Kyoung-Woon On, Simeng Han, Arman Cohan, and Julia Hockenmaier. 2025. Evaluating legal reasoning traces with legal issue tree rubrics. *arXiv preprint arXiv:2512.01020*.
- Jinu Lee, Kyoung-Woon On, Sophia Simeng Han, Arman Cohan, and Julia Hockenmaier. 2026. [Evaluating legal reasoning traces with legal issue tree rubrics](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3299–3322, San Diego, California, United States. Association for Computational Linguistics.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.
- Ang Li, Yiquan Wu, Yifei Liu, Ming Cai, Lizhi Qing, Shihang Wang, Yangyang Kang, Chengyuan Liu, Fei Wu, and Kun Kuang. 2025. [UniLR: Unleashing the power of LLMs on multiple legal tasks with a unified legal retriever](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11953–11967, Vienna, Austria. Association for Computational Linguistics.
- Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of psychology*.
- Dugang Liu, Weihao Du, Lei Li, WeiKe Pan, and Zhong Ming. 2022. [Augmenting legal judgment prediction with contrastive case relations](#). In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2658–2667, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Antoine Louis, Gijs Van Dijck, and Gerasimos Spanakis. 2024. Interpretable long-form legal question answering with retrieval-augmented large language models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 22266–22275.
- Bingfeng Luo, Yansong Feng, Jianbo Xu, Xiang Zhang, and Dongyan Zhao. 2017. [Learning to predict charges for criminal cases with legal basis](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2727–2736, Copenhagen, Denmark. Association for Computational Linguistics.
- Yougang Lyu, Jitai Hao, Zihan Wang, Kai Zhao, Shen Gao, Pengjie Ren, Zhumin Chen, Fang Wang, and Zhaochun Ren. 2023. [Multi-defendant legal judgment prediction via hierarchical reasoning](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2198–2209, Singapore. Association for Computational Linguistics.
- Shengjie Ma, Qi Chu, Jiaxin Mao, Xuhui Jiang, Haozhe Duan, and Chong Chen. 2024. Leveraging large language models for relevance judgments in legal case retrieval. *arXiv preprint arXiv:2403.18405*.
- Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D Manning, and Daniel E Ho. 2025. Hallucination-free? assessing the reliability of leading ai legal research tools. *Journal of empirical legal studies*, 22(2):216–242.
- Robert Mahari, Dominik Stammbach, Elliott Ash, and Alex Pentland. 2024. [LePaRD: A large-scale dataset of judicial citations to precedent](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9863–9877, Bangkok, Thailand. Association for Computational Linguistics.
- Vijit Malik, Rishabh Sanjay, Shubham Kumar Nigam, Kripabandhu Ghosh, Shouvik Kumar Guha, Arnab Bhattacharya, and Ashutosh Modi. 2021. [ILDC for CJPE: Indian legal documents corpus for court judgment prediction and explanation](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4046–4062, Online. Association for Computational Linguistics.
- Mihai Masala, Traian Rebedea, and Horia Velicu. 2024. [Improving legal judgement prediction in Romanian with long text encoders](#). In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pages 126–132, Torino, Italia. ELRA and ICCL.
- Masha Medvedeva and Pauline McBride. 2023. [Legal judgment prediction: If you are going to do it, do it right](#). In *Proceedings of the Natural Legal Language*

- Processing Workshop 2023*, pages 73–84, Singapore. Association for Computational Linguistics.
- Masha Medvedeva, Michel Vols, and Martijn Wieling. 2018. Judicial decisions of the european court of human rights: Looking into the crystal ball. In *Proceedings of the conference on empirical legal studies*, page 24.
- Stuart Nagel. 1960. Weighting variables in judicial prediction. *MULL Mod. Uses Log. L.*, 2:93.
- Cheol Ryu, Seolhwa Lee, Subeen Pang, Chanyeol Choi, Hojun Choi, Myeonggee Min, and Jy-Yong Sohn. 2023. Retrieval-based evaluation for llms: a case study in korean legal qa. In *Proceedings of the Natural Legal Language Processing Workshop 2023*, pages 132–137.
- TYSS Santosh, Mahmoud Aly, and Matthias Grabmair. 2024a. Lexabsumm: Aspect-based summarization of legal decisions. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10422–10431.
- TYSS Santosh, Mohamed Hesham Elganayni, Stanisław Sójka, and Matthias Grabmair. 2024b. Incorporating precedents for legal judgement prediction on european court of human rights cases. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3743–3750.
- TYSS Santosh, Youssef Tarek Elkhayat, Oana Ichim, Pranav Shetty, Dongsheng Wang, Zhiqiang Ma, Armineh Nourbakhsh, and Xiaomo Liu. 2025a. Co-colex: Confidence-guided copy-based decoding for grounded legal text generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 19002–19018.
- TYSS Santosh, Youssef Farag, and Matthias Grabmair. 2025b. Coperlex: Content planning with event-based representations for legal case summarization. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1016–1032.
- TYSS Santosh, Isaac Misael Olguín Nolasco, and Matthias Grabmair. 2025c. Lecopr: Legal concept-guided prior case retrieval for european court of human rights cases. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1654–1661.
- TYSS Santosh, Cornelius Weiss, and Matthias Grabmair. 2024c. Lexsumm and lext5: Benchmarking and modeling legal summarization tasks in english. In *Proceedings of the Natural Legal Language Processing Workshop 2024*, pages 381–403.
- Weizhe Shi, Qiqi Wang, Yihong Pan, Qian Liu, and Kaiqi Zhao. 2026a. [LegalChainReasoner: Grounding criminal judicial opinion generation via structured legal chains](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23844–23863, San Diego, California, United States. Association for Computational Linguistics.
- Yuzhen Shi, Huanghai Liu, Yiran HU, Song Gaojie, Xu Xinran, Yubo Ma, Tianyi Tang, Li Zhang, Qingjing Chen, Feng Di, Wenbo Lv, Weiheng Wu, Kexin Yang, Sen Yang, Wei Wang, Rongyao Shi, Qiu Yuanyang, Yuemeng Qi, Zhang Jingwen, and 11 others. 2026b. [PLAWBENCH: A rubric-based benchmark for evaluating LLMs in real-world legal practice](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10067–10116, San Diego, California, United States. Association for Computational Linguistics.
- Dong Shu, Haoran Zhao, Xukun Liu, David Demeter, Mengnan Du, and Yongfeng Zhang. 2024. Lawllm: Law large language model for the us legal system. In *Proceedings of the 33rd ACM International Conference on information and knowledge management*, pages 4882–4889.
- Ruihao Shui, Yixin Cao, Xiang Wang, and Tat-Seng Chua. 2023. [A comprehensive evaluation of large language models on legal judgment prediction](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7337–7348, Singapore. Association for Computational Linguistics.
- Santosh T.y.s.s, Mahmoud Aly, Oana Ichim, and Matthias Grabmair. 2025a. [LexGenie: Automated generation of structured reports for European court of human rights case law](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 672–683, Vienna, Austria. Association for Computational Linguistics.
- Santosh T.Y.S.S, Kevin Ashley, Katie Atkinson, and Matthias Grabmair. 2024. [Towards supporting legal argumentation with NLP: Is more data really all you need?](#) In *Proceedings of the Natural Legal Language Processing Workshop 2024*, pages 404–421, Miami, FL, USA. Association for Computational Linguistics.
- Santosh T.y.s.s, Mohamed Hesham Elganayni, Stanisław Sójka, and Matthias Grabmair. 2024. [Incorporating precedents for legal judgement prediction on European court of human rights cases](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3743–3750, Miami, Florida, USA. Association for Computational Linguistics.
- Santosh T.y.s.s, Youssef Tarek Elkhayat, Oana Ichim, Pranav Shetty, Dongsheng Wang, Zhiqiang Ma, Armineh Nourbakhsh, and Xiaomo Liu. 2025b. [Co-CoLex: Confidence-guided copy-based decoding for grounded legal text generation](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 19002–19018, Vienna, Austria. Association for Computational Linguistics.

