Ontology-Driven Clustering for Extracting Legal Reasoning Graphs

Anonymous ACL submission

Abstract

Extracting structured reasoning from long court judgments is a persistent challenge in legal NLP: large language models can identify individual legal concepts but struggle to consistently link them across thousands of tokens, producing graphs that are incomplete, structurally invalid, or non-reproducible. We present LEXCGRAPH, a pipeline that decomposes long-document graph extraction into local inference problems by clustering extracted nodes around canonical legal concepts defined in a compiled ontology; within each cluster, an LLM infers edges under ontology-provided constraints (definitions, AND/OR requirements, defeaters, relation whitelists), while all cross-cluster structure is constructed deterministically. On 2,517 IL-TUR Indian Supreme Court cases, ontology-driven graphs improve zero-shot outcome prediction by +11.6pp on Grok-4.1 and +11.1pp on Sonnet 4.5 (McNemar $p < 0.05$) with substantially improved calibration (ECE: 0.110→0.027 on Grok), exceeding fine-tuned InLegalBERT (82.1% vs. 84.0% macro-F1) on the same evaluation set in zero-shot. A controlled Stage 4 ablation reveals a *separation of concerns*: ontology-driven node extraction drives prediction, while adding edges, justification sets, and reasoning chains to the prediction prompt does not improve it; relational structure instead drives retrieval (+0.100 nDCG@10 over BM25) and counterfactual analysis. Cross-jurisdictional evaluation on 297 ECHR Turkey-respondent cases under a separately compiled 55-concept ontology and a non-frontier extractor (GPT-4.1 mini) replicates the gains, with macro-F1 improving by +18.3pp on GPT-5.4 and the pattern further replicating on the open-weights Kimi K2.6 model (+13.7pp macro-F1, −0.032 Brier). We release the ontology, extraction code, and evaluation scripts.[1]

[1]Code, ontologies, and evaluation data provided in supplementary materials.

1 Introduction

Court judgments are long, multi-issue documents whose reasoning is distributed across factual findings, statutory interpretation, doctrine-specific tests, and citations to precedent. Capturing this reasoning in structured form is essential for legal case retrieval, explainability, and corpus-level legal analytics; yet the very properties that make court reasoning rich also make it difficult to extract: legally relevant spans are separated by thousands of tokens, doctrinal tests involve conjunctive and disjunctive requirements that must be tracked jointly, and the same abstract legal concept manifests differently across jurisdictions.

Large language models can extract individual nuggets of information from legal text, but they struggle to consistently link distant items across long inputs. When tasked with end-to-end graph extraction from a 15,000-token judgment, LLMs frequently produce graphs that are incomplete, structurally invalid, or non-reproducible. These failures are not primarily linguistic (the models understand the text) but combinatorial: the space of possible edges grows quadratically with the number of extracted nodes, and without structural constraints, the model must navigate this space with only implicit knowledge of legal reasoning patterns.

We propose LEXCGRAPH, an ontology-driven approach that addresses this problem by decomposing long-document graph extraction into a set of *local* inference problems. The key mechanism is **ontology-driven clustering**: extracted nodes are assigned to clusters defined by canonical legal concepts in a compiled ontology, and edge inference is restricted to within-cluster pairs using ontology-provided constraints. All cross-cluster structure is constructed deterministically. This design yields three properties: (1) *reproducibility*: the same input produces the same graph, con-

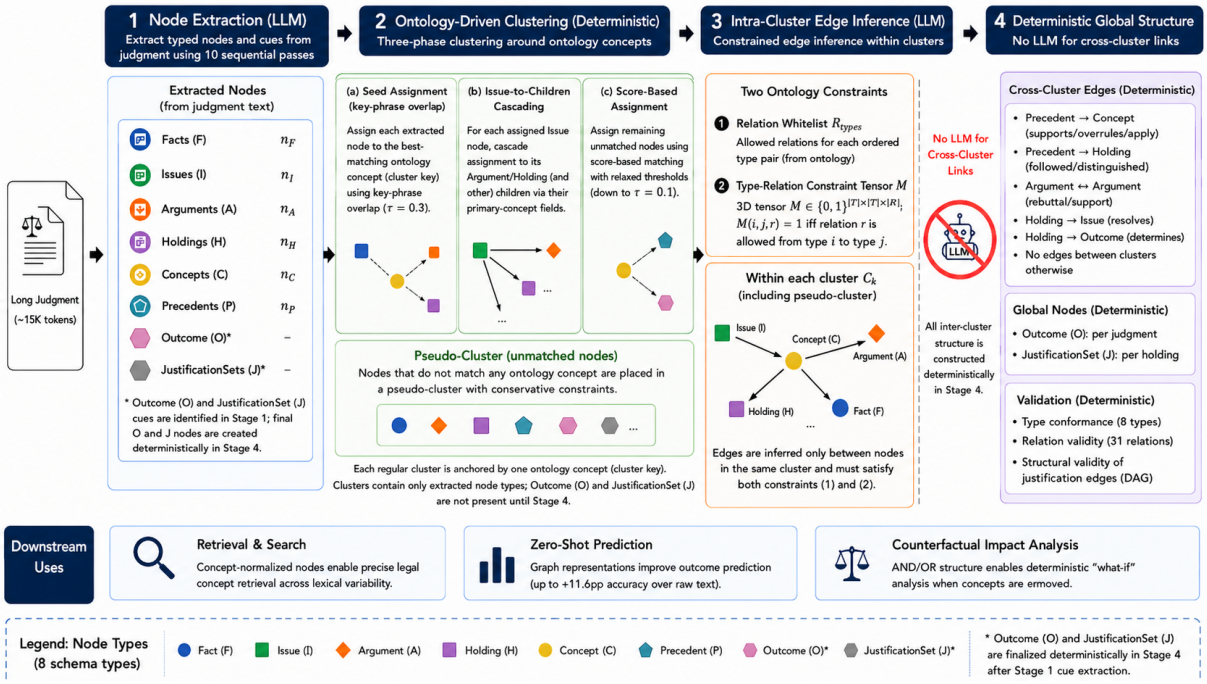

Figure 1: Overview of the LEXCGRAPH pipeline. A compiled ontology provides concept definitions, key phrases, edge patterns, AND/OR requirements, and defeaters. The four-stage pipeline decomposes long-document graph extraction into local inference problems, producing validated legal reasoning graphs for concept retrieval, zero-shot outcome prediction, and counterfactual impact analysis.

ditional on fixed LLM extraction outputs (deterministic for Stages 2 and 4); (2) *structural validity*: every edge satisfies type-level constraints; and (3) *counterfactual traceability*: the AND/OR justification structure enables deterministic impact analysis when a concept node is removed.

Our evaluation reveals four findings with implications beyond legal NLP. First, ontology-driven graphs substantially outperform raw text on zero-shot prediction (+11.6pp on IL-TUR Grok-4.1; +18.3pp macro-F1 on ECHR GPT-5.4) and concept retrieval (+0.100 nDCG@10 over BM25), and exceed fine-tuned InLegalBERT (84.0% vs. 82.1% macro-F1) on the same eval set. Second, naive single-pass extraction performs *worse* than raw text while ontology-driven multi-pass extraction exceeds it: *extraction methodology*, not relational structure, drives prediction gains. Third, a Stage 4 ablation on ECHR shows that adding ontology-constrained edges, justification sets, and reasoning chains to the prediction prompt does not improve over nodes alone (a matched-blinding control finds them statistically indistinguishable; App L), establishing a separation of concerns: nodes drive prediction; relational structure drives retrieval and counterfactual analysis. Fourth, a facts-only control plus a length-matched summarization baseline rule out information advantage and mere compression as alternative explanations.

2 Related Work

Legal NLP and judgment prediction. LexGLUE (Chalkidis et al., 2022), LegalBench (Guha et al., 2023), IL-TUR's CJPE task (Joshi et al., 2024) (from ILDC (Malik et al., 2021)), the ECHR baselines of Chalkidis et al. (2019), and COLIEE (Kim et al., 2023) standardize legal NLP evaluation. Structured prediction approaches model subtask dependencies as DAGs (Zhong et al., 2018), apply relational learning (Dong and Niu, 2021), or survey the area (Feng et al., 2022). Our work differs in extracting the reasoning structure itself as a first-class output rather than using implicit structure to improve a classifier.

Legal knowledge graphs and ontologies. LKIF Core (Hoekstra et al., 2007) provides foundational legal concepts; Küçük and Can (2025) survey computational-law datasets and ontologies. Jain et al. (2022) build a KG from Indian judgments with rule-based NER, and Dhani et al. (2021) combine legal KGs with GNNs for case similarity. We use the ontology not merely as a schema

but as a *constraint* on extraction, scoping which edges the LLM may propose and enforcing structural validity at each step.

Structure-aware legal retrieval. Li et al. (2023), Ma et al. (2023), and Tang et al. (2024) show structural information improves legal case retrieval; Donabauer and Kruschwitz (2025) flag reproducibility issues with graph-based retrieval, which our deterministic cross-cluster construction directly addresses.

Argumentation mining and graph reasoning. Toulmin (1958) provides the claim/grounds/warrant abstraction; Mochales Palau and Moens (2011) and Wyner et al. (2010) develop automatic argumentation extraction from legal text. Graph-of-Thoughts (Besta et al., 2024) models LLM reasoning as graph traversal without domain constraints. Unlike edge-perturbation counterfactual explainers (Tan et al., 2022), our AND/OR justification structure enables domain-grounded counterfactual analysis (§J).

3 Method

LexCGraph extracts a legal reasoning graph from a court decision in four stages: (1) multi-pass node extraction, (2) ontology-driven clustering, (3) constrained local edge inference, and (4) deterministic global structure construction.

3.1 Graph Schema

Each decision is represented as a typed, anchored graph $G = (V, E, A)$, where V is a set of typed nodes, E is a set of typed directed edges, and A maps each node and edge to source-text anchors (character offsets and text hashes) and metadata. The schema defines eight node types (Fact, Concept, Issue, Argument, Holding, Precedent, Outcome, JustificationSet) and 31 edge relations derived from ontology edge patterns and legal argumentation literature (Toulmin, 1958; Mochales Palau and Moens, 2011). A type-level validation matrix $M \in \{0,1\}^{|T|\times|T|\times|R|}$ specifies permitted relations between each pair of node types. Edges are normalized through an alias mapping layer ($\sim$200 aliases $\rightarrow$ 31 canonical relations) and deterministically deduplicated.

3.2 Compiled Ontology

The compiled ontology serves dual roles: defining cluster keys and providing extraction constraints. Each concept includes: (1) label and definition, (2) key phrases for node-to-cluster assignment, (3) typical edge patterns constraining intra-cluster inference, (4) AND/OR requirements encoding conjunctive/disjunctive test structures, and (5) defeaters specifying blocking conditions.

