Definitional Prefix Lifts Contextual Retrieval on Long Legal Documents

Anonymous ACL submission

Abstract

Law is built on definitions. A single legal document enumerates hundreds of operative definitions across extreme lengths and the task demands retrieval of the exact language they govern. Chunked retrieval, however, inevitably divorces a definition from the exact language it governs. This paper therefore introduces the Definitional Prefix for contextual retrieval, an indexing-time refinement of contextual chunk embedding that reconnects each chunk to its governing definitions, canonical parties, and containing section without any query-time language-model call. Evaluated on the LegalBench-RAG benchmark spanning merger agreements, commercial contracts, privacy policies, and non-disclosure agreements, this paper first finds that the general-purpose contextual prefix partially lifts retrieval on legal text and then shows that the Definitional Prefix substantially outperforms all published baselines and the general-purpose prefix under every experimental condition. A factorial decomposition further reveals what helps and what hurts on legal documents. HyDE query reformulation amplifies the Definitional Prefix to reach the strongest configuration reported, while BM25 hybrid retrieval and cross-encoder reranking consistently subtract value. These findings generalize beyond law. Any high-stakes long-document domain that encodes its operative meaning through structured internal vocabularies and tolerates no paraphrase of the source language stands to benefit from the same domain-aware indexing-time conditioning. Code is available at https://anonymous.4open.science/r/definitional-prefix/.

1 Introduction

Context length continues to scale. Even a million tokens of context, however, does not necessarily confer the ability to sustain reasoning across it or to surface the right span from it. Attention attenuates over facts buried in mid-input (Liu et al., 2024), and performance collapses catastrophically once a critical length threshold is crossed, even in models explicitly trained for long context (Wang et al., 2026). Language models are also particularly poor at the more basic task of verbatim extraction. Even state-of-the-art long-context models still retrieve only approximate, coarse-grained spans on grounded generation benchmarks (Zhang et al., 2024).

Where an approximate answer or paraphrase suffices, such failures are tolerable. Where the task demands the verbatim span, however, they are fatal. Few domains punish these failures more severely than law. Two structural properties of the domain explain why.

The first reason is length. Legal instruments are among the longest documents in any natural-language corpus because they must enumerate every operative clause, defined term, representation, and covenant to be enforceable. A central driver of this length is the dense network of defined terms on which legal drafting relies, since each operative concept is typically introduced through an explicit in-document definition and reused verbatim throughout the instrument, and the precision and consistency of those definitions is itself load-bearing for enforceability (Mellinkoff, 1963; Tiersma, 1999). Individual merger agreements, for example, can exceed 200K tokens (Wang et al., 2023). These documents sit squarely in the regime where attention degradation and long-context collapse are most severe.

The second reason is that legal work demands verbatim span retrieval, not paraphrased answers. Practitioners must cite the exact language of a contract or statute, not the restatement of a model. The task, however, proves difficult even for the best models. On benchmarks that strip generation away and ask only whether the model can locate the right span, the best-performing systems only reach 63% accuracy (Yu et al., 2025). Asking the model to find the relevant span is no substitute for retrieving

it. Retrieval in law is, therefore, not a convenience layer that long context renders unnecessary. It is the load-bearing primitive on which citation, audit, and professional reliance are built.

These constraints motivate a retrieval-first approach as well as raise a question of which retrieval augmentations actually help on long legal documents. Prior work on general-domain corpora improves dense retrieval by augmenting the chunk, the query, or both with language-model-generated context (Nogueira et al., 2019; Gao et al., 2023; Mao et al., 2021; Günther et al., 2024). The most promising of these is contextual chunk embedding, which augments each chunk's representation with document-level context at indexing time (Anthropic, 2024; Conti et al., 2025; Günther et al., 2024). On a general technical corpus, the prefix-based variant composes additively with BM25 hybrid retrieval (Robertson and Zaragoza, 2009; Cormack et al., 2009) and cross-encoder reranking (Cohere, 2024).

Whether such augmentations carry over to the legal domain, however, has not been tested thus far. This paper therefore makes three contributions. First, it establishes the first rigorous test bed for transferring contextual retrieval methods from general-domain corpora to legal span retrieval, evaluating on the structurally diverse LegalBench-RAG benchmark (Pipitone and Alami, 2024) under a controlled factorial design. Second, it introduces a domain-aware refinement, the Definitional Prefix, and shows that it outperforms both the baseline and the Generic Prefix (Anthropic, 2024) without any query-time language model call. Third, it factorially decomposes the remaining augmentations and reports that BM25 hybrid retrieval and cross-encoder reranking degrade performance on legal text across every sparse and multi-vector channel tested under both the Generic Prefix and the Definitional Prefix, while a HyDE query-side rewrite composes super-additively with both prefix variants and reaches the strongest composition reported in this paper when stacked on the Definitional Prefix.

2 Experimental Setup

2.1 Datasets

This paper evaluates on the four datasets of LegalBench-RAG (Pipitone and Alami, 2024), which adapts select tasks from LegalBench (Guha et al., 2023) into a character-span retrieval benchmark. Each query is paired with ground-truth character intervals in the source document. MAUD (Wang et al., 2023) contains M&A merger agreements, with queries that reference specific defined terms and party tickers. CUAD (Hendrycks et al., 2021) contains commercial contracts drawn from a broad range of contract types such as services, distribution, joint venture, endorsement, and franchise agreements, with short interrogative queries against heterogeneous clause language. PrivacyQA (Ravichander et al., 2019) contains mobile-app privacy policies with consumer-facing questions about data handling. ContractNLI (Koreeda and Manning, 2021) contains NDAs with templatic entailment queries that often quote source language verbatim. Average document length ranges from $\sim$10K characters in ContractNLI to $\sim$356K in MAUD, a roughly $35\times$ span across the four datasets. Together these corpora span heterogeneous legal domains with substantially different contractual syntax, provision structure, and operative language, thereby providing a rigorous testbed for assessing retrieval robustness.

2.2 Baselines

The published LegalBench-RAG baselines use two chunkers at 500 characters, a Naive fixed-window splitter and LangChain's RecursiveCharacterTextSplitter. Dense retrieval uses `text-embedding-3-large` at 3072 dimensions, with optional cross-encoder reranking using `rerank-english-v3.0` (Cohere, 2024). All four baseline cells in rows 1–4 of Table 1 are independently reproduced on the official codebase of Pipitone and Alami (2024) under the published settings.

2.3 Models

All experiments retain the chunkers, embedding model, and reranker of the published baselines as controlled variables. The language model for both the chunk-side contextual prefix and the query-side HyDE rewrite is `gpt-5-nano`, chosen as the smallest and cheapest available frontier model to minimize the influence of language model capability on the results. Both calls use `reasoning_effort=minimal`. For the channel-swap robustness check, BGE-M3 sparse (Chen et al., 2024) and ColBERT v2 (Santhanam et al., 2022) serve as drop-in replacements for BM25 in the hybrid pipeline, both used off-the-shelf with no fine-tuning.

2.4 Metrics and Significance

Retrieval quality is measured by character-overlap Precision@k and Recall@k. Let R_k be the union of retrieved character intervals at truncation depth k and G the union of ground-truth intervals. Then $P@k = |R_k \cap G|/|R_k|$ and $R@k = |R_k \cap G|/|G|$, where $|\cdot|$ denotes total interval length in characters. Each method returns up to 64 retrieved snippets per query, truncated at $k \in \{1, 2, 4, 8, 16, 32, 64\}$. Statistical robustness is assessed via 95% percentile bootstrap confidence intervals (5,000 resamples) on per-query Recall@k.

3 Methodology

3.1 Contextual Prefix

A contextual prefix is a short, language-model-generated token sequence prepended to a chunk before its dense embedding is computed, so that the chunk's representation in the retrieval index conditions on document-level context in addition to its local content (Anthropic, 2024; Conti et al., 2025; Günther et al., 2024). The motivation is that dense passage retrievers encode each chunk through an independent forward pass of a bi-encoder (Karpukhin et al., 2020), so when a long document is split into many short chunks the resulting embeddings shed the document-level signal that would otherwise disambiguate sibling sections sharing the same lexical surface (Merola and Singh, 2025). The prepended tokens typically name the document type, the parties involved, and the topic of the section in which the chunk sits, and the embedder folds them into the chunk's coordinates at index time.

3.1.1 Generic Prefix

The contextual retrieval procedure of Anthropic (2024) is applied as published. One prefix is generated per chunk using `gpt-5-nano` under the prompt shown in Appendix A. The prompt receives the full document together with the chunk and returns a short situating snippet, which is prepended to the chunk so that the prefixed text supersedes the raw chunk as the embedding input. The original character span is preserved, and the prefix itself is excluded from the character-overlap evaluation.

3.1.2 Definitional Prefix

The generic procedure lets the language model choose what document-level context to surface, but on legal text it leaves two recurring sources of signal under-exploited. The first is the defined-term gloss. A contract defines a term such as "Material Adverse Effect" once in its definitions section and reuses it verbatim throughout, but queries about clauses governed by that term typically paraphrase rather than quote it ("events causing major economic harm"), so the chunk and the query share no surface tokens unless the prefix carries the gloss across, and the generic prompt rarely does. The second is canonical identifiers. Party names, agreement type, and section labels recur verbatim within a document, but the language model paraphrases them differently across chunks (writing "Avangrid", "the Parent", and "Avangrid, Inc." for the same entity in different prefixes), injecting spurious variance into the embedded text that is unrelated to substantive content. The Definitional Prefix addresses both with a two-stage prompt pipeline that produces, for each chunk, a tight prefix anchored on the agreement type, canonical party identifiers, verbatim section label, and document defined-term glosses.

Figure 2 diagrams the construction. Stage A is invoked once per document and emits a structured JSON record holding the agreement type, the parties under canonical role names paired with their full legal entities, the governing law, a section outline, and the top fifteen defined terms each with a short gloss. Stage B is invoked once per chunk and synthesizes the prefix from the Stage A metadata, the chunk text, and the chunk's containing section, which is resolved deterministically by a regex parser merged with the Stage A outline rather than left to the language model. Six behavioral rules listed in Appendix A bind Stage B to verbatim copying of party names, mandatory inclusion of any defined terms it cites, verbatim section-label transcription, a hard length cap, and a prohibition on inventing material not already in the metadata or chunk.

The language model, the

Raw chunk (Avangrid / PNM merger, §3.8 carve-out)
"…a decline in the trading price, or adverse credit events resulting in deterioration in the credit markets generally and not specifically relating to such party…"

↓ Generic: one `gpt-5-nano` call (Appendix A)

Generic prefix (Anthropic, 2024)
"This chunk is from a merger agreement between Avangrid, Inc. and PNM Resources, Inc. It discusses an exception to the Material Adverse Effect carve-out, specifying that declines in trading price and broader credit-market deterioration do not constitute an MAE on the agreement parties."

Figure 1: Generic Prefix (Anthropic, 2024) on an Avangrid / PNM merger chunk.

