CORA: Causal Outcome Retrieval Augmentation for Corpus-Level Causal Question Answering

Anonymous ACL submission

Abstract

When analysts ask *why* an outcome occurs across a labelled document corpus, standard retrieval fails them: the documents that *explain* an outcome differ systematically in vocabulary from those that *describe* it, and similarity-based retrieval finds the latter. We demonstrate this empirically : semantic retrieval achieves Precision@10 of 49.5% for outcome-relevant chunks across 60 causal queries, below the corpus base rate of 55.9%, meaning it actively deprioritises causal evidence relative to random sampling.

We propose CORA (**C**ausal **O**utcome **R**etrieval **A**ugmentation): estimate topic-level outcome uplift from corpus labels, fuse with any retriever via reciprocal rank fusion. No annotations, no graph construction, no training. Applied to BM25, nDCG@10 improves from 0.568 to 0.622; applied to dense retrieval, MRR improves from 0.734 to 0.836. Gains are consistent across retrieval paradigms, outcome uplift is complementary to both lexical and semantic similarity, and their naive fusion is not. We release CCQA, the first benchmark for corpus-level causal QA: 60 queries across three tiers with annotated ground truth.

1 Introduction

Analysts working with labelled document collections routinely ask causal questions. A platform trust team seeks conversational patterns preceding rule violations. A contact centre wants to identify drivers of customer escalations. A clinical researcher aims to determine which consultation dynamics correlate with adverse outcomes. In each case, the question takes the same form: *why does outcome Y occur across this corpus?* Answering it requires identifying conditions associated with the outcome, not merely documents that describe it.

This matters because documents explaining an outcome differ from those describing it. Explanations often lie in antecedent signals (e.g., early frustration) rather than the outcome's occurrence. Standard retrieval (Lewis et al., 2020) relies on query similarity, therefore retrieving outcome descriptions while missing explanatory antecedents lacking vocabulary overlap.

We demonstrate this mismatch empirically: across 60 causal queries on a 19,578-conversation CGA-CMV corpus, semantic retrieval yields 49.5% Precision@10 for outcome-relevant chunks, worse than the 55.9% random baseline. Similarity-based retrieval thus actively deprioritises causal evidence due to structural misalignment.

Instead, we propose optimising *outcome relevance*, defined using labels already present in corpora without additional annotation. We compute *outcome uplift*, the statistical association between latent topics and outcome labels, positive or negative used as a ranking signal via $|\mathrm{Uplift}(t)|$.

We introduce CORA (**C**ausal **O**utcome **R**etrieval **A**ugmentation), a training-free framework fusing uplift scores with any base retriever via reciprocal rank fusion. Lexical/semantic retrieval efficiently surfaces candidates; outcome uplift reranks them to promote causal evidence. With BM25, CORA improves nDCG@10 from 0.568 to 0.622; with dense retrieval, MRR rises from 0.734 to 0.836. Our contributions are:

- **Outcome-relevance gap.** Empirical proof that similarity retrieval deprioritises causal evidence.
- **CORA framework.** Training-free, graph-free retrieval augmentation fusing topic uplift with base retrievers.
- **Paradigm complementarity.** Evidence that outcome uplift complements both lexical and semantic retrieval where naive semantic-lexical fusion fails.
- **CCQA benchmark.** 60 causal queries with LLM-annotated ground truth for future causal QA research.

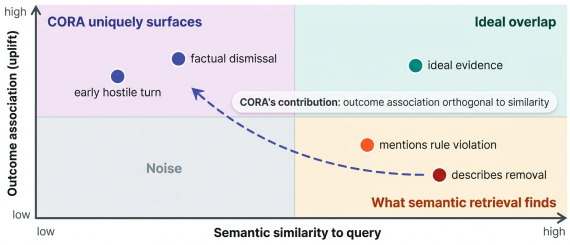

Figure 1: Semantic retrieval optimises along the horizontal axis, returning chunks that *describe* the outcome (bottom-right). CORA introduces the vertical axis: outcome uplift from corpus labels, surfacing causally informative chunks that similarity-based retrieval cannot reach (top-left).

2 Related Work

Causal retrieval and RAG. Recent works incorporate causal structure into retrieval by extracting causal relation triples for graph traversal (CausalRAG (Wang et al., 2025), GraphRAG (Edge et al., 2024)) or by training dense retrievers to capture causal language markers (Shin and Hwang, 2025). These approaches depend on explicit causal language for graph construction or encoder training, and they typically require extensive offline LLM preprocessing. In contrast, CORA estimates outcome relevance statistically directly from corpus labels without requiring language models, graph construction, or explicit causal language dependencies.

Retrieval augmentation and topic modelling. Retrieval augmentation improves base rankers using auxiliary signals such as cross-encoders (Nogueira and Cho, 2019), LLM rerankers (Sharifymoghaddam et al., 2025), reciprocal rank fusion (Cormack et al., 2009), or hypothetical document expansion (Gao et al., 2022). Unlike these heavily semantic or lexical enhancements, CORA targets a complementary relevance dimension. While topic models (Blei et al., 2009; Miao et al., 2017; Grootendorst, 2022; Angelov, 2020) often organise retrieved evidence semantically, CORA uniquely uses topic-level outcome uplift as a ranking criterion, directly linking latent topic structures to outcome prediction labels.

Uplift modelling, inference, and conversational forecasting. Originating in clinical and marketing analytics (Radcliffe and Surry, 2012), uplift modelling estimates treatment effects. In NLP, causal inference aids debiasing (Feder et al., 2021) and the estimation of observational effects. We apply uplift to latent topics as a novel retrieval ranking signal. We also depart from the traditional derailment forecasting task on the CGA-CMV corpus (Zhang et al., 2018; Chang and Danescu-Niculescu-Mizil, 2019; Altarawneh et al., 2023), which predicts future outcomes for single conversations, focusing instead on retrieving corpus-level aggregated explanations. Finally, while BRIGHT (Su et al., 2024) demonstrates that dense retrievers fail on reasoning-heavy queries, our diagnostics show a stronger failure: on corpus-level causal questions, semantic retrieval can underperform random sampling. CORA addresses this gap entirely training-free.