The Indian ontology contains 54 schema-complete concepts: 11 constitutional doctrines (e.g., Basic Structure, Maneka Gandhi proportionality framework, Wednesbury Unreasonableness) with complete AND/OR reasoning structures, and 43 offense, procedural, and evidence concepts (e.g., Murder, Regular Bail, Electronic Evidence, Dying Declaration) with legal elements, defenses, establishing cases, and key phrases. All concepts were audited against primary legal sources, with landmark case citations validated by independent audit; detailed audit outcomes are reported in the Limitations section. Appendix A provides representative examples spanning each concept category and jurisdiction.

Concrete example: proportionality. The Maneka Gandhi proportionality framework is encoded as a six-part AND test: procedure by law must exist; it must be "right, just and fair"; not "arbitrary, fanciful or oppressive"; must satisfy Article 14 (non-arbitrariness); Article 19 (reasonableness); and conform to natural justice. Defeaters include express statutory exclusion and emergency situations. Key phrases include "just, fair and reasonable" and "Golden Triangle." This structure scopes LLM edge inference: within a proportionality cluster, the LLM may only propose edges from the concept's relation whitelist, and each proposed edge is validated against the type-level matrix M. Figure 2 in Appendix A provides a visual representation of this ontology entry, illustrating how the five components (AND/OR requirements, defeaters, edge patterns, key phrases, and establishing cases) compose a single concept definition.

3.3 Extraction Pipeline

Stage 1: Multi-pass extraction. Ten sequential passes with Grok-4.1 (xAI, 2025), targeting facts, concepts, issues, arguments, holdings, precedents, outcome, intra-cluster edges, long-distance link discovery, and justification sets, respectively. Each pass is validated before commitment: the type-aware matrix rejects invalid node types, anchors are repaired via text hashes, and du-

plicates are merged.

Stage 2: Ontology-driven clustering. Nodes are assigned to concept clusters via three phases: (1) seed assignment from key-phrase overlap ($\tau = 0.3$), (2) cascading of Issue nodes to children via primary-concept fields, and (3) score-based assignment for remaining nodes. Unmatched nodes enter pseudo-clusters with conservative constraints. Each cluster is summarized into a structured context record containing concept definition, AND/OR requirements, defeaters, and candidate node mentions.

Stage 3: Constrained local edge inference. Within each cluster, an LLM proposes edges under two constraints: a relation whitelist from the ontology's edge patterns, and type-level validation via matrix M. Proposed edges are normalized and deduplicated.

Stage 4: Deterministic global structure. All cross-cluster structure is constructed without LLM involvement: `resolves` edges from holdings to issues, `determines` edges from holdings to outcomes via keyword scoring, justification sets compiling intra-cluster evidence under AND/OR logic, and reasoning chains via breadth-first traversal. Each graph receives a quality tier (gold/silver/bronze/reject).

4 Experimental Setup

4.1 Dataset and Task

We evaluate on 2,517 IL-TUR CJPE cases (Joshi et al., 2024) (838 accepted, 1,679 rejected; stratified, seed=42). For zero-shot outcome prediction, we compare two input representations, both identically blinded: (i) *Raw*: decision text with outcome-revealing content removed via regex filtering; and (ii) *Graph*: a structured linearization of the ontology-normalized concept nodes (facts, concepts, issues, party arguments, precedents) with all outcome-bearing fields (holdings, court responses, issue answers, outcome) withheld. *The Graph prompt is node-only*: Stage 3 intra-cluster edges and Stage 4 outputs (cross-cluster edges, justification sets, reasoning chains) are not serialized in the prediction prompt; they serve retrieval (Table 3) and counterfactual analysis (§J). A dedicated ablation isolating their contribution to prediction appears in §5.5. For each case the LLM receives either the Raw or Graph prompt and outputs a binary prediction (accepted/rejected for IL-TUR; violation/no-violation for ECHR) together with a confidence score that drives the calibration metrics below.

ECHR cross-jurisdiction set. For cross-jurisdictional evaluation we use 297 full-judgment ECHR merits decisions scraped from HUDOC (Turkey-respondent cases) (Chalkidis et al., 2019): 262 violation, 35 no-violation (88% imbalance). Cases were retrieved from HUDOC filtered by respondent state (`TUR`), language (English), and document collection (Chamber, Grand Chamber, judgments and decisions); friendly-settlement and struck-out dispositions were excluded, and importance was restricted to levels 1–3. Turkey was selected for its high case volume and doctrinal breadth across Convention articles. A separately compiled 55-concept ECHR ontology covers Convention articles, doctrinal tests (margin of appreciation, proportionality, prescribed-by-law), and procedural principles (exhaustion of domestic remedies, victim status, four-month rule). Extraction is performed by GPT-4.1 mini (OpenAI, 2025) rather than a frontier model, deliberately testing pipeline robustness under weaker extraction. The blinding protocol is identical to IL-TUR (outcome-bearing graph fields withheld; regex-based outcome removal from raw text, plus the HUDOC metadata banner stripped), and the raw-text character budget is set to $\approx 1.05\times$ the average graph summary length to match information access at the prompt level.

4.2 Models and Metrics

For IL-TUR we evaluate on two architecturally distinct frontier models: Grok-4.1 and Claude Sonnet 4.5.[2] Since Grok also serves as the extraction model, Sonnet provides an independent cross-model control. For the ECHR set we use GPT-4.1 mini (OpenAI, 2025) (which is both extractor and a non-frontier predictor) and GPT-5.4 (OpenAI, 2026) as a frontier predictor on the same graphs. To test whether the pipeline's benefits also extend to open-weights predictors, we additionally run Moonshot's Kimi K2.6 (Moonshot AI, 2026) in two configurations on the same 297 cases: a

[2] API slugs: `grok-4-1-fast-reasoning` (xAI), `claude-sonnet-4-5-20250929` (Anthropic), `gpt-4.1-mini` and `gpt-5.4` (OpenAI), `kimi-k2.6` (Moonshot). All Grok-4.1 extraction runs completed before May 15, 2026 (pre-retirement). Prediction runs for all models completed by May 25, 2026.

Table 1: Zero-shot CJPE outcome prediction on 2,517 IL-TUR cases. Graph inputs outperform raw text on both models across all metrics.

Model	Input	Acc.	F1	ECE↓	Brier↓
Grok-4.1	Raw	.739	.718	.110	.196
	Graph	**.855**	**.840**	**.027**	**.119**
Sonnet 4.5	Raw	.663	.631	.091	.228
	Graph	**.774**	**.754**	**.014**	**.174**
InLegalBERT-Flat†	Full text	.847	.821	.053	.121

†Fine-tuned end-to-end on `single_train` excluding the 2,517 eval cases (n=2,565). Published InLegalBERT-Hier on the canonical IL-TUR test split reports 81.31% macro-F1 (Joshi et al., 2024); the frozen-encoder hierarchical variant achieves 68.5% on our sample, confirming end-to-end fine-tuning is needed for the hierarchical architecture.

Table 2: Extraction methodology ablation: ontology-driven multi-pass vs. naive single-pass vs. raw text. Both graph conditions present concept nodes to the predictor; the contrast isolates extraction method, not edges (Stage 4 edge contrast: Table 4, + Stage 4 column). Raw uses the full 2,517-case sample (Table 1); Ontology-Driven and Naive use the 2,508-case structured-ablation subset (8 cases failed naive single-pass extraction and were excluded uniformly across all three conditions). †McNemar $p < 0.05$.

Model	Metric	Ontology-Driven	Naive Extr.	Raw
Grok-4.1	Acc.	**0.864**†	0.718	0.739
	Brier↓	**0.113**	0.211	0.196
Sonnet 4.5	Acc.	**0.771**†	0.683	0.663
	Brier↓	**0.175**	0.219	0.228

non-reasoning variant (*thinking off*) comparable to GPT-4.1 mini and a reasoning-enabled variant (*thinking on*) comparable to GPT-5.4. Using a non-frontier extractor stress-tests whether the ontology constraints can compensate for weaker upstream extraction. We measure accuracy, macro-F1, calibration (ECE (Guo et al., 2017), Brier),[3] and use McNemar's test with bootstrap 95% CIs. For retrieval, we report nDCG@10, MAP, and P@10 with paired t-tests. Our IL-TUR evaluation across 10,068 model runs produced zero parse or API errors.

Throughout the tables, "Graph" (Table 1), "Ontology-Driven" (Table 2), and "Nodes Only" (Table 4) refer to the same ontology-extracted concept-node linearization; in Table 4, "Nodes Only" is used specifically to contrast with "+ Stage 4 outputs" (cross-cluster edges, justification sets, and reasoning chains).

5 Results

5.1 Outcome Prediction

Table 1 shows that graphs consistently outperform raw text. Accuracy improves by +11.6pp on Grok and +11.1pp on Sonnet (McNemar χ^2=119.97 and 92.34; $p < 0.05$), with bootstrap 95% CIs excluding zero. Calibration improves markedly: ECE drops from 0.110→0.027 on Grok, with raw text systematically overconfident in the 0.70–0.90 range while graph inputs track the reliability diagonal closely.

Supervised baseline. A fine-tuned InLegalBERT (Paul et al., 2023) trained on IL-TUR `single_train` excluding the 2,517 eval cases (n=2,565) achieves 82.1% macro-F1 on the eval set (competitive with the published 81.31% InLegalBERT-Hier benchmark) but is surpassed by Grok-4.1 with graph input at 84.0% macro-F1 zero-shot. The supervised model also shows a substantial generalization gap (56.7% macro-F1 on `single_dev`; class distribution shifts from train's 38/62 to dev's 50/50), and worse calibration (ECE .053 vs. .027), characteristic of fine-tuning on limited data.

5.2 Extraction Methodology Ablation

Table 2 reveals that *extraction methodology* (not the presence of relational structure in the prompt) drives the gain. Both graph columns present extracted concept nodes to the predictor; they differ in *how* those nodes are extracted. Ontology-driven multi-pass extraction (type-validated, AND/OR-constrained, ontology-clustered) achieves 0.864/0.771 accuracy on Grok/Sonnet, while naive single-pass extraction achieves 0.718/0.683, *below* raw text's 0.739/0.663. Naive extraction destroys implicit textual signals (discourse cues, paragraph structure, narrative flow) without compensating organization; ontology-driven extraction replaces those signals with cleaner, type-validated, doctrinally-grounded concept nodes that the predictor can rea-

[3] ECE uses 15 equal-width bins on $[0, 1]$ (first bin closed-closed, others left-open right-closed), binning by maximum-class probability. For supervised InLegalBERT predictions this is the softmax over the 2-class logits; for LLM predictions it is the self-reported confidence value the model returns in the structured JSON output (a single `"confidence": 0.0-1.0` field accompanying each prediction). Implementation in `recompute_ece.py` mirrors `supervised_baseline_compare.py`'s `expected_calibration_error`.