Raw chunk (same chunk as Figure 1)

"…a decline in the trading price, or adverse credit events resulting in deterioration in the credit markets generally and not specifically relating to such party…"

↓ Stage A of the Definitional Prefix pipeline: one `gpt-5-nano` call per document

Document metadata (extracted JSON, abbreviated)

`agreement_type`: Agreement and Plan of Merger
`parties`: {Parent = Avangrid, Inc.; Company = PNM Resources, Inc.}
`defined_terms` (top 15, each with ≤25-word gloss):
"Material Adverse Effect" → "events with material harm"
"Acquisition Proposal" → "third-party offer to acquire 15%+ of equity"
"Merger Consideration" → "consideration to Company shareholders"
…
`section_outline`: [§2.1 Merger, §3.8 Material Adverse Effect, …]

↓ Stage B of the Definitional Prefix pipeline: one `gpt-5-nano` call per chunk + SectionMap(§3.8)

Definitional Prefix output (prepended to raw chunk before embedding)

"In an Agreement and Plan of Merger between Parent (Avangrid, Inc.) and Company (PNM Resources, Inc.). Section: §3.8 Material Adverse Effect. Terms: `"Material Adverse Effect"` (events with material harm). Topic: MAE carve-out for credit-market events."

Figure 2: Definitional Prefix construction on the same chunk as Figure 1. The Definitional Prefix injects canonical party names (M1), defined-term glosses (M2), and a verbatim section label (M3) into fixed slots that the Generic Prefix lacks.

`reasoning_effort=minimal` setting, the `max_doc_chars=300,000` cap, the chunker, the embedder, the retrieval pipeline, and the query sample are held identical to the Generic Prefix. Length is tight by construction, with the empirical Definitional Prefix distribution on CUAD at mean 294 / max 374 chars, shorter than the Generic Prefix at mean 316 / max 619 chars, pre-empting any length-confound objection at comparison time.

3.2 Compositional Augmentations

The contextual prefix constitutes the primary intervention. The following two augmentations are evaluated compositionally to assess whether additional retrieval layers yield further gains or introduce interference when layered on the prefix. Each composition is applied to both prefix families, the Generic Prefix and the Definitional Prefix, holding all other components constant.

3.2.1 Sparse and Multi-Vector Hybrid Retrieval with Reranking

Prior work reports that composing the contextual prefix with BM25 hybrid retrieval and cross-encoder reranking, a procedure known as Contextual BM25, produces further additive gains on general technical corpora (Anthropic, 2024). This paper evaluates whether the same composition generalizes to legal text. The BM25 channel indexes the same prefix-augmented chunk text that the dense path embeds, scored with BM25 (Robertson and Zaragoza, 2009) at default parameters (k_1=1.5, b=0.75). The dense and BM25 top-64 lists are fused via Reciprocal Rank Fusion (Cormack et al., 2009):

$$\mathrm{RRF}(c) = \frac{1}{k + r_{\text{dense}}(c)} + \frac{1}{k + r_{\text{bm25}}(c)}, \quad (1)$$

with k=60. To verify that negative results are not attributable to BM25 specifically, the sparse channel is substituted with BGE-M3 sparse (Chen et al., 2024) and ColBERT v2 (Santhanam et al., 2022).

3.2.2 Query Reformulation via HyDE

The preceding augmentation operates on the chunk side. HyDE (Gao et al., 2023) operates on the query side by replacing the raw query embedding with a convex combination of the original query and a hypothetical answer passage:

$$e_{\text{query}} = \alpha \cdot e(q) + (1 - \alpha) \cdot e(\mathrm{HyDE}(q)), \quad (2)$$

followed by L2-renormalization and cosine retrieval over the same prefix-augmented chunk pool. The hypothetical passage is generated by a single `gpt-5-nano` call per query, prompted to draft a short passage in declarative legal-document register (Appendix A). The mixing weight α is selected per chunker by maximizing ALL R@1 on a 9-point sweep (Table 8), yielding $\alpha = 0.7$ for Naive and $\alpha = 0.6$ for RCTS.

4 Results

Table 1 reports Precision@k and Recall@k on the four LegalBench-RAG datasets. All findings are statistically significant by 95% bootstrap CI on per-query Recall@k.

4.1 Contextual Prefix

The Definitional Prefix lifts ALL R@1 from 3.37% to 18.29% on the Naive chunker and from 4.94% to 17.70% on RCTS, with both prefix variants outperforming every baseline. The Definitional Prefix further outperforms the Generic Prefix on every dataset and chunker combination except PrivacyQA on Naive, where the two are tied.

On MAUD, the longest dataset in the benchmark and the one most saturated with templatic defined-term clauses, the Generic Prefix lifts R@1 from 2.54% to 4.72% on Naive and from 1.65% to 3.28% on RCTS. The Definitional Prefix then more

Dataset	Method	Precision@k							Recall@k						
		1	2	4	8	16	32	64	1	2	4	8	16	32	64
MAUD	Naive	3.36	2.65	2.18	1.89	1.48	1.06	0.75	2.54	3.12	4.53	8.75	13.16	18.36	25.62
	RCTS	2.65	1.77	1.96	1.40	1.39	1.15	0.82	1.65	2.09	4.59	6.18	12.93	21.04	28.28
	Naive + Rerank	2.64	1.77	1.96	1.40	1.38	1.15	0.82	1.65	2.09	5.59	6.18	12.93	21.04	28.28
	RCTS + Rerank	1.94	2.63	2.05	1.77	1.79	1.55	1.12	0.52	2.48	4.39	7.24	14.03	22.60	31.46
	Naive + Generic	5.84	4.30	3.53	2.72	2.24	1.67	1.22	4.72	6.34	8.50	11.35	18.42	25.95	37.03
	RCTS + Generic	5.64	5.19	4.44	3.35	2.45	2.23	1.76	3.28	4.37	6.77	8.42	10.69	18.28	25.74
	Naive + Definitional	**19.66**	**15.96**	**12.07**	**7.63**	**5.15**	**3.20**	**1.92**	**8.39**	**13.64**	**20.75**	**25.08**	**33.11**	**40.04**	**49.42**
	RCTS + Definitional	**14.96**	**11.80**	**11.24**	**7.36**	**5.16**	**3.42**	**2.19**	**5.35**	**7.86**	**13.09**	**15.78**	**20.59**	**27.09**	**33.48**
CUAD	Naive	9.27	8.05	5.98	4.33	2.77	1.77	1.09	12.60	19.47	27.92	40.70	51.02	64.38	75.71
	RCTS	1.97	4.03	4.83	4.20	2.94	1.99	1.25	1.62	8.11	17.72	31.68	44.38	60.04	74.70
	Naive + Rerank	1.97	4.03	4.83	4.20	2.94	1.99	1.25	1.62	8.11	17.72	31.68	44.38	60.04	74.70
	RCTS + Rerank	3.53	4.18	6.18	5.06	3.93	2.74	1.66	3.17	7.33	18.26	28.67	42.50	55.66	70.19
	Naive + Generic	9.07	7.67	6.26	4.69	3.56	2.21	1.36	10.14	17.87	29.44	40.61	56.54	68.71	81.66
	RCTS + Generic	8.90	8.46	7.88	6.88	5.96	4.00	2.03	7.66	12.39	18.47	25.64	40.84	59.29	72.89
	Naive + Definitional	**16.85**	**13.19**	**8.94**	**5.53**	**4.20**	**2.57**	**1.50**	**21.14**	**31.97**	**41.14**	**48.29**	**66.76**	**79.04**	**90.11**
	RCTS + Definitional	**22.72**	**19.69**	**13.16**	**8.43**	5.72	3.54	**2.14**	**19.63**	**30.94**	**36.88**	**45.15**	**56.05**	**68.57**	**83.05**
PrivacyQA	Naive	7.86	7.31	6.41	5.06	3.58	2.41	1.54	7.45	12.53	20.88	32.38	42.45	54.27	66.07
	RCTS	14.38	13.55	12.34	9.03	6.06	4.17	2.81	8.85	15.21	27.92	42.37	55.12	71.19	84.19
	Naive + Rerank	14.38	13.55	12.34	9.02	6.06	4.17	2.81	8.85	15.21	27.92	42.37	55.12	71.19	84.19
	RCTS + Rerank	13.94	15.91	13.32	9.57	6.88	4.68	3.28	7.32	16.12	25.65	35.60	51.87	64.98	79.61
	Naive + Generic	28.98	24.69	17.75	12.42	8.47	5.64	3.26	19.33	30.22	41.29	53.98	66.74	79.55	88.41
	RCTS + Generic	26.29	23.84	17.87	11.77	8.74	6.01	3.76	12.73	21.26	29.25	40.71	56.53	72.32	82.99
	Naive + Definitional	**30.02**	**25.86**	**19.44**	**13.43**	**9.17**	**6.06**	**3.54**	19.01	**32.81**	**46.09**	**59.84**	**76.76**	**91.21**	**98.61**
	RCTS + Definitional	**29.99**	21.31	**18.67**	**13.98**	**10.00**	**6.94**	**4.61**	**19.08**	**28.53**	**41.42**	**51.98**	**65.97**	**83.27**	**95.65**
ContractNLI	Naive	16.45	14.80	12.53	9.73	6.70	4.65	3.04	11.32	19.10	29.79	45.59	56.75	69.88	86.57
	RCTS	6.63	5.29	3.89	2.81	1.98	1.29	0.90	7.63	11.33	17.34	24.99	35.80	46.57	61.72
	Naive + Rerank	6.63	5.28	3.89	2.81	1.98	1.29	0.90	7.63	11.34	17.34	24.99	35.80	46.57	61.72
	RCTS + Rerank	5.08	5.59	5.04	3.67	2.52	1.75	1.17	4.91	9.33	16.09	25.83	35.04	46.90	62.97
	Naive + Generic	12.10	10.94	7.28	4.97	3.09	1.95	1.19	14.82	25.99	32.64	43.96	53.49	66.57	80.81
	RCTS + Generic	14.87	13.02	9.11	6.18	4.41	2.76	1.62	16.57	24.76	33.00	42.92	57.28	67.52	78.29
	Naive + Definitional	**19.84**	**16.80**	**11.65**	**7.80**	**4.74**	**2.60**	**1.36**	**24.62**	**38.21**	**50.29**	**66.29**	**78.70**	**86.27**	**89.71**
	RCTS + Definitional	**28.15**	**22.86**	**15.33**	**10.37**	**6.35**	**3.65**	**1.92**	**26.73**	**38.43**	**47.77**	**59.97**	**71.63**	**80.66**	**86.08**
ALL	Naive	2.40	3.76	4.97	4.33	3.39	2.17	1.29	3.37	8.44	21.30	34.51	48.88	64.47	76.39
	RCTS	6.41	6.16	5.76	4.36	3.09	2.15	1.45	4.94	9.19	16.90	26.30	37.06	49.71	62.22
	Naive + Rerank	6.41	6.16	5.76	4.36	3.09	2.15	1.45	4.94	9.19	16.90	26.30	37.05	49.71	62.22
	RCTS + Rerank	6.13	7.08	6.65	5.02	3.78	2.68	1.81	3.98	8.82	16.10	24.34	35.86	47.54	61.06
	Naive + Generic	14.00	11.90	8.71	6.20	4.34	2.87	1.76	12.25	20.10	27.97	37.48	48.80	60.19	71.98
	RCTS + Generic	13.93	12.63	9.82	7.05	5.39	3.75	2.29	10.06	15.70	21.87	29.42	41.33	54.35	64.98
	Naive + Definitional	**21.59**	**17.95**	**13.03**	**8.60**	**5.82**	**3.61**	**2.08**	**18.29**	**29.16**	**39.57**	**49.88**	**63.83**	**74.14**	**81.96**
	RCTS + Definitional	**23.95**	**18.91**	**14.60**	**10.04**	**6.81**	**4.39**	**2.71**	**17.70**	**26.44**	**34.79**	**43.22**	**53.56**	**64.90**	**74.56**