- Santosh Tyss, Rashid Haddad, and Matthias Grabmair. 2024. Ecthr-pcr: A dataset for precedent understanding and prior case retrieval in the european court of human rights. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5473–5483.
- Santosh T.y.s.s., Rashid Haddad, and Matthias Grabmair. 2024. [ECtHR-PCR: A dataset for precedent understanding and prior case retrieval in the European court of human rights](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5473–5483, Torino, Italia. ELRA and ICCL.
- Santosh T.y.s.s and Elvin Quero Hernandez. 2025. [LexKeyPlan: Planning with keyphrases and retrieval augmentation for legal text generation: A case study on European court of human rights cases](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 425–436, Vienna, Austria. Association for Computational Linguistics.
- Santosh T.y.s.s, Isaac Misael Olguín Nolasco, and Matthias Grabmair. 2025c. [LeCoPCR: Legal concept-guided prior case retrieval for European court of human rights cases](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1654–1661, Albuquerque, New Mexico. Association for Computational Linguistics.
- Santosh T.y.s.s, Shanshan Xu, Oana Ichim, and Matthias Grabmair. 2022. [Deconfounding legal judgment prediction for European court of human rights cases towards better alignment with experts](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1120–1138, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- S Sidney Ulmer. 1963. Quantitative analysis of judicial processes: Some practical and theoretical applications. *Law and Contemporary Problems*, 28(1):164–184.
- Rohit Upadhy and TYSS Santosh. 2025. Lexclipr: Cross-lingual paragraph retrieval from legal judgments. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13971–13993.
- Fanyu Wang, Xiaoxi Kang, Paul Burgess, Aashish Srivastava, Chetan Arora, Adnan Trakic, Lay-Ki Soon, Md Khalid Hossain, and Lizhen Qu. 2026. [LePREC: Reasoning as classification over structured factors for assessing relevance of legal issues](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7701–7736, San Diego, California, United States. Association for Computational Linguistics.
- Xuran Wang, Xinguang Zhang, Vanessa Hoo, Zhouhang Shao, and Xuguang Zhang. 2024. [Legalreasoner: A multi-stage framework for legal judgment prediction via large language models and knowledge integration](#). *IEEE Access*, 12:166843–166854.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.
- Korbinian Q Weidinger, TYSS Santosh, Oana Ichim, and Matthias Grabmair. 2025. Aquaechr: Attributed question answering for european court of human rights. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 1418–1447.
- Yiquan Wu, Yifei Liu, Weiming Lu, Yating Zhang, Jun Feng, Changlong Sun, Fei Wu, and Kun Kuang. 2022. [Towards interactivity and interpretability: A rationale-based legal judgment prediction framework](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4787–4799, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Yiquan Wu, Siying Zhou, Yifei Liu, Weiming Lu, Xiaozhong Liu, Yating Zhang, Changlong Sun, Fei Wu, and Kun Kuang. 2023. [Precedent-enhanced legal judgment prediction with LLM and domain-model collaboration](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12060–12075, Singapore. Association for Computational Linguistics.
- Shanshan Xu, TYSS Santosh, Oana Ichim, Isabella Risini, Barbara Plank, and Matthias Grabmair. 2023. From dissonance to insights: Dissecting disagreements in rationale construction for case outcome classification. *arXiv preprint arXiv:2310.11878*.
- Shanshan Xu, Santosh T.y.s.s, Oana Ichim, Barbara Plank, and Matthias Grabmair. 2024. [Through the lens of split vote: Exploring disagreement, difficulty and calibration in legal case outcome classification](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 199–216, Bangkok, Thailand. Association for Computational Linguistics.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.
- Rujing Yao, Yang Wu, Chenghao Wang, Jingwei Xiong, Fang Wang, and Xiaozhong Liu. 2025. [Elevating legal LLM responses: Harnessing trainable logical structures and semantic knowledge with legal reasoning](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5630–5642, Albuquerque, New Mexico. Association for Computational Linguistics.

- Hai Ye, Xin Jiang, Zhunchen Luo, and Wenhan Chao. 2018. Interpretable charge predictions for criminal cases: Learning to generate court views from fact descriptions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1854–1864.
- Fangyi Yu, Lee Quartey, and Frank Schilder. 2023. Exploring the effectiveness of prompt engineering for legal reasoning tasks. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13582–13596, Toronto, Canada. Association for Computational Linguistics.
- Li Zhang, Matthias Grabmair, Morgan Gray, and Kevin Ashley. 2026. Thinking longer, not always smarter: Evaluating llm capabilities in hierarchical legal reasoning. In *Proceedings of the Symposium on Computer Science and Law*, pages 71–85.
- Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Chaojun Xiao, Zhiyuan Liu, and Maosong Sun. 2018. Legal judgment prediction via topological learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3540–3549, Brussels, Belgium. Association for Computational Linguistics.
- Haoxi Zhong, Yuzhong Wang, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. Iteratively questioning and answering for interpretable legal judgment prediction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 1250–1257.
- Youchao Zhou, Heyan Huang, and Zhijing Wu. 2023. Boosting legal case retrieval by query content selection with large language models. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region, SIGIR-AP ’23*, page 176–184, New York, NY, USA. Association for Computing Machinery.

A Extended Related Work

Legal judgment prediction and explanation: Legal judgment prediction has a long history, beginning with early rule-based and quantitative approaches that sought to formalize judicial decision-making (Lawlor, 1963; Nagel, 1960; Ulmer, 1963; Keown, 1980). After these early efforts, research increasingly turned to machine learning methods, spanning traditional classifiers (Aletras et al., 2016; Medvedeva et al., 2018), deep neural architectures (Luo et al., 2017; Zhong et al., 2018; Khatri et al., 2023; Lyu et al., 2023), and more recently transformer-based models, including BERT-style encoders and legal-domain pretraining (Huang et al., 2024; de Arriba-Pérez et al., 2022; Liu et al.,

2022; Chalkidis et al., 2019, 2020; Masala et al., 2024). Although these methods substantially improved predictive performance, much of this line of work remained centered on outcome prediction rather than the generation of faithful legal reasoning. This motivated efforts to pair prediction with explanation, including rationale extraction from case facts (Malik et al., 2021; Chalkidis et al., 2021), phrase-level rationale generation (Feng et al., 2022), iterative question-answering frameworks (Zhong et al., 2020), and court-view generation (Wu et al., 2022). However, subsequent work showed that legal prediction models can exploit spurious correlations and shortcut features (T.y.s.s et al., 2022; Xu et al., 2023), raising doubts about whether such explanations genuinely reflect legal reasoning.

LLM-based LJP The emergence of LLMs shifted attention from prediction alone toward generation of richer legal analyses (Shui et al., 2023; Shu et al., 2024). Recent work has explored prompting strategies designed to better scaffold legal reasoning, including Chain-of-Thought prompting (Wei et al., 2022; Yu et al., 2023), IRAC-style prompting (Kang et al., 2023), syllogistic reasoning (Jiang and Yang, 2023b; Deng et al., 2023), and prompts enriched with legal definitions or rhetorical-role structure (Khatri et al., 2025). These approaches reflect the broader intuition that legal reasoning benefits from explicit structure. However, even when prompted carefully, LLMs remain prone to hallucination and legal-sounding but unsupported analysis (Dahl et al., 2024). Our work shares the motivation for structure, but focuses not on prompting alone; instead, it externalizes article-specific legal structure to guide retrieval and generation.

Retrieval and grounding in legal reasoning:

To address hallucination and improve grounding, recent work has increasingly adopted retrieval-augmented generation in legal settings (Lewis et al., 2020). Existing approaches retrieve prior cases or legal materials to support reasoning, including precedent-aware pipelines (T.y.s.s et al., 2024; Wu et al., 2023; Magesh et al., 2025), unified legal retrievers (Li et al., 2025), and decoding strategies aimed at improving faithfulness (T.y.s.s et al., 2025b). More recent work has also begun to move beyond raw similarity-based retrieval toward concept-guided or anticipatory retrieval. For example, LeCoPCR retrieves prior cases using generated legal concepts, while LexKeyPlan improves retrieval by planning with forward-looking

keyphrases (T.y.s.s et al., 2025c; T.y.s.s and Hernandez, 2025). These methods highlight an important shift: effective legal retrieval should be guided by the reasoning needs of the task rather than by surface similarity alone. Our work builds on this insight, but addresses a further challenge: in specialized legal domains, the model may not reliably infer the relevant issues and factual considerations on its own. We therefore externalize that structure through an explicit *Issue–Factor Hierarchy* derived from Case Law Guides and use it to support issue-aligned, factor-specific retrieval.

Multi-agent and deliberative legal reasoning:

A parallel line of work models legal reasoning as interaction among multiple specialized components or agents, including simulated judicial panels (Jiang and Yang, 2024), adversarial lawyer agents (Chen et al., 2025a), and debate-feedback architectures (Chen et al., 2025b). These approaches emphasize deliberation, role specialization, and multi-step refinement. Our work is related in its emphasis on decomposition, but differs in focus: rather than simulating institutional roles or adversarial processes, SCALE decomposes retrieval and reasoning around explicit legal issues and factual factors grounded in precedent summaries.

Evaluation of legal reasoning: Compared with prediction and retrieval, evaluation of legal reasoning remains relatively underexplored. Much prior work still emphasizes predictive performance or retrieval quality, even though both can be misleading proxies for reasoning quality. Prior studies have shown disagreement not only among legal experts (Xu et al., 2023) but also among judges themselves (Xu et al., 2024), underscoring the difficulty of treating any single output as an uncontested gold standard. Earlier evaluation practices also often relied on text-overlap metrics or coarse agreement with ground-truth outcomes, which are poorly suited to assessing whether a generated explanation identifies the right issues, uses the right authority, and applies that authority correctly. Work on fine-grained gap analysis likewise argues for more detailed evaluation of legal analysis beyond text similarity (Hou et al., 2024). More recent work has therefore shifted toward evaluating *legal reasoning traces* directly. CourtReasoner studies judicial-style legal reasoning generation using dimensions such as citation relevance, constraint extraction, and argument validity (Han et al., 2025), while LEGIT introduces issue-tree rubrics for assessing reasoning traces in terms of issue coverage and cor-

rectness (Lee et al., 2025). Our work follows this emerging direction by evaluating generated reasoning along multiple legally meaningful dimensions through expert review, and by comparing those judgments with LLM-as-a-Judge scores to assess how far automated evaluation can serve as a scalable proxy in precedent-grounded legal reasoning.

