ARGUS: Role-Aware Event Knowledge Graphs for U.S. Employment-Discrimination Complaints

Sriram Kannan[1] **Swetha Saseendran**[1] **Vishnu Vardhan Reddy Kandi**[1]
Leslie Barrett[2] Madhavan Seshadri[2] Enrico Santus[2]
[1]University of Massachusetts Amherst
[2]Bloomberg
{sriramkannan, ssaseendran, vkandi}@umass.edu
{lbarrett4, mseshadri, esantus}@bloomberg.net

Abstract

U.S. employment-discrimination complaints describe complex event sequences that are not explicitly captured by lexical or embedding-based representations alone. We present ARGUS, a source-grounded pipeline that combines a 5W1H-inspired schema, legal-domain models, and LLM-based structured generation to construct document-level Event Knowledge Graphs (EKGs) from CourtListener complaints. ARGUS extracts fact-bearing statements, builds chunk-level event graphs with participant, temporal, and causal structure, and merges them into document-level representations. We evaluate graph quality through human and multi-model assessment and test downstream utility on claim classification and legal QA. The graph-structured classifier outperforms raw and linearized baselines on the held-out set, and EKG-only retrieval improves document-scoped QA, while open-retrieval gains remain limited by low first-stage candidate recall. These results suggest that EKGs are most useful for organizing and reasoning over evidence once relevant material has been retrieved.

1 Introduction

Legal practitioners often need to interpret and compare cases whose underlying events are similar even when described differently. Lexical and embedding-based retrieval methods can identify semantically relevant passages, but return largely unstructured text and do not explicitly represent *who did what to whom, when, and why*. This structure is important in legal narratives, where participant roles and temporal or causal relations can determine whether two cases are meaningfully comparable.

We study this problem in U.S. employment-discrimination complaints. We present ARGUS, a pipeline that converts complaint narratives into source-grounded Event Knowledge Graphs (EKGs). Rather than representing a case only as text or a set of entities, ARGUS models events together with their participants, legal roles, source evidence, and supported temporal and causal relations.

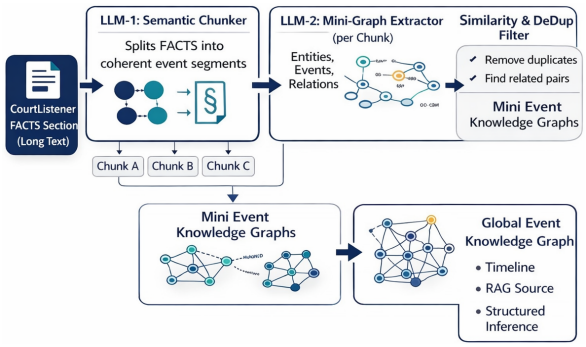


Figure 1: Multi-stage pipeline for EKG construction from federal employment-discrimination complaint narratives.

ARGUS follows a staged construction process. Starting from CourtListener complaints,[1] it identifies fact-bearing statements, groups them into semantic chunks, extracts chunk-level event graphs using a 5W1H-inspired representation (Hamborg et al., 2019), and merges them into a document-level EKG through deterministic and LLM-assisted cross-chunk resolution. Figure 1 provides an overview; implementation and evaluation details are given in Section 3 and the Appendix.

We investigate three questions: **(i)** Can the staged pipeline produce reliable, source-grounded legal event graphs? **(ii)** Do these graphs improve legal classification and document-scoped question answering relative to text-based representations? **(iii)** Do these benefits extend to external and open-retrieval legal QA settings?

We evaluate graph construction at multiple stages, including fact extraction, chunk-level graph extraction, and document-level merging, using au-

[1]https://www.courtlistener.com/

Proceedings of the Natural Legal Language Processing Workshop 2026, pages 142–155
October 28, 2026 ©2026 Association for Computational Linguistics

tomatic measures, human annotation, and multi-model judging. We then assess downstream utility through claim-type classification and document-scoped FAQ question answering, followed by bar-exam experiments testing generalization to multiple-choice QA and open retrieval. Exploratory case-clustering analyses are reported in Appendix C.

Our contributions are threefold. **(1)** We introduce a source-grounded, role-aware EKG pipeline for U.S. employment-discrimination complaints that represents events with participant, temporal, and causal structure. **(2)** We introduce a multi-stage evaluation protocol for legal event graphs, combining human annotation with complementary automatic and LLM-based evaluation.[2] **(3)** We show that the graph-structured classifier outperforms raw and linearized baselines on held-out claim classification and that EKG-based retrieval improves document-scoped QA; in open retrieval, gains remain limited when relevant evidence is absent from the first-stage candidate set.

2 Related Work

ARGUS intersects three lines of work: legal retrieval and representation, event extraction and temporal reasoning, and graph-augmented retrieval.

Legal retrieval and representation. Legal case retrieval has progressed from lexical matching toward neural interaction and representation learning (Feng et al., 2024). BERT-PLI models paragraph-level interactions between cases (Shao et al., 2020), while legal-domain models such as LEGAL-BERT (Chalkidis et al., 2020) and models trained on LeXFiles (Chalkidis et al., 2023) provide domain-adapted representations for legal NLP. The reasoning-focused legal retrieval benchmark used in our bar-exam experiments separates retrieval quality from downstream answer selection (Zheng et al., 2025).

Event extraction and temporal structure. General information-extraction systems such as DyGIE++ (Wadden et al., 2019), MetaGraph (Pedinotti et al., 2026) and OneIE (Lin et al., 2020) jointly model entities, relations, and/or events. In the legal domain, Filtz et al. (2020) extract event types, participants, and temporal information from ECHR decisions, while LexTime evaluates temporal ordering of event pairs in U.S. federal complaints (Barale et al., 2025). The 5W1H framework provides a compact representation of core event information (Hamborg et al., 2019). REGen combines exact, relaxed, and LLM-based matching for generative event-argument evaluation and validates these measures against human judgments (Sharif et al., 2025). ARGUS applies a 5W1H-inspired schema to source-grounded mini-graphs and merges them into document-level EKGs with legal participant roles and temporal and causal relations.

Legal knowledge graphs and graph-augmented retrieval. Li et al. (2024) construct a Chinese legal knowledge graph from criminal-law materials using a knowledge-enhanced LLM. More broadly, GraphRAG constructs entity-based graph representations for query-focused summarization (Edge et al., 2024), while KG^{2}RAG uses knowledge-graph relations to expand and organize retrieved chunks (Zhu et al., 2025). EventRAG builds event knowledge graphs for multi-event reasoning (Yang et al., 2025). In the legal domain, Kordjamshidi et al. (2026) investigate structured representations and their efficacy in tax-law reasoning for QA, and LegalGraphRAG (Chen et al., 2026) combines hierarchical legal graphs with multi-agent retrieval and verification. ARGUS instead focuses on source-grounded, document-level EKGs for U.S. employment-discrimination complaints, preserving legal participant roles and temporal and causal relations through staged document-level merging.

