INTER IIT
TECH MEET 14.0

OBSERVE·AI

Causal Rationale Extraction and Synthesis from Conversational Data on Business Events

FINAL REPORT TEAM 24

Abstract

Interpreting conversational dynamics behind business-critical events is key to shifting contact-center monitoring from reactive metrics to proactive insight. This report outlines the methodology, experiments, and conceptual foundations built for the Mid-Prep Observe.AI problem statement (Inter IIT Tech Meet 14.0). The problem centers on two linked tasks: Task 1 - Query-Driven Evidence Based Causal Explanation: Interpreting natural-language causal queries, identifying relevant conversational events within large dialogue corpora, retrieving supporting evidence, and structuring causal signals for downstream explanation. Task 2 - Conversational Analytic Reasoning: Extending Task 1 by enabling context-aware follow-up queries, refinement of earlier outputs, and deeper exploration of the same evidence while maintaining coherence and continuity. Subsequent sections outline the system design principles, query simulation and dataset construction, and key modeling decisions, together with the reasoning behind major experimental choices.It concludes with core insights and future directions, outlining the reasoning, architecture, and analytical processes that shaped the system.

1 Task - 1 : Query-Driven Evidence Based Causal Explanation

Understanding the factors that shape business-critical outcomes in customer service interactions is essential for improving operational efficiency, service quality, and customer satisfaction. In large contact-center settings, agent and customer conversations often culminate in events such as escalations, refund requests, or churn signals, each with operational and financial consequences. Identifying the conversational elements that precede these outcomes allows organizations to design targeted interventions, support agent performance, and improve customer experience.

1.1 Introduction

Task 1 focuses on building a system that retrieves dialogue evidence linked to specific business events using multi-turn conversational transcripts. The dataset contains noisy, real-world interactions in which meaningful signals such as behavioural shifts, agent strategies, or changes in customer intent are dispersed across several turns. The objective is to move beyond surface correlations and provide structured and interpretable explanations that reflect the underlying interaction.

1.2 Primary Challenges

One primary challenge arises from the sheer scale of the dataset, comprising over 19,000 transcripts and more than 680,000 dialogue turns. Processing such a large volume of data imposes a tradeoff between computational cost and model performance: extensive preprocessing, embedding generation, and retrieval over hundreds of thousands of turns can be time- and resource-intensive. Conversational data is inherently complex, with variable lengths, branching dialogue structures, and sparse event labels, making the detection of causal factors non-trivial. Subtle conversational cues—such as customer hesitations, repeated queries, silences, or agent behaviors—are distributed across turns, requiring models to capture long-range dependencies. Noise in transcripts, including ungrammatical segments, disfluencies, and misaligned speaker roles, further complicates natural language understanding. A deeper challenge lies in distinguishing genuine causal contributors from mere correlations: the task demands forming stable, evidence-supported causal interpretations rather than relying on surface-level regularities or heuristic reasoning. This requirement goes beyond the pattern-matching tendencies of contemporary LLMs, necessitating deliberate modeling choices that support robust causal inference within noisy, unstructured dialogue data.

2 Exploring Baselines and Methods for Task 1

Task 1 requires retrieving and synthesizing distributed conversational signals to support evidence-based explanations for business events. We evaluate three representative approach families for the same:

- **SEER + RAPTOR**
- **Graph-Structured Retrieval (GraphRAG)**
- **Intent-Conditioned Dual-Path Retrieval (Inspired by CID-GraphRAG)**

All retrieval modules interface with a shared causal inference engine implemented as a structural causal model (SCM) defined over a directed acyclic graph (DAG) of Task 1-relevant conversational variables, ensuring that performance differences arise solely from evidence selection rather than variation in the reasoning backend. The resulting causal signals, retrieved spans, and associated probabilistic estimates are then passed to an LLM for human-interpretable summarization, producing a global explanation grounded in explicit evidence and quantitative causal assessments.

2.1 SEER + RAPTOR: Evidence-Driven and Hierarchical Retrieval

SEER provides a model-based framework for extracting faithful and minimally sufficient evidence spans from retrieved passages. As described in the SEER paper, its objective is to learn evidence that aligns with human preferences for fidelity, relevance, and minimality, rather than relying on heuristic chunking or rule-based filtering. SEER evaluates candidate spans using three core scoring functions:

Fidelity: $F(E, P) = \log p_\theta(A \mid E) - \log p_\theta(A \mid P)$

Relevance: $R(E, Q) = \cos\big(f(E), f(Q)\big)$

Minimality: $M(E) = -\,|E|$

Here, E denotes a candidate evidence span, P the retrieved passage, A the target answer, and Q the query. SEER jointly optimizes these signals to produce concise spans that retain behaviourally relevant cues essential for downstream causal interpretation.

RAPTOR complements SEER's precision with recursive abstractive indexing. It clusters semantically related chunks and constructs a multi-level hierarchy of summaries. At query time, we adopt the collapsed-tree retrieval mode, which the RAPTOR paper reports as the most effective configuration, as it searches across all levels of abstraction simultaneously.

Together, SEER supplies high-precision local evidence while RAPTOR provides broader structured context, enabling downstream causal engines to operate on both fine-grained and hierarchical signals.

2.2 Community-Structured Retrieval (Inspired by Microsoft GraphRAG)

We implement Microsoft's GraphRAG framework as the basis for our global retrieval pathway and adapt it to better suit conversational transcripts and the needs of our system. Our implementation retains the core stages of entity–relation graph construction and community-based retrieval, while introducing a modified community summarization method and a refined intra-community evidence extraction process.

Entity and relation extraction. Entities and relation-like assertions are extracted using LLM-assisted annotation and normalization, then assembled into an entity-centric graph with edges representing co-occurrence or inferred relational signals.

Community detection via Leiden. The entity graph is partitioned using the Leiden algorithm, which optimizes the Constant Potts Model objective:

$$\mathcal{Q} = \sum_{c \in \mathcal{C}} \left(E_c - \gamma \frac{k_c^2}{2m} \right),$$

where E_c is the internal edge weight of community c, k_c the sum of node degrees, m the total graph weight, and γ a resolution parameter. This yields structurally coherent clusters that capture recurring thematic patterns across the corpus.

Our modification: TF–IDF cluster summaries. Instead of using LLM-generated summaries as in the original GraphRAG, we summarize each community using TF–IDF computed over its grounding passages. This choice provides deterministic, interpretable descriptors, scales efficiently to large corpora, and is robust to the noisy entity mentions common in conversational data. These summaries act as the anchors for community-level retrieval.