B Case Study

We present two qualitative case studies from expert review to illustrate how different retrieval and reasoning strategies fail—and improve—on legally difficult ECHR cases. Together, these examples clarify the kinds of gains reflected in the aggregate results. The first case (Case 237239) is a difficult Article 6 dispute involving ambiguous domestic procedural law and a threshold question about whether the applicants’ situation falls within the scope of a Convention right at all. The second case (Case 243777) is doctrinally cleaner, but requires a careful proportionality analysis under Article 11 and close attention to the domestic reasons for refusing registration of an association. In both cases, the progression from *PR* to *SCALE* illustrates the cumulative value of (i) adding retrieval, (ii) replacing raw-fact querying with anticipatory abstractions, (iii) replacing dispersed historical cases with article-specific Case Law Guides, and (iv) constraining retrieval through an explicit *Issue–Factor Hierarchy*.

Throughout the analysis, we refer to the same qualitative dimensions used in the main paper: **Legal Issue Identification**, **Factual Faithfulness**, **Logical Coherence**, **Judicial Alignment**, **Precedent Relevance**, and **Precedent Application**. The two precedent dimensions are especially useful in showing where *SCALE* differs most sharply from weaker baselines. In earlier systems, precedent is often generic, topically similar, or even unrelated to the actual legal issue. In stronger systems, especially *SCALE*, precedent is not merely cited but used to structure the legal argument, including support for and against the applicant’s position.

Case 237239: Civil-party constitution in criminal proceedings

Case 237239 is a particularly complex Article 6 case involving the capacity to constitute as a civil party in criminal proceedings. That capacity is not protected *per se* under Article 6, but under certain circumstances a failure to secure effective access

to such a claim may nevertheless engage Convention protection. The legal difficulty is therefore not simply whether the investigation was delayed, but whether the applicants' procedural position, taken together with domestic law and the way they used it, formed the core of a right protected by Article 6. The Court ultimately found *no violation*: despite shortcomings in domestic legislation and significant periods of inactivity, the applicants' use of domestic law meant that the threshold conditions for an Article 6 violation were not met.

This case exposes a recurring failure mode in legal reasoning generation: weaker systems over-generalize from delay or inactivity and infer a violation without first determining whether the underlying civil claim was protected in the circumstances. Stronger systems improve not merely by adding more precedent, but by identifying the threshold issue, selecting the relevant factual factors, and distinguishing supportive from adverse authority.

PR. PR recognizes the broad object of the complaint but does not isolate the threshold issue. Its **Factual Faithfulness** is limited to a surface restatement of investigative inactivity:

In the present case, the applicants lodged complaints with the domestic authorities, which triggered an investigation. However, the investigation was not conducted with the necessary diligence, and no investigative acts were undertaken within the stipulated time-limits.

The reasoning does not explain which factual circumstances were legally decisive. **Logical Coherence** is weak for the same reason: the output restates facts and jumps directly to unfairness without analyzing the domestic procedural setting or the conditions under which constitution as a civil party could engage Article 6. As a result, **Judicial Alignment** is incorrect: the model concludes to a violation because it treats the domestic authorities' inactivity as sufficient in itself, rather than asking whether a protected civil right arose in the first place. **Precedent Relevance** is also weak, since the model invokes only broad and generic authority:

The Court has previously held that the failure to conduct an effective investigation can breach the right to a fair trial ... In the present case, the applicants'

right to a fair trial was compromised by the domestic authorities' failure to conduct a timely and effective investigation.

The citation is not linked to the specific legal question in dispute, and **Precedent Application** is therefore almost non-existent.

CaseRAG. Retrieval over historical judgments introduces some external grounding, but because the query is the full factual narrative, the retrieved context remains noisy. **Factual Faithfulness** improves slightly: the system now identifies expiry of time-limits and discontinuance of proceedings as potentially relevant. However, those facts are still treated in isolation:

The Court considers that this failure to investigate and the discontinuance of the proceedings without any significant steps being taken by the authorities constitutes a violation of the applicants' right to a fair trial under Article 6.

Logical Coherence remains limited because the reasoning does not explain how time-limits, applicant diligence, and the possibility of civil-party constitution interact. **Judicial Alignment** therefore remains incorrect. The system still infers a violation because it does not understand that time-limits and due diligence operate cumulatively, not independently, in determining whether a protected civil claim arose.

A-CaseRAG. Replacing raw-fact querying with anticipatory abstractions improves the retrieval signal. The model begins to retrieve material more closely tied to the legal core of the dispute and shows better **Factual Faithfulness**, for example by considering whether additional conduct by the parties or the judge impaired the applicants' actual ability to constitute themselves as civil parties:

The authorities failed to take steps to ensure that the applicants were informed of the progress of the investigation and the discontinuance of the case, and they were not given the opportunity to participate in the proceedings.

But **Logical Coherence** remains unstable. The system still tends to treat facts one by one instead of analyzing their cumulative effect on the threshold issue. It also retrieves more relevant precedent than **CaseRAG**, but the abstraction step is still free-form,

so the reasoning remains prone to emphasizing a broad fairness narrative rather than the specific conditions under which Article 6 applies.

GuideRAG. A more substantial improvement appears when the retrieval source shifts from dispersed historical cases to article-specific Case Law Guides. Here the model begins to focus on the combination of legislative dysfunction and the diligence of both applicants and domestic authorities. This improves **Precedent Relevance** and **Logical Coherence**, since the guides already summarize article-specific doctrine instead of leaving the model to synthesize it from many scattered cases. However, **Factual Faithfulness** is still incomplete because the system does not yet fully explain how legislative dysfunction may have *de facto* hindered the applicants' ability to bring a civil claim. **Judicial Alignment** remains imperfect: the reasoning becomes more sophisticated, but still supports a violation-oriented reading of the case.

A-GuideRAG. Combining guide-based retrieval with anticipatory querying produces another step change. The model now pays closer attention to due diligence on the part of both the applicants and the judges, and the reasoning becomes more critical and better structured. For instance, it starts to analyze whether the applicants' inability to obtain a decision on the merits was caused by their own conduct, by judicial inactivity, or by structural defects in the domestic system:

The applicants were not required to actively cooperate with the judicial authorities, nor could they be blamed for making full use of the remedies available to them under domestic law ...

This improves **Factual Faithfulness**, **Logical Coherence**, and especially **Precedent Relevance**. Precedent is no longer merely cited; it begins to support a more structured argument. Even so, the system still arrives at a violation. The interpretation is legally plausible and much closer to the Court's reasoning than earlier baselines, but because the query abstraction remains unconstrained, the model still does not fully resolve the threshold issue in the same way as the Court.

SCALE. The strongest performance is obtained with **SCALE**. Here, guide-based retrieval and anticipatory reasoning are further constrained by the explicit *Issue-Factor Hierarchy*. This makes the

threshold issue visible to the model as an issue to be resolved, rather than something it must infer implicitly. As a result, **Factual Faithfulness** reaches its highest level: facts are linked directly to the due diligence standards that determine whether the applicants' situation falls within Article 6.

The Court must examine whether the applicants could have made use of other channels for asserting their civil rights ... the failure to examine a civil-party application on the merits does not ipso facto amount to an unjustified restriction of the right of access to a court.

Logical Coherence is also strongest here, because the reasoning moves through the relevant issues in the same order as the Court's analysis: domestic inactivity, legislative dysfunction, alternative channels for asserting civil rights, and the legal significance of the applicants' own conduct. **Judicial Alignment** is correspondingly highest. The key difference, however, lies in **Precedent Relevance** and **Precedent Application**. Because the IFH links issues and factors to particular guide passages, the system is able not only to retrieve *Nicolae Virgiliu Tănase v. Romania* as controlling authority, but to explain why it governs the dispute. Precedent is thus no longer decorative support; it is integrated into the reasoning architecture itself.

Precedent progression in Case 237239 and related cases. The change in precedent use is especially revealing across this sequence of systems. In *PR*, precedent is broad and generic, with little explanation of why it applies. In *CaseRAG* and *A-CaseRAG*, precedent is sometimes retrieved from cases with only partial or topically superficial overlap. For example, the expert observed instances where the model selected precedents tied to public-interest disclosure or property disputes without adequately explaining how those authorities governed the dispute at hand. From *GuideRAG* onward, precedent becomes more focused because the model reasons over article-specific summaries rather than dispersed case fragments. By the time we reach *A-GuideRAG* and especially **SCALE**, precedents are no longer just thematically relevant; they are justified as either supporting or constraining a particular legal conclusion. This is precisely where the IFH matters most: it helps the model distinguish which precedent governs which issue, and why.

Case 243777: Refusal to register an association

Case 243777 concerns the refusal to register an association on the ground that its name could mislead the public. The factual record is less complex than in Case 237239, but the legal reasoning is still demanding because the Court’s finding of a violation turns on whether the domestic authorities provided *relevant and sufficient* reasons for the refusal. In other words, the task is not merely to identify the aim invoked by domestic courts, but to understand why that aim, as articulated and applied domestically, fails the proportionality analysis under Article 11.