3 Problem Formulation

Let $\mathcal{C} = \{d_1, \ldots, d_n\}$ be a corpus of n documents, each associated with a binary outcome label $y_i \in \{0, 1\}$, where $y_i = 1$ indicates the occurrence of the outcome of interest. Each document d_i is decomposed into overlapping chunks $\{c_{i1}, \ldots, c_{im_i}\}$ spanning k consecutive utterances with stride s.

A **corpus-level causal query** Q asks *"why does outcome Y occur?"*, seeking an aggregated pattern-level explanation. We distinguish three query types with different retrieval requirements, motivating the CCQA benchmark tiers:

- **Outcome queries** seek general explanations (*e.g., "What early conversational dynamics lead to rule-violation removals?"*). Relevant evidence comes primarily from outcome-positive documents.
- **Factor queries** target hypothesised contributing conditions (*e.g., "How do implicit accusations affect subsequent rule violations?"*). Relevant evidence may arise from either outcome class.
- **Contrastive queries** ask for differential patterns (*e.g., "How do OP responses to counter-arguments differ between delta-awarding and non-awarding threads?"*). Relevant evidence necessarily spans both classes.

This taxonomy demands that retrieval surfaces evidence from both sides of the outcome distribution, as retrieving only outcome-positive documents leaves factor and contrastive queries unanswered.

Definition 1 (Outcome vs. Semantic Relevance)**.** *Chunk c_{ij} is **semantically relevant** to Q if $\cos(\text{enc}(c_{ij}), \text{enc}(Q))$ is high. Chunk c_{ij} is*

outcome-relevant *to Q if it belongs to a topic statistically associated with outcome Y in the corpus, regardless of semantic similarity to Q.*

We do not claim outcome relevance identifies true causal effects, as ignorability cannot be assumed from observational data. Our claim is narrower: augmenting retrieval with outcome-associated topics surfaces better evidence for causal queries than similarity alone. This is an operationalisation of outcome relevance, not a ground-truth definition: the experimental evaluation in §5.2 uses independently constructed LLM-annotated ground truth that does not reference uplift scores, ensuring the evaluation is not circular.

The goal of CORA is to produce a ranked chunk list $\mathcal{E}$ balancing outcome relevance, query specificity, and diversity for LLM-based causal explanation generation.

4 CORA: Causal Outcome Retrieval Augmentation

CORA augments a base retriever with a corpus-derived outcome-uplift signal via reciprocal rank fusion. The framework takes two inputs: a query-specific chunk ranking produced by a base retriever $\mathcal{R}$ and topic-level uplift scores estimated offline from corpus labels. These signals are fused symmetrically; neither gates the other. Algorithm 1 summarises the procedure.

Algorithm 1: CORA: Causal Outcome Retrieval Augmentation

Input: Corpus $\mathcal{C}$, labels $\{y_i\}$, query Q, base retriever $\mathcal{R}$
Input: Parameters $k_{\text{rrf}}, K_e, \lambda, n_{\min}$
Offline: Outcome Uplift Estimation
Chunk documents: $d_i \to \{c_{i1}, \ldots, c_{im_i}\}$
Embed chunks; fit BERTopic $\to$ topic labels $\{t_{ij}\}$
$\mathcal{T}(d_i) \leftarrow \{t : \exists j,\ t_{ij} = t\}$
for *each topic* $t \in \mathcal{T}$ *with* $\mathit{support}(t) \geq n_{\min}$ **do**
  $\text{Uplift}(t) \leftarrow P(Y{=}1 \mid t \in \mathcal{T}(d)) - P(Y{=}1 \mid t \notin \mathcal{T}(d))$
end
Online: Retrieval Augmentation
$\text{rank}_{\mathcal{R}}(c) \leftarrow$ base-retriever rank of c for query Q
$\text{rank}_{\text{Uplift}}(c) \leftarrow$ rank by $|\text{Uplift}(t_{ij})|$; $t_{ij} = -1$ ranked last
$\text{score}(c, Q) \leftarrow (k_{\text{rrf}} + \text{rank}_{\mathcal{R}}(c))^{-1} + (k_{\text{rrf}} + \text{rank}_{\text{Uplift}}(c))^{-1}$
Rank chunks by $\text{score}(c, Q)$
$\mathcal{L} \leftarrow$ top-$K_e \cdot r$ chunks
$\mathcal{E} \leftarrow \text{MMR}(\mathcal{L}, Q, K_e, \lambda)$
return $\mathcal{E}$

4.1 Chunk-Level Topic Modelling

Naturalistic documents are rarely topically homogeneous. A conversation may begin with factual disagreement, transition through frustration, and conclude with direct hostility. Representing such documents with a single vector conflates segments that may exhibit distinct outcome associations. We therefore represent each document d_i as overlapping chunks $\{c_{i1}, \ldots, c_{im_i}\}$, where each chunk spans k consecutive utterances with stride s.

We embed all chunks using a pretrained sentence encoder and fit BERTopic (Grootendorst, 2022), assigning each chunk c_{ij} a topic label $t_{ij} \in \mathcal{T} \cup \{-1\}$, where -1 denotes the outlier class. The topic set of document d_i is

$$\mathcal{T}(d_i) = \{t \in \mathcal{T} : \exists j,\ t_{ij} = t\} \tag{1}$$

We compare four topic modelling approaches in §6.2: LDA (Blei et al., 2009), NMF (Lee and Seung, 1999), Top2Vec (Angelov, 2020), and BERTopic (Grootendorst, 2022). All methods operate on an identical chunk embeddings; differences in topic quality therefore reflect the modelling algorithm rather than the underlying representation.

4.2 Outcome Uplift Estimation

Uplift as an observational association measure. For each topic $t \in \mathcal{T}$ appearing in at least $n_{\min}$ documents, we estimate outcome uplift as

$$\begin{aligned}\text{Uplift}(t) = {} & P\big(Y{=}1 \mid t \in \mathcal{T}(d)\big) \\ & - P\big(Y{=}1 \mid t \notin \mathcal{T}(d)\big)\end{aligned} \tag{2}$$

$\text{Uplift}(t) > 0$ indicates a positive association with the outcome, whereas $\text{Uplift}(t) < 0$ indicates a protective or contrasting association.