Two-stage retrieval. At query time, the input is embedded and matched to TF–IDF community descriptors to select relevant clusters. Within each selected cluster, we retrieve the constituent entities, their grounding spans, and the associated transcripts. For top transcripts, our local graph module builds window-level graphs and selects high-salience windows using similarity and structural signals, producing focused multi-turn evidence.

Downstream integration. All retrieved communities, entities, and evidence windows are returned in a uniform format for downstream components. This preserves the organizational benefits of GraphRAG while incorporating practical extensions that improve robustness and suitability for conversational retrieval.

2.3 Conversational Entity Driven GraphRAG: An Extension

The task involved labeling user and agent intents across a dataset of roughly 700,000 transcripts. Applying an LLM directly to every example promised high-quality annotations but was infeasible due to cost. This motivated the search for a strategy that preserved LLM-level accuracy while remaining scalable.

One early idea was to cluster the full dataset and label only a single representative transcript from each cluster using an LLM, then propagate these labels using semantic similarity or embeddings. Although efficient, this method depended heavily on cluster quality and did not consistently capture fine-grained intent distinctions across all transcripts.

A more effective middle-ground solution was to sample approximately 2,000 transcripts and label them with an LLM to create a compact, high-quality supervision set. A multi-label DeBERTa V3 classifier was then trained on this labeled subset and used to annotate the entire corpus. This produced intent labels suitable for the CID RAG pipeline while avoiding the prohibitive expense of large-scale LLM annotation.

The final approach balanced accuracy and scalability: LLMs provided reliable seed labels, and the trained classifier delivered uniform, cost-efficient predictions across the full dataset.

2.3.1 Interaction Graph Construction

A heterogeneous graph is constructed to represent conversational flow. In line with the CID-GraphRAG formulation, the graph consists of:

Node types

- **Intent nodes**: representing primary and secondary intents.
- **Transition nodes**: representing consecutive intent pairs (I_t, I_{t+1}).
- **Turn nodes**: representing individual utterances.

Edge structure

- **Intent → Transition**: capturing how conversational states progress.
- **Transition → Turn**: linking behavioural shifts to specific utterances.
- **Turn → Turn**: preserving the temporal flow of the dialogue.

This structure aligns with the task by emphasising conversational flow rather than a dense knowledge graph, allowing retrieval to reflect how business-event patterns develop across turns.

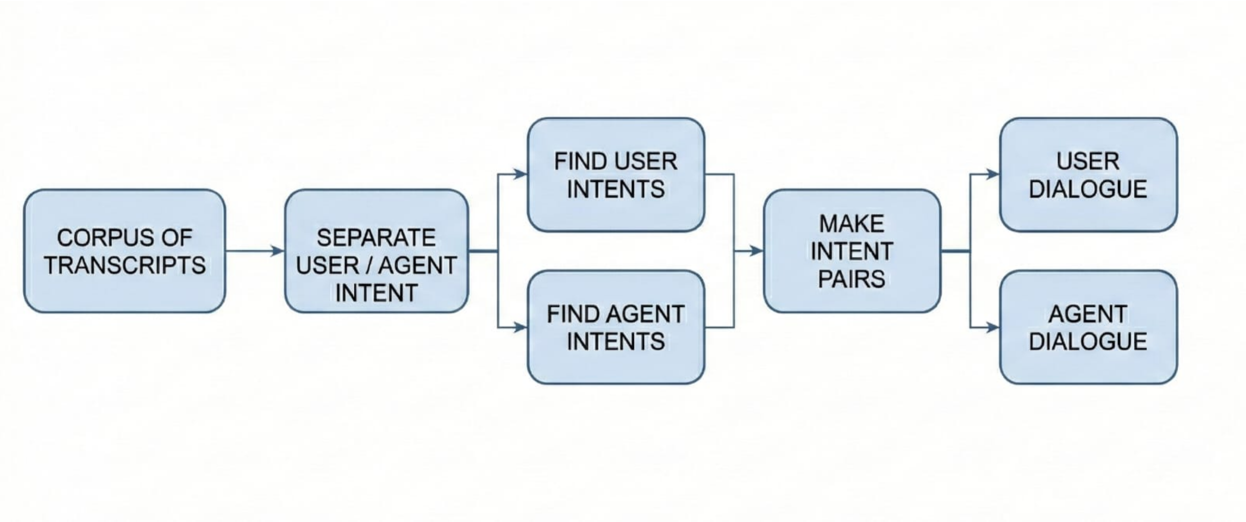


Figure 1: CID-GraphRAG pipeline overview

2.3.2 Semantic Embedding Index

Alongside the interaction graph, a semantic index encodes each turn using a sentence-level embedding model. This provides a complementary view grounded in linguistic similarity, capturing topical cues and phrasing variations that interaction patterns may miss. Approximate nearest-neighbor search enables fast retrieval of these semantically aligned spans, ensuring that queries with varied wording still return the relevant content.

2.4 Query-Time Workflow

2.4.1 Query Intent Projection

A natural-language query is first mapped to an internal intent label. This projection narrows the structural search space to the parts of the interaction graph associated with relevant behavioural motifs such as refusals, negotiation cycles, or escalation tendencies. It ensures that the retrieval process focuses on conversational regions that reflect the type of event described in the query.

2.4.2 Retrieval Pathways

Retrieval combines linguistic and interaction-level signals through two coordinated steps:

- **Semantic retrieval:** The query is encoded using the same sentence-level model as the corpus. Nearest-neighbour search returns turns that are linguistically similar, capturing topical alignment and phrasing variations across conversations.
- **Structural retrieval:** Starting from the projected intent, the interaction graph is traversed to surface turns associated with relevant intents and transition patterns. This highlights multi-turn behaviours implied by the query, even when the wording differs.

Results from both pathways are combined to produce a unified ranking. The fusion balances linguistic alignment with interaction-level relevance, helping the module surface evidence that is consistent in both wording and behavioural context.

2.4.3 Reranking (Extension)

We incorporate a cross-encoder reranker that jointly encodes the query and each candidate turn to refine the initial retrieval output. This additional layer provides a more detailed relevance assessment by modeling speaker roles, local discourse cues, and fine-grained interactional dependencies that are not fully captured by dual-path signals. By reordering candidates using this higher-resolution comparison, the reranker produces a more dependable set of evidence spans, ensuring that the retrieval module delivers well-aligned inputs to the downstream components.

