Structure-Aware Counterargument Generation via GraphRAG in DNA Evidence Cases*

Jee Won Park[1], Sungmi Park[1*]

[1] *Hallym University, 1 Hallymdaehak-gil, Chuncheon, Gangwon-do, 24252, Republic of Korea*

Abstract

We propose an argument structure-aware framework for generating legal counterarguments by retrieving and reconstructing rebuttal structures in DNA evidence cases. Existing legal retrieval and retrieval-augmented generation approaches primarily rely on semantic similarity between documents or arguments, limiting their ability to reuse how courts attack and reject competing claims. To address this limitation, judicial reasoning is represented as Argument Interchange Format (AIF)-based argument graphs, and structurally similar rebuttal patterns are retrieved through inference, conflict, and preference relations. Focusing on South Korean rape cases involving DNA evidence, we construct issue-centered argument structures covering disputes such as DNA contamination, secondary transfer, mixed profile interpretation, and consent-related evidentiary reasoning. Retrieved rebuttal structures are then adapted by a large language model to generate court-side counterarguments for new defendant claims. Experimental results show that the proposed framework can reconstruct many observed judicial rebuttals, particularly when rebuttals are grounded in direct forensic or testimonial findings. These results indicate that rebuttal structures can serve as transferable units of legal reasoning, enabling structure-aware retrieval and counterargument generation that move beyond semantic similarity-based legal RAG.

Keywords

Counterargument generation, Argument Interchange Format, GraphRAG, DNA evidence, legal AI

Content Note

This paper discusses rape and sexual assault cases and includes descriptions of sexual acts and forensic evidence.

1. Introduction

In criminal adjudication, judicial conclusions depend not only on the existence of evidence but on its interpretation and evaluation within competing argumentative contexts [1]. Since evidentiary significance varies based on surrounding facts and alternative explanations [2], judicial decisions contain structured reasoning that reveals how claims are supported, challenged, and rejected. Despite this, existing legal AI research has primarily focused on document- or sentence-level analysis for tasks like retrieval and judgment prediction [3][4], often failing to model how courts resolve competing arguments through structured rebuttal processes [5]. While retrieval-augmented generation (RAG) has extended legal AI toward counterargument generation, current approaches rely largely on textual or semantic similarity [6]. This limits their capacity to utilize prior argumentative structures as formal access points

*RELAF'26: Reasoning with Evidence in Law Enforcement and Forensics, Workshop at ICAIL 2026, Singapore.

*Corresponding author.

jeewonpark@hallym.ac.kr (J. W. Park); sungmi.park@hallym.ac.kr (S. Park)

0009-0007-7841-5387 (J.W. Park); 0000-0003-2265-3013 (S. Park)

for navigating and retrieving the specific judicial reasoning through which previous claims were challenged and rejected.

This limitation is particularly significant in disputes involving scientific evidence, such as DNA analysis in rape adjudication. While DNA may prove physical contact [7], it does not inherently establish the commission of a crime [8]. Its probative value depends instead on how the evidence is interpreted in relation to surrounding facts, such as contamination, secondary transfer, or the timing of deposition. Consequently, defendants frequently challenge not the DNA match itself but its legal significance by proposing alternative explanations [9]. In such disputes, the rebuttal process is essential for determining how these competing interpretations are evaluated and reconciled. However, current AI methodologies often lack the structural framework necessary to consistently retrieve and apply the specific justificatory logic required for forensic legal reasoning [2][4][10].

To address this, we represent judicial reasoning as Argument Interchange Format (AIF) graphs within a GraphRAG framework. The proposed system moves beyond semantic similarity by incorporating structural signals, including inference (RA), conflict (CA), and preference (PA) relations, to reconstruct rebuttal structures. Using South Korean rape cases involving DNA evidence, AIF graphs are organized by legal issue. This organization allows the system to align existing attack structures with new defendant claims to generate court-side counterarguments. The contributions of this study are as follows:

1. **Knowledge Modeling and DNA-Evidence Taxonomy:** We present a systematic workflow for formalizing DNA-evidentiary reasoning as AIF-based argument graphs by developing a domain-specific taxonomy of recurring forensic conflict types and incorporating it into an automated GPT-5 graph-construction pipeline.
2. **Cross-case Architecture:** We propose a Neo4j-based schema that uses shared LegalIssue nodes derived from the taxonomy as cross-case anchors to bridge isolated cases, enabling the retrieval of rebuttal structures through issue-constrained graph traversal.
3. **Structure-aware Counterargument Generation:** We introduce a GraphRAG-based pipeline that reconstructs retrieved attack structures for new defendant claims. We evaluate its performance against gold-standard court attacks across three large language models, including two open-source models, to demonstrate the practical utility of structural reuse.

2. Related Work

2.1. Legal Argument Representation

Legal adjudication is fundamentally a process of structured reasoning, where judicial conclusions emerge from the logical interaction of competing arguments. Early legal argument mining established that judicial decisions can be decomposed into functional units such as claims, premises, and rebuttals [4][5]. While these methods effectively identify argumentative components, representing the systematic interplay of these units across diverse cases requires more formal models like AIF [10][14].