Table 3: Concept retrieval (48 queries, 2,517 cases). Dense baselines: Legal-BERT (Chalkidis et al., 2020), InLegalBERT (Paul et al., 2023), E5-large (Wang et al., 2022), BGE-large (Xiao et al., 2023); late-interaction: ColBERTv2 (Santhanam et al., 2022); sparse: BM25 (Robertson and Zaragoza, 2009). $\dagger p < 0.05$ vs. BM25.

Method	Type	nDCG@10	MAP	P@10
Legal-BERT	Bi-enc.	0.189	0.082	0.175
InLegalBERT	Bi-enc.	0.235	0.098	0.215
E5-large	Bi-enc.	0.346	0.160	0.323
BGE-large	Bi-enc.	0.350	0.158	0.333
ColBERTv2	Late-int.	0.425	0.191	0.390
BM25	Sparse	0.732	0.461	0.702
ConceptSet†	Structured	**0.832**	**0.510**	**0.831**

son over. The consistent performance ordering using identical models and cases makes LLM memorization unlikely as the sole explanation: if models recognized cases from pretraining, performance would be independent of input representation. A direct ablation isolating the contribution of relational structure *on top of* ontology-driven node extraction appears in §5.5.

Further ablations. Two additional analyses, detailed in Appendices H and I, support the main findings: an *ontology completeness ablation* shows that prediction accuracy saturates at 11 core doctrinal concepts (.855 vs. .853 with 54 concepts, $p > 0.05$) while structural quality improves dramatically (gold tier 77%→90%; pseudo-cluster nodes drop 93%), and a *facts-only control* plus a 30-case leakage audit (0.225 vs. 0.233 outcome-signal density per 1K chars) attribute the prediction gap to structural organization rather than information advantage: facts-only graphs at 62% match raw facts at 64% and fall far below full graph at 90%.

5.3 Concept Retrieval

ConceptSet is not a separate retrieval model; it is a structured retrieval index constructed from the ontology-normalized concept annotations extracted by the LexCGraph pipeline, scored against queries using exact and fuzzy token Jaccard similarity with graded relevance weighting. Table 3 shows that ConceptSet substantially outperforms all baselines including BM25 (+0.100 nDCG@10, +0.129 P@10; paired t-test, p=0.016). A clear four-tier ranking emerges: domain-adapted legal models perform worst, general-purpose bi-encoders improve, ColBERTv2 partially recovers lexical signal, and BM25 dominates all dense baselines, reflecting concept-level legal retrieval's dependence on exact term matching. ConceptSet additionally identifies semantically relevant cases lacking explicit surface markers (Jaccard overlap between regex-derived and annotation-derived relevance: 0.43). Relevance labels in this evaluation are silver labels derived by regex-based concept tagging over the case corpus, not gold human judgments; their limitations are discussed in the Limitations section.

5.4 Cross-Jurisdictional Evaluation

Table 4 shows that the Nodes-Only graph beats raw text on accuracy, macro-F1, Brier, and ECE across all three predictors. Macro-F1 is the appropriate primary metric given the 88% class imbalance: graphs improve macro-F1 by +22.2pp on GPT-4.1 mini, +18.3pp on GPT-5.4, and +13.7pp on Kimi K2.6 think-off (+9.5pp on think-on), driven by minority-class no-violation recall (62.9% vs. 22.9% on GPT-5.4; +40pp). The accuracy gap is compressed by the 88% majority baseline; GPT-5.4's +3.4pp aggregate accuracy gain has bootstrap 95% CI $[+.003, +.064]$ and McNemar χ^2=3.68 (marginal). **Open-weights replication.** Kimi K2.6 (Moonshot AI, 2026) reproduces Graph > Raw in both configurations, ruling out closed-model artifacts; the think-on variant matches GPT-5.4 on Brier (.052 vs. .055), and the calibration advantage widens with reasoning capability. Extended capability-scaling, extraction-robustness, and raw-blinding-robustness analyses appear in Appendix E.

5.5 Stage 4 Ablation: Relational Structure vs. Nodes Alone

The Graph linearization in Tables 1, 4 presents only ontology-normalized concept nodes; Stage 3 edges and Stage 4 outputs (cross-cluster edges, justification sets, reasoning chains) are not serialized. To isolate relational structure's contribution to prediction, we augmented the ECHR linearization with Stage 4 outputs (Table 4, + Stage 4 column). On macro-F1 the unmatched ordering is Nodes Only > +Stage 4 > Raw at all three scales, but this + Stage 4 column is *not* blinding-matched: it ran `no_scrub`=False vs. Nodes Only's `no_scrub`=True, trading ≈26% of its fact/concept text for equal-length symbolic scaffolding (App L). Under a matched-blinding control (GPT-5.4), + Stage 4 is statistically indistinguishable from Nodes Only (.826 vs. .845

Table 4: ECHR multi-condition comparison: 297 Turkey-respondent HUDOC cases (88% violation; extraction by GPT-4.1 mini; 296 valid for Kimi K2.6 after one parse error). Each cell reports **Acc/Macro-F1/Brier**↓; **bold** marks the highest macro-F1 per row. Conditions: *Raw* = outcome-blinded judgment text (4,000-char window); *Summary* = length-matched GPT-4.1 mini free-form prose summary of the full blinded judgment (30–100k chars source); *Nodes Only* = ontology-extracted concept-node linearization; + *Stage 4* = Nodes Only augmented with cross-cluster edges, justification sets, and reasoning chains (`no_scrub`=False, *not* matched to Nodes Only's True; control in App L). ECE (n_bins=15) per cell. GPT-4.1 mini: .072/.056/.069/.079; GPT-5.4: .045/.054/.027/.057; Kimi K2.6 (think off): .116/.101/.102/.121. McNemar tests, paired-win counts, bootstrap CIs, and the Kimi K2.6 (think on) variant (Brier .052 matching GPT-5.4's .055) are discussed in §5.4–§5.6.

Predictor	Raw	Summary	Nodes Only	+ Stage 4
GPT-4.1 mini	.886/.522/.103	.892/.623/.096	**.912/.744/.081**	.875/.589/.104
GPT-5.4	.909/.662/.077	.909/.710/.067	**.943/.845/.055**	.909/.738/.079
Kimi K2.6	.885/.642/.108	.916/.757/.080	**.902/.779/.076**	.892/.720/.098

macro-F1; paired McNemar p=.72). The implication holds: *ontology-driven node extraction is sufficient for prediction*; relational structure is redundant, not harmful. The Stage 3–4 structure remains valuable in the artifact: it powers the +0.100 nDCG@10 retrieval gain (Table 3) and the 15,276-justification-set counterfactual analysis (§J), but not for the prediction prompt itself. The graph's value is *distributed across tasks* (§6).

5.6 Summarization Baseline: Structure vs. Compression

A reviewer could object that the Graph > Raw gap reflects *compression* rather than ontology-driven structure. To isolate, we constructed a length-matched GPT-4.1 mini prose summary of the entire blinded judgment (30–100k chars source; target ≈5,600 chars to match the Stage 4 linearization; identical Raw-condition prompt; no labels/bullets/dispositions). Table 4 (Summary vs. Nodes Only columns) shows a consistent Nodes-Only > Summary > Raw ordering on macro-F1 across all three predictors: Summary beats Raw by +4.8–11.5pp, but Nodes-Only further beats Summary by +2.2–13.5pp. Compression captures part of the gain; ontology-grounded concept normalization captures the rest. Critically, this comparison is biased *in favor* of Summary (it saw ≈6–25× more blinded source text than Raw's 4,000-char window, with target length matching the longer Stage 4 linearization rather than the shorter Nodes-Only one), yet Nodes-Only still wins. A free-form summary is also a *single-task artifact*: it provides no concept index for retrieval, no AND/OR justification structure for counterfactual analysis, and no typed edges for structural queries (§G).

Graph quality validation. An independent LLM-as-a-judge evaluation (Claude Opus 4.7, n=100 graphs across 5 dimensions; Appendix D) confirms Reasoning Coverage (the dimension the pipeline was designed to optimize) as the strongest dimension (3.50/5; IL-TUR 3.75, ECHR 2.76), and a 45-case human calibration study finds significant rank correlation on that dimension (ρ=0.33, p=.026) with 91.6% within-±1 agreement pooled across dimensions. Approximate per-case graph quality (Overall Faithfulness 3.03/5) coexists with strong downstream gains, indicating that structured organization of judicial reasoning contributes value well below the threshold of flawless extraction.

Counterfactual capability. The AND/OR justification structure also enables deterministic counterfactual impact analysis: 77.5% of holdings across 2,446 graphs (15,276 justification sets) are supported exclusively by conjunctive justification, and the mean fragility score is 0.344. Per-concept load-bearing counts, doctrinal asymmetries, and corpus statistics appear in Appendix J.

6 Discussion

The graph as reasoning structure, not information advantage. The graph's value derives specifically from structuring judicial reasoning. Three findings converge: (1) facts-only graphs provide no advantage (62% vs. 64% raw facts), confirming the signal comes from the court's analytical sections; (2) the leakage audit shows comparable information access (1.0× density ratio); and (3) naive single-pass extraction nodes perform *worse* than raw text while ontology-driven multi-pass extraction substantially exceeds it (§5.2), ruling out information advantage and isolating ex-

traction quality as the driver.

The graph's value is distributed across tasks. The Stage 4 ablation (§5.5) isolates *which* part of the graph drives *which* task: ontology-driven *node extraction* drives prediction, while *relational structure* (edges, justification sets, reasoning chains) drives retrieval and counterfactual analysis. Under matched blinding, adding relational structure to the prediction prompt does not improve over nodes alone (App L). Practical implication: prediction applications use node-only linearization; retrieval and explainability use the full Stage 3–4 structure: the same graph, linearized differently per task.

Extended discussion appears in Appendix G.

7 Conclusion

We introduced LexCGraph, an ontology-driven pipeline that decomposes long-document graph extraction into local inference problems via concept clusters and deterministic cross-cluster structure. On IL-TUR CJPE, graphs improve zero-shot prediction by +11.6pp on Grok-4.1 (ECE 0.110→0.027), exceeding fine-tuned InLegalBERT (84.0 vs. 82.1 macro-F1). Cross-jurisdictional ECHR evaluation replicates the gains (+18.3pp macro-F1 on GPT-5.4) under a separately compiled ontology and a non-frontier extractor, and a Stage 4 ablation establishes a *separation of concerns*: node extraction drives prediction, while relational structure drives retrieval (+0.100 nDCG@10) and counterfactual analysis (77.5% holdings fragile, mean fragility 0.344). Robustness checks (open-weights Kimi K2.6 replication and a length-matched summarization baseline) rule out closed-model artifacts and compression as alternative explanations. The approach may generalize to other domains with long documents and structured diagnostic ontologies.

Limitations

Ontology coverage. The 54-concept ontology reduced pseudo-cluster assignments by 93%, but novel or emerging doctrines may still fall back to pseudo-clusters and be under-linked.