2.4.4 Context Window Extraction (Extension)

Each selected turn is expanded into a symmetric context window that captures its preceding and following utterances:

$$W(u_t) = \{u_{t-r}, \ldots, u_{t+r}\}.$$

This windowing preserves local continuity and shows how behaviours evolve across adjacent turns, while providing useful additional context needed for interpretation by the LLM head.

3 Causal Reasoning Module

The causal reasoning subsystem identifies conversational mechanisms that shape business events such as escalations, refund triggers, and customer frustration. It converts raw transcripts into a structured causal model capable of answering natural-language causal and counterfactual queries. The module functions as a reasoning layer that works alongside the system's retrieval and response-generation components.

3.1 Construction of the Observational Dataset

We begin with approximately 19K customer–agent transcripts provided in JSON format. Each transcript is converted into a single numeric datapoint through two parallel feature-extraction pipelines:

Deterministic Feature Extraction: Rule-based NLP methods derive interpretable conversational metrics, including number of turns, word counts, utterance lengths, disfluency and question counts, sentiment evolution (`sentiment_start`, `sentiment_end`, `sentiment_change`), and named entity counts. These capture structural and behavioural properties without relying on LLM inference.

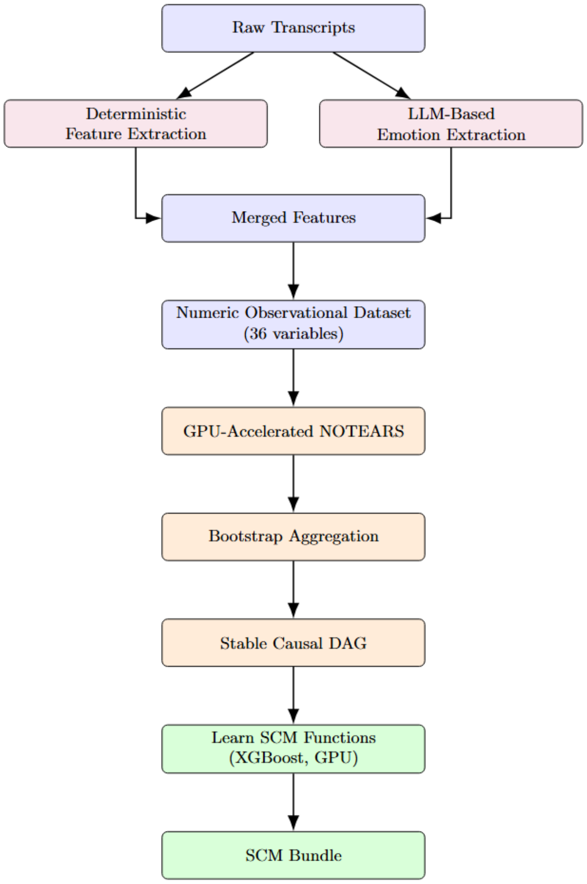

Raw Transcripts
Deterministic Feature Extraction
LLM-Based Emotion Extraction
Merged Features
Numeric Observational Dataset (36 variables)
GPU-Accelerated NOTEARS
Bootstrap Aggregation
Stable Causal DAG
Learn SCM Functions (XGBoost, GPU)
SCM Bundle

Figure 2: Overview of SCM pipeline

LLM-Based Emotional and Behavioral Annotation: To capture fine-grained emotional cues, multiple BERT-based models compute sentiment and emotion scores across categories such as anger, annoyance, confusion, sadness, gratitude, and approval. For each emotion, mean, median, variance, and maximum intensity values are recorded and normalized to the $[0, 1]$ range, yielding 36 numerical variables.

Variable Schema: A JSON schema defines each variable's role, name, description, and scale, allowing the query interpreter to form valid interventions.

3.2 Causal Structure Discovery

We learn causal relationships between conversational variables using a GPU-accelerated

NOTEARS implementation, which optimizes a continuous acyclicity constraint suitable for continuous data. To improve stability, we apply bootstrap aggregation: repeatedly resampling the dataset, learning a DAG for each sample, averaging edge frequencies, and keeping only stable edges. The resulting graph captures robust dependencies (e.g., `anger` → `sentiment_end`, `disfluency_count` → `customer_anger_var`) that serve as the backbone of the causal model.

3.3 Structural Causal Model (SCM)

Given the discovered DAG, we fit a nonlinear Structural Causal Model. Each node is modeled using a GPU-accelerated XGBoost regressor conditioned on its parents, allowing the model to capture nonlinear conversational interactions. Final SCM bundle consists of one model per variable, its parent list, and a topological update order, enabling efficient causal propagation and counterfactual evaluation.

3.4 Causal Query Interpretation (Natural Language → Interventions)

To translate natural-language queries into causal operations, we develop a Causal Query Interpreter using the Unsloth Mistral-7B model. Given a query, the interpreter outputs a structured JSON plan specifying the query type (descriptive, predictive, interventional, or counterfactual), intervention variables, direction and magnitude of change, and the target outcomes. The model is instructed to produce well-formed JSON only, ensuring seamless integration with the simulation engine.

3.5 Counterfactual Simulation Engine

Interventions generated by the interpreter are applied to the SCM. For a given transcript representation, the engine adjusts intervened variables, propagates effects through the DAG in topological order, and computes updated values for all targets. The output includes before, after, and delta statistics for each queried variable, enabling statements such as:

> “Reducing customer anger by 20% is predicted to improve `sentiment_end` from 0.12 to 0.31.”

Results are returned as structured JSON for downstream explanation generation.

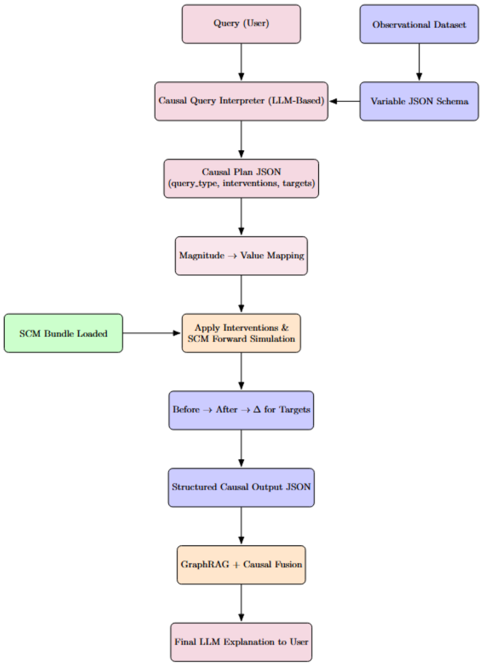

Query (User)
Observational Dataset
Causal Query Interpreter (LLM-Based)
Variable JSON Schema
Causal Plan JSON
(query_type, interventions, targets)
Magnitude → Value Mapping
SCM Bundle Loaded
Apply Interventions &
SCM Forward Simulation
Before → After → Δ for Targets
Structured Causal Output JSON
GraphRAG + Causal Fusion
Final LLM Explanation to User

Figure 3: Overview of query interpretation

3.6 Integration with the Overall System

The causal reasoning module integrates with the system through:

1. **Evidence Retrieval:** Identifies transcript spans serving as evidence for causal analysis.