AIF provides a unified framework by distinguishing information nodes (I-nodes) for propositions from scheme nodes (S-nodes) for inference, conflict, and preference relations. By integrating formal argumentation theory, specifically concepts of attack and acceptability [1],

AIF effectively maps the layered and interdependent nature of judicial logic. Despite these advancements, existing research remains largely focused on representing internal case structures, leaving rebuttal frameworks underutilized as reusable units for cross-case retrieval and generation.

Structured legal representations have also been combined with machine-learning-based text processing. Prior work has used H-BERT to map case facts onto factors and issues in a structured legal knowledge model [19], while another approach extracts observations from crime reports and uses them as premises in a formal argumentation system [20]. Unlike these approaches, which primarily use machine learning to populate predefined reasoning models, our work retrieves and reuses inference, conflict, and preference relations observed across judicial decisions.

2.2. Retrieval Based Legal Reasoning

Recent retrieval-augmented generation (RAG) approaches have extended legal AI toward explanation and counterargument generation. For instance, ALEX [6] generates Korean criminal counterarguments by combining RAG with formal argumentation schemes, utilizing Toulmin-style structures to build argument networks. While ALEX is closely related to our work in its aim to reuse prior judicial reasoning, its retrieval remains primarily centered on semantic similarity and textual content. Consequently, underlying rebuttal structures, specifically conflict and attack relations, are not explicitly treated as reusable units. In contrast, our approach performs retrieval directly at the level of argument structures by focusing on recurring rebuttal patterns and attack relations observed across judicial decisions.

Related research on legal knowledge graphs and GraphRAG similarly integrates structured knowledge into generation pipelines [17][18][19]. These methods connect precedents, norms, and factual elements through graph structures to improve integration with language models. However, GraphRAG approaches generally operate at the level of documents, entities, or factual relations rather than argumentation structures. Similarly, case-based reasoning (CBR) retrieves prior cases based on factual or outcome similarity [2], but because it typically operates at the level of entire cases, it prevents the internal argumentative processes from being explicitly modeled as reusable reasoning structures.

3. Argument Graph Construction for DNA Evidence

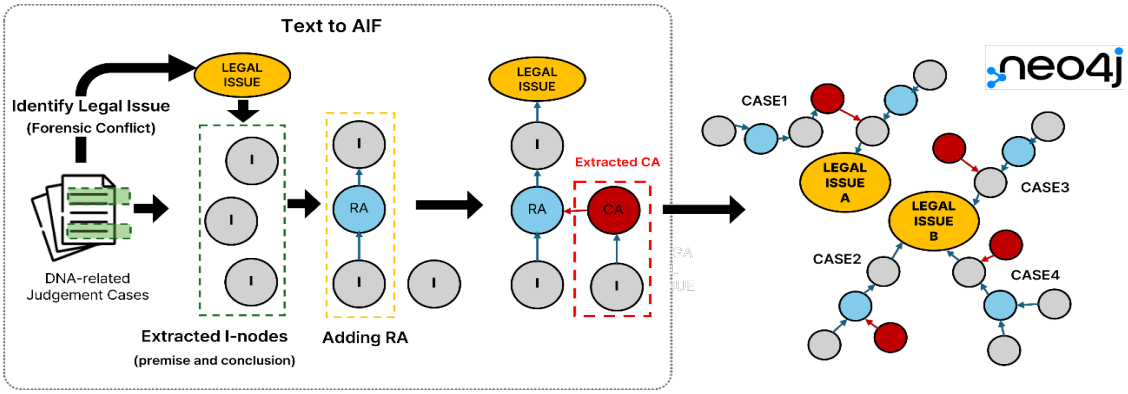

Figure 1: Workflow for AIF-based Argument Graph Construction and Cross-case Integration.

Figure 1 illustrates the workflow for transforming judicial decisions into a structured database. The automated Text-to-AIF process extracts propositions as information nodes (I-nodes) and maps their relationships through rule application (RA), conflict application (CA) nodes. These case-specific graphs are then integrated into Neo4j, using shared LegalIssue nodes as central anchors. This architecture converts isolated decisions into a unified network, enabling the structural retrieval of rebuttal patterns across the forensic domain.

3.1. Data

The dataset consists of 109 South Korean first-instance rape case rulings where DNA evidence was substantively evaluated, covering disputes such as contamination, secondary transfer, and statistical interpretation. Judicial texts were automatically processed through OCR and converted into structured JSON according to the section structure of each judgment. Only the court's reasoning sections were used for subsequent argument-graph construction. Of these, 79 cases were used to construct the argument graph database, with 30 reserved for evaluation. The evaluation set was selected to ensure a diverse representation of forensic disputes, allowing for a robust assessment of whether the framework can retrieve and reconstruct patterns focusing on structural reasoning rather than surface-level textual similarity.