Table 1: Precision@k and Recall@k at `chunk_size=500`. Rows 1–4 reproduce the LegalBench-RAG baselines from Pipitone and Alami (2024). Rows 5–6 are the Generic Prefix (§3.1.1, Anthropic, 2024); rows 7–8 are the Definitional Prefix (§3.1.2). Bold and underlined cells exceed every previous row in that column.

than doubles that lift, reaching 8.39% on Naive and 5.35% on RCTS. MAUD is the cleanest empirical demonstration of the Definitional Prefix's defined-term gloss mechanism, since merger agreements introduce dozens of operative defined terms such as Material Adverse Effect, Acquisition Proposal, and Merger Consideration, and queries on those clauses paraphrase the terms rather than quote them. The Generic Prefix surfaces parties and agreement type but rarely the verbatim defined-term gloss, leaving the bridging work undone, whereas the Definitional Prefix's Stage A metadata injects the gloss directly into the embedded text. The compounding extends to higher k, with Naive R@64 climbing from 25.62% on the baseline to 49.42% with the Definitional Prefix.

On CUAD, the Generic Prefix shows a small R@1 decrease on Naive that traces to a short-amendment artefact analysed in Appendix F, while on RCTS it outperforms the baseline on every cell. The Definitional Prefix far surpasses the Generic Prefix, delivering the largest single-dataset Definitional Prefix lift in the benchmark, with R@1 climbing from 10.14% to 21.14% on Naive and from 7.66% to 19.63% on RCTS. The commercial-contract corpus is the heterogeneous-template regime where the Definitional Prefix's canonical party slot and defined-term gloss have the most discriminative weight.

On PrivacyQA, the Generic Prefix produces its cleanest single-dataset win, with Naive R@1 jumping from 7.45% to 19.33% and R@64 saturating near 88%. The 7-document corpus is short and only loosely templated, so the Generic Prefix already captures most of the available signal by folding in topic and section markers that informal subheadings do not make explicit. The Definitional Prefix is approximately tied with the Generic Prefix on Naive R@1 but lifts R@1 from 12.73% to 19.08% on RCTS, where the Generic Prefix started lower

and the Definitional Prefix's structural slots have more headroom to recover.

On ContractNLI, the Generic Prefix lifts R@1 from 11.32% to 14.82% on Naive with most of the gain at higher k. NDA templates already carry distinctive topic labels that the dense embedding handles well, leaving the Generic Prefix limited headroom. The Definitional Prefix then lifts R@1 from 14.82% to 24.62% on Naive and from 16.57% to 26.73% on RCTS, the second-largest absolute Definitional Prefix lift in the benchmark, by injecting the NDA-templatic defined-term frame as a verbatim gloss inside the embedded prefix.

4.2 Compositional Augmentations

4.2.1 Sparse Hybrid Retrieval and Reranking

BM25 hybrid retrieval and cross-encoder reranking actively degrade retrieval on legal text under both prefix families (Table 3, Appendix B). Adding BM25 reduces ALL R@1 by 5.97 / 5.40 pp under the Generic Prefix and by 7.33 / 7.69 pp under the Definitional Prefix. The reranker on top partially recovers but never exceeds prefix-only (-3.92 / -2.66 pp under the Generic Prefix; -0.78 / -1.68 pp under the Definitional Prefix). The negative composition therefore tracks the strength of the prefix rather than substitutes for it, and is structural to the legal corpus rather than specific to the prefix family.

4.2.2 Channel-Swap Robustness

Substituting the BM25 channel with BGE-M3 sparse (Chen et al., 2024) or ColBERT v2 (Santhanam et al., 2022) still degrades retrieval under both prefix families, with no combination reaching prefix-only on either chunker (full 12-cell grid in Appendix C). The sparse channel's failure to compose with the prefix is therefore structural to the legal corpus rather than channel-specific or prefix-specific.

4.2.3 Query Reformulation via HyDE

HyDE composes super-additively with both prefix families at the per-chunker peak α (Table 4, Appendix B, with the full α sweep in Appendix E). The Generic Prefix with HyDE reaches ALL R@1 of 15.85 / 12.75 (+3.60 / +2.69 pp over prefix-only), and the Definitional Prefix with HyDE reaches 21.07 / 20.13 (+2.78 / +2.43 pp), the strongest composition reported in this paper.

	Naive			RCTS		
	−H	+H	Δ	−H	+H	Δ
−Pref	7.06	8.72	+1.66	7.69	9.12	+1.42
+Generic	12.25	15.85	+3.60	10.06	12.75	+2.69
+Definitional	18.29	21.07	+2.78	17.70	20.13	+2.43
Gen.−−Pref Δ	+5.20	+7.13		+2.36	+3.63	
Def.–Gen. Δ	+6.04	+5.22		+7.64	+7.38	

Interaction (joint − sum-of-marginals over no-prefix baseline): Generic + HyDE Naive +1.93 / RCTS +1.27; Definitional + HyDE Naive +1.12 / RCTS +1.01.

Table 2: 3×2 factorial across prefix mode and ±HyDE (ALL R@1, %). Each cell uses the per-prefix peak α (Generic Naive 0.7, RCTS 0.6; Definitional Naive 0.7, RCTS 0.8).

4.3 Factorial Ablation

To isolate how much the chunk-side prefix and the query-side HyDE rewrite each contribute on their own, a 3×2 factorial is run across prefix mode $\in$ {NONE, GENERIC, DEFINITIONAL} and ±HyDE on both chunkers (Table 2). The parallel no-prefix α-sweep is in Appendix E (Table 10).

The chunk-side prefix is the dominant intervention. HyDE alone barely moves R@1 above the no-prefix baseline, whereas the Definitional Prefix alone already exceeds every HyDE-on-no-prefix configuration by a wide margin (Table 2). HyDE then composes super-additively with both prefix families, but the joint lift over the prefix-only condition is smaller than the prefix's own lift over no-prefix. HyDE supplies complementary query-side signal without substituting for the chunk-side prefix.

5 Discussion

5.1 Why the Contextual Prefix Transfers to Legal Retrieval

The contextual prefix is an index-time intervention on the chunk-side dense encoder. Dense bi-encoders encode each chunk through an independent forward pass and therefore discard the document-level signal once a long instrument is split into many short chunks. With a contextual prefix, however, that loss is repaired by injecting document-level context into the same forward pass that produces the dense vector. Since the prefix is written into the chunk before encoding, it changes the geometry of the embedding space that the cosine retrieves over rather than reordering candidates after retrieval has already taken place.

Originally validated on general-corpus retrieval, the contextual prefix transfers cleanly to long le-

gal documents as well. The Generic Prefix more than triples Naive-chunker ALL R@1 over the published baseline ($3.37\% \rightarrow 12.25\%$, Table 1). The retrieval bottleneck on long legal instruments therefore resides inside the chunk representation itself rather than in any subsequent reordering stage. Prefix-only retrieval already surpasses every BM25-and-reranker configuration evaluated in this paper, so the parsimonious lever in this regime is to write document-level context into the chunk before encoding rather than to engineer a richer sparse channel or reranker after the fact.

5.2 Why the Definitional Prefix Improves on the Generic Prefix

The Definitional Prefix sits on the same index-time axis as the Generic Prefix but is constructed against three properties of legal drafting that a general-purpose prefix systematically under-exploits. The three mechanisms below are conceptually independent and additive at index time.

Canonical party surface form (M1). The Generic Prefix is produced from a single per-chunk language-model call with no document-wide consistency constraint, and the model paraphrases the same legal entity differently across chunks of one document, for example writing "Avangrid", "the Parent", and "Avangrid, Inc." across chunks of the same merger agreement. Each paraphrastic variant becomes surface-form variance inside the embedded text that is unrelated to the substantive content of the chunk and competes with the discriminative signal the embedder is supposed to encode. M1 collapses this variance by binding every chunk of a document to a single canonical party string. The variance disappears, and the embedder no longer spends representational capacity discriminating against paraphrased identifiers.

Defined-term gloss bridging (M2). The most legal-specific mechanism. The writer of a legal query phrases the clause topic in the ordinary English of the practitioner rather than quoting the defined term verbatim, since the practitioner does not normally have the source document open at query time. A query about "events causing major economic harm" is targeting a chunk that discusses "Material Adverse Effect", but the two strings share no surface tokens and the dense encoder has very little signal to bridge them. The Definitional Prefix writes the defined-term gloss into the prefix at index time, so the bridging tokens live inside the embedded text and cosine similarity rises on overlap that a general-purpose prefix strips out. This mechanism exploits the self-defining-vocabulary property of legal drafting (Mellinkoff, 1963; Tiersma, 1999), in which each operative concept is introduced through an explicit in-document definition and then reused verbatim throughout the instrument. The property distinguishes legal text most sharply from general technical prose, and it is the reason the Definitional Prefix is a domain-specific contribution rather than a marginal refinement of the general-purpose procedure. The mechanism is most visible on MAUD, where defined-term density is highest among the four sub-corpora and the Definitional Prefix's gain over the Generic Prefix peaks.

Deterministic section anchoring (M3). An earlier prototype let the language model identify the chunk's containing section by topic alignment, and the model produced wrong labels on roughly half of operative chunks, which dragged the chunk's vector toward the wrong neighborhood of the embedding space. M3 resolves the section deterministically through a regex parser merged with the document outline, and binds every chunk to copy the resolved label verbatim. Chunks whose containing section cannot be resolved are tagged with an explicit unresolved marker rather than guessed. The deterministic anchor matters because legal instruments encode operative meaning in their section hierarchy (Article numbers, Section labels, subclause letters), and a chunk vector that is anchored to the wrong location in that hierarchy is worse than one that admits ignorance. M3 is the mechanism by which the procedure refuses to hallucinate structural information that the document does not unambiguously provide.