Information access. Evaluation uses full judgment text with outcome blinding, not facts alone (the pipeline's value derives from structuring judicial reasoning, requiring analytical sections), limiting applicability to post-decision analysis.

Concept matching. Node-to-cluster assignment uses keyword overlap rather than learned embeddings, prioritizing reproducibility over semantic precision; regex-based relevance labels for retrieval may under-count semantically relevant cases.

Proprietary extraction. IL-TUR uses Grok-4.1 for extraction; ECHR uses GPT-4.1 mini. The pipeline ports across predictors (closed and open-weights, §5.4) but extraction itself still uses closed models; fully open-weights extraction with Llama/Qwen-class models remains future work.

Ontology verification. Of 43 LLM-assisted offense/procedural concepts, 22 passed cleanly, 6 had minor imprecisions, 2 had errors (corrected), and 13 contained case citations not independently verified. Practicing-expert verification would strengthen confidence.

Graph quality and human evaluation. The LLM-as-a-judge reports moderate per-dimension scores (Appendix D); single-annotator calibration (n=45) shows significant Reasoning Coverage rank correlation (ρ=0.33, p=.026). Multi-annotator human evaluation remains future work.

ECHR coverage. Restricted to Turkey-respondent HUDOC cases with severe class imbalance (88% violation); broader respondent-state coverage and balanced pools would strengthen the generalizability claim.

Ethics Statement

All data used in this work are publicly available court decisions from the IL-TUR corpus (Joshi et al., 2024) and ECHR cases (Chalkidis et al., 2019). The dataset contains no non-public personal data; published judgments may contain personal or sensitive facts as matters of public record, and we use them only for retrospective aggregate analysis.

Our system performs *post-decision* structural analysis of judgments that have already been decided and published. We do not deploy outcome predictions on pending or real-world cases; the counterfactual analysis is restricted to retrospective examination of decided judgments. We therefore see no direct risk of automated judicial decision-making arising from this work.

Limitations of the ontology verification process are discussed in the Limitations section.

References

Anthropic. 2026. Claude Opus 4.7 [Large language model, API model ID: `claude-opus-4-7`]. Accessed via the Anthropic Messages API, May 2026. Model documentation: https://docs.anthropic.com/en/docs/about-claude/models.

Maciej Besta, Nils Blach, Ales Kubicek, et al. 2024. Graph of Thoughts: Solving Elaborate Problems with Large Language Models. In *Proc. AAAI*, Vol. 38, 17682–17690.

Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. 2019. Neural Legal Judgment Prediction in English. In *Proc. ACL*, 4317–4323.

Ilias Chalkidis, Manos Fergadiotis, et al. 2020. LEGAL-BERT: The Muppets Straight out of Law School. In *Findings of EMNLP*, 2898–2904.

Ilias Chalkidis, Abhik Jana, et al. 2022. LexGLUE: A Benchmark Dataset for Legal Language Understanding in English. In *Proc. ACL*, 4310–4330.

Jaspreet Singh Dhani, Ruchika Bhatt, Balaji Ganesan, Parikshet Sirohi, and Vasudha Bhatnagar. 2021. Similar Cases Recommendation using Legal Knowledge Graphs. *arXiv:2107.04771*.

Gregor Donabauer and Udo Kruschwitz. 2025. A Reproducibility Study of Graph-Based Legal Case Retrieval. In *Proc. SIGIR*, 3135–3144.

Qian Dong and Shuzi Niu. 2021. Legal Judgment Prediction via Relational Learning. In *Proc. SIGIR*, 983–992.

Yi Feng, Chuanyi Li, and Vincent Ng. 2022. Legal Judgment Prediction: A Survey of the State of the Art. In *Proc. IJCAI*, 5461–5469.

Neel Guha, Julian Nyarko, et al. 2023. LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models. *arXiv:2308.11462*.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On Calibration of Modern Neural Networks. In *Proc. ICML*, 1321–1330.

Rinke Hoekstra, Joost Breuker, et al. 2007. The LKIF Core Ontology of Basic Legal Concepts. In *Proc. LOAIT*, CEUR Vol. 321, 43–63.

Sarika Jain, Pooja Harde, Nandana Mihindukulasooriya, Sudipto Ghosh, Ankush Bisht, and Abhinav Dubey. 2022. Constructing a Knowledge Graph from Indian Legal Domain Corpus. In *Text2KG Workshop at ESWC*, CEUR Vol. 3184, 80–93.

Abhinav Joshi, Shounak Paul, et al. 2024. IL-TUR: Benchmark for Indian Legal Text Understanding and Reasoning. In *Proc. ACL*, 11460–11499.

Mi-Young Kim, Juliano Rabelo, Randy Goebel, Masaharu Yoshioka, Yoshinobu Kano, and Ken Satoh. 2023. COLIEE 2022 Summary: Methods for Legal Document Retrieval and Entailment. In *JSAI-isAI 2022 Workshop*, LNCS Vol. 13859. Springer, 51–67.

Dilek Küçük and Fazli Can. 2025. Computational Law: Datasets, Benchmarks, and Ontologies. *arXiv:2503.04305*.

Haitao Li, Qingyao Ai, et al. 2023. SAILER: Structure-aware Pre-trained Language Model for Legal Case Retrieval. In *Proc. SIGIR*, 1035–1044.

Yixiao Ma, Yueyue Wu, et al. 2023. Incorporating Structural Information into Legal Case Retrieval. *ACM TOIS*, 42(2), Art. 40.

Vijit Malik, Rishabh Sanjay, et al. 2021. ILDC for CJPE: Indian Legal Documents Corpus for Court Judgment Prediction and Explanation. In *Proc. ACL-IJCNLP*, 4046–4062.

Raquel Mochales Palau and Marie-Francine Moens. 2011. Argumentation Mining. *AI and Law*, 19(1):1–22.

Moonshot AI. 2026. Kimi K2.6 [Open-weights large language model, API model ID: `kimi-k2.6`]. Accessed via the Moonshot Chat Completions API (https://api.moonshot.ai/v1), May 2026.

OpenAI. 2025. GPT-4.1 mini [Large language model, API model ID: `gpt-4.1-mini`]. Accessed via the OpenAI Chat Completions API, May 2026. Model documentation: https://platform.openai.com/docs/models.

OpenAI. 2026. GPT-5.4 [Large language model, API model ID: `gpt-5.4`]. Accessed via the OpenAI Chat Completions API, May 2026. Model documentation: https://platform.openai.com/docs/models.

Shounak Paul, Arpan Mandal, et al. 2023. Pre-trained Language Models for the Legal Domain: A Case Study on Indian Law. In *Proc. ICAIL*, 187–196.

Stephen Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *FnTIR*, 3(4):333–389.

Keshav Santhanam, Omar Khattab, et al. 2022. ColBERTv2: Effective and Efficient Retrieval via Lightweight Late Interaction. In *Proc. NAACL*, 3715–3734.

Juntao Tan, Shijie Geng, et al. 2022. Learning and Evaluating Graph Neural Network Explanations based on Counterfactual and Factual Reasoning. In *Proc. The Web Conference*, 1018–1027.

Yanran Tang, Ruihong Qiu, et al. 2024. CaseLink: Inductive Graph Learning for Legal Case Retrieval. In *Proc. SIGIR*, 2199–2209.

Stephen E. Toulmin. 1958. *The Uses of Argument*. Cambridge University Press.

Liang Wang, Nan Yang, et al. 2022. Text Embeddings by Weakly-Supervised Contrastive Pre-training. *arXiv:2212.03533*.

Adam Wyner, Raquel Mochales-Palau, et al. 2010. Approaches to Text Mining Arguments from Legal Cases. In *Semantic Processing of Legal Texts*. Springer, 60–79.

xAI. 2025. Grok-4.1 [Large language model, API model ID: `grok-4-1-fast-reasoning`]. Accessed via the xAI API (`https://api.x.ai/v1`), February 2026. Note: model retired May 15, 2026; all experiments predate retirement.

Shitao Xiao, Zheng Liu, et al. 2023. C-Pack: Packaged Resources to Advance General Chinese Embedding. *arXiv:2309.07597*.

Haoxi Zhong, Zhipeng Guo, et al. 2018. Legal Judgment Prediction via Topological Learning. In *Proc. EMNLP*, 3540–3549.

A Ontology Examples

We present representative examples from each concept category and jurisdiction to illustrate the ontology's structure. Each concept encodes the legal test as AND/OR requirements, blocking conditions as defeaters, and surface-level triggers as key phrases. The full ontologies are released with our code.

A.1 Indian Concepts

Maneka Gandhi Framework (Doctrine; Golden Triangle of Articles 14, 19, 21). *Definition:* State action affecting Article 21 must satisfy a six-part substantive test interlinking the Article 14, 19, and 21 guarantees, beyond the bare procedure-established-by-law requirement.

AND Procedure established by law exists
AND Procedure must be "right, just and fair"
AND Procedure must not be "arbitrary, fanciful or oppressive"
AND Must satisfy Article 14 (non-arbitrariness)
AND Must satisfy Article 19 (reasonableness if applicable)
AND Must conform to principles of natural justice

Defeaters: express statutory exclusion of natural justice; emergency situations; clear legislative intent excluding procedural requirements.
Key phrases: "just, fair and reasonable"; "right and just and fair"; "not arbitrary, fanciful or oppressive"; "Golden Triangle".
Establishing cases: Maneka Gandhi v. UOI (1978); R.C. Cooper v. UOI (1970).
Edge pattern: (State_Action) $\xrightarrow{\text{must satisfy}}$ (Maneka_Gandhi_Test) $\xrightarrow{\text{if failed}}$ (Unconstitutional).

Murder (Offense; BNS §103 / IPC §302). *Definition:* Causing the death of a human being with the mens rea specified in IPC 300 / BNS 101 and not falling within any of the five exceptions.

AND Actus reus: act causing death of a human being
AND Mens rea (any one of four IPC 300 / BNS 101 alternatives): intention to cause death; *or* intention to cause bodily injury known to be likely to cause death; *or* intention to cause injury sufficient in the ordinary course of nature to cause death; *or* knowledge that the act is so imminently dangerous that it must in all probability cause death
AND Death actually results from the act
AND Act does not fall within any of the five exceptions to IPC 300 / BNS 101

Defeaters: private defence (IPC 96–106 / BNS 34–44); insanity (IPC 84 / BNS 22); accident without criminal intent (IPC 80 / BNS 20); act of a child under 7 (IPC 82 / BNS 17); sudden and grave provocation (Exception 1); sudden fight without premeditation (Exception 4); consent of victim above 18 (Exception 5).
Key phrases: "committed murder"; "Section 302"; "BNS 103"; "intention to kill"; "culpable homicide amounting to murder"; "imminently dangerous act".
Establishing cases: Virsa Singh v. State of Punjab (1958); Rajwant Singh v. State of Kerala (1966); State of AP v. R. Punnayya (1977).
Edge pattern: (Accused_Act) $\xrightarrow{\text{causes}}$ (Death_of_Victim) $\xrightarrow{\text{with}}$ (Mens_Rea) $\xrightarrow{\text{not within exceptions}}$ (Conviction_BNS103).