3.2. Argument Graph Construction and Issue Structuring

To represent the court's reasoning, each decision is converted into an AIF-based graph through a procedure that combines a manually developed issue taxonomy with automated graph construction. GPT-5 extracts the defendant's primary claims, assigns each case to the most appropriate predefined issue category, and identifies additional I-nodes representing judicial findings, forensic evidence, and supporting propositions. It then clusters related I-nodes and constructs the corresponding RA-, CA-, and PA-nodes. All node extraction, clustering, issue assignment, and relation construction are performed automatically without manual annotation.

DNA-Evidence Issue Analysis and Taxonomy. We manually developed a domain-specific taxonomy by organizing recurring defendant-side DNA disputes into canonical categories. Based on the extracted defendant-side I-nodes, GPT-5 automatically assigns each case to the most specific applicable category in the predefined taxonomy. The assigned category guides subsequent extraction by limiting nodes and relations to the relevant forensic issue, such as DNA reliability, secondary transfer, or negative test results. In the Neo4j database, each category is represented as a shared LegalIssue node. The LegalIssue is an indexing node rather than a standard AIF I-node or S-node; it connects cases involving similar disputes but does not participate directly in inference, conflict, or preference relations. Table 1 presents the most frequent categories, and the complete taxonomy is provided in Appendix A.

Table 1: List of most frequently occurring DNA Evidence-Related Legal Issues

Issue Type	Count
Defendant denial of forensic fact	33
Dna absence or negative result	20
Dna consent or non-consent relevance	10

Information Node (I-Node) Extraction. The extraction process identifies defendant claims, judicial conclusions, and forensic findings from the judgment reasoning to represent them as I-nodes. Each node is assigned structured metadata, including its connections to AIF scheme nodes and a domain-specific forensic role (e.g., expert interpretation, defendant forensic claim) to define its function within the DNA evidence discourse. Deterministic validation removes duplicate or out-of-scope propositions and rejects outputs that violate required fields, controlled labels, or node-identifier constraints. The final stage of I-node processing involves grouping semantically related nodes into thematic clusters using GPT-5. Each cluster is assigned a temporary marker, based on its dominant argumentative side and function, indicating whether it is likely to participate in inference or conflict relations. The marker provides logical orientation and structural cues for subsequent RA- and CA-node construction but is not retained in the final graph.

Rule Application Node (RA-Node) Construction. RA-nodes represent the logical connections between forensic evidence, witness statements, and judicial conclusions. Using the validated I-node set, GPT-5 identifies supporting reasoning patterns and annotates each RA-node using Walton's argumentation schemes [20]. The process employs seventeen schemes, including argument from expert opinion, sign, and circumstantial evidence. To prioritize coherent reasoning over fragmented chains, the construction process groups multiple supporting statements into a single RA-node when the court evaluates them jointly.

Conflict Application Node (CA-Node) Construction. CA-nodes represent rebuttal relations where the court weakens or contradicts defendant-side claims and alternative explanations. The framework prioritizes a court-rebuttal orientation, positioning judicial conclusions as attacking elements against defendant arguments to facilitate the retrieval of court-endorsed rebuttal patterns.

Preference Application Node (PA-Node) Construction. PA-nodes capture instances where a court explicitly evaluates one interpretation as more persuasive than another. Unlike CA-nodes, PA-nodes are only created when the judgment contains an evaluative comparison of credibility, reliability, or probative strength. In this domain, these nodes typically represent judicial preferences regarding *forensic reliability, alternative explanations for DNA transfer, and comparative evidentiary* strength between prosecution-side forensic findings and defendant-side alternative accounts.

3.2.1. AIF-Based Reconstruction of Judicial Rebuttal

Figure 2 illustrates an AIF-based argument structure reconstructed from a sexual assault case to demonstrate the framework's decomposition of judicial reasoning. The graph maps the court's rejection of the defendant's claim (I_1, I_2) regarding the nature of the physical contact. The court's conclusion (I_6) is supported by the victim's firsthand testimony (RA_1) and forensic DNA evidence (RA_2). Conflict relations are captured through CA-nodes, representing the explicit judicial rejection of the defendant's alternative accounts, while a PA-node encodes the court's preference for the prosecution's interpretation based on its higher probative value. By anchoring this structure to the *Defendant Denial of Forensic Fact* issue category, the framework transforms complex reasoning into a reusable pattern of support, conflict, and preference relations for cross-case retrieval.