The empirical lift is largest precisely where defined-term density and template repetition are highest, with the Definitional Prefix nearly doubling the Generic-Prefix R@1 on MAUD ($4.72\% \rightarrow 8.39\%$ on Naive) and delivering the single largest dataset-level R@1 lift on CUAD ($10.14\% \rightarrow 21.14\%$ on Naive). The Definitional Prefix demonstrates that domain-specific document structure, when written into the chunk representation at indexing time, recovers retrieval signal that a general-purpose prefix discards. The lift is obtained entirely through structured prompting at indexing time without any fine-tuning of the embedding model, positioning domain-aware prefix design as a training-free axis for retrieval improvement on specialized corpora.

5.3 Why Compositions Split Along the Chunk-Query Axis

BM25 hybrid retrieval and the cross-encoder reranker subtract value under both prefix families. This inverts the additive composition reported on general technical corpora by Anthropic (2024) and surfaces a domain-specific failure mode not previously documented for legal text. Legal boilerplate at near-uniform density across chunks collapses the IDF signal that BM25 (Robertson and Zaragoza, 2009) depends on, and the prefix tokens themselves repeat across chunks of a single instrument, raising sparse-channel scores uniformly. BGE-M3 sparse and ColBERT v2 inherit the same failure mode, and the reranker rescues only when the sparse channel mis-ranks rather than mis-truncates the correct chunk out of the top-K candidate pool.

HyDE is the one composition that works, because it acts on the query side, rewriting the raw query as a hypothetical passage in the chunk's register. It is the mirror image of the chunk-side prefix on the opposite side of the cosine, and the super-additive composition with both prefix families confirms complementary signal. The cleanest empirical signature is the ContractNLI sign-flip in Table 4, in which HyDE hurts the Generic Prefix on the dataset's quote-heavy NDA queries but helps the Definitional Prefix, whose chunk-side gloss already supplies the surface alignment. The resulting Definitional Prefix with HyDE is the strongest configuration reported in this paper.

6 Conclusion

Legal text is a paradigmatic instance of a high-stakes long-document setting in which the system's reliability is bounded by the quality of its retrieval step. The same is true of related high-stakes domains such as medical records, regulatory filings, and financial disclosures, where downstream professional reliance depends on returning the verbatim source span rather than a paraphrase. Long-context language models do not change this requirement, and their reported accuracy on verbatim span tasks remains substantially below the threshold these domains demand.

This paper shows that the right axis of intervention in such settings is index-time chunk-side conditioning. The contextual prefix lifts cross-dataset Precision@k and Recall@k substantially over the LegalBench-RAG baselines at every truncation depth k. The Definitional Prefix then takes the further step of writing legal-drafting structure into the prefix at index time, anchoring on canonical party identifiers, in-document defined-term glosses, and deterministically-resolved section labels, and lifts retrieval again.

Law is only one such domain. A Generic Prefix establishes a substantial floor on any retrieval setting in which chunks of a long instrument share lexical surface across positions, and a domain-aware refinement of that prefix supplies further headroom whenever the target domain has structural conventions that a general-purpose prefix systematically under-encodes. Legal drafting exhibits this property unusually clearly through its in-document defined-term system and its Article and Section hierarchy. The same chunk-side recipe carries directly to other high-stakes long-document settings whose retrieval target is the verbatim span, and the broader significance of the work is that domain-aware refinement of the prefix unlocks retrieval headroom that a general-purpose procedure leaves on the table.

7 Limitations

All experiments use a single closed-source embedding model. All prefix-effect deltas are computed against matched no-prefix reproductions to control for pipeline-level confounds.

Ethics Statement

This paper abides by the ACL Code of Ethics. The data resources used in this study are publicly available.

References

Anthropic. 2024. Introducing contextual retrieval. Anthropic Engineering Blog.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-Embedding: Multilinguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*.

Cohere. 2024. Rerank 3 – Cohere's state-of-the-art reranker. Cohere Blog.

Max Conti, Manuel Faysse, Gautier Viaud, Antoine Bosselut, Celine Hudelot, and Pierre Colombo. 2025. Context is gold to find the gold passage: Evaluating and training contextual document embeddings. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages

22594–22608, Suzhou, China. Association for Computational Linguistics.

Gordon V. Cormack, Charles L. A. Clarke, and Stefan Büttcher. 2009. Reciprocal rank fusion outperforms Condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 758–759. ACM.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023. Precise zero-shot dense retrieval without relevance labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023)*.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, and 1 others. 2023. LegalBench: A collaboratively built benchmark for measuring legal reasoning in large language models. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023), Datasets and Benchmarks Track*.

Michael Günther, Isabelle Mohr, Daniel J. Williams, Bo Wang, and Han Xiao. 2024. Late chunking: Contextual chunk embeddings using long-context embedding models. *arXiv preprint arXiv:2409.04701*.

Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. 2021. CUAD: An expert-annotated NLP dataset for legal contract review. *arXiv preprint arXiv:2103.06268*.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781. Association for Computational Linguistics.

Yuta Koreeda and Christopher D. Manning. 2021. ContractNLI: A dataset for document-level natural language inference for contracts. *arXiv preprint arXiv:2110.01799*.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12.

Yuning Mao, Pengcheng He, Xiaodong Liu, Yelong Shen, Jianfeng Gao, Jiawei Han, and Weizhu Chen. 2021. Generation-augmented retrieval for open-domain question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL 2021)*, pages 4089–4100.

David Mellinkoff. 1963. *The Language of the Law*. Little, Brown and Company, Boston, MA.

Carlo Merola and Jaspinder Singh. 2025. Reconstructing context: Evaluating advanced chunking strategies for retrieval-augmented generation. In *Second Workshop on Knowledge-Enhanced Information Retrieval at ECIR 2025*.

Rodrigo Nogueira, Wei Yang, Jimmy Lin, and Kyunghyun Cho. 2019. Document expansion by query prediction. *arXiv preprint arXiv:1904.08375*.

Nicholas Pipitone and Ghita Houir Alami. 2024. LegalBench-RAG: A benchmark for retrieval-augmented generation in the legal domain. *arXiv preprint arXiv:2408.10343*.

Abhilasha Ravichander, Alan W. Black, Shomir Wilson, Thomas Norton, and Norman Sadeh. 2019. Question answering for privacy policies: Combining computational and legal perspectives. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4949–4959, Hong Kong, China. Association for Computational Linguistics.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2022. ColBERTv2: Effective and efficient retrieval via lightweight late interaction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Peter M. Tiersma. 1999. *Legal Language*. University of Chicago Press, Chicago, IL.

Steven H. Wang, Antoine Scardigli, Leonard Tang, Wei Chen, Dimitry Levkin, Anya Chen, Spencer Ball, Thomas Woodside, Oliver Zhang, and Dan Hendrycks. 2023. MAUD: An expert-annotated legal NLP dataset for merger agreement understanding. *arXiv preprint arXiv:2301.00876*.

Weiwei Wang, Jiyong Min, and Weijie Zou. 2026. Intelligence degradation in long-context LLMs: Critical threshold determination via natural length distribution analysis.

Yifei Yu, Qian-Wen Zhang, Lingfeng Qiao, Di Yin, Fang Li, Jie Wang, Zengxi Chen, Suncong Zheng, Xiaolong Liang, and Xing Sun. 2025. Sequential-NIAH: A needle-in-a-haystack benchmark for extracting sequential needles from long contexts. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 29450–29468. Association for Computational Linguistics.

Jiajie Zhang, Yushi Bai, Xin Lv, Wanjun Gu, Danqing Liu, Minhao Zou, Shulin Cao, Lei Hou, Yuxiao Dong, Ling Feng, and Juanzi Li. 2024. LongCite: Enabling LLMs to generate fine-grained citations in long-context QA. *arXiv preprint arXiv:2409.02897*.

Appendix

A Prompts **11**

B Compositional Augmentation Results **11**

C Channel-Swap Robustness with BGE-M3 and ColBERT v2 **11**

D Factorial Ablation of Contextual Prefix, BM25, and Reranker **11**

E HyDE Weight Sweep across Per-Chunker α **11**

F Per-Dataset Observations **19**

G Qualitative Analysis of Contextual BM25 **20**

A Prompts

This appendix specifies the four language-model prompts that drive the indexing-time and query-time language-model calls of the pipeline. Figure 3 reports the Generic Prefix prompt of Anthropic (2024), applied as published in §3.1.1. Figures 4 and 5 report the two stages of the Definitional Prefix construction introduced in §3.1.2: Stage A extracts document-level metadata under a strict JSON schema, and Stage B synthesizes the chunk-level prefix under the six behavioral rules of §3.1.2. Figure 6 reports the HyDE query-side rewrite prompt of §3.2.2. All four prompts are invoked through `gpt-5-nano` at `reasoning_effort=minimal`, with no other prompt engineering or post-processing intervening between the model output and the downstream embedding or retrieval step.

B Compositional Augmentation Results

This appendix reports the per-dataset, per-k tables behind the compositional findings of §4.2. Table 3 sweeps BM25 hybrid retrieval $\pm$ cross-encoder reranking on both prefix families; Table 4 sweeps HyDE at the chunker-specific peak α.

C Channel-Swap Robustness with BGE-M3 and ColBERT v2

This appendix reports the full per-cell Precision@k and Recall@k tables for the channel-swap robustness check supporting Finding 2 (§4.2.2). The ALL-row summary in §4.2.2 establishes the channel-swap conclusion, and this appendix reports the full per-cell tables. Per-dataset structure observations are in Appendix F.

Per-dataset structure. The ALL-row generalization of the +BM25 / +Rerank negative result hides per-dataset shifts when the channel is swapped to a modern alternative. On **MAUD**, the three channels span a narrow band ($\sim$1–3 pp R@1) below prefix-only on both chunkers. On **CUAD**, modern channels dominate BM25 by 4–7$\times$ at R@1 in the no-rerank cells (Naive ColBERT v2 6.41% vs BM25 0.96%, RCTS ColBERT v2 3.47% vs BM25 0.73%, BGE-M3 sparse 3.73% / 3.00%). Both modern channels avoid the title-page entity trap to which the qualitative analysis (§G) attributes the BM25 CUAD failure. On **PrivacyQA**, the three channels again span a narrow band below prefix-only. On **ContractNLI**, channel performance is heterogeneous and chunker-dependent. BGE-M3 sparse and BM25 swap rank on RCTS NLI (14.90% vs 11.70%), while ColBERT v2 *underperforms* BM25 on Naive NLI (4.30% vs 11.15%). Two patterns generalize across datasets. The reranker on top of the swapped channel still does not reach prefix-only ALL R@1 in any cell. R@64 ceilings of both modern channels also sit below BM25's, indicating that the modern channels narrow rather than widen the candidate pool, the wrong direction for the small-target span-retrieval setting (§4.2.2, §5).