Anticipatory Bail (Procedural; BNSS §482 / CrPC §438). *Definition:* Pre-arrest direction for release on bail granted by the High Court or Court of Session to a person apprehending arrest on accusation of a non-bailable offense.

AND Applicant has reason to believe arrest may be made on a non-bailable accusation
AND Application filed before the High Court or Court of Session (CrPC 438 / BNSS 482)
AND Court considers: nature/gravity of accusation; applicant's antecedents; flight risk; whether the accusation is to humiliate or injure
AND Court may impose conditions: availability

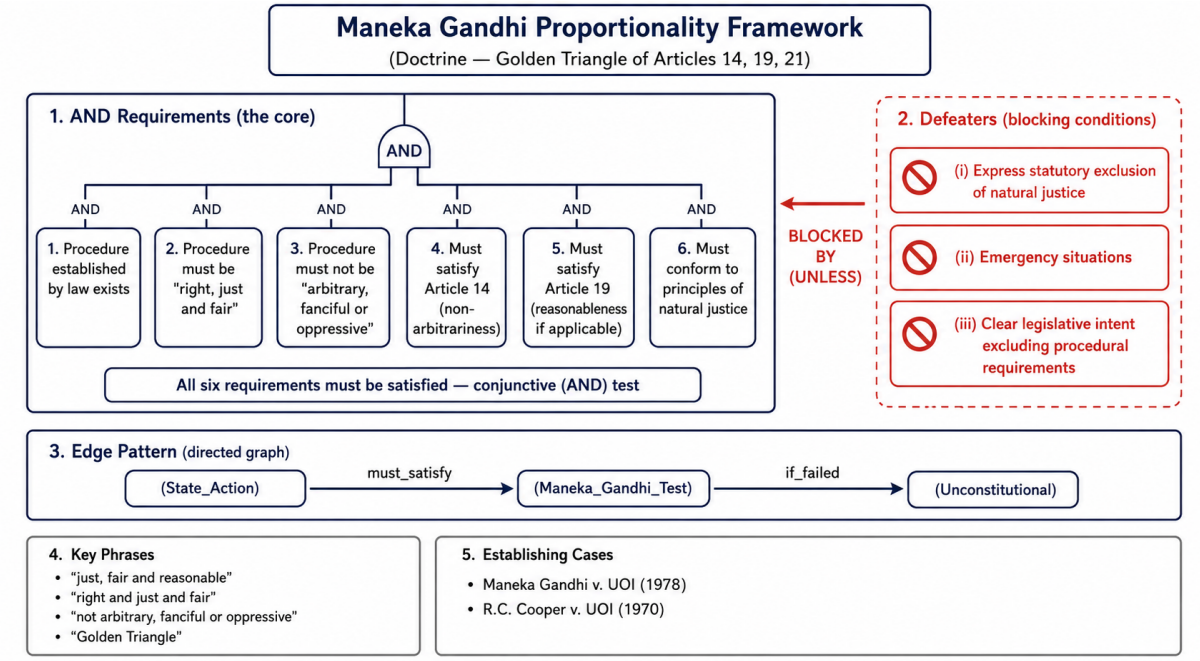

Figure 2: Anatomy of a compiled ontology entry: the five components (AND/OR requirements, defeaters, edge patterns, key phrases, and establishing cases) that compose a single concept definition. Subsequent subsections (§A.1, §A.2, §A.3) present text-based instantiations of this schema across the Indian, ECHR, and Turkish jurisdictions.

for interrogation; no inducement/threat to witnesses; no leaving India without permission

AND BNSS 482(4) carve-out: anticipatory bail does *not* apply to BNS 65 (rape of a minor under 16) or BNS 70(2) (gang rape of a minor)

Defeaters: accusation falls under the BNSS 482(4) statutory exclusion; prima facie case requiring custodial interrogation; flight risk; likelihood of witness tampering; apprehension of arrest not bona fide.

Key phrases: "anticipatory bail"; "Section 438"; "BNSS 482"; "reason to believe arrest may be made"; "pre-arrest bail"; "custodial interrogation"; "humiliate or injure".

Establishing cases: Gurbaksh Singh Sibbia v. State of Punjab (1980); Sushila Aggarwal v. State (NCT of Delhi) (2020); Siddharam Satlingappa Mhetre v. State of Maharashtra (2011).

Edge pattern: (Apprehension_of_Arrest) $\xrightarrow{\text{HC/Sessions}}$ (Court_Evaluates_BNSS482) $\xrightarrow{\text{if no exclusion}}$ (Anticipatory_Bail_Granted).

Electronic Evidence (Evidence; BSA §63 / IEA §65B). *Definition:* Admissibility of computer outputs and digital records requires a certificate authenticating the production process and the integrity of the data.

AND Evidence is an electronic record (as defined in the IT Act 2000)

AND Accompanied by a certificate under IEA 65B(4) / BSA 63(4) identifying the record, describing its production, and giving particulars of the device used

AND Under BSA 63: certificate in prescribed format; hash value of the record is *mandatory*; signed by both the person in charge of the device *and* an expert

AND Computer was regularly used for lawful activities during the relevant period; information was fed in the ordinary course of business

Defeaters: no IEA 65B(4) / BSA 63(4) certificate (per Anvar P.V.); certificate fails prescribed format or hash requirement; hash mismatch indicating tampering; broken chain of custody; computer malfunction in the relevant period; certificate signed by an unauthorized person.

Key phrases: "electronic evidence"; "Section 65B"; "BSA 63"; "certificate"; "hash value"; "electronic record"; "Anvar P.V. compliance"; "prescribed format".

Establishing cases: Anvar P.V. v. P.K. Basheer (2014); Arjun Panditrao Khotkar v. Kailash Kushanrao Gorantyal (2020); Shafhi Mohammad v. State of HP (2018, overruled by Khotkar).

Edge pattern: (Electronic_Record)

$\xrightarrow{\text{valid 65B/63 cert + hash}}$ (Admissible) $\xrightarrow{\text{court evaluates authenticity}}$ (Evidence_Weighed).

A.2 ECHR Concepts

Article 3 — Prohibition of Torture (Absolute Right). *Definition:* Absolute prohibition on torture and inhuman or degrading treatment; non-derogable under Article 15§2; no balancing permitted.

AND Treatment crosses the *minimum level of severity* threshold (relative: duration, effects, victim characteristics)

AND Classification into a tier: torture (deliberate, purposive) > inhuman treatment (intense suffering) > degrading treatment (humiliation)

Defeaters: none — the right is absolute. No exceptions, no derogation, no balancing.

Key phrases: "minimum level of severity"; "absolute prohibition"; "no derogation"; "non-refoulement"; "substantial grounds for believing"; "real risk"; "most fundamental values".

Establishing cases: Ireland v. UK (1978); Soering v. UK (1989); Chahal v. UK (1996); Selmouni v. France [GC] (1999); Saadi v. Italy [GC] (2008); M.S.S. v. Belgium and Greece [GC] (2011).

Edge pattern: (Treatment) $\xrightarrow{\text{min severity?}}$ (Torture/Inhuman/Degrading) $\xrightarrow{\text{if met}}$ (Automatic_Violation).

Article 8 — Right to Respect for Private and Family Life (Qualified Right). *Definition:* Protection of private life, family life, home, and correspondence; restrictions are permitted only when the four-step interference test under Article 8§2 is satisfied.

AND State action constitutes an *interference* with private life, family life, home, or correspondence

AND Interference is *prescribed by law* (accessible and foreseeable)

AND Interference pursues a *legitimate aim* from the Article 8§2 closed list

AND Interference is *necessary in a democratic society* (pressing social need plus proportionate response)

Defeaters: interference is in fact prescribed by law; pursues a legitimate aim; is necessary and proportionate; State acted within its margin of appreciation.

Key phrases: "private life"; "family life"; "home"; "correspondence"; "personal data"; "physical and psychological integrity"; "surveillance safeguards".

Establishing cases: Dudgeon v. UK (1981); López Ostra v. Spain (1994); Von Hannover v. Germany (No. 2) [GC] (2012); S. and Marper v. UK (2008); Zakharov v. Russia (2015).

Edge pattern: (Interference_with_Art8_right) $\xrightarrow{\text{four-step test}}$ (Prescribed + Aim + Necessary) $\xrightarrow{\text{if fails}}$ (Violation).

A.3 Turkish Concepts

Right to Personal Liberty and Security (Art. 19, 1982 Constitution; Kişi Özgürlüğü ve Güvenliği). *Definition:* Constitutional protection against unlawful or arbitrary detention. The same ontology schema is adapted to Turkish criminal procedure (*Ceza Muhakemesi Kanunu*, CMK) and Anayasa Mahkemesi (AYM) bireysel başvuru jurisprudence; structurally parallel to ECHR Article 5.

AND *Kuvvetli suç şüphesi* — strong suspicion based on concrete evidence (CMK m.100)

AND At least one of: *kaçma şüphesi* (flight risk) **or** *delilleri karartma* (evidence tampering)

AND Maximum detention period under CMK m.102 not exceeded: 1.5 years (non-heavy penal cases), 5 years (heavy penal cases), 7 years (enumerated offenses including the Anti-Terror Law list)

AND Alternative measures (*adli kontrol*, CMK m.109) considered before detention

Defeaters: CMK m.100 detention requirements not met; CMK m.102 maximum period exceeded; no concrete evidence supporting strong suspicion; *adli kontrol* alternatives not considered; automatic detention without individualised assessment.

Key phrases: "kişi özgürlüğü"; "tutuklama"; "kuvvetli suç şüphesi"; "kaçma şüphesi"; "delilleri karartma"; "gözaltı"; "adli kontrol"; "tutukluluk süresi".

Establishing cases: AYM, Mehmet Haberal [GK], B. No. 2012/849, 4/12/2013 (foundational Art. 19 bireysel başvuru); AYM, Murat Narman, B. No. 2012/1137, 2/7/2013; AYM, Hikmet Kopar ve diğerleri [GK], B. No. 2014/14061, 8/4/2015 (detention lawfulness criteria).

Edge pattern: (Detention_Order) $\xrightarrow{\text{CMK m.100 test}}$ (Kuvvetli_Şüphe + Tutuklama_Nedeni) $\xrightarrow{\text{if not met}}$ (Art. 19_Violation).

B Edge Relation Vocabulary

Table 5 enumerates the 31 canonical edge relations in the LexCGraph schema together with the node-type pairs each relation is permitted to connect under the type-level validation matrix M (§3.1). Variant surface forms emitted by the extraction LLM are normalized to these canonical names through the alias mapping layer; only canonical names appear here. Node-type abbreviations: F=Fact, C=Concept, I=Issue, A=Argument, H=Holding, P=Precedent, O=Outcome.