The expert noted that this case reveals a different but equally important progression across systems. In weaker baselines, the model identifies the broad issue—registration refusal—but fails to explain why the authorities’ reasons were inadequate. In stronger baselines, the reasoning gradually becomes more proportionality-sensitive and more attentive to the specific way in which domestic authorities framed the risk of confusion.

PR and CaseRAG. In the weaker systems, the reasoning remains general. *PR* mentions that the refusal rested on a narrow interpretation of the facts, but does not explain how the domestic concern with misleading identity was actually framed. *CaseRAG* improves **Factual Faithfulness** by introducing the purpose behind the refusal, but still does not explain why that purpose mattered legally or how it linked to the actual identity of the association. **Judicial Alignment** is therefore only partial: the system reaches the right broad direction but misses the core reason why the Court rejected the domestic justification.

A-CaseRAG. With anticipatory retrieval over historical cases, the model begins to incorporate domestic law more explicitly. It links the aim of the refusal to factual elements that might justify that aim under domestic legislation, though not yet systematically. This improves **Factual Faithfulness** and **Logical Coherence**, but the reasoning still drifts. The expert noted, for example, that the model sometimes introduced Article 14 considerations that were invoked by the applicants and legally relevant in the background, but not actually analyzed by the Court. This is a good example of why free-form anticipation can improve focus while still remaining unstable.

GuideRAG and A-GuideRAG. Once the retrieval source shifts to Case Law Guides, the argument becomes more proportionality-oriented. The model starts to analyze whether the domestic aim, as framed in domestic law, was compatible with the Court’s case law and whether the concrete facts actually justified refusal. This yields clearer improvements in **Factual Faithfulness** and **Logical Coherence**. In the expert’s reading, *A-GuideRAG* in particular improves because it introduces proportionality analysis rather than treating the refusal as a binary matter of legality:

The domestic courts found that the association’s name was objectively likely to create a misleading image and cause confusion regarding the identity of its members. However, the Court must consider whether this justification is sufficient to outweigh the importance of the right to freedom of association.

This is much closer to the actual reasoning task. Still, the model falls slightly short of fully explaining why identity mattered in this case and why the domestic reasons were insufficient.

SCALE. *SCALE* comes closest to the Court’s reasoning. By structuring the analysis around article-specific issues and factors, it is able to examine the domestic aim, the applicants’ stated purpose, the proportionality framework, and the absence of concrete evidence supporting the refusal. This produces the strongest **Factual Faithfulness** and **Logical Coherence**:

The Government argued that the name was misleading and could cause confusion, as it referred to Turkish nationality, whereas the members were Greek nationals. However, the Court notes that the association’s charter clearly stated that membership was open only to adult women of Greek nationality who were residents of the Prefecture of Xanthi.

The reasoning then closes the argumentative gap that earlier systems leave open:

The applicants had the right to choose a name that reflected their cultural and ethnic identity, and the name ‘Cultural Association of Turkish Women of the Prefecture of Xanthi’ did not necessarily imply that the members were Turkish nationals.

Here, **Judicial Alignment** is strongest because the model does not merely cite the Court’s principles; it applies them to the actual domestic justification. The same improvement appears in **Precedent Relevance** and **Precedent Application**: the system selects precedents tied to association naming, public order, and proportionality, and explains why those authorities are controlling in this case instead of merely topically similar.

Overall observations from the case studies

Taken together, the two case studies illustrate the same broader progression that appears in the aggregate results. First, moving from *PR* to retrieval-based systems improves grounding, but retrieval over historical judgments remains noisy and often retrieves authority that is topically similar without being legally decisive. Second, replacing raw-fact queries with anticipatory abstractions helps by removing irrelevant detail and focusing retrieval on the latent legal content of the case, but free-form anticipation alone remains unstable in difficult cases. Third, replacing historical judgments with article-specific Case Law Guides reduces the synthesis burden by providing denser, more authoritative legal summaries. Finally, adding the explicit *Issue–Factor Hierarchy* yields the clearest gains: it stabilizes legal issue identification, improves the selection of controlling authority, and enables precedent to be applied as part of the legal argument rather than appended as support.

The expert’s comments also help explain why the largest improvements in the quantitative analysis appear in **Legal Issue Identification**, **Precedent Relevance**, and **Precedent Application**. The strongest systems do not simply cite better cases or write more fluent prose. They identify the legally decisive issue more accurately, connect the relevant facts to that issue, and then select precedent in a way that supports or constrains the argument in line with the Court’s own reasoning. This is precisely the behavior that *SCALE* is designed to induce.

C Expert Evaluation of Issue–Factor Hierarchies

We evaluate the quality of the automatically constructed Issue–Factor Hierarchies (IFHs) at both the issue and factor levels using expert assessment on a five-point Likert scale (Likert, 1932), where 1 denotes poor quality and 5 denotes high quality.

Criterion / Article	5	6	8	10
Issues (1–5)				
Completeness	5	5	5	5
Precision	4	5	5	4
Redundancy	5	4	4	4
Factors (1–5)				
Completeness	5	5	5	4
Correctness	4	5	5	4
Redundancy within issue	4	4	4	3
Redundancy across issues	4	4	4	3

Table 4: Expert evaluation of automatically constructed Issue–Factor Hierarchies for four representative Convention articles. Scores are on a 1–5 Likert scale, where higher is better. For redundancy criteria, higher scores indicate lower redundancy.

Issue-level criteria. Issues are evaluated along three dimensions: **Completeness**, which assesses whether all issues required under the article are identified without omissions; **Precision**, which evaluates whether issues are accurate, clearly defined, and directly relevant to the article; and **Redundancy**, which examines whether issues are unnecessarily repeated and, if so, whether they are adequately distinguished.

Factor-level criteria. Factors are evaluated along four dimensions: **Completeness**, which assesses whether all relevant factors for each issue are identified without omissions; **Correctness**, which evaluates whether factors are conceptually valid and correctly assigned to their respective issues; **Redundancy within issues**, which examines duplicate or overlapping factors under the same issue; and **Redundancy across issues**, which assesses whether similar factors appear under multiple issues and whether they are consistently distinguished, cross-referenced, or linked.

Table 4 reports expert evaluation of the constructed IFHs for four representative Convention articles (Articles 5, 6, 8, and 10). Overall, the hierarchies receive strong scores across both issue- and factor-level dimensions.

At the issue level, completeness is rated at the maximum score for all four articles, indicating that the extracted issues adequately cover the legal scope represented in the Case Law Guides. Precision scores are also high, suggesting that the extracted issues are well defined and aligned with

article-specific legal structure. Redundancy remains low, indicating that the merging procedure successfully consolidates chunk-level hierarchies without introducing substantial duplication.

At the factor level, completeness and correctness are likewise strong, indicating that the extracted factors are both substantively appropriate and well assigned to the corresponding issues. Slightly lower redundancy scores, particularly across issues for Articles 8 and 10, reflect overlap between factors that naturally recur across related proportionality analyses rather than major structural errors in extraction.

These findings support the reliability of the IFH construction process and suggest that the resulting hierarchies provide a coherent and legally meaningful intermediate representation for downstream reasoning.

D Implementation Details

We provide the prompt templates used in the *SCALE* pipeline and the baselines, and the LLM-as-a-Judge evaluation. For *SCALE*, prompting is used at three stages. In stage one, for creation of Issue-Factor hierarchies, we use the **Local Issue-Factor Hierarchy creation prompt** (Figure 2) extracts candidate legal issues, factors, and supporting paragraph references from each Case Law Guide chunk. The **Local Issue-Factor Hierarchy merge prompt** (Figure 3) then consolidates chunk-level hierarchies into a single article-level structure. At inference time, the **Issue selection prompt** (Figure 4) identifies which issues from the hierarchy are implicated by the facts of a new case, while the **Factor selection prompt** (Figure 5) selects the applicable factors under those issues. Finally, the **Reasoning prompt** (Figure 6 and 7) generates the precedent-grounded legal analysis from the case facts and the retrieved guide passages associated with the selected issues and factors.

The baselines use subsets of these prompts. *PR* uses the **Parametric Reasoning prompt** (Figure 8 & 9) without external retrieval. *CaseRAG* and *GuideRAG* use the same **Reasoning Prompt** (Figure 6 & 7) which is used by *SCALE*, but with retrieved historical-case paragraphs or guide paragraphs appended as context. *A-CaseRAG* and *A-GuideRAG* additionally use the **Anticipation Extraction prompt** (Figure 10) to convert the raw case facts into anticipatory legal-principle queries before retrieval. The value for K in top-K retrieval

is 20 for each RAG baseline.

For evaluation, criterion-specific LLM-as-a-Judge prompts are used: **Issue Identification** (Figure 11), **Factual Accuracy** (Figure 12), **Coherence** (Figure 13), two precedent-specific prompts for **Precedent Relevance** (Figure 14) and **Precedent Application** (Figure 15) and **Judicial Alignment** (Figure 16)

For generating reasoning, to turn off sampling, we set *do_sample* parameter of the models to *False* for deterministic outputs. For LLM-as-a-Judge model, GPT-4.1, we set temperature to 0. to reduce stochasticity.