D Factorial Ablation of Contextual Prefix, BM25, and Reranker

This appendix reports the full $2\times2\times2$ factorial ablation behind §4.2, crossing (Generic Prefix on/off) $\times$ (BM25 hybrid on/off) $\times$ (cross-encoder reranker on/off) on the Naive chunker at `chunk_size=500`. The 8-cell cube is in Table 7. Two of the prefix-on cells are promoted to the main paper table, `naive_ctx_bm25_norerank` (Table 1 row 7) and `naive_ctx_bm25_rerank` (row 9). The remaining 6 cells are reported here so the BM25 / reranker subtraction effect can be compared across the no-prefix counterparts.

E HyDE Weight Sweep across Per-Chunker α

This appendix reports the full HyDE weight-sweep grid (Table 8). The per-chunker peaks reported as Table 4 are the bolded rows of Table 8.

No-prefix HyDE α-sweep. The 2×2 factorial result in §4.3 uses the prefix-on peak α for each chunker. Table 10 reports the parallel no-prefix HyDE α-sweep (6 points) so the lift trajectory can be compared cell-by-cell against Table 8. The no-prefix sweep peaks at $\alpha = 0.6$ on both chunkers, with HyDE-alone reaching ALL R@1 of 9.44 on Naive and 9.12 on RCTS, both well below the prefix-on-no-HyDE baselines of 12.25 and 10.06. Contextual retrieval carries the bulk of the lift on its own.

```
<document>
{full document text}
</document>
Here is the chunk we want to situate within the whole document
<chunk>
{chunk text}
</chunk>
Please give a short succinct context to situate this chunk within the overall document for the
purposes of improving search retrieval of the chunk. Answer only with the succinct context and
nothing else.
```

Figure 3: Generic Prefix prompt for the chunk-side intervention (§3.1.1), following (Anthropic, 2024). One `gpt-5-nano` call per (document, chunk) pair at indexing time.

```
You are a legal document analyst.  Read the following legal document and extract structured
metadata. Return ONLY a valid JSON object that conforms exactly to this schema (no markdown fences,
no commentary):
{
  "agreement_type": "<short string>",
  "parties": [{"canonical": "<role>", "full_name": "<entity>"}, ...],
  "governing_law": "<jurisdiction or null>",
  "section_outline": [{"number": "<X.Y>", "heading": "<text>"}, ...],
  "defined_terms": [{"term": "<exact term>", "gloss": "<short>"}, ...]
}
Include at most 15 defined_terms, ordered by how frequently they are reused later in the document
(most reused first). Include only top-level sections in section_outline, not sub-clauses. If a
field cannot be determined from the document, use null or an empty list, but do not omit the
field. Use canonical names that exactly match how the document refers to each party (e.g. 'Parent',
'Disclosing Party').
<document>
{doc}
</document>
```

Figure 4: Definitional Prefix Stage A prompt (§3.1.2). One `gpt-5-nano` call per document extracts structured metadata.

```
You write a one-line legal-document context prefix for a chunk. Six STRICT rules:
R1. TOTAL LENGTH under 350 characters. Single line, plain text, no markdown.
R2. PARTIES: copy each party from the parties list character-for-character. Do not retype from
memory, do not abbreviate, do not change capitalization.
R3. DEFINED TERMS: exactly one of two output forms.
  FORM_A (1-5 terms): ‘"Term1" (glossA); "Term2" (glossB)‘
  FORM_B (zero terms): ‘none‘
Never mix. Each term verbatim from metadata; each gloss is a ≤30-char substring of the metadata
gloss.
R4. SECTION: copy the ‘containing_section‘ argument verbatim. Do not modify, shorten, reformat,
or guess.
R5. NO INVENTION: no parties, terms, sections, governing law, or facts not in the metadata or
chunk.
R6. NO META-COMMENTARY: no parenthetical remarks about these rules.
DOCUMENT METADATA:
  agreement_type: {agreement_type}
  parties: {parties_str}
  governing_law: {governing_law}
  defined_terms (use these verbatim, with their glosses):
{defined_terms_with_gloss_block}
CHUNK'S CONTAINING SECTION (deterministic, from regex parser — copy verbatim):
  {containing_section}
CHUNK TEXT:
<chunk>
{chunk}
</chunk>
OUTPUT FORMAT (one line, target ∼300 chars, no surrounding quotes):
In an {agreement_type} between {parties_list_for_template}. Section: {containing_section}. Terms:
FORM_A_or_FORM_B_per_R3. Topic: <one short clause>.
```

Figure 5: Definitional Prefix Stage B prompt (§3.1.2). One gpt-5-nano call per chunk synthesizes the prefix from Stage A metadata plus a deterministic section label.

```
You are a legal AI assistant. Given a question about a legal document, write a single hypothetical
passage (3-5 sentences, written in declarative legal-document style) that, if it appeared in the
source document, would directly answer the question.
Do NOT answer the question yourself. Do NOT add commentary. Just write the hypothetical passage as
if it were extracted verbatim from a contract, agreement, privacy policy, or NDA. Use legal-register
vocabulary and phrasing typical of the document type implied by the question.
Question: {query}
Hypothetical passage:
```

Figure 6: HyDE prompt for the query-side rewrite (§3.2.2). One gpt-5-nano call per query generates a hypothetical answer passage in declarative legal-document register.

Dataset	Method	Precision@k							Recall@k						
		1	2	4	8	16	32	64	1	2	4	8	16	32	64
MAUD	Naive + Generic + BM25	3.29	3.02	2.14	1.87	1.57	1.25	0.95	2.04	3.36	4.82	8.36	12.94	19.97	29.68
	RCTS + Generic + BM25	2.31	2.03	1.54	1.60	1.69	1.59	1.37	0.80	0.99	1.36	4.67	7.84	11.94	19.85
	Naive + Generic + BM25 + Rerank	6.40	4.02	2.99	2.84	2.09	1.51	0.95	3.66	4.62	6.35	10.48	15.39	23.63	29.68
	RCTS + Generic + BM25 + Rerank	5.44	4.07	3.66	3.06	2.32	1.91	1.37	2.33	2.91	5.17	8.56	10.85	16.51	19.85
	Naive + Definitional + BM25	12.97	12.15	9.38	6.94	4.87	3.09	1.85	5.66	10.37	15.62	22.95	31.62	39.65	48.06
	RCTS + Definitional + BM25	10.43	10.78	8.33	7.19	5.07	3.40	2.20	3.47	6.41	8.80	14.60	20.02	25.10	31.81
	Naive + Definitional + BM25 + Rerank	17.80	13.45	10.36	7.42	4.89	3.16	1.85	7.70	12.13	17.62	24.24	31.72	40.76	48.06
	RCTS + Definitional + BM25 + Rerank	18.92	13.86	10.55	7.21	4.80	3.23	2.20	7.37	10.29	13.68	17.97	22.02	26.80	31.81
CUAD	Naive + Generic + BM25	1.37	3.69	3.61	2.97	2.74	2.00	1.24	0.96	6.86	15.94	26.63	46.49	61.88	75.55
	RCTS + Generic + BM25	1.02	1.36	2.46	2.61	3.33	2.70	1.82	0.73	3.01	7.83	15.08	32.42	47.23	65.08
	Naive + Generic + BM25 + Rerank	5.38	4.73	4.74	3.63	2.98	2.08	1.24	6.81	12.98	24.32	35.12	50.68	66.59	75.55
	RCTS + Generic + BM25 + Rerank	8.29	6.06	5.04	4.53	4.76	2.95	1.82	6.83	9.51	15.39	23.51	37.81	51.62	65.08
	Naive + Definitional + BM25	6.18	6.17	5.77	4.50	3.47	2.42	1.48	6.78	14.04	25.24	38.01	56.51	74.44	89.71
	RCTS + Definitional + BM25	9.29	7.57	5.97	6.13	5.31	3.53	2.13	5.52	9.35	14.47	28.61	49.82	64.74	79.83
	Naive + Definitional + BM25 + Rerank	16.92	13.35	8.74	5.34	3.35	2.39	1.48	21.19	31.74	40.01	48.03	57.08	75.75	89.71
	RCTS + Definitional + BM25 + Rerank	18.16	14.41	9.82	6.58	4.54	3.03	2.13	15.40	23.32	33.19	43.48	55.14	66.44	79.83
PrivacyQA	Naive + Generic + BM25	15.26	11.79	10.31	8.12	6.43	4.75	3.07	10.97	17.24	24.92	35.00	53.26	70.08	84.90
	RCTS + Generic + BM25	11.83	11.33	9.48	8.13	6.44	4.99	3.41	5.41	9.51	15.65	27.10	42.46	61.74	77.39
	Naive + Generic + BM25 + Rerank	22.75	17.66	15.32	12.10	8.55	5.51	3.07	12.95	18.83	30.52	47.28	62.87	77.88	84.90
	RCTS + Generic + BM25 + Rerank	19.55	18.30	16.99	12.87	9.03	6.04	3.41	9.77	15.60	27.06	37.56	55.64	70.48	77.39
	Naive + Definitional + BM25	14.09	15.52	12.50	10.51	7.86	5.36	3.48	10.81	22.67	33.30	50.33	66.97	83.98	97.58
	RCTS + Definitional + BM25	17.12	16.17	12.80	10.56	8.23	6.05	4.14	10.41	18.31	26.15	41.47	57.55	75.76	90.07
	Naive + Definitional + BM25 + Rerank	25.25	22.77	19.49	14.07	9.15	5.76	3.48	16.97	28.28	45.48	61.81	75.82	87.90	97.58
	RCTS + Definitional + BM25 + Rerank	26.68	23.44	20.35	16.02	10.96	6.77	4.14	15.12	25.75	39.84	56.59	69.95	82.24	90.07
ContractNLI	Naive + Generic + BM25	9.35	7.89	6.29	4.36	2.89	1.83	1.12	11.15	18.15	28.58	39.20	51.60	63.40	75.64
	RCTS + Generic + BM25	10.68	8.16	7.62	5.15	3.55	2.43	1.46	11.70	17.02	29.37	35.54	48.16	63.16	73.62
	Naive + Generic + BM25 + Rerank	8.53	7.51	5.42	4.12	2.85	1.84	1.12	9.89	17.35	24.37	35.21	47.41	61.12	75.64
	RCTS + Generic + BM25 + Rerank	10.63	8.27	6.21	4.89	3.39	2.33	1.46	10.66	14.98	22.22	33.16	44.78	59.35	73.62
	Naive + Definitional + BM25	17.90	13.28	9.91	6.77	4.37	2.51	1.34	20.60	29.80	44.13	60.10	74.73	83.56	89.62
	RCTS + Definitional + BM25	21.82	15.91	11.41	7.75	5.53	3.25	1.81	20.66	28.74	39.44	50.91	67.94	76.96	83.49
	Naive + Definitional + BM25 + Rerank	20.44	15.41	11.07	7.21	4.45	2.49	1.34	24.16	35.18	48.47	61.63	74.02	83.80	89.62
	RCTS + Definitional + BM25 + Rerank	26.92	19.29	12.56	8.33	5.38	3.29	1.81	26.19	35.89	44.95	54.41	65.58	74.45	83.49
ALL	Naive + Generic + BM25	7.32	6.60	5.59	4.33	3.41	2.46	1.59	6.28	11.40	18.57	27.30	41.07	53.83	66.44
	RCTS + Generic + BM25	6.46	5.72	5.27	4.37	3.75	2.93	2.02	4.66	7.64	13.55	20.60	32.72	46.02	58.98
	Naive + Generic + BM25 + Rerank	10.77	8.48	7.12	5.67	4.12	2.74	1.59	8.33	13.45	21.39	32.02	44.09	57.31	66.44
	RCTS + Generic + BM25 + Rerank	10.98	9.17	7.98	6.34	4.88	3.30	2.02	7.40	10.75	17.46	25.70	37.27	49.49	58.98
	Naive + Definitional + BM25	12.78	11.78	9.39	7.18	5.15	3.34	2.04	10.96	19.22	29.57	42.85	57.46	70.41	81.24
	RCTS + Definitional + BM25	14.66	12.61	9.63	7.91	6.03	4.06	2.57	10.01	15.70	22.22	33.89	48.83	60.64	71.30
	Naive + Definitional + BM25 + Rerank	20.10	16.25	12.42	8.51	5.46	3.45	2.04	17.51	26.83	37.90	48.93	59.66	72.05	81.24
	RCTS + Definitional + BM25 + Rerank	22.67	17.75	13.32	9.53	6.42	4.08	2.57	16.02	23.81	32.92	43.11	53.17	62.48	71.30