Table 5: Canonical edge relations and their permitted source/target node types.

Relation	Source(s)	Target(s)
triggers	F	C, I
negates	F, H	C
partially_satisfies	F	C
satisfies	F	C
claims_satisfies	F, A	C
supports	F, C, A, H, P	A, H, C, I
grounds	F, C, A, H, P	A, H, C
rebuts	F, C, A	A, H, C
undercuts	F, C, A, H	A, H, C
addresses	F, C, A, H, P, I	I, C, H, A, P
requires	C, I	C, I, H, A
excludes	C	C
specializes	C, H, I	C, H, I
conflicts_with	C, H, I	C, H, I
constrains	C, H	H
enables	C	H
concedes	A	I
attacks	A	A
supports_arg	A	A
responds_to	A	A
resolves	H	I
partially_resolves	H	I
determines	H	O
contributes_to	H	O
follows	H	P
applies	H	P
distinguishes	H	P
overrules	H	P
doubts	H	P
explains	H	P
establishes	P	C

C Quality Tier Distribution

Ontology	Gold	Silver	Bronze	Reject
11 concepts	77.1%	20.7%	1.7%	0.6%
54 concepts	90.3%	7.5%	1.8%	0.4%

D Graph Quality Assessment: LLM-as-a-Judge

Setup. To obtain an automated estimate of graph faithfulness, we conducted an independent LLM-as-a-judge evaluation using Claude Opus 4.7 (Anthropic, 2026), architecturally distinct from both Grok-4.1 (IL-TUR extractor) and GPT-4.1 mini (ECHR extractor), ruling out within-family stylistic bias. The judge scored 100 graphs against their source judgments on five dimensions (Concept Identification, Edge Accuracy, AND/OR Structure, Reasoning Coverage, Overall Faithfulness) using a 5-point anchored scale; cases were stratified across automated quality tiers (25 gold, 25 silver, 25 bronze IL-TUR; 10 gold, 10 silver, 5 bronze ECHR), and the judge never saw the tier label.

Table 6: LLM-as-a-judge evaluation (Claude Opus 4.7, anchored 5-point rubric, n=100: 75 IL-TUR + 25 ECHR). The IL-TUR/ECHR gap reflects the frontier (Grok-4.1) vs. non-frontier (GPT-4.1 mini) extraction-model difference.

Dimension	Overall	IL-TUR	ECHR
Concept Identification	2.66	2.72	2.48
Edge Accuracy	2.83	2.97	2.40
AND/OR Structure	2.57	2.64	2.36
Reasoning Coverage	**3.50**	**3.75**	2.76
Overall Faithfulness	3.03	3.16	2.64

Findings. Table 6 surfaces three findings. (1) *Reasoning Coverage is the strongest dimension (3.50/5)*, exactly the property the pipeline was designed to optimize; the judge confirms the graphs reliably capture the chain of reasoning from issues through arguments to holdings, even when individual concept labels are imperfect. (2) *Approximate graphs are valuable.* Overall Faithfulness at 3.03/5 indicates moderate per-case quality, yet these same graphs drive the +11.6pp IL-TUR and +18.3pp ECHR macro-F1 gains documented in §5: structured organization of judicial reasoning contributes downstream value well below the threshold of flawless extraction, and Concept Identification at 2.66/5 reflects the legitimate cost of compressing a 50-page judgment to 15–40 nodes rather than an extraction failure. (3) *Schema-conformance tiers and semantic faithfulness capture orthogonal properties*: no dimension exhibits the expected gold > silver > bronze ordering (within IL-TUR the per-tier means cluster within 0.2 points of each other and are non-

monotonic), suggesting that structural validation alone is an incomplete proxy for graph quality and that future automatic graph-QA work should pair conformance checks with semantic evaluation. The cross-extractor gap is largest on Reasoning Coverage (IL-TUR 3.75 vs. ECHR 2.76), reinforcing §5.4's finding that frontier extractors produce measurably better reasoning chains.

Rubric. The rubric below is the one used in this graph quality assessment. Anchored level definitions calibrate the scale across dimensions and across cases. The judge is explicitly told that the graph is a *compressed structural summary*, not an exhaustive transcription, to avoid penalizing legitimate compression as omission.

Transparency note on rubric calibration. A pilot v1 of the rubric (without the compressed-summary framing and without anchored level definitions) is preserved alongside the released code. It produced uniformly harsh and uninformative scores (e.g., Concept Identification 2.13/5 across all tiers, AND/OR Structure 2.16/5) because the judge penalized any concept missing from the source rather than evaluating whether extracted concepts were correct. The calibrated v2 below is the version reported in Table 6; comparing v1 and v2 shows that the calibration moved exactly the dimensions whose questions had a well-defined compression mismatch (Concept Identification +0.53, AND/OR Structure +0.41), while leaving Edge Accuracy (-0.05) and Reasoning Coverage ($+0.07$) essentially unchanged. The selective movement is consistent with anchored-scale calibration correcting a known rubric defect rather than uniform score inflation.

Prompt framing (verbatim).

> You are evaluating the quality of an automatically extracted legal reasoning graph against its source court judgment.
>
> **Context:** This graph is a compressed structural representation of the judgment's reasoning, not meant to be exhaustive. It extracts key legal concepts, their relationships, and the chain of reasoning leading to holdings. Evaluate the graph's quality relative to its purpose as a structured reasoning summary, not as a complete transcription of the judgment.

Anchored scale definitions (1–5 per dimension).

CONCEPT IDENTIFICATION.
1: extracted concepts mostly wrong or irrelevant to the case.
2: some relevant concepts captured but major gaps or misidentifications.
3: core legal concepts correctly identified; some secondary concepts missing.
4: all major concepts correct; minor gaps in secondary concepts.
5: comprehensive and accurate concept identification.

EDGE ACCURACY.
1: most edges spurious or unsupported by the judgment.
2: some edges grounded in text but many are inaccurate.
3: majority of edges reflect real relationships; some questionable.
4: nearly all edges well-grounded; minor inaccuracies.
5: all edges clearly supported by the source text.

AND/OR STRUCTURE.
1: conjunctive/disjunctive requirements ignored or seriously misrepresented.
2: some logic structure present but key requirements misrepresented.
3: core conjunctive/disjunctive requirements captured; some nuance lost.
4: logic structure largely accurate; minor issues.
5: conjunctive/disjunctive requirements precisely captured.

REASONING COVERAGE.
1: graph misses most of the chain of reasoning leading to holdings.
2: partial chain captured but major reasoning steps missing.
3: core chain from issues to holdings captured; some intermediate steps missing.
4: reasoning chain largely complete with minor gaps.
5: full reasoning chain from issues through arguments to holdings represented.

OVERALL FAITHFULNESS.
1: graph would mislead a legal professional about the case.
2: significant inaccuracies but core outcome recognizable.
3: faithful representation of core reasoning; some imperfections acceptable.
4: highly faithful structured summary with minor issues.
5: a legal professional would consider this an accurate structured summary.

Sampling and blinding. The judge scored 100 graphs sampled with fixed seed 42: 75 from IL-TUR (25 gold, 25 silver, 25 bronze, drawn from the 1939/518/46 tier pools) and 25 from ECHR Turkey (10 gold, 10 silver, 5 bronze, drawn from the 135/102/33 pools). Each prompt contains the full source judgment (truncated to 80K characters only if it exceeds 100K) plus the linearized graph; the automated quality tier is never included. All 100 prompts were scored by Claude Opus 4.7 (Anthropic, 2026) via the `claude -p` CLI in a single batch; all 100 responses parsed as well-formed JSON with valid 1–5 integer scores across all five dimensions.

E Extended Cross-Jurisdictional Analysis

This appendix extends §5.4 with the capability-scaling, extraction-robustness, and raw-blinding-robustness analyses summarized there.

Capability scaling. The graph macro-F1 advantage is large at both predictor scales (+22.2pp on GPT-4.1 mini, +18.3pp on GPT-5.4). No-violation recall rises with predictor capability (+37.1pp $\rightarrow$ +40.0pp), and absolute graph macro-F1 climbs from .744 to .845, consistent with the §5 finding that the graph's value lies in relational structure requiring active inference: stronger reasoners more effectively exploit the conjunctive/disjunctive justification structure that distinguishes a no-violation finding (e.g., margin-of-appreciation deference, prescribed-by-law satisfied, legitimate aim served) from a violation finding.

Extraction robustness. Graphs were extracted by GPT-4.1 mini, a non-frontier model, with 80% gold+silver quality tier (vs. 90% gold on IL-TUR with Grok-4.1). Despite this lower extraction quality, the prediction advantage is large at both predictor scales, indicating that ontology-provided edge whitelists and type-level validation compensate for weaker extraction. This evidence (together with the calibration improvement replicating across both jurisdictions) supports the claim that the pipeline's benefits derive from the ontology-driven structural constraints, not from frontier-model extraction.

Raw blinding robustness. A leak audit identified ECHR-specific outcome-revealing patterns (Registry headnote bullet chains, Grand Chamber narrations of prior Chamber rulings, voting splits, admissibility narrations) that the Indian-SC regex missed and that were 2$\times$ more prevalent in no-violation cases (54% vs. 28%); ECHR-specific strips were added before this evaluation. The graph's structural blinding (withholding outcome fields by schema rather than regex) is robust to this failure mode by construction.

F Human Calibration Details

This appendix extends the human calibration summary in §D with full methodology and per-dimension results.

Sampling methodology. From the same 75 IL-TUR graphs Opus scored, we drew a stratified subsample of 45 cases (15 per automated quality tier) across two rounds. Round 1 sampled 15 cases (5 per tier) with `random.Random(42)` on sorted per-tier pools (matching the seed used for the original 75-case Opus selection in §D). Round 2 sampled an additional 30 cases (10 per tier) from the remaining 60 with `random.Random(43)` on sorted per-tier pools, to extend the calibration to n=45 without overlap. An independent legal annotator re-scored all 45 graphs on the identical 5-dimension rubric (Appendix D), blinded to the Opus scores; the manifest given to the annotator stripped the `opus_result_path` field, and the bundle contained only the prompts and graphs.

Per-dimension agreement. Table 7 reports per-dimension means and Spearman rank correlations. Reasoning Coverage is the only dimension whose human/LLM-judge rank correlation reaches significance (ρ=0.33, p=.026); Overall Faithfulness means are nearly identical between the two judges (3.24 vs. 3.16); and the human rates Concept Identification roughly one point higher than Opus on average (3.76 vs. 2.78), suggesting that the automated judge applies a stricter concept-labeling standard. Pooled across the 225 score pairs (45 cases $\times$ 5 dimensions), exact agreement is 50.2%, within-$\pm$1 agreement is 91.6%, and the mean absolute error is 0.58: the two judges co-locate cases on the rubric (over 90% within one rubric point) even where per-dimension rank correlations are modest, consistent with a shared signal observed through different scoring styles rather than disagreement on underlying case quality.