Equation 2 defines an observational association measure rather than a structural causal quantity. Topics co-occur within documents; no adjustment is made for confounding among them, and no causal identification assumptions are imposed. We therefore use uplift solely as a ranking signal and evaluate its utility empirically for causal retrieval.

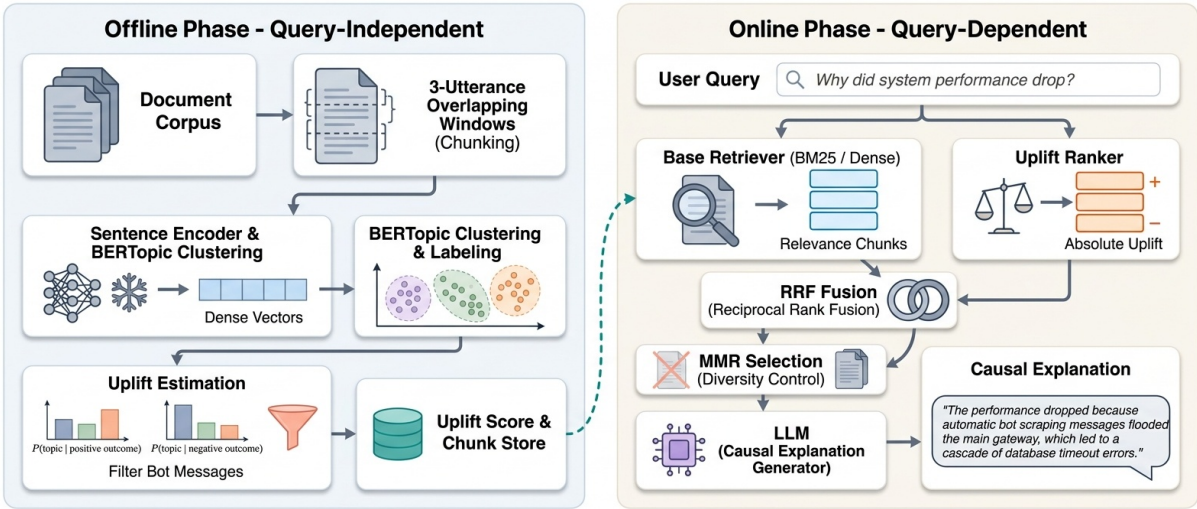

Figure 2: **CORA pipeline.** *Offline* (left): documents are chunked, embedded, and clustered via BERTopic; topic-level outcome uplift is estimated from corpus labels after artifact removal. *Online* (right): a base retriever and an uplift ranker produce independent rankings, RRF fuses them, and MMR selects a diverse evidence set for causal explanation.

Uplift ranking. Chunks are ranked by the uplift magnitude of their assigned topic:

$$\text{rank}_{\text{Uplift}}(c_{ij}) \propto -|\text{Uplift}(t_{ij})| \tag{3}$$

Using uplift magnitude rather than signed uplift promotes both strongly outcome-positive and strongly outcome-negative (protective) topics. This is particularly important for contrastive queries, which require evidence from both sides of the outcome distribution. Chunks assigned to the outlier topic ($t_{ij} = -1$) receive the lowest uplift rank.

For outcome queries, the base retriever naturally suppresses purely protective chunks. A chunk from a low-outcome topic may receive a strong uplift score, but will typically rank poorly under the base retriever; symmetric RRF fusion therefore demotes it without requiring query-specific ranking rules.

Practical considerations. Two practical issues arise. First, in near-balanced corpora the uplift dynamic range is limited in practice; on CGA-CMV ($\rho = 0.559$) the largest observed content topic uplift is 28.4pp, constraining the multiplicative PMI signal relative to rank-based RRF fusion.

Second, system-message artifacts (delta-bot utterances and moderation messages) can dominate high-uplift rankings because of their mechanical association with outcomes. We remove these during preprocessing before uplift estimation.

4.3 Retrieval Augmentation via RRF

Given a base retriever $\mathcal{R}$ that ranks chunks for query Q, CORA augments this ranking by fusing it with the uplift ranking using reciprocal rank fusion (RRF) (Cormack et al., 2009):

$$\text{score}(c, Q) = \frac{1}{k_{\text{rrf}} + \text{rank}_{\mathcal{R}}(c, Q)} + \frac{1}{k_{\text{rrf}} + \text{rank}_{\text{Uplift}}(c)} \tag{4}$$

where $k_{\text{rrf}} = 60$ is the standard RRF constant (Cormack et al., 2009). The two rankings are treated symmetrically: neither gates the other, and neither requires score normalisation. RRF rewards chunks that rank well under both signals: a chunk ranking first under one signal but last under the other receives a moderate combined score, while a chunk ranking highly under both receives the maximum combined score.

RRF is well suited to this setting for three reasons. First, apart from the standard constant k_{rrf}, it introduces no mixing parameter, avoiding the need to tune weights between heterogeneous signals. Second, it is robust to scale differences between retrieval scores and uplift magnitudes, which occupy different numerical ranges. Third, it treats retrieval relevance and outcome association as peer ranking signals rather than treating one as a primary model and the other as a correction, consistent with our empirical finding that outcome uplift complements both lexical and semantic retrieval.

We evaluate CORA with two base retrievers: BM25 (Robertson and Zaragoza, 2009) and dense retrieval using sentence embeddings (Reimers and Gurevych, 2019). We evaluate CORA as an augmentation of Hybrid retrieval for completeness; empirically, Hybrid + Uplift achieves nDCG@10 of 0.537 vs. Hybrid alone at 0.536, a negligible gain (§6.4). We include Hybrid + Uplift in Table 2 and discuss this result in §7.

4.4 Diversity-Aware Evidence Selection

Ranking by fused RRF score may over-concentrate retrieved evidence around the same conversation or topic. We therefore apply Maximal Marginal Relevance (MMR) (Carbonell and Goldstein-Stewart, 1998) to promote evidential diversity:

$$\begin{aligned}\mathrm{MMR}(c_{ij}) = \lambda\, \mathrm{score}(c_{ij}, Q) \\ - (1-\lambda) \max_{c_{kl} \in \mathcal{E}} \cos(c_{ij}, c_{kl})\end{aligned} \tag{5}$$

The evidence set $\mathcal{E}$ is constructed greedily until $|\mathcal{E}| = K_e$, with λ controlling the relevance–diversity trade-off. The final set contains chunks that are jointly outcome-associated, query-specific, and diverse across the retrieved evidence space.