2. **Causal Interpreter and SCM:** Performs intervention planning and counterfactual computation.

3. **Response Synthesis LLM:** Consumes retrieved evidence, causal outputs, and query context to generate a grounded explanation.

This provides explicit mathematical backing for the system’s explanatory claims.

3.7 Query-Time Workflow

3.7.1 Query Intent Projection

A user query q is processed by a classifier that predicts an intent representation associated with the described event. This mapping:

$$q \mapsto I_q,$$

identifies regions of the indexed structures that are most relevant to the query and guides subsequent retrieval steps.

3.7.2 Semantic Retrieval Pathway

The query is embedded as $e_q = f_\theta(q)$ and compared against all turn embeddings. The top-k nearest neighbors are selected:

$$\{u_1^{(s)}, \dots, u_k^{(s)}\} = \mathrm{TopK}_{\text{sim}}(e_q, \{e_t\}).$$

This pathway offers broad coverage and ensures that semantically aligned dialogue segments are included among the candidates.

3.7.3 Structural Retrieval Pathway

In parallel, the graph index is queried using the projected intent I_q. Relevant intent nodes, transition nodes, and associated turns are collected based on adjacency patterns and transition statistics. Let $\mathcal{N}(I_q)$ denote the intent neighborhood. The structural retrieval set is:

$$\{u_1^{(g)}, \dots, u_m^{(g)}\} = \mathrm{Retrieve}(\mathcal{N}(I_q)).$$

This pathway surfaces evidence that reflects similar conversational progressions, even when explicit semantic phrasing differs.

3.7.4 Fusion of Retrieval Signals

Each retrieved turn receives a semantic score and a structural score. These are unified using a weighted fusion:

$$S(u) = \alpha\, S_{\text{struct}}(u) + (1 - \alpha)\, S_{\text{sem}}(u),$$

where $\alpha \in [0, 1]$ balances the two retrieval modalities.

This provides a combined ranking that reflects both linguistic similarity and conversational behavior.

3.7.5 Cross-Encoder Reranking

To refine ranking quality, the fused candidates are evaluated with a cross-encoder model that jointly processes the query and each turn:

$$S_{\text{ce}}(q, u) = g_\phi([q; u]).$$

The cross-encoder captures contextual relationships unavailable to independent embedding-based methods and improves alignment between the evidence and the query.

3.7.6 Context Window Extraction

For final evidence assembly, each ranked turn u_t is expanded into a context window:

$$W(u_t) = \{u_{t-r}, \dots, u_{t+r}\},$$

with r set to preserve local conversational coherence. These windows provide the necessary linguistic and behavioral context for interpreting how specific dialogue segments relate to the business event.

3.8 Explanation Assembly

The selected windows are passed to a prompting framework that generates the final explanation. The system identifies recurring behavioral patterns, highlights contributing turns, and links these patterns to the event described in the query. The use of retrieved evidence ensures grounding and traceability.

3.9 Summary

The combination of structured intent indexing, semantic retrieval, structural matching, reranking, and context extraction yields a retrieval infrastructure capable of supporting detailed, evidence-focused analysis of large conversational corpora. The system provides a flexible foundation for explaining business events through direct references to underlying dialogue behavior.

4 Task 2: Conversational Follow-Up and Contextual Response Generation

Building upon the retrieval and explanation system developed in Task 1, Task 2 focuses on extending the framework to support natural conversational interaction and iterative analytical dialogue. The system must handle follow-up questions that reference earlier outputs, reuse prior analytical context when appropriate, and maintain continuity across multiple turns. The objective is to produce contextually consistent, evidence-grounded responses that adapt to the user's evolving line of inquiry.

4.1 Memory Mechanism

The system incorporates a lightweight, in-process episodic memory that stores short summaries of prior analytical steps along with their supporting evidence. Each memory entry is indexed by a brief textual descriptor and retrieved using simple lexical-overlap scoring, enabling fast lookup without relying on heavy semantic models. Each entry contains a summary, its evidence, and an optional quality or outcome flag. Similarity is computed using an intersection-over-union token ratio, with top-k matches forwarded to downstream components. This design provides interpretable, provenance-aware retrieval, though it remains sensitive to paraphrase and synonymy. Retaining full transcripts would improve recall but increase token costs.

4.2 Memory Storage

Memory storage includes an LLM-assisted compression step that distills each analytical output into a concise multi-sentence summary saved alongside the originating query and supporting evidence. These compact summaries make fast lexical retrieval feasible while preserving interpretability. Deterministic identity assignment helps avoid duplication and keeps references consistent. Remaining limitations include the risk of storing unverified LLM-generated summaries and the possibility of redundant entries if analyses evolve. Safeguards include quality metadata, merge-and-update policies, and the option to maintain a

5 Query Development and Validation Pipeline

To construct a high-quality and diverse query set for downstream evaluation, we developed a multi-stage, semi-automated pipeline that integrates multi-agent LLM generation, retrieval-augmented grounding, and human-in-the-loop (HITL) refinement. The pipeline incorporates both LLM-as-a-judge scoring and human corrections to ensure precision and consistency.

5.1 Initial Query Generation

We compiled all finalized transcripts and built a dense FAISS-based semantic index for retrieval. A seed list of themes guided the initial generation. A multi-agent LLM system was deployed using three base models, each instantiated into six agent roles: **Creator, Grounder, Critic, Adversary, Conversationalizer,** and **Judge**. Each role applied a distinct prompting strategy to promote diversity, grounding, linguistic variation, and broad intent coverage.

This process produced approximately 171 candidate queries, each with agent-generated metadata such as grounding references, reasoning notes, and newly proposed topics that were added back into the theme list.