3 Method

ARGUS converts the FACTS narrative of a U.S. employment-discrimination complaint into a source-grounded, document-level Event Knowledge Graph (EKG). The pipeline has four stages: **(1)** sentence-level fact extraction, **(2)** semantic chunking, **(3)** two-stage mini-graph extraction, and **(4)** deterministic and LLM-assisted document-level merging. Figure 1 summarizes the pipeline. We evaluate each component separately to distinguish errors introduced during fact selection, event and relation extraction, and graph merging.

Fact extraction. We define facts as event-bearing statements describing concrete actions, communications, decisions, or outcomes, distinguishing them from procedural or rhetorical text. Each sentence s is classified as *Fact* or *Non-Fact* using a fine-tuned LEGAL-BERT or LexLM encoder

[2] The human annotations will be released upon publication.

(Chalkidis et al., 2020, 2023). Let $p(s)$ denote the classifier's maximum softmax probability. Predictions above a confidence threshold τ are retained directly; sentences with $p(s) < \tau$ are passed to an LLM, which returns the identifiers of sentences judged to be facts. Only sentences ultimately labeled as facts are passed to graph construction.

Semantic chunking. Retained fact sentences preserve their original identifiers and are formatted as sentence-labeled input (`[S1]`, `[S2]`, ...). An LLM groups them into coherent event sequences, topics, or narrative phases. The prompt favors boundaries associated with changes in theme, time, or actor, avoids splitting an event across chunks, requires coverage of all retained sentences, and introduces approximately two to three sentences of overlap.

Event representation. ARGUS uses a 5W1H-inspired representation (Hamborg et al., 2019) within a structured event-extraction formulation (Srivastava et al., 2025). Entity records contain an identifier, name, entity kind, canonical legal role, and optional aliases. Event records contain an identifier, event type, main verb, trigger, participants, temporal information, source evidence, and confidence. Participants reference entity identifiers and carry event-specific roles, while evidence retains supporting sentence identifiers.

Temporal information may be explicit, relative, or unknown, following the broader problem of event-pair temporal relation modeling (Ning et al., 2018). The extraction schema supports BEFORE, AFTER, OVERLAP, and SAME-TIME temporal relations and CAUSES, ENABLES, and PREVENTS causal relations. ARGUS retains relations only when supported by the source text. For graph-based classification, events and participant entities become nodes; event–entity edges represent participation, and temporal and causal relations remain event–event edges. An example extracted mini-graph is shown in Appendix A, Figure 4; the full downstream GNN architecture is provided in Appendix B.3.

Two-stage mini-graph extraction. Each chunk is processed through two structured-generation stages. Stage 1 extracts grounded entities and events and is explicitly instructed not to produce temporal or causal edges. Stage 2 receives the original chunk and the complete Stage 1 JSON output and adds only source-supported temporal and causal relations. Both stages require strict JSON, preservation of sentence identifiers, and no unsupported entities, events, dates, or relations.

The extraction configuration is optimized with GEPA (Agrawal et al., 2026). A candidate specifies the Stage 1 and Stage 2 models, prompts, retry instructions, and extraction settings. Candidates are evaluated using structural checks—including schema validity, reference integrity, temporal and causal consistency, and coverage—together with an LLM judge that scores factual accuracy, completeness, relevance, faithfulness, coherence, and overall quality. Cost is measured using token counts when available and API-call count otherwise. GEPA uses the metrics and observed failures to propose mutations and retains non-dominated configurations on a quality–cost Pareto frontier. The two-stage extractor and GEPA optimization loop are shown in Appendix Figures 2 and 3.

We also evaluate a Generative FrameNet-guided configuration (Tayyar Madabushi et al., 2025), using frame hints such as hiring, retaliation, and termination to guide event typing.

Document-level graph merging. Mini-graphs are first combined through deterministic merging. Entity identifiers are remapped into a document-level namespace, and mentions with the same lowercased, whitespace-normalized name are assigned to the same entity. Events are deduplicated using a structured key containing event type, normalized main verb, temporal value, participant entity–role pairs, and supporting sentence identifiers. Duplicate records accumulate their source references, and existing intra-chunk edges are remapped to the resulting document-level event identifiers.

ARGUS then performs LLM-assisted cross-chunk resolution. Candidate pairs are restricted to events originating from different chunks and ranked by similarity over their event descriptions and contextual evidence. For each candidate, the LLM determines whether the records describe the same occurrence and whether a supported temporal or causal relation should be added. The prompt favors precision and returns no relation when evidence is insufficient. Duplicate events may be collapsed, after which existing edges are remapped and accepted cross-chunk relations are added.

Component-level evaluation. Fact extraction is evaluated by comparing LEGAL-BERT and LexLM with a prompted GPT-5 baseline, using Fact-class F1 as the primary metric and accuracy,

precision, and recall as supporting measures. Mini-graph extraction is evaluated on a fixed 24-chunk set using automatic structural and semantic measures and human annotation. Four annotators inspect each graph with its source text and assign *Aligned*, *Partially Aligned*, or *Misaligned*, together with an error tag and explanation when applicable. We report percent agreement, Fleiss' κ, and Krippendorff's α.

Document-level construction is compared with a one-shot Claude Sonnet 4.5 baseline that generates a complete EKG directly from the full FACTS narrative; the exact prompt is provided in Appendix B.1. Three LLM judges—DeepSeek-R1, Qwen3-Coder, and Mistral-Large—compare the staged and one-shot graphs on event grounding, evidence grounding, edge correctness, and overall quality. We additionally report inter-judge agreement, Gwet's AC1 (Gwet, 2008), and stability under temperature variation (Appendix B.2, Table 11). Multiple judges complement human annotation and reduce reliance on a single evaluator (Zheng et al., 2023). U.S. holdout results are reported in Appendix B.4, Table 12.

4 Experimental Setup

4.1 Datasets

We collect 153 employment-discrimination complaints (Nature of Suit 442, Civil Employment Cases) from CourtListener,[3] an open legal-search repository maintained by the Free Law Project.

A complaint is the initiating pleading filed by a plaintiff and describes allegations rather than facts established by a court. Accordingly, the events represented by ARGUS reflect the allegations contained in the complaint and should not be interpreted as adjudicated findings.