5 Experimental Setup

5.1 Dataset

We evaluate on **CGA-CMV-Large** (Chang et al., 2020), an approximately balanced corpus of 19,578 Reddit ChangeMyView conversations labelled by moderator intervention (derailment). Automated delta-awarding and moderation utterances are removed before chunking to prevent artifact dominance. We use CGA-CMV because observable platform interventions provide reliable outcome labels for naturalistic conversational dynamics.

5.2 The CCQA Benchmark

CCQA (**C**orpus-level **C**ausal **QA**) is a new retriever-agnostic benchmark of 60 natural language queries over CGA-CMV (20 outcome, 20 factor, 20 contrastive). We release the queries, relevance judgements, and evaluation scripts at `[anonymous link]`.

Query construction. Authors drafted queries based on a 50-conversation stratified sample and high-uplift topic structures. Each query was validated for answerability, causal intent, and non-triviality.

Ground truth construction. Top-50 retrieved chunks from all evaluated methods are independently pooled. Deduplicated pairs are then judged by `gemini-3.1-flash-lite` (Gemini (Team, 2024)) via a prompt distinguishing outcome-explaining from outcome-describing evidence. A second pooling round exclusively targets CORA variants to minimise evaluation bias. Two authors independently annotate a random sample of 100 (query, chunk) pairs; Cohen's κ between human annotators and between each human and the LLM judge is reported in Appendix D, confirming annotation reliability.

5.3 Topic Modelling Selection

We compare LDA (Blei et al., 2009), NMF (Lee and Seung, 1999), Top2Vec (Angelov, 2020), and BERTopic (Grootendorst, 2022), all using identical `all-MiniLM-L6-v2` chunk embeddings (Reimers and Gurevych, 2019). Model selection is based on coherence CV (Röder et al., 2015); full configurations and results appear in Appendix A. BERTopic on chunk-level representations achieves the highest coherence (CV $= 0.830$) and is used throughout.

5.4 Retrieval Baselines

We evaluate against **Random** (lower bound), **BM25** (Robertson and Zaragoza, 2009), **Dense** (cosine similarity over `all-MiniLM-L6-v2` embeddings), and **Hybrid** (BM25+Dense RRF, k=60) as production-strength baselines, plus **Semantic-in-Topics** (dense retrieval within query-similar topics, no uplift) as an ablation isolating topic filtering without outcome association. **CORA-Hard5** and **CORA-Hard100** apply hard uplift filters (top-5 and top-100 topics) with dense retrieval within the filtered set; **CORA-PMI** replaces RRF with multiplicative $\exp(\mathrm{PMI})$ scoring. We exclude CausalRAG (Wang et al., 2025) and GraphRAG (Edge et al., 2024) as direct baselines. Both require offline LLM preprocessing over the full corpus (approximately 127,004 LLM calls on CGA-CMV) and target settings in which causal structure is explicitly expressed in text rather than inferred from outcome labels.

5.5 Evaluation Metrics

We report **nDCG@5** and **nDCG@10** (graded relevance, with @5 more sensitive to early-rank precision), **MRR** (speed of first relevant chunk), **MAP** (consistent precision throughout the list),

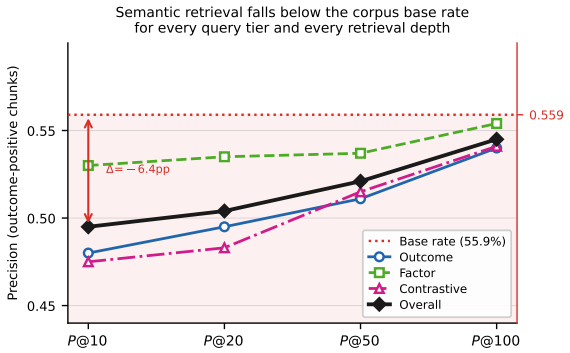

Figure 3: Semantic retrieval precision (proportion of top-K chunks from derailed conversations) across query tiers and retrieval depths. All tiers fall below the corpus base rate of 55.9% (dotted red line) at every K, confirming that semantic similarity actively deprioritises outcome-relevant evidence. Gap at $P@10$: -6.4pp overall.

Query tier	P@10	P@20	P@50	P@100
Outcome	0.480	0.495	0.511	0.540
Factor	0.530	0.535	0.537	0.554
Contrastive	0.475	0.483	0.515	0.541
Overall	0.495	0.504	0.521	0.545
Base rate	0.559	0.559	0.559	0.559
Δ vs. base rate	-0.064	-0.055	-0.038	-0.015

Table 1: Precision@K: proportion of top-K semantically retrieved chunks from derailed conversations, by query tier. Semantic retrieval falls below the 55.9% corpus base rate at every K and for every query tier, indicating systematic deprioritisation of outcome-relevant evidence.

Recall@50 (coverage of the causally relevant evidence set), and **Success@1/@10** (proportion of queries with at least one relevant chunk in the top-K). All metrics are computed against CCQA ground truth overall and by tier.

5.6 Implementation Details

Chunks span k=3 utterances with stride s=1. We use `all-MiniLM-L6-v2` embeddings (Reimers and Gurevych, 2019), $n_{\min}$=30, k_{rrf}=60, K_e=50, and λ=0.5. All hyperparameters are fixed using a 10% validation split. Full implementation details appear in Appendix E. Code, queries, ground truth, and evaluation scripts are released at `[anonymous link]`.

6 Results

6.1 Diagnostic: Semantic Retrieval and Outcome Relevance

Table 1 reports mean Precision@K for semantically retrieved chunks across all 60 CCQA queries, measured as the proportion of top-K retrieved chunks drawn from derailed conversations (outcome label $= 1$). Under a retrieval system sensitive to outcome relevance, this precision should exceed the corpus base rate of 55.9%.