Table 3: BM25 hybrid retrieval and optional cross-encoder reranking, layered on top of the Generic Prefix (§3.1.1, top 4 rows per dataset) and on top of the Definitional Prefix (§3.1.2, bottom 4 rows per dataset). Reference values for the prefix-only rows are in Table 1. Channel-swap variants (BGE-M3 sparse, ColBERT v2) are in Appendix C.

Dataset	Method	Precision@k							Recall@k						
		1	2	4	8	16	32	64	1	2	4	8	16	32	64
MAUD	Naive + Generic + HyDE	7.15	5.84	5.64	3.92	2.99	2.29	1.63	4.45	8.10	12.13	16.46	25.64	37.91	49.77
	RCTS + Generic + HyDE	8.51	6.80	5.28	4.19	3.34	2.62	2.02	3.07	4.84	8.30	14.64	22.60	33.11	45.44
	Naive + Definitional + HyDE	**22.06**	**17.12**	**13.07**	**8.85**	**5.89**	**3.72**	**2.26**	**9.02**	**13.25**	**20.84**	**27.44**	**37.28**	**46.97**	**58.48**
	RCTS + Definitional + HyDE	**16.89**	**14.82**	**11.90**	**8.16**	**5.78**	**3.78**	**2.47**	**7.26**	**10.41**	**14.28**	**18.37**	**25.91**	33.06	40.47
CUAD	Naive + Generic + HyDE	19.26	13.00	9.80	6.61	4.13	2.48	1.46	23.92	30.98	42.63	52.03	62.97	74.59	86.35
	RCTS + Generic + HyDE	24.82	18.97	14.31	10.43	8.44	4.45	2.30	23.44	29.08	37.58	45.16	55.62	69.14	82.14
	Naive + Definitional + HyDE	**22.44**	**16.92**	**10.36**	**6.95**	**4.53**	**2.77**	**1.54**	**29.07**	**41.52**	**46.57**	**57.62**	**71.76**	**82.24**	**90.43**
	RCTS + Definitional + HyDE	**28.77**	**22.02**	**14.55**	9.25	6.10	3.71	2.21	**24.99**	**35.68**	**42.49**	**50.20**	**60.72**	**71.74**	**84.83**
PrivacyQA	Naive + Generic + HyDE	33.39	29.61	21.46	14.10	9.53	5.84	3.30	21.79	34.11	46.50	56.00	70.59	81.93	89.99
	RCTS + Generic + HyDE	30.37	25.85	20.36	15.95	10.27	6.52	3.84	15.62	22.89	32.27	47.24	61.20	76.14	87.29
	Naive + Definitional + HyDE	**37.44**	**30.78**	**21.84**	**15.15**	**10.19**	**6.22**	**3.55**	**23.00**	**37.40**	**50.81**	**62.81**	**80.17**	**92.12**	**98.77**
	RCTS + Definitional + HyDE	**32.14**	24.63	**21.27**	15.79	**11.19**	**7.35**	**4.72**	**20.16**	**31.22**	**44.40**	**56.16**	**70.83**	**85.71**	**96.30**
ContractNLI	Naive + Generic + HyDE	11.90	8.95	6.52	4.33	3.04	1.97	1.25	13.26	18.95	28.73	38.05	53.51	68.60	84.13
	RCTS + Generic + HyDE	9.60	8.52	7.59	5.65	4.19	2.79	1.63	8.95	15.31	26.31	37.91	54.67	70.28	80.92
	Naive + Definitional + HyDE	**19.13**	**15.75**	**11.13**	**7.50**	**4.45**	**2.45**	**1.34**	**23.17**	**36.59**	**49.02**	**62.09**	**73.32**	**81.56**	**88.74**
	RCTS + Definitional + HyDE	**30.38**	**22.53**	**15.91**	**9.74**	**5.89**	**3.48**	**1.88**	**28.10**	**36.77**	**50.53**	**58.76**	**69.38**	**78.96**	**85.56**
ALL	Naive + Generic + HyDE	17.93	14.35	10.85	7.24	4.93	3.14	1.91	15.85	23.04	32.50	40.63	53.18	65.76	77.56
	RCTS + Generic + HyDE	18.33	15.03	11.89	9.05	6.56	4.09	2.45	12.77	18.03	26.12	36.24	48.52	62.17	73.95
	Naive + Definitional + HyDE	**25.27**	**20.14**	**14.10**	**9.61**	**6.27**	**3.79**	**2.17**	**21.07**	**32.19**	**41.81**	**52.49**	**65.63**	**75.72**	**84.10**
	RCTS + Definitional + HyDE	**27.04**	**21.00**	**15.91**	**10.74**	**7.24**	**4.58**	**2.82**	**20.13**	**28.52**	**37.92**	**45.87**	**56.71**	**67.37**	**76.79**

Table 4: HyDE query-side rewrite layered on top of the Generic Prefix and on top of the Definitional Prefix at the chunker-specific peak α (Generic Naive $\alpha = 0.7$, RCTS $\alpha = 0.6$; Definitional Naive $\alpha = 0.7$, RCTS $\alpha = 0.8$). Bold and underlined cells mark the Definitional + HyDE rows that exceed the corresponding Generic + HyDE row for that (dataset, k).

Sparse / multi-vector channel	Precision@k (%)							Recall@k (%)						
	1	2	4	8	16	32	64	1	2	4	8	16	32	64
Generic Prefix														
BM25	7.32	6.60	5.59	4.33	3.41	2.46	1.59	6.28	11.40	18.57	27.30	41.07	53.83	66.44
BM25 + Rerank	10.77	8.48	7.12	5.67	4.12	2.74	1.59	8.33	13.45	21.39	32.02	44.09	57.31	66.44
BGE-M3 sparse	6.69	6.13	5.39	4.41	3.61	2.56	1.64	6.65	11.73	18.96	28.17	43.04	55.64	67.29
BGE-M3 sparse + Rerank	10.83	8.68	7.36	5.69	4.10	2.78	1.64	8.26	13.55	21.93	32.12	44.13	58.12	67.29
ColBERT v2	7.48	6.80	5.87	4.70	3.50	2.38	1.51	5.94	10.68	18.28	28.64	39.63	51.07	62.69
ColBERT v2 + Rerank	10.91	8.70	7.48	5.96	4.34	2.81	1.51	8.58	13.56	22.51	33.67	46.38	59.45	62.69
Definitional Prefix														
BM25	12.78	11.78	9.39	7.18	5.15	3.34	2.04	10.96	19.22	29.57	42.85	57.46	70.41	81.24
BM25 + Rerank	20.10	16.25	12.42	8.51	5.46	3.45	2.04	17.51	26.83	37.90	48.93	59.66	72.05	81.24
BGE-M3 sparse	13.31	11.99	9.90	7.35	5.19	3.40	2.04	12.30	21.01	31.40	44.47	58.63	72.03	80.58
BGE-M3 sparse + Rerank	20.03	16.51	12.61	8.62	5.54	3.44	2.04	17.49	27.10	38.27	49.32	59.85	71.68	80.51
ColBERT v2	11.51	11.39	9.16	7.12	4.75	3.09	1.86	10.18	18.53	29.44	42.35	53.09	66.26	75.94
ColBERT v2 + Rerank	19.97	16.37	12.48	8.64	5.67	3.53	1.86	17.55	27.10	38.02	49.32	61.06	73.11	75.94

Table 5: Naive chunker @ `chunk_size=500`. BM25, BGE-M3 sparse (Chen et al., 2024), and ColBERT v2 (Santhanam et al., 2022) as the sparse / multi-vector channel, layered with the Generic Prefix (top half) and the Definitional Prefix (bottom half), with optional cross-encoder reranking. The chunker, dense embedding (`text-embedding-3-large`), and RRF combine (k_{rrf}=60, equal weights) are held identical across rows. Reference values for the prefix-only conditions are Generic ALL R@1 = 12.25 and Definitional ALL R@1 = 18.29 (Table 1 rows 5 and 7). Per-dataset structure is in Appendix F.

Sparse / multi-vector channel	Precision@k (%)							Recall@k (%)						
	1	2	4	8	16	32	64	1	2	4	8	16	32	64
Generic Prefix														
BM25	6.46	5.72	5.27	4.37	3.75	2.93	2.02	4.66	7.64	13.55	20.60	32.72	46.02	58.98
BM25 + Rerank	10.98	9.17	7.98	6.34	4.88	3.30	2.02	7.40	10.75	17.46	25.70	37.27	49.49	58.98
BGE-M3 sparse	8.20	7.21	6.00	4.86	4.17	2.95	2.00	6.84	11.40	16.77	23.93	36.20	48.42	61.12
BGE-M3 sparse + Rerank	10.92	9.29	8.06	6.37	5.00	3.30	2.00	7.34	10.75	17.30	25.30	37.88	50.62	61.12
ColBERT v2	7.29	6.51	5.79	4.51	3.75	2.68	1.82	5.15	9.01	15.45	21.48	32.11	44.30	56.37
ColBERT v2 + Rerank	11.23	9.36	8.30	6.55	5.14	3.54	1.82	7.57	10.91	17.66	26.00	38.84	52.98	56.37
Definitional Prefix														
BM25	14.66	12.61	9.63	7.91	6.03	4.06	2.57	10.01	15.70	22.22	33.89	48.83	60.64	71.30
BM25 + Rerank	22.67	17.75	13.32	9.53	6.42	4.08	2.57	16.02	23.81	32.92	43.11	53.17	62.48	71.30
BGE-M3 sparse	16.14	13.10	9.93	7.72	5.58	3.85	2.44	12.60	19.19	27.30	38.82	50.49	62.92	72.96
BGE-M3 sparse + Rerank	22.91	18.09	13.42	9.60	6.38	3.99	2.44	16.09	23.96	33.00	43.28	53.43	63.36	72.94
ColBERT v2	14.94	13.31	10.13	8.16	5.52	3.63	2.26	9.86	17.53	25.50	37.31	47.50	58.18	67.34
ColBERT v2 + Rerank	22.83	18.32	13.61	9.85	6.49	4.25	2.26	16.08	24.41	33.51	43.42	54.14	64.12	67.40

Table 6: RCTS chunker @ `chunk_size=500`, same construction as Table 5. Reference values for the prefix-only conditions are Generic ALL R@1 = 10.06 and Definitional ALL R@1 = 17.70 (Table 1 rows 6 and 8). Per-dataset structure is in Appendix F.