Local Issue-Factor Hierarchy Creation Prompt

You are a **legal expert** specializing in understanding, interpreting, and applying the **European Convention on Human Rights (ECHR)** and related case law. You will be provided with **sets of paragraphs** from a case law guide. Each paragraph will be labeled with a **unique paragraph number** (e.g., ¶1, ¶2, etc.). For each set, your task is to:

- Identify all Legal Issues and Legal Factors** presented in the text.
- Construct a self-contained, hierarchical structure** that organizes:
 - Legal Issues** (top-level nodes)
 - Legal Factors** (nested beneath the corresponding issue they relate to)

Each hierarchy should stand read coherently based **only on the paragraph(s) provided** — do not introduce background or specific information unrelated to the context.

For **each issue and factor**, include the **paragraph number(s)** that support or mention it. This allows for easy traceability and review.

Use the following definitions:

- Legal Issue:** A specific question of law that arises from facts and is disputed by the parties and must be resolved by the judge. It determines the applicable legal principles or precedents.
- Legal Factor:** Criteria or considerations established by law or precedent that guide the resolution of a legal issue.

Output Format Example:

Issue 1: Whether a valid contract was formed (¶1, ¶3)
└ Factor A: Existence of offer and acceptance (¶1)
└ Factor B: Presence of consideration (¶3)

Issue 2: Whether the contract was enforceable under public policy (¶2)
└ Factor C: Agreement not contrary to law or morals (¶2)

Issue 3: Whether slapping a child constitutes degrading or inhuman treatment?
└ Factor D: What were the circumstances under which the child was slapped?

Notes:

- If an issue or factor is supported by multiple paragraphs, list **all relevant paragraph numbers**.
- If no clear issue or factor is identifiable in a paragraph, simply note: “*No identifiable issue or factor in ¶[x]*”.

This task repeats for each new paragraph set.

Figure 2: **Local Issue-Factor Hierarchy creation prompt** takes a chunk of tokens from Case-Law Guides and creates IFH out of it.

Local Issue-Factor Hierarchy Merge Prompt

You are a **legal expert** specializing in understanding, interpreting, and applying the **European Convention on Human Rights (ECHR)** and related case law. You will be provided with several independently created **issue-factor hierarchies**, each of which has been derived from a different set of paragraphs from a case law guide. Your task is to **merge these hierarchies into a single, coherent master hierarchy** that organizes and consolidates the legal issues and their associated factors across all the provided sources taking into account their similarity and reducing redundancy.

Each hierarchy will include:

- One or more **legal issues**, and under each, a set of relevant **legal factors**.
- References to the **paragraph numbers** from which each issue or factor was extracted.

Please follow these definitions when analyzing and merging the hierarchies:

- **Legal Issue:** A specific question of law that arises from facts and is disputed by the parties and must be resolved by the judge. It determines the applicable legal principles or precedents.
- **Legal Factor:** Criteria or considerations established by law or precedent that guide the resolution of a legal issue.

Instructions for merging:

- Identify and consolidate **identical or closely related legal issues** across the hierarchies.
- Under each merged issue, combine the **unique and overlapping legal factors** from the various hierarchies.
- Retain the **original paragraph references** for each issue and factor (they may appear in multiple places).
- Ensure the **final structure is logically organized**, with clear nesting of issues and factors.
- Do not omit any distinct legal issue or factor unless it's truly redundant or duplicative in both meaning and reference.

Your output should be a single, integrated “*issue-factor hierarchy*”, clearly reflecting the consolidated legal issues and their associated factors.

Figure 3: **Local Issue-Factor Hierarchy Merge** step takes local IFHs iteratively from **Local Issue-Factor Hierarchy Creation** step and merge them into each other till a single IFH is formed based on rules laid down by **Local Issue-Factor Hierarchy Merge Prompt**.

Issue Selection Prompt

You are a **legal expert** specializing in understanding, interpreting, and applying the **European Convention on Human Rights (ECHR)** and related case law. You will be provided with the following, step by step:

1. An **independently created issue-factor hierarchy**, derived from a separate set of case law guide paragraphs.
2. A **master hierarchy**, updated in the previous step to reflect the last merge.

Your task is to **merge the new hierarchy into the master hierarchy**, creating a single, coherent structure that consolidates legal issues and their associated factors across all sources, while minimizing redundancy.

Structure of Each Hierarchy

- One or more **legal issues**, each with relevant **legal factors**.
- References to the **paragraph numbers** from which each issue or factor was extracted.

Key Definitions

- **Legal Issue**: A distinct legal question that arises from the facts, disputed by the parties, and resolved by the judge. It determines which legal principles or precedents apply.
- **Legal Factor**: A criterion or consideration established by law or precedent that informs how a legal issue is resolved.

Merging Instructions

- Identify and consolidate **identical or closely related legal issues** across the hierarchies.
- Under each merged issue, combine the **unique and overlapping legal factors** from the various hierarchies.
- Retain the **original paragraph references** for each issue and factor (they may appear in multiple places).
- Ensure the **final structure is logically organized**, with clear nesting of issues and factors.
- Do not omit any distinct legal issue or factor unless it's truly redundant or duplicative in both meaning and reference.
- All the issues in the hierarchy should be numbered.

Critical Output Instructions

“This is the most important part. Follow these formatting rules precisely.”

- **Output ONLY the Changed or New Issues**: Your response must only contain the full text of the issues from the Master Hierarchy that were modified or newly added in this step.
- **Show the Complete Issue**: When an issue is updated (e.g., a new factor is added), you must output the **entire issue** from top to bottom, including all its old and new factors.
- **NO Placeholders or Summaries**: **NEVER** use shortcuts like [No change], [Same as before], ..., or other placeholders. Always write out the full content of the updated issue.
- **Continue Numbering**: When adding new issues, continue the numbering from where the Master Hierarchy left off. (like Issue 1:, Issue 2: and so on)

Example:

Imagine the **Master Hierarchy** has:

Issue N: Right to a Fair Trial (¶15)
* Access to a court (¶16)
* Equality of arms (¶17)

And the **Independent Hierarchy** has:

Issue: Right to a Fair Trial (¶99)
* Equality of arms (¶100)
* Right to legal assistance (¶101)

Your **Correct Output** would be:

Issue N: Right to a Fair Trial (¶15, ¶99)
* Access to a court (¶16)
* Equality of arms (¶17, ¶100)
* Right to legal assistance (¶101)

Your **Incorrect Output** would be:

Issue N: Right to a Fair Trial (¶15, ¶99)
* [No change]
* Equality of arms (¶17, ¶100)
* Right to legal assistance (¶101)

Figure 4: The LLM is prompted with the **Issue Selection Prompt** when based on the case facts, the model needs to select the relevant issues from the IFH obtained.

Factor Selection Prompt

You are a **legal expert** specializing in understanding, interpreting, and applying the **European Convention on Human Rights (ECHR)** and related case law. You will be given:

1. A list of **Legal Issues**.
2. The **Legal Factors** corresponding to each issue.
3. The **ECHR Case Facts**.

Your task is to **identify and output only the Legal Factors along with given page numbers in brackets** that are **applicable** to the given case facts.

- Carefully read the case facts.
- Examine the legal issue.
- Output **the applicable factor names along with paragraph numbers in brackets**, separated by commas or line breaks.
- If **no factor** applies, output exactly: **NO** (in uppercase, with no extra text).
- Do **not** output issues, explanations, or reasoning — only the matching factor(s) with its full text and page numbers.

Output Format:

Clearly list the chosen relevant factors only along with mentioned paragraph numbers without any explanation.

Figure 5: After the **issue selection** is done, we prompt the model with **Factor Selection prompt** and each selected issue along with its corresponding factors, to get contributing factors iteratively for a particular case.

Reasoning Prompt (1st Half)

You are acting as a “*judge of the European Court of Human Rights (ECtHR)*”. You will be provided with (A) **case facts** relating to an application submitted to the Court, (B) selected **Case Law Guide** paragraphs (legal authority excerpts) and (C) — for iterations after the first — the **previous iteration’s LAW section** produced by you. Your task is to produce the Court’s **LAW** section for each alleged violation set out in the facts, updating prior reasoning when new guide paragraphs are supplied.

Overall Rules (apply every iteration)

1. **Output only the LAW section(s)** and nothing else. Do not provide any explanation of what changed since the previous iteration. Do not include private chain-of-thought. Produce no introductory or closing remarks beyond the LAW sections requested below.
2. **When this is not the first iteration**, you will be provided with the prior iteration’s LAW section as the input labelled PREVIOUS_LAW. Start from that text and **update it** to reflect the new Case Law Guide paragraphs and any new facts provided — preserving structure, but changing wording and conclusions where warranted.
3. **Treat applicant allegations as allegations** — don’t assume facts are established unless the case facts state them as such.
4. Maintain a **formal judicial tone** consistent with ECtHR practice.
5. **Do not consolidate multiple Articles** into a single analysis — produce a separate LAW subsection for each Article alleged.