We use outcome-class membership as a necessary-condition proxy: a retriever that cannot exceed the base rate on class membership cannot be recovering the finer-grained outcome-relevance signal of Definition **??**. Semantic retrieval falls below the corpus base rate at every retrieval depth and for every query tier. At K=10, overall precision is 49.5%, 6.4 percentage points below base rate. The gap is largest for outcome queries (48.0%) and smallest for factor queries (53.0%), consistent with the fact that factor queries reference mechanisms appearing in both derailed and civil conversations. As K approaches 100, precision converges toward the base rate, indicating that semantic retrieval remains largely uninformative about outcome relevance even at deeper retrieval depths.

These findings support the central premise of this work: semantic similarity optimisation deprioritises causally relevant evidence relative to random sampling, reflecting a misalignment between causal queries (seeking antecedents) and semantic retrieval (favouring vocabulary overlap).

6.2 Topic Modelling Quality

Table 4 compares the four topic modelling methods on chunk-level representations of CGA-CMV.

BERTopic on chunk-level representations achieves the highest coherence across all methods (CV $=$ 0.830, versus 0.668 for NMF, the next strongest baseline), while producing interpretable topics without outlier-noise contamination in uplift ranking. Full results appear in Appendix A.

6.3 Uplift Validity

Of the 355 topics passing the minimum-support filter, 31 (8.7%) are nominally significant at $p <$ 0.05. After Benjamini–Hochberg correction, 13 remain significant, 11 of which are system-message artifacts with extreme effect sizes (up to +46.0pp) driven by mechanical associations with persuasion outcomes.

After artifact removal, two content topics remain: Topic 67 (refugees and immigration, $q = 0.005$) and Topic 81 (meta-argumentation, $q = 0.029$).

Method	nDCG@5	nDCG@10	MRR	MAP	R@50	S@1	S@10
Random	0.481	0.476	0.585	0.054	0.092	0.417	0.933
BM25	0.558	0.568	0.695	0.085	0.131	0.517	0.983
Dense	0.549	0.547	0.734	0.075	0.119	0.633	0.983
Hybrid	0.512	0.536	0.615	0.073	0.124	0.400	1.000
Sem-in-Topics	0.508	0.509	0.653	0.068	0.109	0.500	0.983
CORA-Hard5	0.208	0.208	0.422	0.011	0.036	0.267	0.817
CORA-Hard100	0.522	0.505	0.639	0.058	0.101	0.400	0.983
CORA-PMI	0.496	0.519	0.731	0.069	0.116	0.617	0.983
CORA + BM25	0.618	**0.622**†	0.805	**0.090**††	**0.135**	0.717	0.967
CORA + Dense	**0.627**	0.577	**0.836**†	0.049	0.092	**0.750**	0.917
CORA + Hybrid	0.512	0.537	0.642	0.049	0.096	0.400	0.967

$^{\dagger}p < 0.05$, $^{\dagger\dagger}p < 0.01$ vs. base retriever, paired bootstrap (n=10,000), applied to all metrics; markers shown only where $p < 0.05$. Large percentage gains without markers reflect high cross-query variance rather than inconsistent testing.

Table 2: Full retrieval metrics on CCQA. R@50 = Recall@50; S@K = Success@K. Best per column in **bold**. CORA-Hard5/Hard100/PMI are ablations; Sem-in-Topics = Semantic-in-Topics.

We use significance descriptively; uplift magnitude remains the operative ranking criterion. The full significance table and top-topic analysis appear in Appendix B.

High-uplift content topics centre on conviction-driven, factually resolvable debates (LGBT rights, taxation, UBI, WWII), whereas low-uplift topics reflect identity-linked debates resistant to persuasion (Israel–Palestine, GMOs, surveillance). Both ends of the uplift spectrum contribute to CORA's ranking through the $|\text{Uplift}|$ criterion, supporting contrastive as well as outcome queries.

6.4 Main Retrieval Results

Table 2 reports the full metric suite on CCQA. CORA consistently improves over the corresponding base retriever across metrics and retrieval paradigms.

Consistent gains across retrieval paradigms. CORA + BM25 improves over BM25 on every metric: nDCG@10 increases from 0.568 to 0.622 (+9.5%), MRR from 0.695 to 0.805 (+15.8%), MAP from 0.085 to 0.090 (+5.9%), and Recall@50 from 0.131 to 0.135 (+3.1%). Likewise, CORA + Dense improves over Dense on precision-focused metrics: nDCG@5 from 0.549 to 0.627 (+14.2%), MRR from 0.734 to 0.836 (+13.9%), and S@1 from 0.633 to 0.750 (+18.5%), at the cost of MAP and Recall@50 (the precision-coverage trade-off discussed below). Gains across both lexical and semantic retrieval confirm that outcome uplift is a complementary signal rather than a correction specific to a single retrieval paradigm.

Outcome uplift, not lexical–semantic fusion. Hybrid retrieval (BM25 + Dense RRF) achieves nDCG@10 of 0.536, below both BM25 (0.568) and Dense (0.547). Adding Dense to BM25 hurts performance, whereas adding uplift to BM25 substantially improves it. This asymmetry identifies outcome uplift, rather than semantic similarity, as the missing complementary signal for causal retrieval. Lexical and semantic retrieval recover broadly similar evidence; outcome uplift surfaces a different class of evidence: chunks from outcome-associated topics regardless of surface similarity to the query.

Precision vs. coverage trade-off. CORA + Dense achieves higher MRR (0.836) and Success@1 (0.750) than CORA + BM25 (MRR 0.805, S@1 0.717), retrieving the first relevant chunk more quickly. CORA + BM25 achieves higher nDCG@10 (0.622), MAP (0.090), and Recall@50 (0.135), distributing relevant evidence more evenly throughout the ranked list.

These results reflect complementary strengths: CORA + Dense is preferable when the analyst needs the single strongest evidence chunk quickly, whereas CORA + BM25 is preferable when broad evidential coverage matters. The MAP and Recall@50 degradation of CORA + Dense relative to Dense is statistically significant ($p < 0.001$, paired bootstrap), confirming that the trade-off is structural rather than measurement noise; CORA + BM25 exhibits no significant degradation on any metric.

6.5 Results by Query Tier

Table 3 reports nDCG@10 by query tier for selected methods.