The raw filings are OCR-derived scanned court documents with no gold labels for facts, events, or case similarity. We retrieve them through a Selenium-based pipeline and preprocess them using OCR-to-text conversion, header/footer and boilerplate removal, section detection, sentence segmentation, and normalization. Section detection isolates the FACTS narrative from jurisdictional and prayer-for-relief text.

The corpus covers five U.S. district courts to include jurisdiction-specific writing variation (Table 1): Eastern District of Pennsylvania (EAPD), Eastern District of Kentucky (KYED), Southern District of New York (NYSD), Western District of Pennsylvania (PAWD), and District of Rhode Island (RID). Documents vary substantially in length and structure. Where sentence-level annotation is complete, the fact rate ranges from 81.7% (KYED) to 93.0% (PAWD). All splits are defined at the document level to prevent boilerplate and near-duplicate leakage; sentence-level fact extraction inherits its parent document split, while mini-graph extraction and merging are inference-only.

[3] https://www.courtlistener.com/

District	Docs	Chunks	Avg/Doc	Sentences	Fact Rate
EAPD	20	87	4.35	–	–
KYED	31	138	4.45	1,155	81.7%
NYSD	46	318	6.91	4,808	92.7%
PAWD	28	165	5.89	1,607	93.0%
RID	28	118	4.21	1,631	90.9%

Table 1: Dataset statistics across districts. Fact Rate is the fraction of sentences labeled Fact; sentence-level annotation is not yet complete for EAPD.

Model	Acc.	Prec.	Rec.	F1
LEGAL-BERT	0.9400	0.8889	0.6154	0.7273
LexLM	0.9300	0.8000	0.6154	0.6957
GPT-5	0.8500	0.4615	0.9231	0.6154

Table 2: Fact-extraction results on a fixed 100-sentence U.S. gold sample. LEGAL-BERT and LexLM are fine-tuned legal-domain encoders; GPT-5 is a prompted baseline. Fact is the positive class.

4.2 Evaluation of Graph Construction

Fact extraction. We fine-tune LEGAL-BERT and LexLM for sentence-level Fact/Non-Fact classification and compare them with a prompted GPT-5 baseline on a fixed 100-sentence U.S. gold sample (Table 2). Fact-class F1 is the primary metric, with accuracy, precision, and recall also reported. LEGAL-BERT achieves the highest F1 (0.7273) and precision (0.8889), while GPT-5 achieves substantially higher recall (0.9231). This precision–recall complementarity motivates the hybrid fallback described in Section 3. On the U.S. holdout, classifier-only and hybrid modes perform similarly (F1 $\approx$ 0.87 for LEGAL-BERT and $\approx$ 0.86 for LexLM), whereas LLM-only extraction is weaker (F1 $\approx$ 0.71–0.72; see Appendix B.4, Table 12).

Mini-graph extraction. We evaluate mini-graph extraction on a fixed 24-chunk set. Claude Sonnet 4.5 + GEPA achieves a quality score of 0.896, with schema validity and coverage of 1.0; causal

Metric	GPT-OSS	Qwen
Quality Score	0.8272	0.8852
Judge Score	0.6642	0.7966
Structural Score	0.9902	0.9739
Temporal Consistency	0.9118	0.7647
Events/Chunk	7.09	7.50
Temporal Edges/Chunk	2.15	3.12
Causal Edges/Chunk	1.00	2.29
Schema Validity	1.00	1.00

Table 3: Open-source model comparison for mini-graph extraction.

Metric	Claude	Qwen
Aligned	78.30%	85.20%
Partially Aligned	17.40%	14.80%
Misaligned	4.30%	0.00%
Percent Agreement	69.60%	91.70%
Fleiss' κ	0.143	0.669
Krippendorff's α	0.153	0.673

Table 4: Human evaluation of mini-graph alignment with the source text.

self-loops are the main residual structural error. Among open-source alternatives (Table 3), Qwen achieves higher semantic quality and produces more events and relations per chunk, while GPT-OSS-120B scores slightly higher on structural and temporal-consistency measures.

For human evaluation, four team members independently review each mini-graph with its source text and assign *Aligned*, *Partially Aligned*, or *Misaligned*, plus a free-text explanation and, when applicable, an error tag such as missing event, wrong role, or hallucinated detail. We report percent agreement, Fleiss' κ, and Krippendorff's α (Table 4). The FrameNet-guided Qwen configuration reaches 85.2% Aligned and Fleiss' $\kappa = 0.669$, compared with 78.3% and $\kappa = 0.143$ for the Claude configuration. Because model choice and FrameNet guidance vary together, this is a configuration-level comparison rather than an isolated FrameNet ablation.

Document-level graph merging. We first measure how document-level merging changes graph size on five validation documents. Table 5 shows that event counts are almost entirely preserved: only one event is removed across the five cases. Entity counts decrease substantially in four cases as repeated mentions are consolidated into document-level entities. This provides a structural check before comparison with a one-shot baseline.

Case	Pre-Ev	Post-Ev	Loss	Pre-En	Post-En
Miczulski v. Alix	27	26	1	31	13
Poole v. Ampler	40	40	0	34	17
Sermarini v. Step Up	35	35	0	26	10
Bleiler v. Chester	27	27	0	24	13
Bacon v. M.A.G.	5	5	0	6	6

Table 5: Event and entity counts before and after document-level merging on five validation cases.

Case	Events	Temporal	Causal
Miczulski v. Alix Inc	21	19	6
Poole v. Ampler Pizza	0	0	0
Sermarini v. A Step Up	16	14	20
Bleiler v. Chester Co.	0	0	0
Bacon v. M.A.G. Ent.	17	16	13

Table 6: One-shot LLM baseline event-graph extraction on the five merger-validation documents.

We then compare ARGUS with a one-shot baseline in which Claude Sonnet 4.5, accessed through an OpenAI-compatible hosted API, generates a complete document-level EKG directly from the full FACTS section, without semantic chunking, intermediate mini-graphs, or staged merging. The exact prompt is provided in Appendix B.1. The baseline extracts no events for two of five validation documents and varies substantially in coverage on the others (Table 6), demonstrating the difficulty of performing event extraction and relational reasoning over the complete narrative in one generation.

Three LLM judges—DeepSeek-R1, Qwen3-Coder, and Mistral-Large—compare the one-shot and staged graphs on event grounding, evidence grounding, edge correctness, and overall quality. ARGUS is preferred on evidence grounding, edge correctness, and overall quality in all comparisons and on event grounding in 93.3% of comparisons (all $p = 0.031$; Table 7). Intra-model stability under temperature variation is 100% for Qwen and Mistral and 80% for DeepSeek-R1 (Appendix B.2, Table 11).