Input You Will Receive (per iteration)

- **FACTS** — case facts (may include applicant and Government submissions)
- **CASE_LAW_GUIDE_PARAGRAPHS** — one or more numbered paragraphs/excerpts from Case Law Guides (these are to be treated as authoritative material to cite and apply)
- (Only for subsequent iterations) **PREVIOUS_LAW** — the full LAW section produced by you in the prior iteration

Required Structure (for each alleged Article) — Exact Format

LAW

[Article X] – Alleged violation

1. “*The Parties’ Submissions*”
Briefly summarise the key submissions made by the applicant and the Government (only if present in the facts).
2. “*The Court’s Assessment*”
Provide a detailed and neutral legal assessment:
 - Interpret the “*scope of rights and obligations under the relevant Article(s) by using relevant case law details from Case Law Guides*”;
 - Analyse whether the applicant’s claims, when assessed against the facts and legal standards, disclose a violation;
 - Rely on “*well-established ECtHR case law and precedents*” to substantiate your reasoning “*and identify relevant Case Law paragraphs to support your argument*”;
 - Apply core ECtHR “*relevant legal concepts*” such as “*proportionality*”, “*margin of appreciation*”, “*necessity in a democratic society*”, and “*procedural safeguards*”, where appropriate and “*identify how they fit the patterns described in the Case Law Guides*”;
 - Reference specific “*past judgments*” with similar facts or legal questions to demonstrate how the current case aligns with or departs from prior jurisprudence by “*referring to Case Law Guides examples and highlight where relevant the most relevant precedent*”;
 - Maintain a formal, judicial tone and structure consistent with ECtHR practice.
3. “*Conclusion*”
State clearly whether the Court finds a **violation** or **no violation** of the relevant Article.

Figure 6: After selection of issues and factors, the corresponding paragraphs are fetched from the Case Law Guide. The paragraphs along with **Reasoning Prompt** and case facts are prompted to the model to generate the reasoning. Remaining prompt is covered in the Figure 7

Reasoning Prompt (cont.)

Critical Guidance

- The case facts may include “*allegations*” by the applicant. **Do not treat these as established violations.**
- Your role is to act as a “*neutral judge*”: independently assess the facts against Convention standards.
- **Do not confirm any violation without thorough legal analysis** grounded in ECtHR jurisprudence.
- Emphasise “*comparative legal reasoning*” by citing relevant ECtHR “*precedents and interpretative standards*.”
- “*When case law details and references are found in several Case Law Guides make sure to identify the relevant perspective in relation to the provided context.*”
- Avoid summarising the facts; focus on legal analysis and the implications of the facts under the Convention and “*rely on relevant Case Law guide paragraphs to support legal assessment of a situation.*”
- If the applicant raises allegations under more than one Article, **structure your response accordingly** by generating a separate “LAW” subsection for **each** Article (e.g., “Article 3 – Alleged violation”, “Article 8 – Alleged violation”, etc.).
- **Do not consolidate multiple Articles into a single analysis.**

Formatting & Length

- Keep each Article subsection self-contained and clearly labelled as above.
- Be thorough in legal reasoning but concise: aim for well-structured paragraphs and precise citations to the supplied Case Law Guide paragraphs.
- Introduce headings **where needed**.

Figure 7: Remaining half of the prompt from Figure 6 After selection of issues and factors, the corresponding paragraphs are fetched from the Case Law Guide. The paragraphs along with **Reasoning Prompt** and case facts are prompted to the model to generate the reasoning.

Parametric Reasoning Prompt (1st half)

You are acting as a “*judge of the European Court of Human Rights (ECtHR)*”. You will be provided with (A) **case facts** relating to an application submitted to the Court, (B) selected **Case Law Guide** paragraphs (legal authority excerpts) and (C) — for iterations after the first — the **previous iteration’s LAW section** produced by you. Your task is to produce the Court’s **LAW** section for each alleged violation set out in the facts, updating prior reasoning when new guide paragraphs are supplied.

Overall Rules (apply every iteration)

1. **Output only the LAW section(s)** and nothing else. Do not provide any explanation of what changed since the previous iteration. Do not include private chain-of-thought. Produce no introductory or closing remarks beyond the LAW sections requested below.
2. **When this is not the first iteration**, you will be provided with the prior iteration’s LAW section as the input labelled PREVIOUS_LAW. Start from that text and **update it** to reflect the new Case Law Guide paragraphs and any new facts provided — preserving structure, but changing wording and conclusions where warranted.
3. **Treat applicant allegations as allegations** — don’t assume facts are established unless the case facts state them as such.
4. Maintain a **formal judicial tone** consistent with ECtHR practice.
5. **Do not consolidate multiple Articles** into a single analysis — produce a separate LAW subsection for each Article alleged.

Input You Will Receive (per iteration)

- FACTS — case facts (may include applicant and Government submissions)
- CASE_LAW_GUIDE_PARAGRAPHS — one or more numbered paragraphs/excerpts from Case Law Guides (these are to be treated as authoritative material to cite and apply)
- (Only for subsequent iterations) PREVIOUS_LAW — the full LAW section produced by you in the prior iteration

Required Structure (for each alleged Article) — Exact Format

LAW

[Article X] – Alleged violation

1. “*The Parties’ Submissions*”
Briefly summarise the key submissions made by the applicant and the Government (only if present in the facts).
2. “*The Court’s Assessment*”
Provide a detailed and neutral legal assessment:
 - Interpret the “*scope of rights and obligations under the relevant Article(s) by using relevant case law details from Case Law Guides*”;
 - Analyse whether the applicant’s claims, when assessed against the facts and legal standards, disclose a violation;
 - Rely on “*well-established ECtHR case law and precedents*” to substantiate your reasoning “*and identify relevant Case Law paragraphs to support your argument*”;
 - Apply core ECtHR “*relevant legal concepts*” such as “*proportionality*”, “*margin of appreciation*”, “*necessity in a democratic society*”, and “*procedural safeguards*”, where appropriate and “*identify how they fit the patterns described in the Case Law Guides*”;
 - Reference specific “*past judgments*” with similar facts or legal questions to demonstrate how the current case aligns with or departs from prior jurisprudence by “*referring to Case Law Guides examples and highlight where relevant the most relevant precedent*”;
 - Maintain a formal, judicial tone and structure consistent with ECtHR practice.
3. “*Conclusion*”
State clearly whether the Court finds a **violation** or **no violation** of the relevant Article.

Figure 8: **Parametric Reasoning prompt**. Case facts along with the alleged article is given to the model to generate the reasoning. Remaining prompt is present in Figure 9

Parametric Reasoning Prompt (cont.)

Critical Guidance

- The case facts may include “*allegations*” by the applicant. **Do not treat these as established violations.**
- Your role is to act as a “*neutral judge*”: independently assess the facts against Convention standards.
- **Do not confirm any violation without thorough legal analysis** grounded in ECtHR jurisprudence.
- Emphasise “*comparative legal reasoning*” by citing relevant ECtHR “*precedents and interpretative standards*.”
- “*When case law details and references are found in several Case Law Guides make sure to identify the relevant perspective in relation to the provided context.*”
- Avoid summarising the facts; focus on legal analysis and the implications of the facts under the Convention and “*rely on relevant Case Law guide paragraphs to support legal assessment of a situation.*”
- If the applicant raises allegations under more than one Article, **structure your response accordingly** by generating a separate “LAW” subsection for **each** Article (e.g., “Article 3 – Alleged violation”, “Article 8 – Alleged violation”, etc.).
- **Do not consolidate multiple Articles into a single analysis.**

Formatting & Length

- Keep each Article subsection self-contained and clearly labelled as above.
- Be thorough in legal reasoning but concise: aim for well-structured paragraphs and precise citations to the supplied Case Law Guide paragraphs.
- Introduce headings **where needed**.

Figure 9: Remaining prompt from Figure 8. **Parametric Reasoning prompt.** This is continuation of the above prompt.

Anticipation Extraction Prompt

You are an expert legal analyst specializing in the case law of the European Convention on Human Rights (ECHR). Your task is to analyze the provided case facts and extract all the relevant legal principles that could be applied to them. The output will be used to match against a database of ECHR legal sections, so it must be clear, concise, and strictly follow the format below.

Core Instructions:

1. Identify the core of a situation as its peers across case law guides/ECHR language.
2. Identify variations of this situation by focusing on what might constitute a new fact that may affect the core of a legal situation.

Output Format Specification

Your entire response must be a plain text list. Each principle must be presented in a two-line block, numbered sequentially. Separate each block with a single blank line.

- **Line 1:** Must begin with Principle X: (where X is the number), followed by the concise statement of the legal principle.
- **Line 2:** Must begin with Article:, followed by the relevant ECHR Article number(s).

Figure 10: **Anticipation Extraction Prompt** is used as pre-cursor to Anticipatory baselines to get only the anticipations from the case facts to be used to fetch Precious Cases/Case Law paragraphs from guides.

Issue Identification Prompt

You are a **legal expert** specializing in understanding, interpreting, and applying the **European Convention on Human Rights (ECHR)** and related case laws but you are not a court judge. Your task is to evaluate the quality of an LLM's legal reasoning by comparing it to authoritative ground-truth reasoning.

You will be given two pieces of reasoning:

1. **Ground-Truth Reasoning (Correct / Expected Reasoning)**
2. **LLM-Generated Reasoning**

Your goal is to assess how accurately the LLM identifies and applies the relevant legal issues.