Method	Outcome	Factor	Contrastive
BM25	0.647	0.500	0.557
Dense	0.646	0.454	0.542
Hybrid	0.630	0.414	0.564
CORA-Hard100	0.650	0.381	0.447
CORA-PMI	0.627	0.394	0.536
CORA + BM25	**0.715**	0.501	**0.650**
CORA + Dense	0.660	**0.514**	0.555

Table 3: nDCG@10 by query tier. CORA + BM25 performs best on outcome and contrastive queries; CORA + Dense leads marginally on factor queries (+0.013). Both CORA variants improve over their corresponding base retrievers across all tiers.

CORA + BM25 achieves 0.715 nDCG@10 on outcome queries, the largest absolute gain over any

baseline (+6.8pp over BM25). This is the setting where uplift augmentation should help most: outcome queries ask broadly why an outcome occurs, and the uplift signal directly identifies outcome-associated topics across the corpus.

CORA + BM25 also leads on contrastive queries (0.650, +9.3pp over BM25), where ranking by $|\text{Uplift}|$ surfaces both outcome-positive and protective evidence. By contrast, CORA + Dense performs best on factor queries (0.514 vs. 0.501), suggesting that Dense retrieval's stronger query specificity is advantageous when a targeted mechanism is explicitly named. The margin is small (+1.3pp) and consistent with the broader precision–coverage trade-off between the two CORA variants.

6.6 Ablation Study

The ablations in Table 2 isolate the contribution of individual design choices in CORA.

CORA-Hard5 (hard filtering to the top-5 topics) achieves nDCG@10 of 0.208, far below Random (0.476). With only 1% corpus coverage, the filter excludes most ground-truth relevant chunks regardless of retrieval quality within the remaining subset. CORA-Hard100 (top-100 topics; 11.4% coverage) improves to 0.505, indicating that coverage, rather than uplift itself, is the primary failure mode of hard filtering.

CORA-PMI (multiplicative fusion of cosine similarity and $\exp(\text{PMI})$) achieves 0.519, outperforming CORA-Hard100 but remaining well below CORA + BM25. In this near-balanced corpus ($\rho = 0.559$), the bounded uplift range limits the multiplicative signal: even the strongest outcome signal yields only a modest multiplicative advantage over a neutral chunk. RRF is more robust because it operates in rank space rather than score space, making it less sensitive to bounded signal magnitudes.

The progression from CORA-Hard5 (0.208) to CORA-Hard100 (0.505) to CORA-PMI (0.519) to CORA + BM25 (0.622) traces a clear pattern: hard filtering fails because of coverage loss, multiplicative scoring improves coverage but remains constrained by bounded uplift values, and RRF fusion avoids both issues by operating in rank space atop a strong lexical base. Hybrid + Uplift achieves nDCG@10 of 0.537, essentially identical to Hybrid alone (0.536), providing direct empirical support for the claim that outcome uplift does not complement a retriever that already mixes lexical and semantic signals.

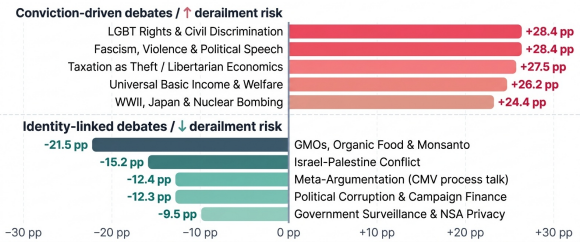

Figure 4: Top-5 content topics by outcome uplift magnitude after artifact filtering. Conviction-driven factually resolvable debates (positive uplift) are associated with derailment; identity-linked debates resistant to persuasion (negative uplift) are not. Both poles contribute to CORA's ranking via $|\text{Uplift}(t)|$.

7 Analysis

Complementary retrieval signals. The contrast between Hybrid retrieval (nDCG@10 = 0.536) and CORA + BM25 (0.622) demonstrates that BM25 and Dense retrieval are not strongly complementary for causal queries, as both prioritise similarity with the query. Outcome uplift is orthogonal, retrieving chunks associated with the outcome regardless of query similarity. RRF succeeds for CORA + BM25 because the two signals operate along different relevance axes. This implies that complementary retrieval signals should differ in objective, not merely implementation. Additional analyses, including topic confounding, qualitative illustrations, and ground-truth pooling details, appear in Appendix C.

8 Conclusion

We introduced CORA, a retrieval augmentation framework for corpus-level causal question answering. Motivated by our finding that semantic retrieval deprioritises causal evidence, CORA fuses topic-level outcome uplift with any base retriever. Applied to BM25, CORA improves nDCG@10 from 0.568 to 0.622; with Dense retrieval, it improves MRR from 0.734 to 0.836. The failure of Hybrid retrieval (nDCG@10 = 0.536) to match baseline performance confirms that causal retrieval mainly lacks outcome association rather than semantic improvements.

We additionally release CCQA, the first benchmark for corpus-level causal QA, as a community resource for future retrieval and QA research.

Limitations

CORA requires binary or categorical outcome labels to estimate topic-level uplift; corpora without such labels fall outside the current scope, although unsupervised outcome proxies are a natural extension. Uplift estimation treats topics as independent binary treatments and does not model co-occurrence confounding. We analyse this empirically in §7 and observe limited impact in our setting, but more rigorous multi-treatment estimation remains an important direction for future work.

Evaluation is conducted on a single English-language corpus; generalisability across domains, languages, and outcome structures requires further validation. The bounded uplift range on CGA-CMV constrains the strength of the augmentation signal; corpora with more imbalanced outcome distributions may exhibit larger gains.

LLM-based relevance annotation introduces annotation cost and potential model biases. We partially validate the labels against human judgements via Cohen's κ, but fully human annotation at scale remains expensive. Finally, Benjamini–Hochberg correction over all 355 topics is dominated by system-message artifacts, leaving content-topic significance without strong formal guarantees. We therefore report significance descriptively and use uplift magnitude as the operative ranking criterion.

Ethical Considerations

We use CGA-CMV-Large, distributed via ConvoKit under a research-use licence. We use gemini-3.1-flash-lite via the Google AI API under Google's standard API terms of service. The all-MiniLM-L6-v2 sentence encoder is used under the Apache 2.0 licence. Usernames are present in the dataset but are not used in retrieval or analysis.