Chunk	+Generic	+BM25	+Rerank	R@1	P@1	R@8	R@64
Naive	✗	✗	✗	7.06	7.42	31.92	65.56
Naive	✗	✗	✓	4.99	6.52	27.58	65.56
Naive	✗	✓	✗	3.01	2.87	19.75	59.65
Naive	✗	✓	✓	4.97	6.50	26.98	59.65
Naive	✓	✗	✗	12.25	14.00	37.48	71.98
Naive	✓	✗	✓	8.37	10.79	32.96	71.98
Naive	✓	✓	✗	6.28	7.32	27.30	66.44
Naive	✓	✓	✓	8.33	10.77	32.02	66.44

Table 7: $2\times2\times2$ factorial ablation cube on the Naive chunker @ `chunk_size=500`. "+Generic" is the Generic Prefix (§3.1.1), "+BM25" is the sparse-dense hybrid via RRF (§3.2.1), and "+Rerank" is the `rerank-english-v3.0` cross-encoder reranker (§3.2.1). All combinations of (+Generic, +BM25, +Rerank) are reported so the BM25 / reranker subtraction effect can be read against the no-prefix baseline.

		Naive @ 500			RCTS @ 500		
α (raw)	$1-\alpha$ (HyDE)	R@1	R@8	R@64	R@1	R@8	R@64
1.00	0.00	12.25	37.48	71.98	10.06	29.42	64.98
0.95	0.05	12.82	39.49	73.02	10.11	30.67	66.33
0.90	0.10	13.60	39.80	74.19	10.30	31.86	67.27
0.85	0.15	14.27	40.13	75.19	11.01	32.79	69.09
0.80	0.20	14.63	39.87	75.99	11.97	33.77	70.20
0.75	0.25	15.09	40.50	76.80	11.73	35.19	71.21
0.70	**0.30**	**15.85**	**40.63**	77.56	12.23	35.48	72.21
0.60	**0.40**	15.66	40.35	**77.78**	**12.75**	**36.39**	**73.89**
0.50	0.50	12.31	39.35	76.59	11.68	34.96	73.11

Table 8: HyDE weight sweep on the convex combination $\alpha \cdot e(q) + (1-\alpha) \cdot e(\text{HyDE}(q))$ at retrieval time, on top of the Generic Prefix. ALL row only; chunker, prefix, embedding, and chunk pool are identical to Table 1 rows 5–6. Bolded rows are the per-chunker peaks (Naive $\alpha = 0.7$, RCTS $\alpha = 0.6$) reported as the Generic + HyDE rows of Table 4. The parallel Definitional Prefix weight sweep is in Table 9.

		Naive @ 500, Definitional			**RCTS @ 500, Definitional**		
α (raw)	$1-\alpha$ (HyDE)	R@1	R@8	R@64	R@1	R@8	R@64
1.00	0.00	18.29	49.84	81.85	17.70	43.22	74.02
0.90	0.10	19.53	51.96	83.28	19.11	44.43	75.33
0.80	0.20	20.85	52.20	83.86	**20.13**	**45.87**	**76.79**
0.70	**0.30**	**21.07**	**52.49**	**84.10**	19.50	46.77	76.83
0.60	0.40	19.71	51.86	83.35	18.31	46.68	76.01
0.50	0.50	17.50	49.80	81.51	15.72	44.25	74.42

Table 9: HyDE weight sweep on top of the Definitional Prefix (parallel to Table 8 which sweeps on top of the Generic Prefix). Chunker, dense embedding, and chunk pool are identical to Table 1 rows 7–8. The $\alpha = 1$ row reproduces the Definitional Prefix alone. Bolded rows are the per-chunker peaks (Naive $\alpha = 0.7$, RCTS $\alpha = 0.8$) reported as the Definitional + HyDE rows of Table 4.

	Naive, no prefix			RCTS, no prefix		
α	R@1	R@8	R@64	R@1	R@8	R@64
1.00	7.06	31.92	65.56	7.69	31.68	64.69
0.90	7.21	33.11	68.15	7.57	34.06	67.11
0.80	7.98	35.43	70.22	7.87	36.09	68.97
0.70	8.72	35.06	71.11	8.52	36.30	70.52
0.60	**9.44**	33.67	70.12	**9.12**	**35.59**	**70.85**
0.50	8.58	31.53	69.79	9.05	32.21	69.81

Table 10: HyDE α-sweep with prefix off (parallels Table 8). Peak at $\alpha = 0.6$ on both chunkers, both below the prefix-on-no-HyDE baselines (Table 2).

F Per-Dataset Observations

MAUD (M&A agreements). NAIVE + GENERIC reaches MAUD R@1 = 4.72% (+2.18 pp over the Naive baseline, 1.9×). The Generic Prefix surfaces the discriminating tokens (parties, defined terms, parent Article) in the embedded representation, partially counteracting the sibling-definition cosine collisions that limit dense embedding alone. Both +BM25 variants drop MAUD R@1 to $< 3\%$ (rows 7 and 8 at 2.04 / 0.80%), and the reranker rescues to 3.66 / 2.33% (rows 9 and 10) but neither configuration recovers prefix-only. HyDE (Table 4) lifts MAUD at all higher k (Naive R@8 11.35 → 16.46, +5.11 pp, and R@64 37.03 → 49.77, +12.74 pp), with R@1 within noise of prefix-only on both chunkers (Naive 4.72 → 4.45, −0.27 pp, and RCTS 3.28 → 3.07, −0.21 pp). MAUD definitional queries reference specific defined terms and party tickers that HyDE's paraphrased passages do not preserve precisely at top-1, but the broader candidate pool is uniformly enriched.

CUAD (commercial contracts). The CUAD corpus is the broadest in document size and template heterogeneity. On short ($<$ 3K chars) license amendments the Naive baseline's 500-char window incidentally captures the answer span at k=1 on ∼25% of the dataset, an artefact the Generic Prefix variant does not match. NAIVE + GENERIC therefore reaches CUAD R@1 = 10.14% (−2.46 pp vs the Naive baseline), and from R@4 onward the Generic Prefix variant dominates (+1.5 pp at R@4, +5.95 pp at R@64). The +BM25 layer crashes CUAD R@1 to 0.96% on Naive + Generic + BM25 (row 7) and 0.73% on RCTS + Generic + BM25 (row 8), a −9+ pp drop from prefix-only, because CUAD's commercial-contract boilerplate is particularly dense and uniform across chunks. The reranker recovers CUAD R@1 to 6.81 / 6.83% on rows 9 and 10, still below prefix-only. HyDE (Table 4) is the largest single-cell improvement of any method reported, lifting CUAD R@1 10.14 → 23.92 (+13.78 pp) on Naive and 7.66 → 23.44 (+15.78 pp) on RCTS, roughly 2× on both chunkers. Short interrogative CUAD queries against heterogeneous commercial-contract clauses face the largest query–chunk distribution gap in the benchmark. HyDE's hypothetical rewrite (e.g., expanding *"Is there a cap on liability?"* into a multi-sentence draft of a liability-cap clause) closes most of that gap and breaks the near-ties between answer chunks and sibling boilerplate sections that prefix-only retrieval was struggling with.

PrivacyQA (privacy policies). PrivacyQA is the cleanest Generic Prefix win. NAIVE + GENERIC reaches PrivacyQA R@1 = 19.33% (+11.88 pp over the Naive baseline, 2.6×) and saturates R@64 at 88.41% (+22.34 pp). The 7-document corpus has small documents and informal subheadings, and the Generic Prefix surfaces topic and section markers that the dense embedding alone misses. The Contextual BM25 / reranker full-procedure rows behave as elsewhere. Row 7 (Naive + Generic + BM25) reaches PrivacyQA R@1 = 10.97% (−8.36 pp vs row 5), and row 9 (+Rerank) climbs back to 12.95% (−6.38 pp), but neither recovers prefix-only. HyDE (Table 4) is uniformly positive, lifting PrivacyQA R@1 19.33 → 21.79 (+2.46 pp) on Naive and 12.73 → 15.62 (+2.89 pp) on RCTS, with bolded gains across every cell of the PrivacyQA block. Short interrogative privacy queries gain consistently from the legal-register hypothetical rewrite.

ContractNLI (NDAs). NAIVE + GENERIC reaches ContractNLI R@1 = 14.82% (+3.50 pp over the Naive baseline). The improvement is moderate at k=1 because NDA templates already carry distinctive topic labels ("CONFIDENTIAL INFORMATION", "RETURN OR DESTRUCTION") that the dense embedding handles well, and the Generic Prefix mostly helps at higher k (+11.18 pp at k=2). The +BM25 layer (row 7) drops ContractNLI R@1 to 11.15% (−3.67 pp), and +BM25+Rerank (row 9) lands at 9.89% (−4.93 pp). ContractNLI also exhibits the sharpest HyDE-by-prefix interaction (§5.3). On the Generic Prefix, HyDE regresses on R@1 (Naive 14.82 → 13.26, −1.56 pp, and RCTS 16.57 → 8.95, −7.62 pp), because NDA queries are highly templatic (*"Does the document state that Confidential Information shall include all technical information?"*) and already lexically aligned with chunk language (*"Confidential Information shall include…"* appears near-verbatim in the source documents), so HyDE's hypothetical rewrite shifts the query away from this naturally tuned alignment. The Definitional Prefix's chunk-side gloss supplies the surface alignment that HyDE undoes for the Generic case, and Definitional + HyDE lands at 24.62 → 23.17 on Naive (within noise) and recovers on RCTS at 26.73 → 28.10 (Table 4).