5 Experiments on Downstream Tasks

5.1 Classification

Setup. We formulate claim-type prediction as multi-label classification over 23 concepts from the SALI Alliance's Legal Matter Specification Stan-

Dimension	Win	Agree.	AC1	p
Event Grounding	93.3%	86.7%	0.848	0.031
Evidence Grounding	100%	100%	1.0	0.031
Edge Correctness	100%	100%	1.0	0.031
Overall	100%	100%	1.0	0.031

Table 7: Multi-model evaluation of ARGUS relative to the one-shot baseline.

Representation	Micro-F1@0.5
Raw-text LEGAL-BERT	0.5897
Linearized-EKG LEGAL-BERT	0.5833
EKG-GNN	**0.6415**

Table 8: SALI claim-type classification on the 22-document held-out test set.

dard (LMSS)[4] across 130 complaints (176 label assignments), using LLM-generated silver labels reviewed by humans. We compare three representations over the same label space: the preprocessed complaint text encoded from a 512-token input, a linearized entity/event EKG, and a graph-structured EKG using participation, temporal, and causal edges. Raw text provides a text baseline; the linearized EKG tests whether graph topology adds value beyond serialization of the extracted graph.

Results. On the 22-document held-out test set (Micro-F1@0.5; Table 8), the raw-text and linearized-EKG LEGAL-BERT models obtain 0.590 and 0.583, respectively, while the graph-structured EKG-GNN reaches 0.642. This is an absolute gain of 0.052 Micro-F1, or 8.8% relative to the raw-text baseline. The graph-structured representation therefore performs best on this held-out set.

GNN architecture. The EKG-GNN represents participant entities and events as nodes, with event–entity participation edges and event–event temporal and causal edges. Each node's natural-language description is encoded with LEGAL-BERT (Chalkidis et al., 2020) and projected to 256 dimensions. Two mean-neighbor graph-convolution layers propagate information over the document graph:

$$h_i^{(l+1)} = \mathrm{ReLU}\left(W_s h_i^{(l)} + W_n \frac{1}{|\mathcal{N}(i)|} \sum_{j \in \mathcal{N}(i)} h_j^{(l)} \right).$$

[4]The LMSS is available at https://sali.org/explore-the-standard/.

Each layer is followed by dropout ($p = 0.1$). Self-loops are excluded from the neighbor average because the node's own representation is modeled separately through W_s. After two layers, node representations are mean-pooled and passed to a 23-label classification head. The reported `ft-BERT` configuration jointly fine-tunes LEGAL-BERT and the graph layers using weighted binary cross-entropy and AdamW. Full architecture and training settings are provided in Appendix B.3.

5.2 Clustering vs. Distance

Setup. Each merged EKG is reduced to Core6D: log-counts of entities, events, temporal edges, and causal edges, plus graph density and temporal-edge ratio. The vectors are L2-normalized and clustered with K-means. We use the silhouette score to compare graph-derived representations with a TF-IDF bag-of-words baseline and to evaluate alternative values of K.

Core6D reaches a silhouette score of 0.649 at $K = 4$, compared with 0.091 for TF-IDF, 0.363 for the full high-dimensional EKG representation, and 0.465 for Legacy7D. Although the $K = 2$–40 sweep favors very small K, we use $K = 4$ to retain stronger granularity with high cohesion. Compact structural summaries therefore produce substantially cleaner cluster geometry than TF-IDF or the unreduced EKG representation at this corpus size. Full results are reported in Appendix C, Table 13.

Feature ablations. We compare three Core6D-based configurations using distance to the assigned centroid and an independent LLM-judged cluster-membership score. Run A combines SALI claim tags with Core6D and produces one dominant cluster (111/14/6/2), with Pearson $r \approx -0.06$ between centroid distance and judged membership. Run B augments Core6D with a 16-dimensional PCA-compressed EKG-context embedding, producing more balanced clusters, with Pearson $r \approx -0.28$ and Spearman $\rho \approx -0.31$. Run C adds SALI tags and a kNN graph-neighbor block to Run B, reaching Pearson $r \approx -0.29$ and Spearman $\rho \approx -0.37$, without clearly outperforming Run B.

The pattern indicates that continuous EKG-derived context contributes more clustering signal than SALI claim tags alone, while the additional Run C components provide limited further benefit. Validation uses geometric and LLM-judged membership diagnostics rather than attorney relevance judgments. Detailed feature definitions, ablations,

and plots are provided in Appendix C (Tables 14–15 and Figures 5–7).

5.3 Document-Scoped Question Answering

Setup. We build a 300-question FAQ benchmark over 10 CourtListener employment-discrimination complaints (30 questions per document), spanning four graph-relevant question types: *temporal order* (what occurred before or after an event), *causal/enablement* (what allegedly caused or enabled an event), *temporal overlap* (whether two events were concurrent), and *party resolution* (who did what to whom), plus a small residual temporal-relation category. Each question has a gold answer grounded in its source complaint.

Holding GPT-4o and the answering prompt fixed, we compare three retrieval conditions:

1. **Raw document** — the FACTS section is chunked and retrieved directly, without graph structure.
2. **Hybrid text + EKG** — linearized document-level EKG context is combined with retrieved raw-text context.
3. **EKG only** — the linearized EKG is used without raw-text context.

Answer quality is measured with Token-F1 against the gold answer. Because each question is paired with its source complaint, this is a within-document retrieval experiment: it evaluates representation and reasoning after the relevant document is known, not first-stage retrieval across CourtListener.

Results. EKG-only retrieval achieves 0.446 mean Token-F1, compared with 0.323 for raw text and 0.363 for hybrid text + EKG (Table 9), absolute gains of 0.123 and 0.083, respectively. Hybrid improves over raw text by 0.040, but adding raw context reduces performance relative to EKG alone. This pattern is consistent with additional unstructured context introducing noise, although the experiment does not isolate that mechanism.

Results by question type. The EKG gain over raw text is largest for *temporal order* (0.381 $\rightarrow$ 0.538, $+0.157$) and *causal/enablement* (0.387 $\rightarrow$ 0.522, $+0.134$), categories directly aligned with graph relations. Temporal-overlap questions also improve substantially (0.241 $\rightarrow$ 0.384, $+0.143$). The smallest improvement is for party resolution (0.152 $\rightarrow$ 0.197, $+0.045$), where participant names and mentions are already accessible in source text. Full question-type results are reported in Appendix D.1, Table 16.