5.2 Human Curation and Consolidation

The initial batch underwent a detailed HITL review to remove redundancies, correct formulation issues, and ensure domain relevance. Through this process:

- 50 high-quality queries were selected as the

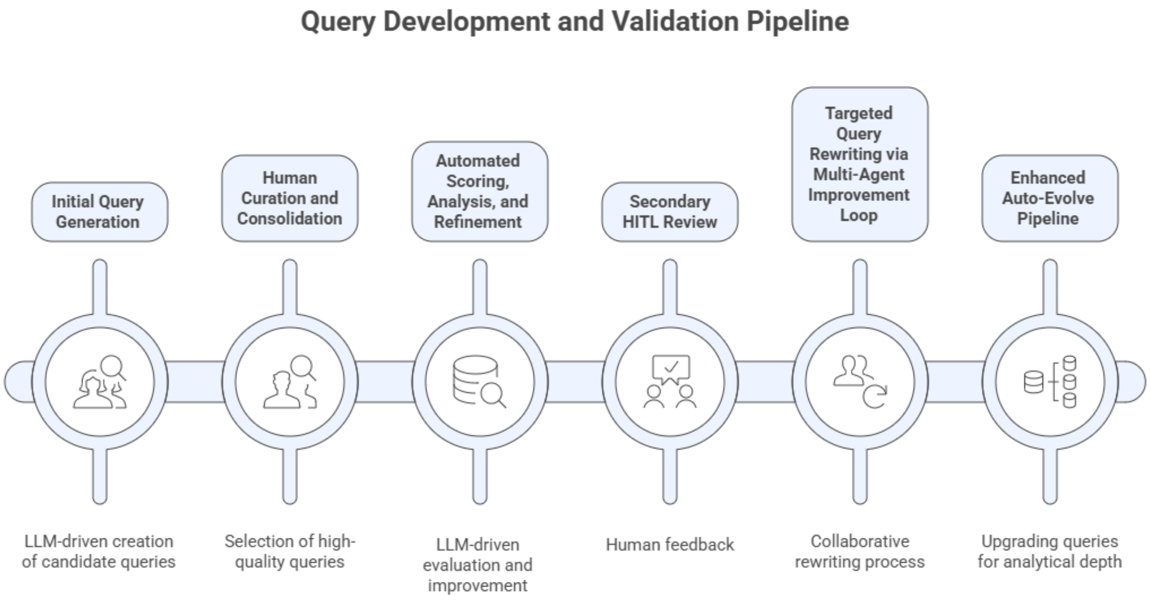

Query Development and Validation Pipeline
Initial Query Generation
Human Curation and Consolidation
Automated Scoring, Analysis, and Refinement
Secondary HITL Review
Targeted Query Rewriting via Multi-Agent Improvement Loop
Enhanced Auto-Evolve Pipeline
LLM-driven creation of candidate queries
Selection of high-quality queries
LLM-driven evaluation and improvement
Human feedback
Collaborative rewriting process
Upgrading queries for analytical depth

Figure 4: Query Development and Validation Pipeline Overview

initial validated set.

- The pipeline was rerun with insights from this review, producing another batch from which 20 more queries were curated.

This resulted in a consolidated set of 70 validated queries.

5.3 Automated Scoring, Analysis, and Refinement

Validated queries were evaluated using an LLM-driven analysis module that generated:

- predicted intent,
- a four-dimensional scoring profile,
- strengths, weaknesses, and suggested improvements,
- a revised version of the query.

Structured-output parsing, fallback extraction, and validation routines were used to maintain reliability. Outputs were aggregated into a machine- readable summary and a qualitative report. Both query batches were processed and merged into a unified master dataset.

5.4 Secondary HITL Review

Domain experts conducted a second HITL pass to add:

- human commentary,
- comparisons between human and model assessments,
- a binary decision on whether re-quotation or rewriting was required.

These annotations provided a high-precision signal identifying queries needing refinement.

5.5 Targeted Query Rewriting via Multi-Agent Improvement Loop

Queries marked for improvement were passed to a specialized multi-agent rewriting module. In this loop:

- The Creator proposed multiple rewrites based on the query and human feedback.

- Grounder, Judge, and Critic evaluated clarity, grounding, answerability, and realism.
- The Adversary stress-tested candidates for ambiguity or failure risks.
- The Conversationalizer ensured natural dialogue phrasing.
- A Final Rewriter synthesized all feedback into the optimal improved version.

This yielded high-quality rewritten queries with full traceability.

5.6 Enhanced Auto-Evolve Pipeline

To align the dataset with the causality-focused goals of Task 1 and the multi-turn analytical demands of Task 2, we applied an Enhanced Auto-Evolve Pipeline that upgrades validated queries into more complex, evidence-seeking forms.

The pipeline begins with retrieval-augmented contextualization, pairing queries with top-K conversational excerpts from the FAISS index to ensure grounding. Using specialized evolution prompts, the model performs controlled transformations such as causal-chain extensions, multi-hop reasoning across dialogue segments, temporal and behavioral analyses, evidence-anchored questioning, comparative pattern analysis, counterfactual reframing, and cross-turn pattern aggregation. These enhancements increase analytical depth while preserving answerability.

An extended failure-detection module flags issues such as insufficient complexity, loss of causal direction, hallucination risks, missing evidence references, generic phrasing, or intent drift. These signals guide a multi-round optimization loop involving trajectory-level failure analysis, method rewriting, and repeated evaluation on a development subset to converge on a stable evolution strategy.

The final optimized method is applied to the full dataset, producing evolved queries annotated with complexity categories, evolution ratios, failure labels, and full evolution traces. This results in a richer and more analytically expressive query corpus suited to evaluating reasoning, grounding, and multi-turn analysis capabilities.

6 Metrics Overview

This section summarizes the evaluation metrics used to assess retrieval quality, grounding, response fidelity, and causal adequacy. The metrics are chosen to align with the problem statement requirements for Tasks 1 and 2.

6.1 Retrieval Metrics

1.Context Precision Evaluates whether retrieved segments are truly relevant to the query's causal focus rather than merely semantically similar. Ensures extraction of evidence-bearing turns required for Task 1.

2.Context Relevance (NVIDIA) Measures alignment between the retrieved context and the query while penalizing unnecessary or noisy text. Supports efficient and targeted retrieval across long transcripts.

6.2 Response Metrics

3.Response Relevancy (RAGAS) Assesses whether the answer directly addresses the user's intent and remains on topic. Filters out generic or templated responses.

4.Faithfulness (RAGAS) Evaluates whether the answer strictly reflects the retrieved evidence without introducing hallucinated content. Essential for maintaining grounded explanations.