CORA is designed for corpus-level causal pattern analysis. Generated explanations reflect statistical associations aggregated across the corpus and should not be interpreted as claims about individual users or individual causation. Uplift scores are observational associations rather than causal estimates and should not be used to support causal claims about specific topics or conversational behaviours outside this retrieval setting. CCQA is released for research purposes only and should not be used to make claims about individual users or conversations.

AI writing assistance (Claude, Anthropic) was used for manuscript editing and LaTeX formatting; all scientific content and experimental results are the authors' own.

References

Enas Altarawneh, Ammeta Agrawal, Michael Jenkin, and Manos Papagelis. 2023. Conversation derailment forecasting with graph convolutional networks. *WOAH*, arXiv:2306.12982.

Dimo Angelov. 2020. Top2Vec: Distributed representations of topics. In *arXiv.org*.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2009. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3.

J. Carbonell and Jade Goldstein-Stewart. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *SIGIR Forum*.

Jonathan P Chang, Caleb Chiam, Liye Fu, Andrew Wang, Justine Zhang, and Cristian Danescu-Niculescu-Mizil. 2020. ConvoKit: A toolkit for the analysis of conversations. In *SIGDIAL Conferences*, pages 57–60. Association for Computational Linguistics.

Jonathan P Chang and Cristian Danescu-Niculescu-Mizil. 2019. Trouble on the horizon: Forecasting the derailment of online conversations as they develop. In *Conference on Empirical Methods in Natural Language Processing*, pages 4742–4753. Association for Computational Linguistics.

G. Cormack, C. Clarke, and Stefan Büttcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 758–759. ACM.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. 2024. From local to global: A graph RAG approach to query-focused summarization. In *arXiv.org*.

Amir Feder, Katherine A. Keith, Emaad A. Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E. Roberts, and 1 others. 2021. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. In *Transactions of the Association for Computational Linguistics*.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2022. Precise zero-shot dense retrieval without relevance labels. In *Annual Meeting of the Association for Computational Linguistics*.

Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. In *arXiv.org*.

Daniel D. Lee and H. Sebastian Seung. 1999. Learning the parts of objects by non-negative matrix factorization. In *Nature*, volume 401, pages 788–791. Springer Science and Business Media LLC.

Patrick Lewis, Ethan Perez, Aleksandara Piktus, F. Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, M. Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Neural Information Processing Systems*.

Hugo Mercier and Dan Sperber. 2017. *The Enigma of Reason*. Harvard University Press.

Yishu Miao, Edward Grefenstette, and Phil Blunsom. 2017. Discovering discrete latent topics with neural variational inference. In *International Conference on Machine Learning*.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with BERT. In *arXiv.org*.

Nicholas J Radcliffe and Patrick D Surry. 2012. Real-world uplift modelling with significance-based uplift trees. *White Paper TR-2011-1, Stochastic Solutions*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Conference on Empirical Methods in Natural Language Processing*, pages 3980–3990. Association for Computational Linguistics.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 4(1-2):1–174.

Paul R Rosenbaum and Donald B Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70.

Michael Röder, A. Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. *Web Search and Data Mining*, pages 399–408.

Sahel Sharifymoghaddam, Ronak Pradeep, Andre Slavescu, Ryan Nguyen, Andrew Xu, Zijian Chen, Yilin Zhang, Yidi Chen, Jasper Xian, and Jimmy Lin. 2025. Rankllm: A python package for reranking with llms. *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, arXiv:2505.19284.

Hyunseo Shin and Wonseok Hwang. 2025. Causal retrieval with semantic consideration. *arXiv.org*, arXiv:2504.04700.

Hongjin Su, Howard Yen, Mengzhou Xia, Weijia Shi, Niklas Muennighoff, Han Wang, H. Liu, Quan Shi, Zachary S. Siegel, M. Tang, and 1 others. 2024. BRIGHT: A realistic and challenging benchmark for reasoning-intensive retrieval. In *International Conference on Learning Representations*.

Gemini Team. 2024. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Nengbo Wang, Xiaotian Han, Jagdip Singh, Jing Ma, and Vipin Chaudhary. 2025. CausalRAG: Integrating causal graphs into retrieval-augmented generation. In *Annual Meeting of the Association for Computational Linguistics*, pages 22680–22693. Association for Computational Linguistics.

Justine Zhang, Cristian Danescu-Niculescu-Mizil, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Nithum Thain, and Dario Taraborelli. 2018. Conversations gone awry: Detecting early signs of conversational failure. In *Annual Meeting of the Association for Computational Linguistics*, pages 1350–1361. Association for Computational Linguistics.

A Topic Modelling: Full Comparison

We compare four topic modelling methods on chunk-level representations of CGA-CMV. BERTopic and Top2Vec use `all-MiniLM-L6-v2` dense chunk embeddings (Reimers and Gurevych, 2019). LDA (Blei et al., 2009) and NMF (Lee and Seung, 1999) use TF–IDF representations of the same chunks and therefore do not consume dense vectors. All methods operate on identical chunked text units but differ in representation and modelling strategy.

BERTopic uses UMAP ($n_{\text{components}}$=5, $n_{\text{neighbours}}$=15, min_dist=0.0, cosine metric) followed by HDBSCAN (min_cluster_size=50, min_samples=10). LDA and NMF select K via coherence CV over $K \in \{20, 30, 50\}$. All methods use `random_state=42`.

Method	Sil.	Coh.	Out.%	#Top.
LDA	0.004	0.524	N/A	50
NMF	−0.025	0.668	N/A	50
Top2Vec	0.046	0.526	N/A	964
BERTopic (full)	**0.074**	0.348	22.3	48
BERTopic (chunk)	0.073	**0.830**	36.3	407

Table 4: Topic modelling comparison. Sil. = silhouette (↑); Coh. = coherence CV (↑); Out.% = outlier rate (↓). BERTopic on chunk-level representations achieves the strongest coherence by a substantial margin.