Table 7: Human vs. Opus 4.7 calibration on 45 IL-TUR graphs (15 per automated quality tier; two rounds, seeds 42 and 43). Human annotator blinded to Opus scores. ρ is Spearman's rank correlation; bold marks $p < .05$.

Dimension	Human	Opus	ρ	p
Concept Identification	3.76	2.78	0.03	.846
Edge Accuracy	2.98	2.96	0.15	.331
AND/OR Structure	3.27	2.67	0.21	.164
Reasoning Coverage	3.24	3.71	**0.33**	**.026**
Overall Faithfulness	3.24	3.16	0.11	.459

G Extended Discussion

This appendix extends §6 with the discussion paragraphs summarized at the end of that section.

Structure, not just compression. The summarization baseline (§5.6) rules out the most-cited alternative explanation for the Graph > Raw gap: that any LLM-driven compression of the blinded judgment would suffice. A length-matched free-form GPT-4.1 mini prose summary of the full blinded judgment (given ≈6–25× more source text than the Raw condition and a target length matched to the *longer* Stage 4 linearization) improves over Raw by 5–12pp macro-F1 across the three predictors but still loses to ontology-driven Nodes-Only extraction by +12.1pp (GPT-4.1 mini) and +13.5pp (GPT-5.4) on macro-F1. Compression captures part of the gain; ontology-grounded concept normalization captures the rest. Moreover, the summary is a single-task artifact: it provides no concept index for retrieval, no AND/OR justification structure for counterfactual analysis, and no typed edges for structural queries, all capabilities the same graph artifact enables from a single extraction pass.

When to invest in ontology expansion. The ontology completeness ablation (Appendix H) provides practical guidance: 11 well-specified doctrinal concepts are sufficient for prediction, but 54 concepts are needed for high-quality graphs (90% gold) suitable for retrieval, explainability, and counterfactual analysis. This suggests a two-stage adoption path: start with core doctrines for immediate prediction benefits, then expand for richer downstream applications.

Ontology compilation cost. The 54-concept Indian ontology was compiled in approximately 16 hours of domain expert effort, with 43 offense/procedural/evidence concepts enriched via LLM-assisted generation followed by independent legal verification (93% of concepts passed verification; 2 errors corrected). This hybrid approach (LLM generation with expert audit) suggests a scalable methodology for adapting to new jurisdictions.

What does and does not generalize across jurisdictions. The ECHR evaluation (§5.4) shows the pipeline ports beyond Indian law using a separately compiled 55-concept ontology built under the same schema, with macro-F1 gains of +18–22pp driven by +37–40pp minority-class recall on both OpenAI predictors. Accuracy gains are real but compressed by class imbalance: at 88% violation, the +40pp minority win yields only +3.4pp aggregate accuracy on GPT-5.4 (bootstrap 95% CI excludes 0; McNemar marginal at χ^2=3.68); macro-F1 is the fairer summary in imbalanced settings. Calibration improvement also replicates (ECE: .045→.027 on GPT-5.4; Brier favors graph at both scales).

Graph quality vs. downstream utility. Our LLM-as-a-judge evaluation (§D) reveals a productive dissociation: graphs score a moderate 3.03/5 on Overall Faithfulness and 2.66/5 on Concept Identification, yet drive +11.1pp prediction improvement on IL-TUR and +18.3pp macro-F1 on ECHR. *Approximate structural organization of judicial reasoning therefore contributes more downstream value than perfect but unstructured text: the graphs need not be flawless to be useful.* The per-dimension scores also surface an architectural alignment: Reasoning Coverage (3.50/5, the highest dimension) is exactly the property the pipeline was designed to optimize, while Concept Identification at 2.66/5 reflects the legitimate cost of compressing a 50-page judgment into a structured summary. Finally, the absence of a gold > silver > bronze ordering in the judge's per-tier scores indicates that schema-conformance tiers and semantic faithfulness capture orthogonal properties; future automatic graph-QA pipelines should pair structural validation with semantic evaluation rather than relying on conformance alone.

Extractor capability shapes graph quality. The frontier (Grok-4.1) vs. non-frontier (GPT-4.1 mini) extractor gap is visible in two independent signals: the judge's Reasoning Coverage score (IL-TUR 3.75 vs. ECHR 2.76, a near one-point gap) and ECHR's lower absolute prediction head-

room relative to IL-TUR. Both signals converge on the same conclusion (more capable extraction models produce measurably better reasoning chains), suggesting that the ontology constraints, while compensating substantially for weaker extraction, do not fully close the capability gap.

H Ontology Completeness Ablation

To assess the marginal value of ontology investment, we expanded the Indian ontology from 11 schema-complete doctrinal concepts to 54 concepts (adding 43 offense, procedural, and evidence concepts with complete legal elements, defenses, and key phrases) and re-ran the full pipeline on the same 2,517 cases.

Table 8: Ontology completeness ablation. Expanding from 11 to 54 concepts dramatically improves graph quality while prediction accuracy remains unchanged.

Metric	11 concepts	54 concepts
Prediction accuracy	0.855	0.853
Gold tier graphs	77.1%	90.3%
UNLISTED nodes/graph	47.4	3.2
Edges/graph	108.8	116.0
Unique concept IDs	14,064	1,634

Table 8 reveals a dissociation between graph quality and prediction accuracy. Graph structural quality improved dramatically: gold-tier graphs increased from 77.1% to 90.3%, pseudo-cluster (UNLISTED) node assignments dropped by 93% (47.4→3.2 per graph), and 14,064 ad-hoc pseudo-concept IDs collapsed into 1,634 canonical ontology concepts. However, prediction accuracy was statistically unchanged (0.855 vs. 0.853, $p > 0.05$). This indicates that the 11 core doctrinal concepts (proportionality, natural justice, reasonable classification, Wednesbury unreasonableness, and others) capture the reasoning structure predictive of case outcomes. The additional 43 offense-level and procedural concepts improve structural completeness for applications requiring fine-grained graph fidelity (retrieval, explainability, counterfactual analysis) without contributing additional predictive signal for binary outcomes.

I Facts-Only Control and Fairness Audits

To test whether the pipeline's value derives from structuring judicial reasoning or merely from reorganizing factual content, we extracted graphs from only the facts section of 50 cases: facts-only graph (62%) matches raw facts (64%) and is far below full graph (90%), confirming the signal comes from the court's analytical sections rather than the facts. A separate 30-case leakage audit measured comparable outcome-signal density in raw vs. graph (0.233 vs. 0.225 per 1K chars), and the consistent ordering full graph > raw > structured nodes makes memorization unlikely as the sole explanation: no memorization hypothesis explains why removing edges would degrade an extracted graph below raw text. Together these controls attribute the gap to structural organization rather than information advantage.

J Counterfactual Impact Analysis

The AND/OR justification structure enables counterfactual impact analysis: given any concept node, the system deterministically traces which holdings lose support upon its removal.

Table 9: Counterfactual structural statistics across 2,446 graphs (15,276 justification sets, 14,483 holdings with justification).

Metric	Value
Total justification sets	15,276
AND logic	11,877 (77.7%)
OR logic	3,399 (22.3%)
Avg. evidence edges per JS	4.8
Fragile holdings (AND only)	11,220 (77.5%)
Robust holdings (≥1 OR path)	2,606 (18.0%)
Mixed holdings	657 (4.5%)
Mean fragility score	0.344
Cases with fragility > 0.5	324 (13.2%)

Table 9 presents corpus-level statistics. 77.5% of holdings are supported exclusively by conjunctive (AND) justification sets, meaning removal of *any* required concept eliminates the holding's justification. The mean fragility score of 0.344 indicates that removing the single most critical concept eliminates approximately one-third of a case's holdings, consistent with the multi-issue nature of appellate decisions. The most frequently load-bearing concepts are Article 14 (146 cases) and Natural Justice (102 cases), aligning with their prevalence in Indian constitutional litigation.

Doctrinal predictive asymmetry. Individual doctrines vary substantially in outcome directionality. When serving as the structurally critical concept, Reasonable Classification cases are rejected

87.9% of the time (reflecting the twin test's deference to the legislature), while Natural Justice cases show near-equal outcome splits (50.8%), indicating that outcomes depend on which prongs of the multi-part test are satisfied rather than on doctrinal presence alone. This asymmetry suggests that the predictive value of graphs derives from capturing *which* elements of doctrinal tests are satisfied, not merely from identifying *which* doctrines appear.

K Worked Example: A Linearized Graph

This appendix presents one extracted graph end-to-end, focused on the Maneka Gandhi cluster from Figure 2. The example complements the abstract pipeline description in §3 with a concrete input–output pair drawn from the IL-TUR evaluation set.

Case. IL-TUR `2019_890` (Supreme Court of India, F. Nariman J.; gold quality tier; outcome *allowed*): the appellant challenged a tender clause requiring a 10% pre-deposit to invoke arbitration; the Court struck the clause down under Article 14 as arbitrary, anchoring the analysis on the Maneka Gandhi due-process line.

Cluster decomposition. The pipeline extracts 22 facts, 15 concepts, 5 issues, 16 arguments, 9 holdings, and 16 precedents, and assigns them to 11 ontology clusters (Table 10). Cluster membership is not a partition: each Fact, Concept, Argument, and Precedent maps to at most one ontology cluster, whereas Issues and Holdings frequently address multiple doctrines and are counted in every cluster they belong to (the issue and holding column sums therefore exceed the per-type totals); a further 12 nodes (7 facts, 2 arguments, 2 holdings, 1 precedent) fall outside the 11 ontology clusters into the pseudo-cluster. We focus on the Maneka cluster, one of three doctrinal clusters in this case, because it is the running ontology example of §3.2 and Figure 2.

Maneka cluster contents. The cluster's anchor concept node is `c10`, encoding the ontology concept `doctrine_maneka_gandhi_framework_due_process`, with anchored surface text "In Maneka Gandhi v. Union of India [1978] 2 S.C.R. 621 it was observed that Art. 14 strikes at arbitrariness in State action and ensure[s] fairness and equality of treatment." The cluster's other members are: companion concept `c5` (TEST_WEDNESBURY); fact `f13` (the High Court's finding that the clause was not arbitrary); issues `i4` (whether the clause is arbitrary) and `i5` (whether tender terms admit judicial scrutiny); holdings `h1`, `h7`, `h9`; and precedents `p8`–`p10`, `p14`, `p15`.

Table 10: Cluster decomposition of case `2019_890`. Cells are distinct nodes per cluster, by type (F=Fact, C=Concept, I=Issue, A=Argument, H=Holding, P=Precedent). Column sums do not equal the extracted totals because Issues and Holdings can belong to multiple clusters and a small pseudo-cluster (12 nodes) absorbs items not matched to any ontology concept.