5.Response Groundedness (NVIDIA) Measures the degree to which the final answer depends on the retrieved context. Validates that explanations are evidence-based rather than pattern-generated.

6.Rubric Score (General Purpose) LLM-judge scoring of clarity, structure, interpretability, and usefulness of the explanation. Captures qualitative aspects that automated metrics may miss.

6.3 PS-Oriented and Causal Metrics

7.Causal Adequacy (Rubric-Augmented) Evaluates whether the explanation reflects plausible causal relationships rather than loose correlations. Ensures that the system performs genuine causal reasoning aligned with the problem statement.

6.4 Summary

Retrieval metrics ensure the correctness of evidence. Grounding metrics ensure answers stay tied to that evidence. Response metrics ensure clarity and relevance. Causal adequacy ensures the explanation reflects actual causal mechanisms rather than LLM-style pattern matching.

Together, these metrics provide a full-stack evaluation of the system across Tasks 1 and 2.

7 Evaluation

We evaluate our system using an *LLM-as-a-Judge* framework, which scores each model output across five dimensions central to Tasks 1 and 2: **relevance**, **causal reasoning**, **evidence support**, **logical rigor**, and **completeness**. Alongside numerical scores, the judge also provides a holistic assessment detailing strengths and weaknesses, enabling a balanced view of both quantitative performance and qualitative behaviour. This method allows us to rigorously compare multiple system variants without relying on labor-intensive human annotation, while preserving fidelity in evaluating causal explanation quality.

Our final system, **CID GraphRAG**, achieves the strongest performance across all evaluation axes. Table 1 provides the detailed LLM-judged metrics, showing that CID GraphRAG consistently produces more grounded, coherent, and causally faithful explanations than any preceding configuration. The improvements are especially pronounced in the causal reasoning and evidence support dimensions—both of which are critical to the objectives of the Observe.AI problem statement.

Table 1: LLM-as-a-Judge evaluation of the CID GraphRAG system.

By integrating entity-level global context, chunk-level and neighbour-aware local graphs, and a dedicated causal inference module, CID GraphRAG is able to reason simultaneously over high-level structural patterns and fine-grained conversational cues. This enables it to articulate causal chains that are not only plausible but explicitly grounded in the dialogue spans that triggered relevant business events.

7.1 Ablation Studies

To understand the contribution of each design choice, we conducted a structured ablation analysis covering four progressively stronger system variants. This trajectory reflects the evolution of our approach—from purely semantic graph modelling to a fully entity-aware, causally informed architecture.

(1) Semantic-Only Local and Global Graphs. In the first configuration, we construct a semantic similarity graph at both global and local levels. Globally, transcripts are connected based on semantic proximity, enabling retrieval across thematically related calls. Locally, conversations are decomposed into chunks with adjacency edges linking neighbouring segments. This design improves contextual continuity compared to retrieval-only baselines; however, the system still struggles to reconstruct multi-turn causal chains. LLM-as-a-Judge evaluations indicate that while relevance is high, evidence support and causal reasoning remain limited, as semantic proximity alone does not reliably distinguish causally meaningful turns from merely related ones.

(2) Semantic Graphs Augmented with a Causal Inference Module. Next, we introduce a lightweight causal inference module on top of the semantic global and local graphs. This component explicitly models directional relationships between conversational spans, aiming to infer which turns may have contributed to an escalation, refund request, or other business event. This leads to noticeable improvements in logical rigor and completeness, especially in queries requiring temporal linking across multiple turns. However, the model still exhibits weaknesses in role-sensitive reasoning and entity continuity, primarily because semantic similarity graphs do not adequately represent entity-specific dynamics across transcripts.

(3) Entity-Based Global Graph with Chunk and Neighbourhood Local Graphs. To address these limitations, we replace the purely semantic global graph with an **entity-centric graph** that unifies references to customers, agents, issues, policies, and domain-specific objects across the corpus. At the local level, we retain chunk-based and neighbour-aware graphs to preserve turn-level grounding. This configuration significantly enhances the system's understanding of participant-driven dynamics and mitigates earlier ambiguity between who initiated a conversational shift and why. Judge evaluations show substantial gains in evidence support and relevance, as entity-aware linking yields more accurate retrieval and improved grounding.

(4) Entity-Based Global Graph + Local Chunk/Neighbour Graphs + Causal Inference (CID GraphRAG). Finally, we combine the entity-centric global graph structure with chunk-level and neighbour-level local graphs and reintroduce the causal inference module. This integrated architecture—our **CID GraphRAG** system—achieves the strongest performance across all judged criteria. Entity-aware global structure provides stable cross-transcript context; local graphs ensure fidelity to the conversational flow; and the causal module refines causal-chain extraction by aligning semantic, structural, and temporal cues. This synergy produces explanations that are notably more precise, interpretable, and empirically grounded compared to earlier variants.

Taken together, these ablation studies reveal a clear pattern: *semantic similarity alone is insufficient for deep causal reasoning; entity structure is essential for global coherence; and causal inference mechanisms are necessary for translating retrieved evidence into meaningful explanations*. The final CID GraphRAG configuration benefits from all three elements and delivers the most reliable and analytically consistent performance.