Context condition	Mean Token-F1
Raw document	0.323
Hybrid text + EKG	0.363
EKG only	**0.446**

Table 9: Document-scoped FAQ answer quality over 300 questions from 10 CourtListener complaints (GPT-4o).

Significance of RAG gains. We test the $n = 299$ questions scored under all three conditions; one failed hybrid LLM call is excluded from all three arms to preserve pairing. We use a paired bootstrap 95% confidence interval over the mean F1 difference (10,000 resamples), Wilcoxon signed-rank test, and sign-flip paired permutation test (10,000 permutations).

All three tests agree in direction and significance. EKG-only exceeds raw text by $+0.123$ F1 ($p < 10^{-30}$) and hybrid by $+0.083$ ($p < 10^{-20}$), while hybrid exceeds raw text by $+0.040$ ($p < 10^{-10}$). Bootstrap confidence intervals and exact Wilcoxon results are reported in Appendix D.1, Table 17.

Per-type tests show significant EKG-versus-raw improvements for every category with $n \geq 12$: temporal order ($p < 10^{-17}$), causal/enablement ($p < 10^{-15}$), temporal overlap ($p = 2.4 \times 10^{-3}$), and party resolution ($p = 8.0 \times 10^{-5}$). The residual temporal-relation category contains three questions and is not significant ($p = 0.25$); full results are reported in Appendix D.1, Table 18.

6 Generalization to External Datasets

Multiple-Choice Legal QA (Bar Exam). To test generalization beyond the employment-complaint FAQ benchmark, we evaluate on Multistate Bar Examination (MBE) multiple-choice questions (RegLab, Stanford University, 2024). We report two experiments with different retrieval setups and corpus scales; they are evaluated separately.

6.1 Experiment A: Retrieval-Condition Comparison

Setup. We evaluate 117 MBE questions, each containing a fact pattern, four answer choices, and

a gold supporting passage drawn from 100 short appellate opinion excerpts. Unlike the FAQ benchmark, the task selects an answer choice (A–D) rather than generating free text, and its passages contain legal holdings and reasoning rather than narrative complaints. We compare raw document, chunked text, and EKG graph retrieval (top 3 passages each), holding Claude Sonnet 5 fixed.

Results. Raw document retrieval reaches 85.47% accuracy, while chunked text and EKG retrieval both reach 86.32%, a 0.85 percentage-point gain (Table 19). Among the 100 questions with an available gold-source EKG, EKG accuracy is 86.00%.

Significance. We use exact McNemar tests, paired bootstrap 95% confidence intervals (10,000 resamples), and Wilcoxon signed-rank tests. No pairwise comparison is significant: all McNemar p-values exceed 0.68 and all bootstrap intervals include zero. Only 6–9 outcomes differ between any pair of conditions. Full statistics are reported in Appendix D.2, Table 20.

Interpretation. The lack of a significant retrieval-condition difference contrasts with the larger gains on document-scoped FAQ (Section 5.3), where improvements are strongest for temporal order, causality, and enablement. Experiment A instead yields nearly identical chunked-text and EKG performance. EKG benefits therefore depend on the information required by the task rather than holding uniformly across legal QA.

6.2 Experiment B: Open-Retrieval and Oracle-Document Analysis

Setup. We evaluate 100 MBE-style multiple-choice questions from the `legal_bench` subset, retrieving from the `reglab/barexam_qa` corpus of 856,835 legal passages. Claude Sonnet 4.5 selects the final answer (A–D). We first measure open-retrieval Recall@10 and then downstream multiple-choice accuracy. BM25 and E5-large-v2 search the full corpus directly; hybrid Graph-RAG reranks the BM25 top-1,000 candidate set using graph-relevance scores where available.

Retrieval results. BM25 and E5 retrieve the gold passage in the top ten for 1% of questions, while hybrid Graph-RAG reaches 8% Recall@10 (Table 10). The gold passage appears in the BM25 top-1,000 pool for eight questions, and graph reranking moves all eight into the final top ten. The Recall@10 gain therefore occurs when first-stage retrieval has already placed the relevant passage in the candidate pool; graph reranking cannot recover passages absent from that pool.

Retrieval method	Recall@10
BM25	0.0100
E5-large-v2	0.0100
Hybrid Graph-RAG	**0.0800**

Table 10: Open-retrieval performance on the 100-question Bar Exam QA subset.

Downstream accuracy. BM25-RAG achieves 73% accuracy and open Graph-RAG 72%, both below the 76% no-passage baseline (Table 21). The gold-passage condition reaches 79%, showing that relevant evidence improves answer selection. In the open setting, however, the gold passage is absent from most retrieved contexts, so first-stage retrieval errors limit the downstream value of graph reranking.

Oracle-document setting. We additionally evaluate Graph-RAG when the relevant source document is known. The original configuration reaches 75% accuracy but produces empty graph retrievals for 18 questions. A revised configuration adds document-scoped retrieval, an MCQ-specific prompt, and hybrid context combining the gold passage with graph evidence, reaching 82%. Because these changes are introduced together, the 82% result is a configuration-level comparison rather than a graph-only ablation.

Interpretation. These experiments separate first-stage retrieval from downstream use of structured evidence. Graph reranking raises open Recall@10 from 1% to 8%, but relevant passages enter the BM25 top-1,000 candidate pool for only eight of 100 questions, limiting downstream gains. In the oracle-document setting, the revised configuration reaches 82%, but it changes retrieval scope, prompting, and context together. The results support EKG structure for organizing and reasoning over already-relevant evidence rather than replacing first-stage full-corpus retrieval.

7 Conclusion

We presented ARGUS, a staged pipeline for converting U.S. employment-discrimination complaints into source-grounded Event Knowledge

Graphs. The strongest gains occur when tasks exploit participant, temporal, or causal structure: the graph-structured classifier achieves higher held-out Micro-F1 than raw and linearized baselines, and EKG-only document-scoped retrieval improves FAQ answer quality. The bar-exam experiments provide a boundary condition: Experiment A shows no significant retrieval-condition difference, while open-retrieval gains occur only when relevant passages enter the first-stage candidate pool. ARGUS therefore supports organizing and reasoning over relevant legal evidence rather than replacing first-stage retrieval.

Limitations

The corpus contains 153 U.S. federal employment-discrimination complaints from five districts and does not establish generalization to other claims, jurisdictions, or legal genres. Several evaluations use small samples: 24 chunks for mini-graph quality, five documents for the merger comparison, and 100–117 questions for the bar-exam studies. The FAQ benchmark is document-scoped, and the clustering analysis has no attorney relevance judgments, so neither establishes end-to-end similar-case retrieval. Human graph evaluation was performed by four project members rather than legal practitioners, and LLM judges may share biases with the systems they evaluate. OCR, extraction, coreference, and inferred temporal or causal errors can propagate. Finally, the configuration yielding the 82% oracle-document result changes retrieval scope, prompting, and context together and should not be interpreted as a graph-only ablation.