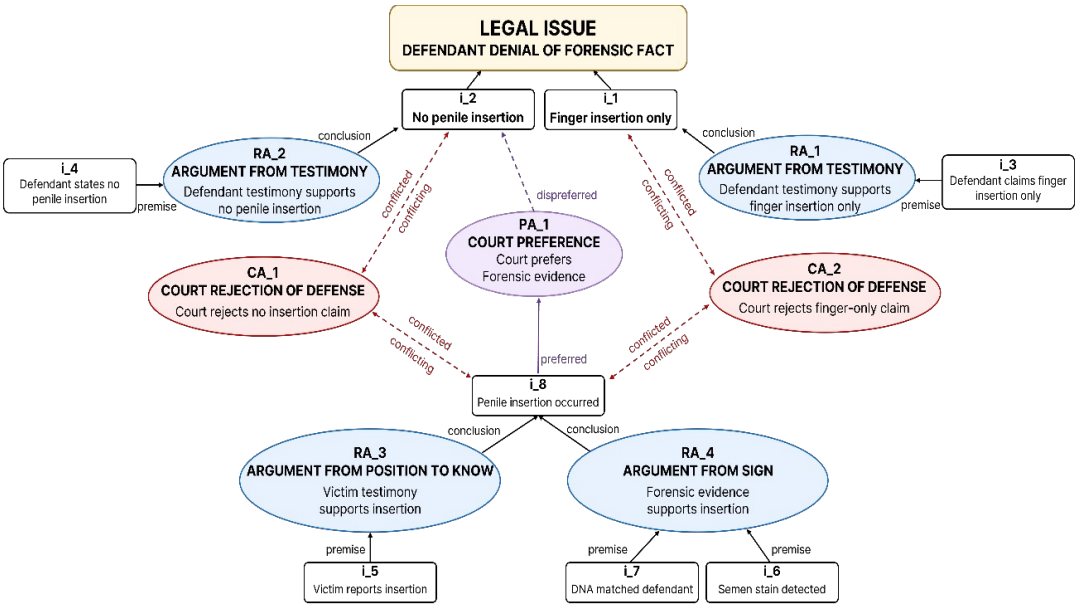

Figure 2: Example of an AIF-based judicial rebuttal structure in a DNA evidence case

3.3. Graph Database Construction

The constructed AIF graphs are stored in a Neo4j database to support graph traversal and issue-based retrieval, with the full database schema and a visualization of an issue-centered subgraph provided in Appendix B.

4. Rebuttal Generation

4.1. Test Data and Input Construction

The rebuttal generation experiment uses test arguments automatically parsed from raw AIF graphs, where each graph represents a single judicial case. Among 30 processed graphs, 3 containing no observed attacks were excluded. The remaining 27 cases produced 36 input arguments and 45 gold attack structures because a single case may contain multiple defendant claims challenged by the court.

The parser groups CA-nodes by target node to define input instances and reconstructs the associated attacked structures. When the target participates in an RA chain, the related premise, RA scheme, and conclusion are included. For each CA-node, the parser identifies the court-side attacking source node and attaches its supporting RA structure when available. Gold attacks are therefore represented as structured units consisting of premises, inference schemes, and conclusions.

Inputs additionally include manually curated evidence lists containing forensic results, collection records, and witness statements. Legal conclusions and court findings are excluded to prevent leakage of the court's final determination, leaving 91 evidence items across 36 inputs (2.5 per input on average).

4.2. GraphRAG-Based Attack Generation

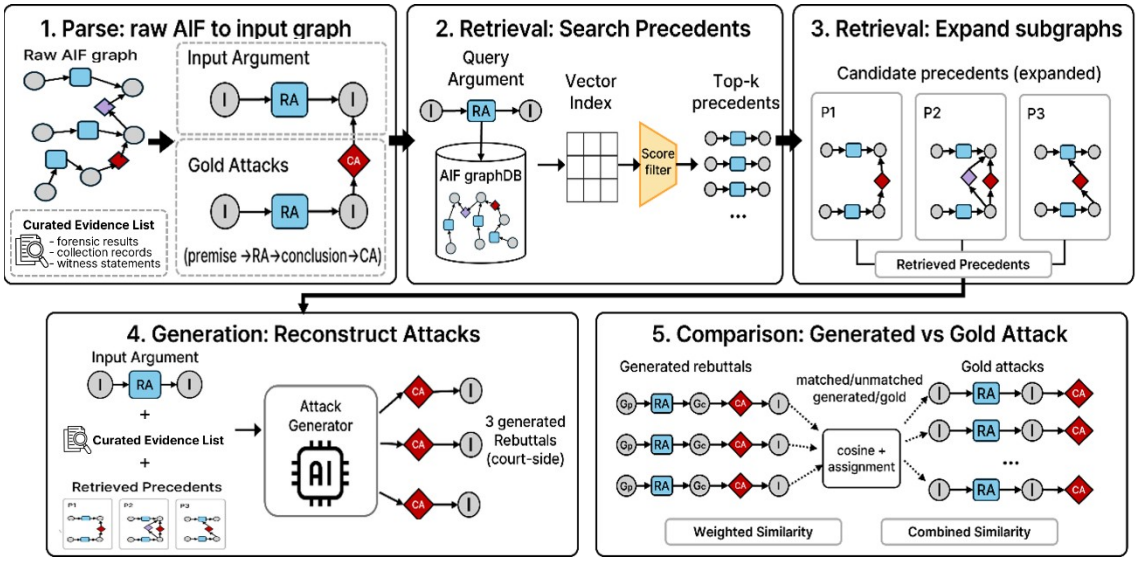

Figure 3: Overview of the Proposed GraphRAG Counterargument Generation Framework

The GraphRAG pipeline matches an input argument against a Neo4j-backed AIF graph of prior cases to retrieve structurally similar precedent arguments and their associated CA attack structures. The input argument, retrieved structures, and manually curated evidence are then provided to the language models, which generate three court-side rebuttals for comparison with the gold attacks. Retrieval operates over argument units consisting of a premise I-node, an RA-node, and a conclusion I-node. The input argument is embedded using *text-embedding-3-small* and used to query a vector index containing embeddings of precedent argument units. Candidate units are ranked using premise similarity (0.35), conclusion similarity (0.35), RA-scheme match (0.10), evidence-type match (0.10), and legal-issue match (0.10), and the top three candidate precedents are retained.