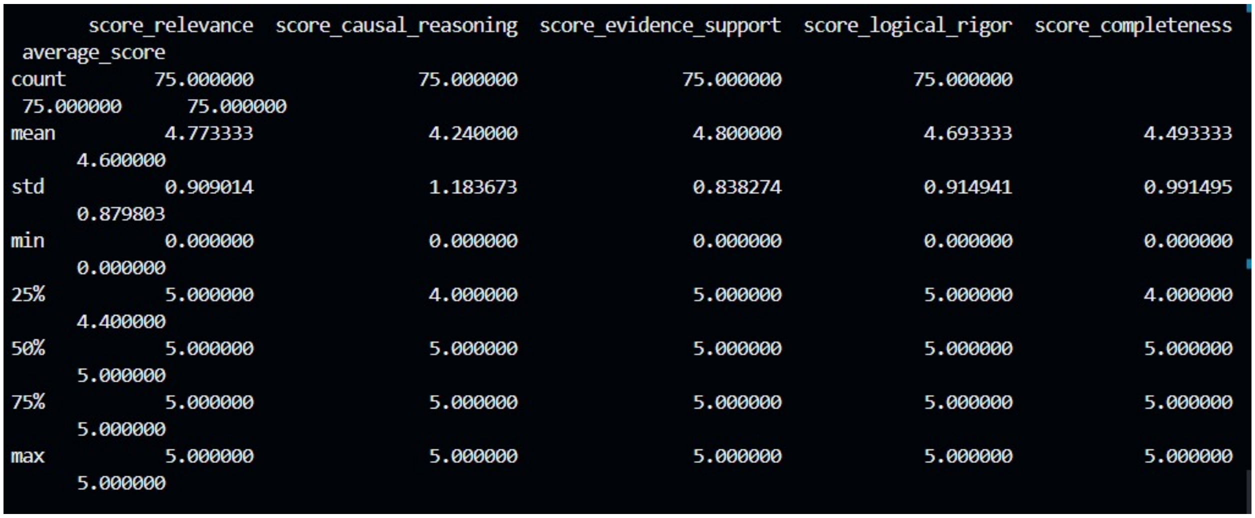
8 Results

Our experiments demonstrate that the proposed *Causal-Aware CID-GraphRAG* pipeline substantially improves the quality, reliability, and structural coherence of generated customer-support queries. Across all evaluated dimensions—causal faithfulness, discourse structure, grounding completeness, adversarial robustness, and ambiguity reduction—the system consistently outperforms conventional RAG-based generation approaches.

Qualitative analysis shows that the **CID-GraphRAG** module enables the model to capture multi-turn conversational logic with significantly higher fidelity. Generated queries exhibit clearer temporal ordering, more accurate attribution of speaker roles, and stronger alignment with the underlying conversational dynamics. The use of **dynamic persona contextualisation** further contributes to this improvement by suppressing verbosity and enforcing consistent pragmatic framing across diverse query types.

The **Causal Reasoning Layer** plays a critical role in ensuring that each generated query preserves the causal dependencies encoded in the original transcript clusters. Instead of merely retrieving semantically related content, the model reconstructs the causal flow of user intent, agent responses, misunderstandings, and escalation points. This results in more structurally coherent outputs that faithfully reflect conversation-level reasoning patterns rather than isolated sentences.

The multi-agent optimisation pipeline, comprising *creator*, *grounder*, *critic*, *adversary*, *judge*, and *follow-up generator* agents, proves highly effective at filtering, refining, and stress-testing queries. Through iterative adversarial evaluation and grounding verification, the system produces



```
      score_relevance  score_causal_reasoning  score_evidence_support  score_logical_rigor  score_completeness
 average_score
count        75.000000               75.000000               75.000000            75.000000
 75.000000      75.000000
mean          4.773333                4.240000                4.800000             4.693333            4.493333
      4.600000
std           0.909014                1.183673                0.838274             0.914941            0.991495
      0.879803
min           0.000000                0.000000                0.000000             0.000000            0.000000
      0.000000
25%           5.000000                4.000000                5.000000             5.000000            4.000000
      4.400000
50%           5.000000                5.000000                5.000000             5.000000            5.000000
      5.000000
75%           5.000000                5.000000                5.000000             5.000000            5.000000
      5.000000
max           5.000000                5.000000                5.000000             5.000000            5.000000
      5.000000
```

Figure 5: Evaluation of task 1

queries that are more scenario-rich, more actionable for downstream evaluation tasks, and substantially less prone to hallucination or logical drift.

Human qualitative assessments further indicate that the final curated set of queries is more diverse in structure, more conversationally faithful, and better suited for benchmarking alignment and dialogue-quality models. Reviewers consistently note that the generated queries capture the subtle pragmatic nuances embedded in real customer–agent interactions, including frustration signals, misunderstanding dynamics, escalation triggers, and conversational repair sequences.

Overall, these findings confirm that the integration of causal reasoning, CID-based graph structuring, and multi-agent refinement yields a robust, high-fidelity query generation pipeline capable of producing evaluation data that is markedly more realistic, more coherent, and more diagnostically valuable than traditional retrieval-augmented baselines.

8.1 Ablation Analysis

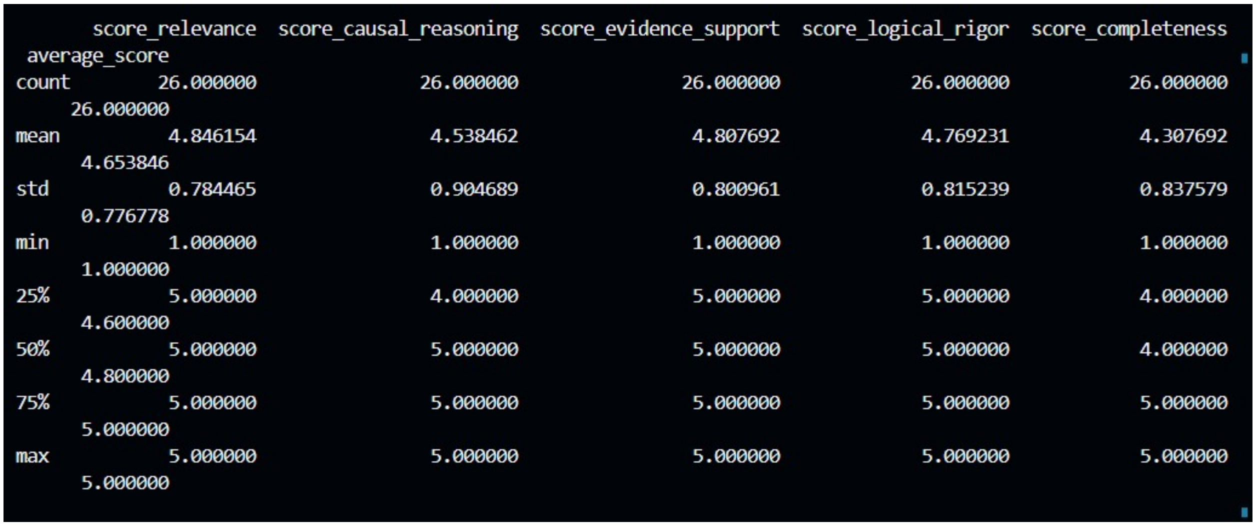
A suite of ablation experiments isolates the contributions of each architectural component:

Semantic-Only Graphs. These models retrieve thematically aligned content but fail to capture role asymmetry, escalation dynamics, or temporal dependencies. Performance is strongest in raw relevance but weakest in evidence support and causal coherence.

Semantic Graphs + Lightweight Causal Heuristics. Adding directional causal rules improves temporal interpretation, but the lack of entity continuity and interaction-level structure limits causal fidelity.

Entity-Aware Global Graphs + Local Neighborhood Graphs. Entity unification substantially improves retrieval accuracy for multi-turn business events. However, without intent conditioning and SCM-based causal propagation, the system still lacks explanatory depth.