G Qualitative Analysis of Why Contextual BM25 Subtracts Value

This appendix presents one worked example per sub-dataset showing how the Contextual BM25 layer (Table 1 row 7) shifts the top-1 retrieval to a wrong chunk in the same document, even though the prefix-only baseline (row 5) had retrieved the ground-truth span correctly. Each figure displays the same query under two retrieval architectures, with the ground-truth answer text in **bold** and the in-chunk text outside the ground truth in *gray italic*. Each example is selected from the set of queries where row 5's per-query Recall@1 exceeds 0.5 but row 7's drops below 0.05, a set that contains 11 PrivacyQA, 10 ContractNLI, 20 CUAD, and 6 MAUD queries (47 total). The four figures together exercise one recurring failure mode. *The BM25 channel is lured by the high token density of document-level identifiers (title block, parties, document type) that the query naturally repeats*, and surfaces the chunk where those identifiers appear most densely (typically the title page or parties block) instead of the substantive clause that contains the answer. The cross-encoder reranker layered on top (row 9) recovers only when the correct chunk survived BM25's depression to within the top-K=64 candidate pool. Figure 8 (Papa John's) is the one rescue case, while Figures 9, 10, 7 all remain mis-retrieved after rerank.

Query. *Consider the Merger Agreement between "Avangrid, Inc." and "PNM Resources, Inc."; What is the Type of Consideration*

Ground truth. `maud/PNM Resources. . . _Avangrid, Inc..txt` characters $[22{,}033, 22{,}883]$ (850 chars), the §2.1(a) Merger Consideration definition.

(a) Naive + Generic (row 5)
prefix-only dense top-1: chunk $[22{,}000, 22{,}500]$

k or Merger Sub Common Stock:
(a) Merger Consideration. Each share of Company Common Stock issued and outstanding immediately prior to the Effective Time (each, a "Company Share") (other than Company Shares owned by Parent, Merger Sub or any other direct or indirect wholly-owned subsidiary of Parent and Company Shares owned by the Company or any of its wholly-owned subsidiaries as treasury stock or otherwise, and in each case not held on behalf of third parties (collectively, the "Cancelled S

500 chars retrieved; 467 of the 850 ground-truth chars (**bold**) are inside the chunk. The dense channel matches the query's "Type of Consideration" onto the §2.1(a) heading "Merger Consideration" and the topical content "share of Company Common Stock".

P@1 = 93.4% R@1 = 54.9%

(b) Naive + Generic + BM25 (row 7)
Generic + BM25 + RRF top-1: chunk $[6{,}500, 7{,}000]$

AGREEMENT AND PLAN OF MERGER
This AGREEMENT AND PLAN OF MERGER, dated as of October 20, 2020 (this "Agreement"), is entered into among Avangrid, Inc., a New York corporation ("Parent"), NM Green Holdings, Inc., a New Mexico corporation and a direct subsidiary of Parent ("Merger Sub"), and PNM Resources, Inc., a New Mexico corporation (the "Company" and, together with Parent and Merger Sub, the "Parties" and each, a "Party").
RECITALS
WHEREAS, the board of directors of the Company (t

500 chars retrieved; *zero* ground-truth overlap. The query string "Merger Agreement between Avangrid, Inc. and PNM Resources, Inc." is a near-verbatim match for this title-and-parties chunk; BM25 ranks it first regardless of the substantive query phrase "Type of Consideration", which has no lexical match here.

P@1 = 0.0% R@1 = 0.0%

Figure 7: **MAUD example.** The query's verbatim mention of both parties' names ranks the title-and-parties chunk above the substantive §2.1(a) Merger Consideration clause. The reranker on top of +BM25 (row 9) returns chars $[15{,}500, 16{,}000]$, a different wrong chunk in the recitals/preamble (R@1 = 0%).

Query. *Consider the Endorsement Agreement between ABG-Shaq, LLC and Papa John's International, Inc.; Is there a cap on liability under this contract?*
Ground truth. `cuad/PapaJohnsInternationalInc. . . _Endorsement Agreement.txt` characters $[58{,}826, 59{,}464]$ (638 chars), the limit-of-liability clause.

(a) Naive + Generic (row 5)
prefix-only dense top-1: chunk $[59{,}000, 59{,}500]$

TAL, PUNITIVE, EXEMPLARY, OR SPECIAL DAMAGES, REGARDLESS OF THE FORM OR ACTION, WHETHER IN CONTRACT OR IN TORT, EVEN IF ABG OR AUTHENTIC BRANDS GROUP LLC HAS BEEN ADVISED OF THE POSSIBLITY OF SUCH DAMAGES OR LOSSES. IN NO EVENT SHALL ABG'S, AUTHENTIC BRANDS GROUP LLC'S, AND CELEBRITY'S TOTAL LIABILITY UNDER THIS AGREEMENT EXCEED THE AMOUNTS ACTUALLY RECEIVED BY ABG (EXCLUSIVE OF REIUMBURSEMENT OF EXPENSES) HEREUNDER, REGARDLESS OF THE NUMBER OR TYPE OF CLAIMS. *9. Arbitration. In the event a dis*

500 chars retrieved; 464 of the 638 ground-truth chars (**bold**) are inside the chunk. The dense channel matches the query's "cap on liability" onto the semantic content "TOTAL LIABILITY . . . EXCEED THE AMOUNTS".
P@1 = 92.8% R@1 = 72.7%

(b) Naive + Generic + BM25 (row 7)
Generic + BM25 + RRF top-1: chunk $[0, 500]$

Exhibit 10.1
ENDORSEMENT AGREEMENT
THIS ENDORSEMENT AGREEMENT ("Agreement") is made and entered into effective March 15, 2019 (the "Effective Date") by and among, on the one hand, ABG-Shaq, LLC, a Delaware limited liability company ("ABG") for the personal services of Shaquille O'Neal, ("CELEBRITY"), and, on the other hand, Papa John's Marketing Fund, Inc., a Kentucky corporation ("PJMF"), and Papa John's International, Inc.
("PJI") (PJMF and PJI are. . .

500 chars retrieved; *zero* ground-truth overlap. The query string "Endorsement Agreement between ABG-Shaq, LLC and Papa John's International, Inc." is verbatim contained in this title-page chunk, so BM25 ranks it first by a wide margin. The query phrase "cap on liability" has no lexical match anywhere in the chunk.
P@1 = 0.0% R@1 = 0.0%

Figure 8: **CUAD example, the one reranker-rescue case.** The reranker on top of +BM25 (row 9) returns the top-1 to chars $[59{,}000, 59{,}500]$, restoring R@1 $= 72.7\%$. The rescue is contingent on the liability clause having survived BM25's depression to within the top-K=64 candidate pool. The aggregate ALL R@1 (row 9 $= 8.33\%$) remains below row 5 ($= 12.25\%$).

Query. *Consider "Groupon"'s privacy policy; is my data safe?*
Ground truth. `privacy_qa/Groupon.txt` characters $[13{,}561, 14{,}100]$ (539 chars), the body of the security-program clause beginning "Groupon has implemented an information security program…".

(a) Naive + Generic (row 5)
prefix-only dense top-1: chunk $[13{,}500, 14{,}000]$

ervices for the Groupon application in iOS privacy settings. **Groupon has implemented an information security program that contains administrative, technical and physical controls that are designed to reasonably safeguardPersonal Information. For example, we use industry-standard encryption technology to secureFinancial Account Information. No method of transmission over the Internet, or method of electronic storage, is 100% secure, however. Therefore, we cannot guarantee its absolute security. I**

500 chars retrieved; 439 of the 539 ground-truth chars (**bold**) are inside the chunk. The dense embedding of the prefix-augmented chunk maps the query phrase "is my data *safe*" onto the semantic match "safeguard" inside the security-program clause.

P@1 = 87.8% R@1 = 81.4%

(b) Naive + Generic + BM25 (row 7)
Generic + BM25 + RRF top-1: chunk $[20{,}000, 20{,}500]$

at privacy@groupon.com with questions. Groupon utilizes various valid data transfer mechanisms for Personal Information at its sole discretion in accordance with applicable data protection and privacy laws. We will continue to monitor and assess the appropriateness of other valid data transfer mechanisms as they become available. Pursuant to Section 1798.83 of the California Civil Code, residents of California can obtain certain information about the types of personal information that companies

500 chars retrieved; *zero* ground-truth overlap. The query's high-IDF tokens "data" and "privacy" co-occur at high frequency in this cross-border data-transfer compliance chunk ("data" $\times 5$, "privacy" $\times 2$), so BM25 ranks it above the security-program chunk. BM25 has no semantic match for "safe / safeguard" and the lexical channel pulls the retrieval the wrong way.

P@1 = 0.0% R@1 = 0.0%

Figure 9: **PrivacyQA example.** Adding Contextual BM25 on top of the prefix-only retrieval re-ranks the data-transfer compliance chunk above the security-program chunk, dropping Recall@1 from 81.4% to 0%. The reranker on top of +BM25 (row 9) returns chars $[0, 500]$ (zero ground-truth overlap).

Query. *Consider Grindrod SA's Non-Disclosure Agreement; Does the document state that Confidential Information shall only include technical information?*

Ground truth. `contractnli/Grindrod. . . Non-Disclosure Undertaking.txt` characters $[757, 1{,}492]$ (735 chars), the §1.1 definition of "Confidential Information".

(a) Naive + Generic (row 5)
prefix-only dense top-1: chunk $[1{,}000, 1{,}500]$

d marketing plans business or personnel information disclosed or otherwise made available in any format and/or physical manner by Grindrod SA or becoming available, before, during and/or after the execution of an interaction, duty or obligation including all information that makes itself known to the Vendor or comes into being as a result of the rendering, production and/or delivery of an agreement/understanding/request for quotation/contract or Purchase Order, or any other interaction. *Confiden*

500 chars retrieved; 492 of the 735 ground-truth chars (**bold**) are inside the chunk. The dense channel pairs the query's topical phrase "Confidential Information shall only include technical information" with the §1.1 enumeration "technical, commercial, procurement. . . ".

P@1 = 98.4% R@1 = 66.9%

(b) Naive + Generic + BM25 (row 7)
Generic + BM25 + RRF top-1: chunk $[3{,}000, 3{,}500]$

indrod SA agree that; 1. The Vendor and Grindrod SA have entered into an agreement / understanding / request for quotation / contract / order or any interaction of whatsoever nature, under which the Vendor may be required to quote on and / or render Deliverables to Grindrod SA and/or Grindrod SA shall be obliged to disclose or otherwise make available to the Vendor, Confidential Information. 2. The Parties are aware that Confidential Information may be disclosed or otherwise be made accessible t

500 chars retrieved; *zero* ground-truth overlap. The query repeats "Grindrod SA" (the party identifier); BM25 ranks the chunk where "Grindrod SA" occurs four times in close proximity above the §1.1 definitional clause where it occurs only once. "Confidential Information" (also in the query) appears in both chunks but doesn't break the tie.

P@1 = 0.0% R@1 = 0.0%

Figure 10: **ContractNLI example.** The query's repetition of the party identifier "Grindrod SA" shifts the BM25 top-1 from the §1.1 definitional clause to a generic obligations chunk, dropping Recall@1 from 66.9% to 0%. The reranker on top of +BM25 (row 9) moves the top-1 to chars $[1{,}500, 2{,}000]$, an adjacent chunk still outside the ground-truth span (R@1 = 0%).