Each retrieved precedent is expanded by attaching its associated CA-nodes and reconstructing the attacker's supporting RA chain up to a depth of one. PA relations linking the precedent and attacking structures are also retrieved as context, although the generator does not create new PA-nodes. Generation is restricted to rebuttal reconstruction. For standalone claims without an RA premise, the conclusion text is reused in the premise field to maintain a uniform retrieval format. Since the same text contributes to both premise and conclusion similarity, this representation may inflate retrieval scores for standalone claims. These scores should therefore be interpreted cautiously, and alternative treatments of missing premises remain a limitation of the current design.

The prompt requires three types of court-side rebuttal: (1) a substantive rebuttal that directly contradicts the defendant's factual claim through a judicial finding, (2) a credibility or procedural rebuttal that challenges the reliability of the defendant's account or evidentiary objection, and (3) an evidentiary rebuttal that relies on forensic or testimonial evidence to undermine the target claim. These categories reflect recurring forms of court-side response observed in the corpus and are intended to encourage functional diversity in the attacks generated. They are generation constraints rather than AIF node types and do not necessarily

correspond one-to-one with the attacks in each gold graph. Three models were evaluated: gpt-5-mini, gpt-oss-120b, and llama-4-maverick. Retrieval settings remained fixed, and each model was tested over five runs.

4.3. Evaluation Protocol

Generated rebuttals are evaluated at the level of individual attack structures against the gold attacks associated with the same input claim. For each input, evaluation is performed with similarity matrix between generated and gold attacks and computes a maximum-weight one-to-one assignment using the Hungarian algorithm. This prevents multiple generated attacks from matching the same gold attack. Since all generated and gold attacks are treated as rebuttals, attack-type labels are not used as matching constraints.

A generated attack is considered matched only if its assigned gold attack reaches a weighted similarity threshold of 0.6. Assigned pairs below this threshold are treated as unmatched. Unassigned generated attacks and gold attacks are recorded as *unmatched generated attacks* and *unmatched gold attacks,* respectively. Two cosine similarity scores are used. The weighted score, which serves as the matching key, is defined as:

$$S_{\text{weighted}}(g, r) = 0.7 \cdot \cos(c_g, c_r) + 0.3 \cdot \cos(p_g, p_r)$$

where g denotes the generated attack and r denotes the gold attacks, c_g and c_r are their attacking conclusions, and p_g and p_r are their premises.

Conclusions receive greater weight because they represent the primary rebuttal stance, while premises capture evidential overlap. If a premise is missing, the score falls back to conclusion-only similarity. An additional combined premise–conclusion similarity is reported for diagnostic purposes but does not affect matching. Precision is calculated as the proportion of generated attacks that receive an eligible match, whereas recall is the proportion of gold attacks that are matched. Mean matched similarity is computed only over matched pairs. The evaluation therefore measures fidelity to observed judicial rebuttal structures rather than the independent legal validity of generated arguments.

4.4. Results and Case Analysis

Table 2 summarizes aggregate results over five runs per model. Each run generates 108 rebuttals from 36 inputs, which are compared against 45 gold attacks. Since all attacks are treated as rebuttals, the table reports only matching and similarity metrics.

Table 2: Aggregate performance of rebuttal generation across five runs per model.

Model	Matched Attacks	Precision	Recall	Weighted Similarity	Combined Similarity
gpt-5-mini	34.8±0.8	0.322 ± 0.008	0.773 ± 0.019	0.753 ± 0.003	0.808 ± 0.009
gpt-oss-120B	34.8±1.1	0.322±0.010	0.773± 0.024	0.758± 0.009	0.803±0.007
llama-4-maverick	34.6±0.9	0.320±0.008	0.769 ± 0.020	0.744±0.005	0.790±0.008

The results are stable across the three generation models. gpt-5-mini and gpt-oss-120b match 34.8 of the 45 gold attacks on average, while llama-4-maverick matches 34.6. Weighted similarities are also comparable across models (0.744–0.758), indicating no clear performance ranking. Precision is bounded by the fixed generation setting. Each run produces 108 rebuttals across 36 inputs, whereas the evaluation set contains 45 gold attacks available for one-to-one matching. Accordingly, at most 45 generated rebuttals can be counted as matches, placing the theoretical upper bound on precision at 45/108 (0.417). The observed precision of approximately 0.32 reflects both unmatched gold attacks and the fixed requirement to generate three rebuttals for each input. Across models, an average of approximately 10–11 gold attacks remained unmatched per run.