Ethical Considerations

Court complaints are public records but contain names and sensitive allegations. Automatically extracted roles, events, and causal relations can be incomplete or wrong and should not be treated as legal findings or used without attorney review. Any artifact release should follow the source repository's terms, minimize unnecessary personal information, and state clearly that ARGUS is a research prototype for retrieval and organization rather than legal advice.

Acknowledgments

This work grew out of a master's capstone project at the University of Massachusetts Amherst conducted in collaboration with Bloomberg mentors. The authors gratefully acknowledge the long-standing research and educational collaboration between Bloomberg and UMass Amherst, including student mentorship, joint research, and scholarly exchange in natural language processing, data science, and related areas.

References

Lakshya A. Agrawal, Shangyin Tan, Dilara Soylu, Noah Ziems, Rishi Khare, Krista Opsahl-Ong, Arnav Singhvi, Herumb Shandilya, Michael J. Ryan, Meng Jiang, Christopher Potts, Koushik Sen, Alexandros G. Dimakis, Ion Stoica, Dan Klein, Matei Zaharia, and Omar Khattab. 2026. GEPA: Reflective prompt evolution can outperform reinforcement learning. In *The Fourteenth International Conference on Learning Representations*. Oral presentation.

Claire Barale, Leslie Barrett, Vikram Sunil Bajaj, and Michael Rovatsos. 2025. LexTime: A benchmark for temporal ordering of legal events. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 5220–5236, Suzhou, China. Association for Computational Linguistics.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904. Association for Computational Linguistics.

Ilias Chalkidis, Nicolas Garneau, Catalina Goanta, Daniel Katz, and Anders Søgaard. 2023. LeXFiles and LegalLAMA: Facilitating english multinational legal language model development. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15513–15535. Association for Computational Linguistics.

Zerui Chen, Qinggang Zhang, Zhishang Xiang, Zhimin Wei, Linfeng Gao, Xiao Huang, Zhihong Zhang, and Jinsong Su. 2026. LegalGraphRAG: Multi-agent graph retrieval-augmented generation for reliable legal reasoning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 37455–37484. Association for Computational Linguistics.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. 2024. From local to global: A graph RAG approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*.

Yi Feng, Chuanyi Li, and Vincent Ng. 2024. Legal case retrieval: A survey of the state of the art. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6472–6485. Association for Computational Linguistics.

Erwin Filtz, María Navas-Loro, Cristiana Santos, Axel Polleres, and Sabrina Kirrane. 2020. Events matter: Extraction of events from court decisions. In *Legal Knowledge and Information Systems: JURIX 2020 – The Thirty-Third Annual Conference*, volume 334 of *Frontiers in Artificial Intelligence and Applications*, pages 33–42. IOS Press.

Kilem Li Gwet. 2008. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48.

Felix Hamborg, Corinna Breitinger, and Bela Gipp. 2019. Giveme5W1H: A universal system for extracting main events from news articles. In *Proceedings of the 7th International Workshop on News Recommendation and Analytics*, volume 2554 of *CEUR Workshop Proceedings*, pages 35–43.

Parisa Kordjamshidi, Samer Aslan, Madhavan Seshadri, Leslie Barrett, and Enrico Santus. 2026. Reasoners or translators? contamination-aware evaluation and neuro-symbolic robustness on tax law. In *Proceedings of the First Workshop on Structured Understanding, Retrieval, and Generation in the LLM Era (SURGeLLM 2026)*, pages 344–360, San Diego, California, United States. Association for Computational Linguistics.

Jun Li, Lu Qian, Peifeng Liu, and Taoxiong Liu. 2024. Construction of legal knowledge graph based on knowledge-enhanced large language models. *Information*, 15(11):666.

Ying Lin, Heng Ji, Fei Huang, and Lingfei Wu. 2020. A joint neural model for information extraction with global features. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7999–8009. Association for Computational Linguistics.

Qiang Ning, Hao Wu, and Dan Roth. 2018. A multi-axis annotation scheme for event temporal relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1318–1328. Association for Computational Linguistics.

Paolo Pedinotti, Peter Baumann, Nathan Jessurun, Leslie Barrett, and Enrico Santus. 2026. MetaGraph: A large-scale meta-analysis of GenAI in financial NLP (2022–2025). In *Proceedings of the Fifth Workshop on Generation, Evaluation and Metrics (GEM)*, pages 848–861, San Diego, California, USA. Association for Computational Linguistics.

RegLab, Stanford University. 2024. Bar exam qa. `https://huggingface.co/datasets/reglab/barexam_qa`.

Yunqiu Shao, Jiaxin Mao, Yiqun Liu, Weizhi Ma, Ken Satoh, Min Zhang, and Shaoping Ma. 2020. BERT-PLI: Modeling paragraph-level interactions for legal case retrieval. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pages 3501–3507.

Omar Sharif, Joseph Gatto, Madhusudan Basak, and Sarah Masud Preum. 2025. REGen: A reliable evaluation framework for generative event argument extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 12146–12168, Suzhou, China. Association for Computational Linguistics.

Saurabh Srivastava, Sweta Pati, and Ziyu Yao. 2025. Instruction-tuning LLMs for event extraction with annotation guidelines. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 13055–13071, Vienna, Austria. Association for Computational Linguistics.

Harish Tayyar Madabushi, Taylor Pellegrin, and Claire Bonial. 2025. Generative FrameNet: Scalable and adaptive frames for interpretable knowledge storage and retrieval for LLMs powered by LLMs. In *Proceedings of Bridging Neurons and Symbols for Natural Language Processing and Knowledge Graphs Reasoning @ COLING 2025*, pages 107–119. ELRA and ICCL.

David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. 2019. Entity, relation, and event extraction with contextualized span representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 5784–5789. Association for Computational Linguistics.

Zairun Yang, Yilin Wang, Zhengyan Shi, Yuan Yao, Lei Liang, Keyan Ding, Emine Yilmaz, Huajun Chen, and Qiang Zhang. 2025. EventRAG: Enhancing LLM generation with event knowledge graphs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16967–16979. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623.

Lucia Zheng, Neel Guha, Javokhir Arifov, Sarah Zhang, Michal Skreta, Christopher D. Manning, Peter Henderson, and Daniel E. Ho. 2025. A reasoning-focused legal retrieval benchmark. In *Proceedings of the 2025 Symposium on Computer Science and Law*, pages 169–193. Association for Computing Machinery.