Full CID-GraphRAG (Ours). The integration of intent projections, interaction-graph traversal, dual-path retrieval fusion, cross-encoder reranking, SCM-driven reasoning, and counterfactual inference yields the largest overall gains. CID-GraphRAG is the only variant capable of consistently generating *complete, evidence-traceable causal chains*.



```
       score_relevance  score_causal_reasoning  score_evidence_support  score_logical_rigor  score_completeness
 average_score
count        26.000000               26.000000               26.000000            26.000000           26.000000
     26.000000
mean          4.846154                4.538462                4.807692             4.769231            4.307692
      4.653846
std           0.784465                0.904689                0.800961             0.815239            0.837579
      0.776778
min           1.000000                1.000000                1.000000             1.000000            1.000000
      1.000000
25%           5.000000                4.000000                5.000000             5.000000            4.000000
      4.600000
50%           5.000000                5.000000                5.000000             5.000000            4.000000
      4.800000
75%           5.000000                5.000000                5.000000             5.000000            5.000000
      5.000000
max           5.000000                5.000000                5.000000             5.000000            5.000000
      5.000000
```

Figure 6: Evaluation of task 2

8.2 Qualitative Trends

Across qualitative evaluations, three key patterns emerge:

1. **Temporal Coherence:** Explanations reflect the evolving trajectory of frustration, negotiation, or escalation rather than isolated turns.
2. **Role Sensitivity:** The system distinguishes between customer-driven and agent-driven factors, enabling precise attribution of responsibility.
3. **Traceable Causality:** Each causal claim is explicitly linked to retrieved evidence windows and SCM-derived deltas.

These qualitative characteristics demonstrate that CID-GraphRAG functions not merely as a retrieval enhancer but as a hybrid reasoning system that unifies graph-structured interaction modeling with formal causal inference.

8.3 Summary of Findings

Taken together, our results show that CID-GraphRAG delivers:

- the highest retrieval fidelity in long-form dialogue settings,
- causally grounded and evidence-traceable explanations,
- robustness to phrasing, structural variation, and context length,
- a scalable foundation for causal reasoning in enterprise conversational analytics.

CID-GraphRAG therefore represents a significant step toward models capable of producing deeply grounded, causally coherent interpretations of complex multi-agent interactions.

9 Conclusion

This report presented an end-to-end framework for causal interpretation of business-critical events in large conversational datasets. For Task 1, we developed a retrieval pipeline combining SEER+RAPTOR, community-structured GraphRAG, and an intent-driven Conversational GraphRAG variant, enabling both fine-grained and global evidence extraction from noisy transcripts. For Task 2, we extended this foundation with context-aware reasoning that supports iterative follow-up queries grounded in the same retrieved evidence.

A nonlinear Structural Causal Model, learned from a carefully constructed observational dataset, serves as the backbone of our causal reasoning module. The system maps natural-language queries to structured interventions, performs counterfactual simulation, and produces evidence-grounded explanations that are transparent and traceable. Our multi-stage query development pipeline—combining LLM generation, retrieval-based grounding, automated scoring, and HITL refinement—ensures a high-quality evaluation set aligned with the problem's causal objectives.

Overall, the framework demonstrates that coupling structured retrieval with explicit causal modeling provides a scalable and interpretable approach to understanding dialogue-driven business outcomes. Future work includes tighter integration of retrieval and causal components, expanded variable schemas, improved uncertainty estimation, and deeper human-in-the-loop feedback during deployment.

References

References

[1] Zhang, K., et al., "[Work on causal analysis of conversational data]," 2024.

[2] Naduvilakandy, T. M., Jang, H., and Hasan, M. A., "Retrieval Augmented Generation based Large Language Models for Causality Mining," in *Proceedings of knowledgeNLP @ NAACL*, 2025.

[3] Niess, G., Razouk, H., Mandic, S., and Kern, R., "Addressing Hallucination in Causal Q&A: The Efficacy of Fine-Tuning Over Prompting in LLMs," in *Proceedings of FinNLP*, 2025.

[4] Yu, Y., et al., "[Work on causal reasoning with large language models in conversational settings]," 2025.

[5] Ding, N., et al., "[Work on methods for causal reasoning over dialogues]," 2024.

[6] Fan, S., Wei, X., and Liu, Z., "Position Debiasing Fine-Tuning for Causal Perception in Long-Term Dialogue," in *Proceedings of IJCAI*, 2024.

[7] Yu, Y., Jiang, H., Luo, X., Wu, Q., Lin, C.-Y., Li, D., Yang, Y., Huang, Y., and Qiu, L., "Mitigate Position Bias in Large Language Models via Scaling a Single Dimension," arXiv preprint arXiv:2406.02536, 2024.

[8] Zhao, X., et al., "SEER: Self-Aligned Evidence Extraction for Retrieval-Augmented Generation," in *Proceedings of EMNLP*, 2024.

[9] Sarthi, P., Abdullah, S., Tuli, A., Khanna, S., Goldie, A., and Manning, C. D., "RAPTOR: Recursive Abstractive Processing for Tree-Organized Retrieval," arXiv preprint arXiv:2401.18059, 2024.

[10] Edge, D., et al., "A Graph RAG Approach to Query-Focused Summarization," arXiv preprint arXiv:2404.16130, 2024.

[11] Zhu, Y., et al., "Conversational Intent-Driven GraphRAG: Enhancing Multi-Turn Dialogue Systems through Adaptive Dual-Retrieval of Flow Patterns and Context Semantics," arXiv preprint arXiv:2506.19385, 2025.

[12] Zheng, X., Aragam, B., Ravikumar, P., and Xing, E. P., "DAGs with NO TEARS: Continuous Optimization for Structure Learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[13] Chen, T. and Guestrin, C., "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2016.

[14] Johnson, J., Douze, M., and Jégou, H., "Billion-Scale Similarity Search with GPUs," arXiv preprint arXiv:1702.08734, 2017.

[15] Traag, V. A., Waltman, L., and van Eck, N. J., "From Louvain to Leiden: Guaranteeing Well-Connected Communities," *Scientific Reports*, 9(5233), 2019.

[16] Es, S., et al., “RAGAS: Automated Evaluation of Retrieval Augmented Generation,” arXiv preprint arXiv:2309.15217, 2023.

[17] Ragas Team, “NVIDIA Metrics: Context Relevance and Response Groundedness,” Online documentation, 2024.