The 2016Gohap36 case illustrates the complete evaluation process. The input consisted of the defendant's claim that only finger insertion occurred and curated evidence showing a semen stain matching the defendant's DNA. The pipeline retrieved related precedent structures and generated three court-side rebuttals. The Hungarian assignment paired one generated rebuttal with the corresponding gold attack at a weighted similarity of 0.79. Since this exceeded the 0.6 threshold, the pair was counted as a successful match.

The main errors arise from mismatches between generated rebuttals and judicial reasoning structures. One recurring failure involves divergence between factual act-level claims and charge-level legal characterization. In the 2017Gohap567 case from the Gwangju District Court, the models generated rebuttals centered on finger insertion because both the claim and evidence focused on DNA recovered from the defendant's hands. However, the gold attack framed the conclusion at the higher legal level of attempted rape, causing the weighted similarity to fall below the matching threshold. Although the generated and gold attacks were semantically related, they differed in argumentative level, illustrating why semantic similarity alone does not ensure reuse of the relevant rebuttal structure. Additional failure cases involving missing evidence grounding and omitted explanatory details are discussed in Appendix C.

Overall, the GraphRAG pipeline reconstructs many court-side rebuttals grounded in forensic and testimonial evidence. Remaining errors are concentrated in charge-level legal characterization, defensive evidentiary reasoning, and cases with insufficient curated evidence.

5. Conclusion

This study proposed a structure-aware framework for legal counterargument generation in DNA evidence cases by retrieving and reconstructing judicial rebuttal structures represented as AIF-based argument graphs. Unlike conventional legal RAG approaches that rely mainly on semantic similarity, the proposed framework retrieves reusable rebuttal patterns through inference, conflict, and preference relations within a Neo4j-based GraphRAG architecture. By organizing judicial reasoning around DNA evidence issue types, the system enables cross-case retrieval and reconstruction of court-side rebuttal structures for new defendant claims.

Experimental results showed that the framework can reconstruct many observed judicial rebuttals, particularly when gold attacks are grounded in direct forensic or testimonial evidence. Performance remained relatively stable across the three generation models, suggesting that structural retrieval contributes more strongly to performance than model-specific variation. The analysis also identified recurring limitations in charge-level legal characterization, defensive evidentiary reasoning, and cases with insufficient factual grounding.

The study is limited by the relatively small dataset size and uneven issue coverage across DNA evidence categories. Expanding the dataset and incorporating more fine-grained forensic issue types will improve the diversity and generalizability of retrievable rebuttal structures. Future work should investigate richer structural retrieval methods, more explicit modeling of implicit judicial reasoning, and stronger integration between evidentiary facts and argument structures.

Declaration on Generative AI

During the preparation of this work, the authors used gpt-5.5 in order to translate Korean court decisions cited in this manuscript into English and to check grammar and spelling. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] P. M. Dung, On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games, Artificial Intelligence 77 (1995) 321–357.

[2] K. D. Ashley, Modelling legal argument: Reasoning with cases and hypotheticals, MIT Press, Cambridge, MA, 1989.

[3] Y. Gu, M. S. Jun, R.-S. Park, Research on crime investigation verification methods through natural language processing-based analysis of court judgments, Journal of Police & Law 21 (2023) 131–162.

[4] I. Habernal, D. Faber, N. Recchia, S. Bretthauer, I. Gurevych, I. Spiecker genannt Döhmann, C. Burchard, Mining legal arguments in court decisions, Artificial Intelligence and Law 32 (2024) 1–38.

[5] G. Grundler, P. Santin, A. Galassi, F. Galli, F. Godano, F. Lagioia, P. Torroni, Detecting arguments in CJEU decisions on fiscal state aid, in: Proceedings of the 9th Workshop on Argument Mining, 2022, pp. 143–157.

[6] Park, S., Choi, A., & Park, R., Objection, your honor!: an LLM-driven approach for generating Korean criminal case counterarguments. *Artificial Intelligence and Law* (2025) 1-47.

[7] K. M. Turman, Understanding DNA Evidence: A Guide for Victim Service Providers, National Institute of Justice, Washington, DC, 2001.

[8] M. Wójcik et al., Current developments in forensic interpretation of mixed DNA profiles, Forensic Science International: Genetics 13 (2014) 1–10.

[9] A. A. Mecikalski, J. M. Golding, K. C. Burke, J. S. Neuschatz, Legal decision-making in an adult rape case involving DNA evidence, Violence Against Women 31 (2025) 1932–1953.

[10] C. Chesnevar, S. Modgil, I. Rahwan, C. Reed, G. Simari, M. South, S. Willmott, Towards an argument interchange format, Knowledge Engineering Review 21 (2006) 293–316.

[11] Argumentation Research Group, The Argument Interchange Format (AIF) Specification, School of Computing, University of Dundee, Dundee, UK, 2011.