Instructions

Analyze the reasoning and provide the following:

1. **Missing Issues**
 - Identify any legal issues, principles, tests, or considerations present in the Ground-Truth Reasoning but “*absent*” from the LLM reasoning.
2. **Extra or Incorrect Issues**
 - Identify any legal issues or considerations introduced by the LLM that are **not** supported by the Ground-Truth Reasoning and **should not** apply under ECHR law.
3. **Correctly Identified Issues**
 - List the issues or legal principles that the LLM reasoning correctly identified and that match the Ground-Truth Reasoning.
4. **Score (0–5)**

Use the rubric below to rate the LLM's reasoning accuracy:

 - **5** – All ground-truth issues correctly identified; no incorrect or irrelevant issues added.
 - **4** – Most issues correctly identified; only minor omissions or one minor incorrect issue.
 - **3** – Some alignment, but with clear omissions or multiple extra issues.
 - **2** – Few correct issues; significant gaps and/or several incorrect applications.
 - **1** – Very poor alignment; most reasoning incorrect.
 - **0** – No alignment; reasoning entirely misses the ground-truth issues.
5. **Summary Judgment**
 - Provide a concise explanation justifying the score, referencing correctness, omissions, and misuse of ECHR legal concepts.

Output Format (Strict):

Missing Issues:

- ...

Extra / Incorrect Issues:

- ...

Correctly Identified Issues:

- ...

Score (0–5):

...

Summary Judgment:

...

Figure 11: **Issue Identification Prompt** used by the LLM judge to assess whether issues have been correctly identified. The judge evaluates this by comparing the generated reasoning against the ground truth reasoning.

Factual Accuracy Prompt

You are a **legal expert** specializing in understanding, interpreting, and applying the **European Convention on Human Rights (ECHR)** and related case law. Your task is to critically analyze whether the **generated reasoning** about an ECHR article is factually faithful to the **case facts** provided.

Your focus is **not** on whether the legal assessment is correct, but strictly on whether the reasoning introduces **hallucinated facts** — i.e., statements, events, dates, motivations, procedural details, or legal elements **not present** in the case facts.

Your Tasks

1. **Read the Case Facts carefully.**
2. **Read the Generated Reasoning carefully.**
3. Compare the reasoning with the facts and determine whether the reasoning:
 - Accurately reflects the facts, **without adding new facts**.
 - Does **not** invent parties, actions, procedural history, motivations, legal elements, or events that are missing from the case facts.
 - Does **not** misstate or distort factual information.
4. Assign a **Factuality Score (1–5)**:
 - **5** – No hallucinations. All reasoning is fully grounded in the provided facts.
 - **4** – Minor imprecision but no meaningful hallucinations.
 - **3** – Some hallucinated or inaccurate details, but the majority reflects the facts.
 - **2** – Frequent hallucinations or invented details that undermine factual accuracy.
 - **1** – Severe hallucinations; the reasoning largely invents facts or contradicts the given case facts.
5. Provide a **brief justification** for your score (2–5 sentences).

Output Format:

Factuality Score: X/5

Explanation:

- [Your brief explanation identifying any hallucinated or unsupported details, or confirming none exist.]

Figure 12: **Factual Accuracy Prompt** used by the LLM judge to assess whether the generated reasoning is factually consistent with the case facts, ensuring that no facts have been hallucinated or fabricated.

Coherence Prompt

You are a **legal expert** specializing in understanding, interpreting, and applying the **European Convention on Human Rights (ECHR)** and related case law. I will provide you with a piece of generated legal reasoning related to a specific legal article. Your task is to evaluate the reasoning based on the criteria below and then provide (1) an assessment and (2) a rating from 1 to 5.

Evaluation Criteria

1. Textual Fidelity

- The reasoning stays grounded in the wording of the legal article.
- Key terms and conditions are used correctly.
- No claims contradict or exceed the text without justification.

2. Logical Structure & Internal Consistency

- The reasoning follows a clear argument structure: premise → reasoning → conclusion.
- No contradictions or circular reasoning.
- No unjustified leaps between statements.

3. Accuracy of Implicatures (Derived Meaning)

- Implicit conclusions genuinely follow from the article.
- Inferences are legally and logically valid.
- No over-interpretation or introduction of unsupported rights/obligations.

4. Completeness of Analysis

- Essential aspects of the article are addressed.
- Relevant conditions or exceptions are considered.
- No major omissions that affect the interpretation.

5. Legal Plausibility

- Reasoning aligns with general legal principles (e.g., consistency, proportionality).
- No absurd or legally implausible interpretations.
- Any analogies or external principles are relevant and appropriate.

6. Clarity & Precision

- Terminology is clear and properly used.
- No ambiguity or vague claims.
- Inferential steps are explicitly and transparently communicated.

7. Evidence of Understanding

- The reasoning demonstrates knowledge of the legal concept, not rote responses.
- Examples or illustrations, if present, apply the rule correctly.

Output Requirements

After reviewing the reasoning, provide:

1. **A short evaluation** (5–7 sentences) addressing the criteria above.
2. **A rating from 1 to 5**, where:
 - **5** = excellent coherence and implicature quality
 - **4** = strong with minor issues
 - **3** = adequate but with noticeable flaws
 - **2** = weak with significant issues
 - **1** = incoherent or fundamentally flawed reasoning

Output Format:

Score: X/5

Explanation:

- [Your evaluation addressing the criteria above.]

Figure 13: **Coherence Prompt** used by the LLM judge to assess the logical coherence of the generated reasoning, evaluated solely on the basis of the generated reasoning itself without reference to any external sources.

Precedent Relevance Prompt

You are a **legal expert** specializing in understanding, interpreting, and applying the **European Convention on Human Rights (ECHR)** and related case law. Your task is to determine whether a cited precedent paragraph is genuinely relevant to the current case and whether the citation appears appropriate.

You will receive three inputs:

1. **Current Case Reasoning** – the legal reasoning generated by the model
2. **Precedent Citation** – the case name and citation (e.g., *A v. B*, 2022, §§138–139).
3. **Cited Paragraph(s) from the Precedent** – only the specific paragraphs referenced (e.g., §§138–139), not the full judgment.

Your objective is to evaluate whether the cited paragraphs **actually support the legal point for which they are cited** in the current case.

Instructions

1. Carefully read the **current case reasoning** to understand the **legal proposition or claim** being supported by the citation.
2. Read the **cited paragraph(s)** from the precedent and identify the **legal rule, principle, or reasoning** expressed in those paragraphs.
3. Determine whether the cited text:
 - Directly supports the legal proposition
 - Partially relates but does not clearly support the proposition
 - Is unrelated or misapplied
4. Focus only on **whether the citation is relevant to the claim**, not whether the precedent ultimately governs the case (that is evaluated separately in precedent applicability).

Rating Scale (1–5)

- **5 – Highly Relevant**
The cited paragraphs clearly and directly support the legal proposition for which the case is cited.
- **4 – Strongly Relevant**
The cited text supports the proposition with minor interpretive gaps.
- **3 – Moderately Relevant**
The cited text relates to the topic but does not clearly establish the exact proposition.
- **2 – Weakly Relevant**
The cited paragraphs are only loosely connected to the legal proposition.
- **1 – Not Relevant / Mis-citation**
The cited paragraphs do not support the proposition or appear to be misapplied.

Output Format

Relevance Rating: <1-5>

Legal Principle in the Cited Paragraph(s):
<Briefly summarize the legal rule or reasoning expressed
in the cited paragraphs>

Proposition in the Current Case:
<Briefly restate the claim or point for which the
precedent is cited>

Assessment of Relevance:
<Explain whether and how the cited paragraphs support
the proposition>

Conclusion:
<A short statement on whether the citation appears
appropriate or misplaced>

Figure 14: **Precedent Relevance Prompt** used by the LLM judge to assess whether a precedent relevant in the generated reasoning, by comparing it against the corresponding paragraph from the cited source.

Precedent Application

You are a **legal expert** specializing in understanding, interpreting, and applying the **European Convention on Human Rights (ECHR)** and related case law. Your task is to evaluate whether a cited legal precedent is applicable to a current case. You will receive three inputs:

1. **Precedent Case Citation** – the cited case
2. **Relevant Paragraphs from the Precedent** – only the specific paragraphs cited in the judgment (e.g., §§138–139). These contain the legal principle relied upon.
3. **Current Reasoning** – the reasoning generated by the model

Your objective is to determine how applicable the cited precedent is to the current case.

Instructions

1. Carefully read the precedent paragraphs and identify the legal principle or rule stated.
2. Analyze the material facts and legal issues in the current case's reasoning.
3. Compare the precedent with the current case by focusing on:
 - Similarity of material facts
 - Similarity of legal issues
 - Whether the legal principle from the precedent logically governs the current case
4. Determine the degree of applicability of the precedent to the current case.

Rating Scale (1–5)

- **5 – Highly Applicable**
The precedent addresses nearly identical legal issues and materially similar facts. The legal rule clearly governs the current case.
- **4 – Strongly Applicable**
The precedent is very relevant with mostly similar facts or legal principles, though some minor distinctions exist.
- **3 – Moderately Applicable**
The precedent is related and may provide guidance, but factual or contextual differences limit direct applicability.
- **2 – Weakly Applicable**
The precedent touches on a related concept but differs significantly in facts or legal context.
- **1 – Not Applicable**
The precedent is largely irrelevant due to different legal issues or materially different facts.