Xiangrong Zhu, Yuexiang Xie, Yi Liu, Yaliang Li, and Wei Hu. 2025. Knowledge graph-guided retrieval augmented generation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8912–8924. Association for Computational Linguistics.

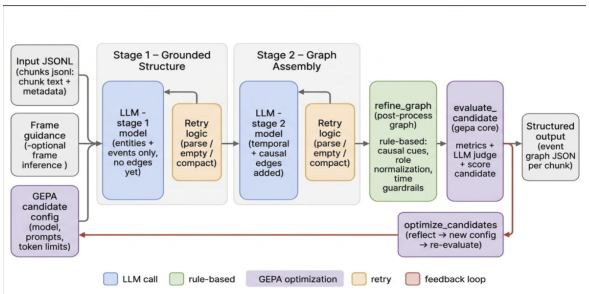


Figure 2: Two-stage mini-graph extractor with GEPA-driven optimization. Stage 1 extracts grounded entities and events; Stage 2 adds temporal and causal relations.

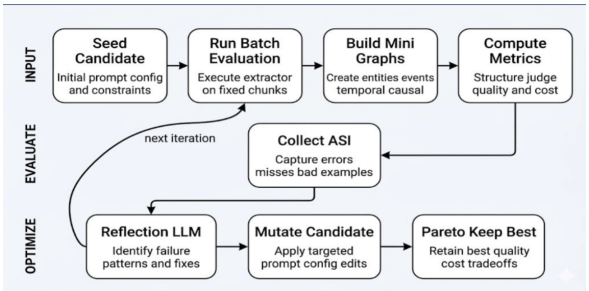


Figure 3: GEPA candidate optimization loop: seed, evaluate, reflect, mutate, and retain Pareto-optimal configurations.

A Additional Implementation Figures

Figure 2 illustrates the two-stage mini-graph extraction procedure, Figure 3 summarizes the GEPA optimization loop, and Figure 4 shows an example extracted mini-graph.

B Additional Experimental Details

B.1 One-Shot Event-Graph Prompt

The document-level baseline receives the complete sentence-labeled FACTS narrative and generates one event graph directly, without semantic chunking, intermediate mini-graphs, or staged document-level merging. The exact prompt is:

> "You are a legal event extraction analyst. You are given one full legal document with sentence-labeled facts. Extract ONE overall document-level event graph, just extract what you think are events. Return strict JSON only. Schema requires: entities, events, temporal_edges, causal_edges with consistent IDs (E1..En, EV1..EVn). Events are the only nodes; entities attach to events as participants. Temporal relations: BEFORE, SAME_TIME. Causal relation: CAUSES, ENABLES."

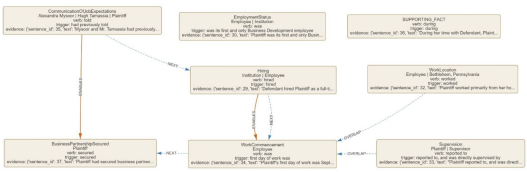


Figure 4: Example of an event mini-graph extracted from a single chunk.

Model	Stability	Flip	AC1
DeepSeek	80%	20%	1.0
Qwen	100%	0%	1.0
Mistral	100%	0%	1.0

Table 11: Intra-model stability under temperature variation.

B.2 Intra-Model Judge Stability

We additionally evaluate the stability of document-level graph judgments under temperature variation. Table 11 reports the fraction of judgments that remain unchanged and the corresponding flip rate.

B.3 Full GNN Architecture

The EKG-GNN represents each merged document-level EKG as a homogeneous, undirected graph that contains both entity and event nodes and includes self-loops. Event–entity edges connect events to participant entities, while event–event edges represent temporal and causal relations. LEGAL-BERT encodes each node's short natural-language description. The resulting `[CLS]` embedding is projected to 256 dimensions and passed through two mean-neighbor graph-convolution layers. Each layer computes

$$h_i^{(l+1)} = \mathrm{ReLU}\Big(W_s h_i^{(l)} + W_n \frac{1}{|\mathcal{N}(i)|} \sum_{j \in \mathcal{N}(i)} h_j^{(l)}\Big).$$

Each layer is followed by dropout ($p = 0.1$). Self-loops are stored in the graph but excluded from the neighbor average because the self contribution is handled separately by W_s. After two layers, node representations are mean-pooled and passed to a 23-label linear classifier.

The reported `ft-BERT` configuration jointly fine-tunes LEGAL-BERT and the graph layers using weighted binary cross-entropy and AdamW with a learning rate of 3×10^{-4} and a weight decay

Mode	LEGAL-BERT	LexLM
Classifier	0.8718	0.8616
Hybrid	0.8711	0.8627
LLM	0.7115	0.7186

Table 12: U.S. holdout Fact-class F1 by fact-extraction mode.

Feature set	Silhouette
TF-IDF baseline	0.091
EKG full (2k dims)	0.363
EKG Legacy7D	0.465
EKG Core6D ($K = 4$)	**0.649**

Table 13: Clustering silhouette by feature configuration.

of 0.01. Training uses eight epochs, a batch size of 2, and gradient clipping at 1.0.

B.4 Fact Extraction Details

Table 12 reports the U.S. holdout F1 scores for classifier, hybrid, and LLM extraction modes.

C Detailed Clustering Analysis

The clustering experiments are exploratory and examine whether graph-derived document representations yield coherent groupings of complaints. Table 13 compares the principal feature configurations using silhouette score.

We further evaluate three Core6D-based feature configurations (Table 14), whose components are summarized in Table 15. Validation compares distance to the assigned centroid with an LLM-judged cluster-membership score at $K = 4$.

Run A combines SALI claim tags with Core6D and produces one dominant cluster (111/14/6/2), with a weak distance–membership association (Pearson ≈ -0.06, $n \approx 17$). Run B combines Core6D with a 16-dimensional PCA-compressed EKG-context embedding and produces more balanced clusters, with Pearson ≈ -0.28 and Spearman ≈ -0.31 ($n = 40$). Run C adds SALI tags and a kNN cross-document graph-neighbor component to Run B; it produces cluster sizes of 31/61/19/22 and correlations of Pearson ≈ -0.29 and Spearman ≈ -0.37 ($n = 40$). The difference between Runs B and C is small, so this exploratory analysis does not establish a clear advantage for the additional Run C features.