[12] I. Rahwan, C. Reed, F. Zablith, On building argumentation schemes using the argument interchange format, in: Proceedings of the 7th International Workshop on Computational Models of Natural Argument (CMNA VII), 2007.
[13] I. Rahwan, C. Reed, The argument interchange format, in: I. Rahwan, G. Simari (Eds.), Argumentation in Artificial Intelligence, Springer, Boston, MA, 2009, pp. 383–402.
[14] F. Bex, S. Modgil, H. Prakken, C. Reed, On logical specifications of the argument interchange format, Journal of Logic and Computation 23 (2013) 951–989.
[15] Mumford, J., Atkinson, K., & Bench-Capon, T. (2023, June). Combining a legal knowledge model with machine learning for reasoning with legal cases. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law* (pp. 167-176).
[16] Schraagen, M. P., Bex, F. J., Odekerken, D., & Testerink, B. J. G. (2018). Argumentation-driven information extraction for online crime reports. In *International workshop on legal data analysis and mining (LeDAM 2018): CEUR workshop proceedings*.
[17] R. Kondo, R. Matsuoka, T. Yoshida, K. Yamasawa, R. Hisano, Capturing legal reasoning paths from facts to law in court judgments using knowledge graphs, in: Proceedings of the 13th Knowledge Capture Conference, 2025, pp. 103–110.
[18] H. Han, Y. Wang, H. Shomer, K. Guo, J. Ding, Y. Lei, J. Tang, Retrieval-augmented generation with graphs (GraphRAG), arXiv preprint arXiv:2501.00309, 2025.
[19] H. de Martim, Graph RAG for legal norms: A hierarchical and temporal approach, arXiv preprint arXiv:2505.00039, 2025.
[20] Walton, D. *Argumentation schemes for presumptive reasoning*. (2013) Routledge.

A. DNA evidence Legal Issue Taxonomy

Issue Category	Issue Type	Description
DNA Result Issue	Dna match	Whether detected DNA matches the defendant's DNA profile.
	Dna mixture	Whether the sample contains DNA from multiple contributors.
	Dna statistical weight	Whether the DNA result has sufficient probative statistical weight.
	Dna partial or degraded profile	Whether a partial or degraded DNA profile can support the inference.
DNA Reliability Issue	Dna contamination	Whether contamination may have made the DNA result unreliable.
	Chain of custody or collection reliability	Whether the sample was properly collected, preserved, and handled.
DNA Interpretation Issue	Dna secondary transfer	Whether DNA was deposited indirectly rather than through direct contact.
	Dna location relevance	Whether the location of DNA detection supports the alleged fact.
	Dna timing	Whether DNA was deposited at the time of the alleged incident.
	Dna mixed profile interpretation	How a mixed profile should be interpreted in relation to the defendant.
	Dna biological material type	Whether the biological source of DNA affects its legal significance.
	Dna absence or	What legal meaning follows from the absence or

	negative result	non-detection of DNA.
Combined Forensic and Non-Forensic Issue	Dna consent or nonconsent relevance	Whether DNA evidence supports an inference about consent or non-consent.
	Dna defendant explanation	Whether the defendant's factual explanation accounts for the DNA evidence.
	Defendant denial of forensic fact	Whether the defendant denies a fact directly supported by forensic evidence.
	Other forensic dna issue	Other DNA-related issues not covered by the above categories.

B. DNA Evidence Argument Graph Database Schema

To prevent identifier conflicts, all node IDs are prefixed with a unique case identifier during ingestion. The schema maps AIF components to specific node labels (I-Node, RA-Node, CA-Node, PA-Node) and preserves metadata such as node type and case ID as properties. For information nodes, vector properties are generated using the *text-embedding-3-small* model to enable semantic similarity matching. Additionally, Evidence and LegalIssue nodes are created and linked to I-Nodes via SUPPORTS and OF_ISSUE relations, allowing retrieval to be explicitly constrained by forensic-specific issue categories.

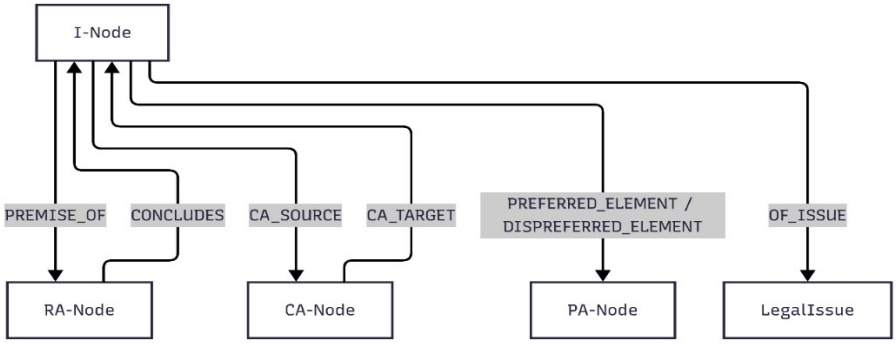

Figure B1: Neo4j Graph Database Schema for AIF-based Argument Representation.