Output Format

Provide your answer in the following structure:

Applicability Rating: <1-5>

Legal Principle from Precedent:

Comparison with Current Case:

Reasoning: <Explain why the precedent is applicable or not, referring to facts and legal issues>

Conclusion:

Figure 15: **Precedent Application Prompt** used by the LLM judge to assess whether a precedent has been correctly applied in the generated reasoning, by comparing it against the corresponding paragraph from the cited source.

Judicial Alignment Prompt

You are a **legal expert specializing in reading, interpreting, and comparing European Court of Human Rights (ECHR) cases**. You will be given two texts:

1. **LLM-Generated Reasoning** – an AI-produced explanation or analysis of an ECHR case.
2. **Ground-Truth Reasoning** – the correct or authoritative explanation/analysis.

Your task is to **evaluate how well the LLM-generated reasoning aligns with the ground truth**.

Consider the following when scoring alignment:

- Accuracy of legal principles
- Correct use of ECHR articles
- Faithfulness to the facts of the case
- Logical coherence and correctness
- Completeness relative to the ground truth

Provide:

1. **A numerical score from 0 to 5**
 - **5** = Perfectly aligned
 - **4** = Mostly aligned, minor issues
 - **3** = Some alignment, but notable inaccuracies or omissions
 - **2** = Poor alignment, major errors
 - **1** = Very poor, barely relates
 - **0** = Completely misaligned
2. **A brief justification** explaining why you gave this score.

Output Format:

Score: X/5,

Explanation:

[short explanation of alignment quality]

Figure 16: **Judicial Alignment Prompt** used by the LLM judge to assess how closely the generated reasoning aligns with the ground truth reasoning.

Chunk (Local) Issue-Factor Hierarchy

I. General considerations

D. Types of prohibited treatment or punishment

1. Torture

- 9. The prohibition of torture has achieved the status of *jus cogens* or of a peremptory norm in international law (Advisory opinion on the applicability of statutes of limitation to prosecution, conviction and punishment in respect of an offence constituting, in substance, an act of torture [GC], § 59, 2022). In order to determine whether a particular form of ill-treatment should be qualified as torture, the Court will have regard to the distinction embodied in Article 3 between this notion and that of inhuman or degrading treatment. It was the intention that the Convention should, by means of this distinction, attach a special stigma to deliberate inhuman treatment causing very serious and cruel suffering; the same distinction is drawn in Article 1 of the United Nations Convention against Torture and Other Cruel, Inhuman or Degrading Treatment or Punishment “UNCAT” (Ireland v. the United Kingdom, 1978, § 167, *Selmouni v. France* [GC], 1999, § 96 and *Ilaşcu and Others v. Moldova and Russia* [GC], 2004, § 426).

- 10. In addition to the severity of the treatment, there is a purposive element, as recognised in the UNCAT, which defines torture in terms of the intentional infliction of severe pain or suffering with the aim, *inter alia*, of obtaining information or a confession, inflicting punishment or intimidation (*Selmouni v. France* [GC], 1999, § 97; *Salman v. Turkey* [GC], 2000, § 114; *Al Nashiri v. Poland*, 2014, § 508 and *Petrosyan v. Azerbaijan*, 2021, § 68).

- 11. Having regard to the fact that the Convention is a living instrument which must be interpreted in the light of present-day conditions, acts which were classified in the past as “inhuman and degrading treatment” as opposed to “torture” could be classified differently in future. The Court has taken the view that the increasingly high standard being required in the area of the protection of human rights and fundamental liberties correspondingly and inevitably requires greater firmness in assessing breaches of the fundamental values of democratic societies (*Selmouni v. France* [GC], 1999, § 101).

- 12. In this respect, the Court has emphasized that the prohibition of torture has achieved the status of *jus cogens* or a peremptory norm in international law.

- 13. For instance, treatment was found to amount to “torture” when:

* the applicant was stripped naked, with his arms tied together behind his back and suspended by his arms (“Palestinian hanging”) by State agents while in police custody in order to extract a confession (*Aksoy v. Turkey*, 1996, § 64; see also *Israilov v. Russia*, 2023, § 165, where the applicant was subjected to severe physical and mental suffering, inflicted intentionally with the aim of extracting information);

* the applicant was raped and subjected to a number acts of other physical and psychological ill-treatment while in custody (*Aydın v. Turkey*, 1997, §§ 83-87, see also *aslova and Nalbandov v. Russia*, 2008, §108 where the applicant was repeatedly raped as well as subjected to a number of acts of physical violence during interrogation and *Zontul v. Greece*, 2012, § 92 where an illegal immigrant was raped by a coastal guard responsible for supervising him);

* the applicants were deprived of sleep, subjected to “Palestinian hanging” and “falaka”, sprayed with water, beaten for several days while in custody in order to extract a confession (*Batı and Others v. Turkey*, 2004, § 110 and §§ 122-124);

* the applicant, a detainee who was on hunger strike, was forced fed, despite the absence of medical necessity and with the use of handcuffs, a mouth-widener, a special rubber tube inserted into the food channel and, in the event of resistance, with the use of force (*Nevmerzhitsky v. Ukraine*, 2005, § 98);

* the applicant was subjected to combined and premeditated measures involving handcuffing, hooding, forcibly undressing, forcibly administering a suppository while held on the ground without any medical necessity, in the framework of “extraordinary rendering”, geared to obtaining information from the applicant or punishing or intimidating him (*El-Masri v. the former Yugoslav Republic of Macedonia* [GC], 2012, § 205);

Figure 17: **Source Chunk from the Case Law Guide.** A ~2,000-token excerpt from the case law guide on the prohibition of torture (Article 3), used as input for local Issue-Factor hierarchy extraction.

Chunk (Local) Issue-Factor Hierarchy (cont.)

* severe beatings by police officers resulting in the death of the applicants' relative (Satybalova and Others v. Russia, 2020, § 76; see also Lutsenko and Verbytsky v. Ukraine, 2021, §§ 79-80 where Mr Verbytsky was beaten to death by non-State agents hired by police in the context of the Maidan protests);

* the applicant was subjected to physical violence, including with PVC pipes as well as psychological abuse such as threat of rape and sexual violence, while held incommunicado at the basement of Chechen police headquarters on account of his homosexuality (Lapunov v. Russia, 2023, §§ 107-110).

- 14. The Court has held that, a particular type of conduct, such as rape of a detainee by an official of the State, must be considered to be an especially grave and abhorrent form of ill-treatment given the ease with which the offender can exploit the vulnerability and weakened resistance of the victim. Moreover, rape leaves deep psychological scars on the victims which do not respond to the passage of time as quickly as other forms of physical and mental violence. The victim also experiences the acute physical pain of forced penetration, which leaves her feeling debased and violated both physically and emotionally (Maslova and Nalbandov v. Russia, 2008, § 105).

- 15. The Court has not ruled out that a threat of torture can also amount to torture, as the nature of torture covers both physical pain and mental suffering. In particular, the fear of physical torture itself may in certain circumstances constitute mental torture. However, it has underlined that the classification of whether a given threat of physical torture amounted to psychological torture or to inhuman or degrading treatment depended upon all the circumstances of a given case, including, notably, the severity of the pressure exerted and the intensity of the mental suffering caused (Gäfgen v. Germany [GC], 2010, § 108).

- 16. In Tunikova and Others v. Russia, 2021, the applicants asked the Court to rule that ill-treatment perpetrated by non-State actors also constituted "torture". While the Court acknowledged that such an additional characterisation

Issue 1: Whether a particular form of ill-treatment constitutes "torture" under international law and Article 3 of the Convention (§9, §10, §11, §13, §14, §15, §16)

- └ Factor A: Distinction between torture and inhuman or degrading treatment, with torture involving deliberate inhuman treatment causing very serious and cruel suffering (§9)
- └ Factor B: Severity of the treatment (§9, §10, §11, §13, §14, §15, §16)
- └ Factor C: Presence of a purposive element-intentional infliction of severe pain or suffering for purposes such as obtaining information, confession, punishment, or intimidation (§10, §13)
- └ Factor D: Evolving interpretation in light of present-day conditions and higher standards of human rights protection (§11)
- └ Factor E: Status of the prohibition of torture as a jus cogens or peremptory norm in international law (§9, §12)
- └ Factor F: Examples of conduct found to amount to torture, including specific acts and circumstances (§13, §14)
- └ Factor G: Special gravity and psychological impact of rape by State officials (§14)
- └ Factor H: Threat of torture, including mental suffering and severity of pressure, may also amount to torture depending on circumstances (§15)
- └ Factor I: Whether ill-treatment by non-State actors can constitute torture, and the necessity of such characterisation in the specific case (§16)

Figure 18: **Extracted Local Issue-Factor Hierarchy.** Continuation of the source chunk (Table 17) and the corresponding local Issue-Factor hierarchy (bottom) extracted from it, illustrating how issues and their supporting factors are identified and linked back to source paragraphs (§).