Cluster	F	C	I	A	H	P
Maneka Gandhi (Due Process)	1	2	2	0	3	5
Manifest Arbitrariness	0	1	1	2	2	0
Proportionality Test	0	1	1	0	2	0
Article 14	2	4	3	6	5	3
Article 226	8	1	0	0	0	1
Contract of Adhesion	1	1	1	2	1	1
Unconscionable Contract	1	1	1	0	1	0
Judicial Review of Tenders	1	1	1	2	1	3
Party Autonomy in Arb.	0	1	0	0	0	0
Arb. & Conciliation §5	0	1	0	0	0	0
Exemplary Costs (Frivolous)	1	1	1	2	2	2

Intra-cluster edges (Stage 3, LLM). Within the cluster, the LLM proposes edges restricted to the ontology's relation whitelist; each is validated against type matrix M. A representative subset of the 24 intra-cluster edges:

```
f13 --[claims_satisfies]--> c10
c10 --[grounds]--> h7
i4  --[addresses]--> c10
h7  --[supports]--> c10
h7  --[resolves]--> i4
h7  --[supports]--> h9
```

Justification set and reasoning chain (Stage 4, deterministic). For holding `h7` ("Clause 25(viii) is arbitrary and violative of Article 14 of the Constitution of India"), the Stage 4 compiler emits the AND-justification set `js6` (label: *Maneka Gandhi Framework (Due Process) support*) and a reasoning chain `rc_i4_h7` linking issue `i4` to `h7`. The chain's 22 critical nodes {`c4`, `c6`, `c10`, `c11`, `c15`, `a4`, `a8`, `a12`–`a14`, `f13`, `f21`, `h1`, `h6`, `h7`, `h9`, `p3`, `p7`–`p11`} span four clusters (Maneka, Article 14, Manifest Arbitrariness, Judicial Review of Tenders). The Maneka cluster supports two further AND-justification sets in this case (`js1`→`h1` and `js7`→`h9`), so all three Maneka-anchored holdings are counterfactually fragile under the Appendix J analysis. The

`addresses`, `resolves`, and `determines` edges joining this chain are produced without LLM involvement; this is the mechanism by which long-distance argumentative structure assembles deterministically from local LLM inferences, while the graph remains reproducible conditional on the Stage 1–3 outputs.

L Does Relational Structure Degrade Prediction? A Targeted Analysis

The Stage 4 ablation (§5.5, Table 4) reports that augmenting the prediction prompt with cross-cluster edges, justification sets, and reasoning chains lowers macro-F1 versus the Nodes-Only linearization at all three predictor scales (-15.5pp GPT-4.1 mini, -10.7pp GPT-5.4, -5.9pp Kimi K2.6). This appendix asks *why*, combining an offline diagnostic (re-linearizing all 297 ECHR graphs and joining the per-case outcomes of the two conditions; it reproduces every Table 4 graph macro-F1 to within 0.001, confirming the condition mapping) with a single controlled re-run that resolves the confound identified below. The bottom line: once blinding is matched, the degradation almost entirely disappears.

A blinding confound, not matched conditions. Inspecting the run configurations reveals that the two graph columns were *not* produced under matched blinding, contrary to the Table 4 caption. Nodes-Only was generated with `no_scrub=True` (extracted fact, concept-interpretation, and precedent-proposition text shown in full), whereas + Stage 4 used `no_scrub=False` (facts regex-scrubbed of court-finding language; concept interpretations and precedent propositions withheld). The + Stage 4 prompt therefore saw *less* lexical content *and* more structure: the two conditions differ along two axes at once. Re-linearizing each graph three ways isolates the effect (Table 11, Figure 3): scrubbing removes a mean of 1,435 characters per case (-26%, from 5,530 to 4,095), and the Stage 4 structure then *refills* 1,375 characters, returning the prompt to essentially the original length (mean ratio 1.00, median 1.00). The net change is a content-for-scaffolding substitution of near-identical length, of which $\approx 30\%$ of the added tokens are bare opaque identifiers (node, edge, justification-set, and reasoning-chain IDs) carrying no standalone lexical signal.

Table 11: Graph-prompt length decomposition, mean per case over 297 ECHR graphs (tiktoken unavailable offline; characters and whitespace words; GPT BPE $\approx$ chars/4). Stage 4 first scrubs the base text, then appends opaque-ID structure, leaving total length unchanged.

Linearization	Chars	Words	Δ chars
Nodes Only (`no_scrub=T`)	5,530	804	—
Stage 4 base (scrubbed)	4,095	590	$-1,435$
+ Stage 4 (scrubbed+struct)	5,470	786	$+1,375$

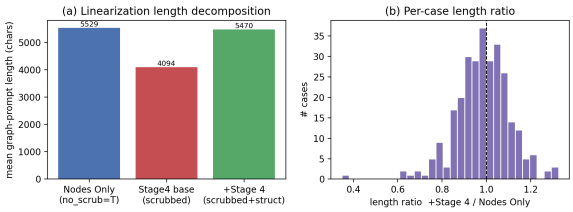

Figure 3: (a) The + Stage 4 prompt is not longer than Nodes-Only: scrubbing removes $\sim$26% of the text, which the opaque-ID structure refills to near-parity. (b) Per-case length ratio is centred on 1.0.

Ruling out the intuitive mechanisms. We tested each candidate explanation for the degradation (Table 12; statistics for GPT-5.4, the clearest signal, replicated in sign on the other predictors). *Length/dilution* is rejected: the conditions are length-matched and, if anything, more added structure correlates with *less* degradation (Kimi point-biserial $r{=}-0.18$, $p{=}.003$). *Edge volume* is rejected: degraded cases carry fewer cross-cluster edges (6.9) than stable-correct ones (7.5). *Specific edge types* are rejected: the cross-cluster relation distribution is nearly identical across groups (the lone apparent spike, `negates` at $8.9\times$, is a single case). *Doctrinal complexity* is rejected: the per-article macro-F1 change is uncorrelated with each concept's ontology requirement count ($\rho{=}0.01$, $p{=}.98$; Figure 4), and the loss is broad-based — every well-populated Convention article degrades (Art. 6 $n{=}182$, Art. 3 $n{=}108$, Art. 5 $n{=}63$), while the only gains fall on small-n articles. *Overconfidence* is rejected: on degraded cases the model almost never raises its confidence in the new (wrong) answer ($\leq 2/14$ cases per predictor); it becomes uniformly *less* confident, though overall worse calibrated (GPT-5.4 Brier $.055{\rightarrow}.079$, ECE $.027{\rightarrow}.057$). The justification-fragility link (App J) is weak and non-significant (degraded mean fragility 0.25 vs. stable 0.14, $p{=}.19$, $n{=}259$).

Table 12: Candidate mechanisms for the + Stage 4 degradation and their verdicts (GPT-5.4; replicated in sign across predictors). Only graph sparsity and the blinding-induced content substitution survive.

Mechanism	Test (GPT-5.4)	Verdict
Prompt length	ratio med. 1.00	rejected
Edge volume	6.9 vs. 7.5 edges	rejected
Edge type	dist. identical	rejected
Complexity	ρ=.01, p=.98	rejected
Confident-wrong	$\leq 2/14$ cases	rejected
JS fragility	.25 vs. .14	weak/NS
Graph sparsity	56 vs. 68, p=.02	**supported**
Content loss	-26%, 30% IDs	**primary**

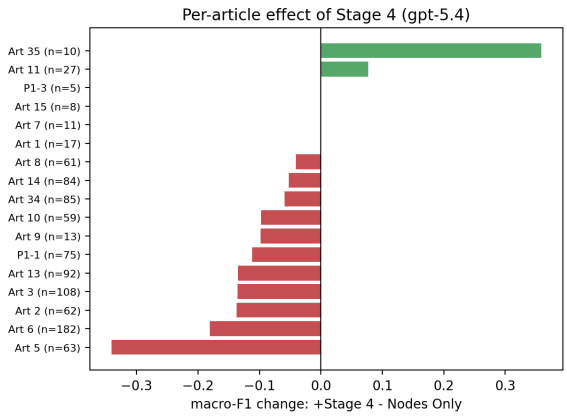

Figure 4: Per-article macro-F1 change (+ Stage 4 − Nodes-Only, GPT-5.4). Degradation is broad-based across all well-populated Convention articles; the only gains are on small-n articles and are uncorrelated with doctrinal complexity.

Table 13: Controlled re-run isolating Stage 4 (GPT-5.4, same 297 cases/settings). Adding Stage 4 structure to the *same full text* Nodes-Only saw (matched `no_scrub=True`) recovers ∼82% of the macro-F1 drop and essentially all of the calibration drop; the residual is non-significant (paired McNemar χ^2=0.13, p=.72).

Condition	Acc	Macro-F1	Brier	ECE
Nodes Only (`no_scrub=T`)	.943	.845	.055	.027
+ Stage 4, `no_scrub=F` (Tab. 4)	.909	.738	.079	.057
+ Stage 4, `no_scrub=T` (matched)	.936	**.826**	.057	.028

What explains the unmatched gap. Two factors drive the Table 4 drop. First, the cases that degrade are the *sparser* graphs — fewer nodes (median 56 vs. 68, Mann-Whitney p=.02), fewer reasoning chains (2 vs. 5, p=.002), and fewer issues per case — where less lexical evidence remains to absorb the scrubbing. Second, and dominant, is the content-for-scaffolding substitution itself: + Stage 4 trades roughly a quarter of its high-signal fact/concept/precedent text for equal-length symbolic cross-references. The pattern is thus most consistent with a measurement artifact of unmatched blinding rather than with relational structure being intrinsically noisy.

The controlled experiment. To isolate Stage 4's marginal effect we ran the decisive test: + Stage 4 with `no_scrub=True` — the same structure layered on the *same full text* Nodes-Only saw — with identical model (GPT-5.4), cases, and settings (Table 13). Under matched blinding, + Stage 4 reaches macro-F1 0.826 versus Nodes-Only's 0.845: a *non-significant* −1.9pp (paired McNemar χ^2=0.13, p=.72; 5 vs. 3 discordant cases), with calibration essentially restored (Brier .057 vs. .055; ECE .028 vs. .027). This recovers ∼82% of the apparent macro-F1 degradation and essentially all of the calibration degradation in Table 4. **Once blinding is matched, adding Stage 4 structure neither significantly helps nor hurts prediction.** This refines the §5.5 reading: the structure does not present "strictly more information," and its apparent harm was overwhelmingly the unmatched-blinding artifact, not relational noise. The separation-of-concerns argument is otherwise intact: relational structure's value for concept retrieval (Table 3) and counterfactual analysis (App J) is computed independently of the prediction blinding; what changes is that structure is *prediction-neutral*

rather than prediction-harmful.