Figure B2 illustrates an issue-centered subgraph where a LegalIssue node acts as a shared anchor for various judicial decisions. Relations are converted into typed edges: PREMISE_OF and CONCLUDES define inference chains, while CA_SOURCE and CA_TARGET identify attack structures. Preference relations are similarly mapped using PREFERRED_ELEMENT and DISPREFERRED_ELEMENT edges through PA-Nodes. This normalized structure enables the system to traverse the graph and recover reusable rebuttal patterns (connecting defendant claims, judicial conclusions, and forensic reasoning) across different cases sharing the same legal issue.

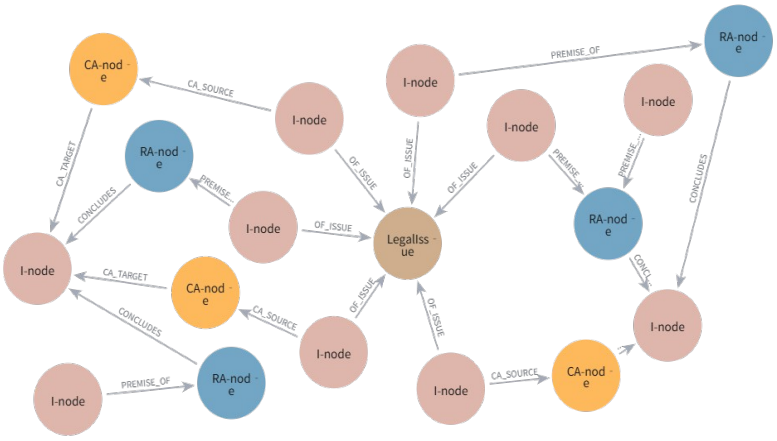

Figure B2: Example of an issue-centered Neo4j subgraph for the legal issue *Defendant Denial of Forensic Fact*

C. Additional Case Studies

The following cases illustrate secondary failure patterns that were excluded from the main discussion because they primarily concern evidence sparsity and omitted explanatory details rather than core structural mismatches in rebuttal reconstruction.

[2018Gohap239, Suwon District Court] Defensive Evidentiary Reasoning
A failure pattern appears in cases where the court's reasoning includes defensive caveats about evidential limitations. In the 2018Gohap239 case from the Seongnam Branch of the Suwon District Court, the defendant denied rape, and the gold graph contains four attacks for this single target. One gold attack states that the offense is sufficiently established. The other three explain why limitations in the genetic testing do not undermine that conclusion, including the variability of genetic test results, the delayed submission of the victim's underwear and related items, and the finding that these circumstances do not prevent recognition of the offense. The generated attacks are relevant but simpler. They state that the defendant committed rape, that the victim's account is credible, or that forensic results corroborate sexual contact. These attacks address the main dispute, but they do not reproduce the court's auxiliary reasoning about why weak or delayed DNA evidence does not defeat the charge. This is not only a similarity-threshold issue. It reflects a mismatch between the prompt's three requested styles and the actual structure of the court's reasoning in this case.

[2020Gohap71, Daejeon District Court] Missing Evidence Grounding

In the 2020Gohap71 case from the Daejeon District Court, the defendant claimed that the intercourse was consensual, but the curated evidence list is empty because the candidate evidence in the raw graph was evaluative rather than objective. The gold attacks include charge-level and evidentiary findings: the defendant committed rape by violence and intimidation, the victim's statements do not conflict with genetic testing results, and genetic testing results were obtained. Without curated factual grounding, the models rely more heavily on retrieved precedent patterns. This produces unstable behavior. gpt-5-mini and llama-4-maverick

sometimes use precedent-like wording such as taking advantage of the victim's inability to resist, which can approach the charge-level gold attack. gpt-oss-120b more often produces a literal negation of consent, such as stating that the intercourse was not consensual. That formulation is topically correct, but it remains too far from the gold's charge-level wording to pass the threshold. This case indicates that evidence curation affects not only factual grounding but also the level at which the generated attack is formulated.

[2019Gohap12, Daegu District Court Western Branch] Missing Explanation Detail]

A related issue arises when the gold attack explains why missing DNA evidence does not support the defendant's denial. In the 2019Gohap12 case from the Western Branch of the Daegu District Court, the gold attack states that the absence of the defendant's DNA from the victim's body and underwear is not enough to conclude that the defendant did not commit the crime. gpt-5-mini and gpt-oss-120b usually match this attack because they preserve the specific explanatory fact that the victim washed afterward. llama-4-maverick sometimes gives a more general rebuttal, stating that the absence of DNA does not negate the crime and that other evidence supports the victim's account. This is legally and semantically close, but it is less stable under the embedding-based threshold because it drops the concrete explanation that connects the forensic absence to the court's reasoning. The case illustrates why defensive evidentiary findings require more than the correct argumentative direction. They require the specific fact that explains why an apparent evidential weakness is not decisive.