Separately, the distance–membership correlation

Run	Representation	Sizes ($K = 4$)	Pearson	Spearman
A	SALI(23) + Core6D	111/14/6/2	≈ -0.06	–
B	Core6D + Hybrid PCA(16)	balanced, 30s–50s	≈ -0.28	≈ -0.31
C	A + Hybrid PCA + kNN(12)	31/61/19/22	≈ -0.29	≈ -0.37

Table 14: Cluster validation across three feature configurations. Correlations compare centroid distance with LLM-judged cluster membership; more negative values indicate stronger association between proximity to the centroid and judged membership. Validation $n = 40$ for Runs B and C and $n \approx 17$ for Run A.

Feature	A	B	C
SALI claim tags	yes	no	yes
Core6D structure	yes	yes	yes
Hybrid PCA context	no	yes	yes
Graph neighbors (kNN)	no	no	yes

Table 15: Features included in each clustering configuration.

for Run B remains similar when the validation sample is increased from five to ten documents per cluster ($n = 20$ versus $n = 40$), providing a check on sensitivity to validation-sample size.

D Additional Evaluation Tables

D.1 FAQ Evaluation

Table 16 reports document-scoped FAQ performance by question type.

For the $n = 299$ questions successfully scored under all three retrieval conditions, we compare methods using paired bootstrap confidence intervals, Wilcoxon signed-rank tests, and sign-flip paired permutation tests. Table 17 reports the paired differences.

Table 18 reports the EKG-versus-raw comparison separately by question type. All categories with at least 12 examples show a significant difference; the three-example temporal-relation category is too small to support a conclusion.

D.2 Bar-Exam Experiment A

Table 19 reports multiple-choice accuracy under the retrieval conditions used in Experiment A.

Pairwise significance is evaluated using exact McNemar tests, paired bootstrap confidence intervals, and Wilcoxon signed-rank tests. No pairwise comparison reaches statistical significance (Table 20).

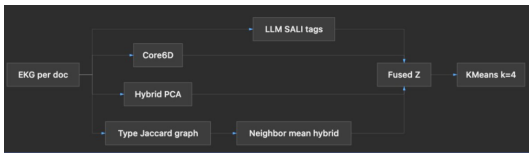


Figure 5: Run C fused representation: Core6D, hybrid PCA context, a type-Jaccard graph-neighbor component, and LLM-assigned SALI tags are combined into a single vector Z before K-means clustering ($K = 4$).

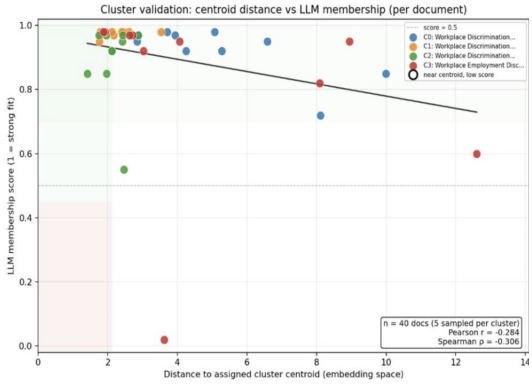
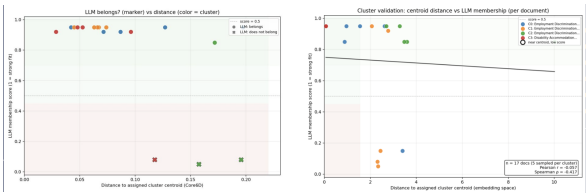


Figure 6: Run C validation: centroid distance versus LLM-judged cluster membership ($n = 40$, Pearson $r \approx -0.29$, Spearman $\rho \approx -0.37$).

D.3 Bar-Exam Experiment B

Table 10 reports Recall@10 for the full-corpus retrieval conditions in Experiment B.

Table 21 reports downstream multiple-choice accuracy under open-retrieval, gold-passage, and oracle-document conditions.



Core6D only (baseline) Run A: SALI + Core6D

Figure 7: Cluster validation by centroid distance and LLM-judged cluster membership for Core6D and Run A. This pairwise analysis preceded the three-configuration comparison in Table 14.

Question type	Raw F1	EKG F1
Temporal order	0.381	0.538
Causal/enablement	0.387	0.522
Temporal overlap	0.241	0.384
Party resolution	0.152	0.197

Table 16: FAQ performance by question type.

Comparison	ΔF1	Bootstrap 95% CI	Wilcoxon p
EKG – Raw	+0.123	[+0.109, +0.137]	3.4×10^{-39}
EKG – Hybrid	+0.083	[+0.069, +0.096]	2.3×10^{-27}
Hybrid – Raw	+0.040	[+0.030, +0.052]	8.5×10^{-11}

Table 17: Paired significance tests for the FAQ benchmark ($n = 299$). All comparisons are significant at $\alpha = 0.05$; sign-flip permutation tests agree in every comparison ($p < 10^{-3}$, not shown).

Question type	n	ΔF1 (EKG−Raw)	Wilcoxon p
Temporal order	119	+0.157	1.0×10^{-18}
Causal/enablement	99	+0.134	2.7×10^{-16}
Temporal overlap	12	+0.143	2.4×10^{-3}
Party resolution	66	+0.045	8.0×10^{-5}
Temporal relation	3	+0.019	0.25

Table 18: EKG-versus-raw FAQ results by question type. Categories with $n \geq 12$ are significant at $\alpha = 0.05$; no conclusion is drawn for the temporal-relation category ($n = 3$).

Retrieval condition	MC accuracy
Raw document (top 3)	0.8547
Chunked text (top 3)	0.8632
EKG graph (top 3)	0.8632
Gold EKG subset ($n = 100$)	0.8600

Table 19: MBE multiple-choice accuracy by retrieval condition in Experiment A ($n = 117$, Claude Sonnet 5).

Comparison	McNemar p	Bootstrap 95% CI	Wilcoxon p
Document vs. chunk	1.000	$[-5.13, +3.42]$ pp	0.813
Document vs. EKG	1.000	$[-5.98, +4.27]$ pp	0.820
Chunk vs. EKG	0.688	$[-4.27, +4.27]$ pp	1.000

Table 20: Pairwise significance tests for MBE accuracy in Experiment A. No comparison is significant at $\alpha = 0.05$.

Context condition	MCQ accuracy
No passage	0.7600
BM25-RAG (open retrieval, top 10)	0.7300
Hybrid Graph-RAG (open retrieval, top 10)	0.7200
Gold passage	0.7900
Original oracle-document Graph-RAG	0.7500
Improved oracle-document Graph-RAG	**0.8200**

Table 21: Downstream Bar Exam multiple-choice accuracy in Experiment B (Claude Sonnet 4.5).