BERTopic (chunk) achieves coherence CV of 0.830 versus 0.668 for NMF, the next strongest baseline. BERTopic (full) attains a marginally higher silhouette score (0.074 vs. 0.073) but incurs a large coherence penalty (0.348 vs. 0.830). We prioritise coherence because it better reflects task utility: silhouette measures geometric compactness in embedding space, whereas coherence measures whether topic keywords co-occur naturally in text (Röder et al., 2015).

The 36.3% outlier rate of BERTopic (chunk) reflects the topical diversity of short three-turn windows. Outlier chunks receive $\text{Uplift} = 0$ and are therefore ranked solely by the base retriever.

B Uplift Validity: Full Tables and Analysis

Threshold	All	Content	Artifact
$p < 0.05$	31	19	12
$p < 0.001$	12	3	9
BH $q < 0.05$	13	2	11

Table 5: Topics passing significance thresholds on CGA-CMV, split into content and system-message artifacts. Artifact topics dominate the Benjamini–Hochberg survivors.

Topic confounding. The binary topic-presence formulation treats topics as independent treatments. To assess confounding severity, we compute pairwise Jaccard similarity between topic-presence indicators for all 13 Benjamini–Hochberg-significant topics.

Across 78 comparisons, mean Jaccard similarity is 0.017 (median 0.011; maximum 0.073 between Topics 20 and 84), indicating that the significant topics are nearly disjoint at the conversation level. The uplift signal therefore captures distinct discourse dimensions with limited mutual contamination.

More rigorous treatment via propensity-score weighting (Rosenbaum and Rubin, 1983) or multi-treatment effect estimation remains a natural direction for future work.

Topic 81: Meta-argumentative disengagement. Topic 81 (keywords: *view, op, change, argument, cmv*) is the only Benjamini–Hochberg-surviving content topic with negative uplift ($q = 0.029$).

The topic captures conversations in which participants discuss the persuasion process itself rather than substantive claims. Its negative association with derailment suggests that reflexive meta-argumentation functions as a marker of disengagement: once participants shift from debating the object of disagreement to debating the process, escalation becomes less likely, consistent with Mercier and Sperber (2017).

C Extended Analysis

Topic confounding and Topic 81. Across 78 pairwise comparisons between the 13 Benjamini–Hochberg-significant topics, mean Jaccard similarity is 0.017, indicating that significant topics are nearly disjoint and confounding is not a first-order concern.

Qualitative illustration. For query O1 (*"what dynamics make hostility-rule removal likely?"*), BM25's top-3 chunks discuss speeding laws, the Scottish independence referendum, and region-locked piracy—because they share surface tokens ("rule", "ban") with the query. By contrast, CORA + BM25's top-3 include two genuine CMV moderator-removal messages citing Comment Rule 2 verbatim, surfaced through high-uplift topic membership in conviction-driven debates sharing discourse properties with derailed threads.

Retrieval coverage and ground-truth pooling. The final pool contains 6,068 unique (query, chunk) pairs, of which 2,013 (33.2%) were retrieved exclusively by CORA variants. Among these, 41.3% were annotated RELEVANT by Gemini, compared with 58.3% among baseline-shared chunks. The lower but substantial relevance rate indicates that CORA surfaces genuine causal evidence inaccessible to similarity-based methods.

D Annotation Protocol

Each (query, chunk) pair is presented to `gemini-3.1-flash-lite` with instructions (abridged):

Mark a chunk as RELEVANT *only if it contains conditions, behaviours, or conversational patterns that contribute to or explain the outcome, not merely if it describes or exhibits it. For contrastive queries, chunks from civil conversations showing antecedent patterns that prevent escalation are also relevant.*

The full prompt, including tier-specific examples, is released with CCQA.

For contrastive queries, Gemini labels 36.4% of civil-conversation chunks as RELEVANT, compared with 47.1% of derailed-conversation chunks, indicating that the annotator recognises protective patterns as relevant evidence rather than defaulting to the outcome label.

Human–human Cohen's κ on the 100-pair validation sample is **0.68**, and human–LLM κ is **0.57**.

Top.	Keywords (top-5)	Uplift	p-value
Positive uplift (conviction-driven debates)			
316	gay, rights, lgbt, civil, discrimination	$+28.4$pp	3.5×10^{-3}
182	nazis, nazi, violence, antifa, speech	$+28.4$pp	8.4×10^{-3}
342	taxation, theft, extortion, voluntary	$+27.5$pp	1.3×10^{-2}
185	ubi, income, welfare, work, basic	$+26.2$pp	1.2×10^{-2}
212	japan, war, surrender, bomb, invasion	$+24.4$pp	2.6×10^{-2}
Negative uplift (identity-linked debates)			
17	government, privacy, nsa, surveillance	-9.5pp	1.9×10^{-2}
66	politicians, campaign, corruption	-12.3pp	3.1×10^{-2}
81	view, op, change, argument, cmv	-12.4pp	1.0×10^{-3}
32	israel, palestinians, land, jews	-15.2pp	1.5×10^{-2}
65	gmo, organic, monsanto, gm, crops	-21.5pp	2.1×10^{-3}

Table 6: Top content topics by uplift magnitude after artifact filtering. p-values are raw chi-square tests. Topic 81 is the strongest Benjamini–Hochberg-surviving content topic ($q = 0.029$).

Disagreements concentrate on chunks describing antecedent conditions that occur in both outcome classes with similar frequency, reflecting the intrinsic difficulty of distinguishing causal evidence from incidental co-occurrence.

E Implementation Details

Conversations are chunked into k=3 consecutive utterances with stride s=1.

Chunk embeddings use `all-MiniLM-L6-v2` (384 dimensions) (Reimers and Gurevych, 2019).

BERTopic configuration follows Appendix A.

Uplift estimation uses minimum support $n_{\min}$=30 documents per topic.

RRF uses the standard constant k_{rrf}=60 (Cormack et al., 2009).

MMR retrieves K_e=50 chunks with λ=0.5.

All hyperparameters are fixed on a held-out 10% validation split before evaluation on the remaining 90%.

Chunk embedding uses `all-MiniLM-L6-v2` (22M parameters). BERTopic clustering and uplift estimation run on CPU and complete in under two hours on CGA-CMV (19,578 conversations). Retrieval evaluation over 60 queries runs in under 30 minutes. All experiments were conducted on a single machine with 16 GB RAM. No GPU training was performed.