Causal Rationale Extraction and Synthesis from Conversational Data on Business Events

Team Name

November 10, 2025

Abstract

This report presents a comprehensive system for extracting causal rationales from large-scale conversational transcripts to identify relationships between dialogue dynamics and business events such as escalations, refunds, and churn. The system implements a two-task framework: (1) query-driven evidence-based causal explanation generation, and (2) contextual follow-up conversation support. Our approach combines retrieval-augmented generation (RAG), semantic search, reranking, and causal pattern detection to provide interpretable, evidence-based explanations. We evaluate the system using a curated dataset of 50-100 queries across multiple event types and demonstrate its effectiveness through quantitative metrics and qualitative analysis.

1 Introduction

1.1 Problem Statement

In large-scale customer-interaction operations, contact centers process tens of thousands to hundreds of thousands of agent-customer dialogues. Many conversations culminate in business-critical outcomes such as customer escalations, refund requests, or signals of churn. These events carry significant cost, risk, and operational overhead. However, current monitoring systems often flag that an adverse event occurred but provide little insight into why it happened.

The challenge lies in identifying which turns or segments of dialogue triggered the event, which conversational cues systematically lead to specific outcomes, and what temporal patterns presage certain events. Without this visibility, organizations cannot systematically perform root-cause analysis, coach agents precisely, redesign processes, or intervene proactively.

1.2 Objectives

This work aims to develop a system that:

- Processes large volumes of transcript data with speaker labels and turn indexing
- Models conversational dynamics and maps dialogue flows to business events
- Surfaces specific dialogue spans that most likely causally contributed to events
- Enables analytic querying across call corpora to identify recurring causal motifs
- Provides evidence-based, interpretable explanations for business events
- Supports contextual follow-up conversations for iterative analysis

2 Related Work

Previous work in conversational analysis has focused on sentiment analysis, topic modeling, and event detection. However, causal rationale extraction from dialogue remains underexplored. Our approach builds upon:

- **Retrieval-Augmented Generation (RAG)**: Combining dense retrieval with language models for grounded generation
- **Causal Inference in NLP**: Methods for identifying causal relationships in text
- **Dialogue Analysis**: Techniques for understanding conversational dynamics
- **Explainable AI**: Approaches for providing interpretable explanations

3 System Architecture

3.1 Overview

Our system consists of six main components:

1. **Data Processing Pipeline**: Transcript ingestion, preprocessing, and vector database indexing
2. **Retrieval System**: Semantic search, reranking, and dialogue span extraction
3. **Causal Analysis Module**: Pattern detection, evidence scoring, and causal span identification
4. **Query Processing**: Natural language understanding and intent classification
5. **Explanation Generation**: LLM-based synthesis with evidence citation
6. **Conversation Management**: Context tracking and follow-up handling

3.2 Data Processing Pipeline

3.2.1 Transcript Loading

The system supports multiple transcript formats:

- **JSON**: Structured format with turns, events, and metadata
- **CSV**: Tabular format with columns for transcript_id, turn_id, speaker, text, timestamp, event_type, event_label
- **TXT**: Simple text format with basic parsing

3.2.2 Preprocessing

The preprocessing module performs:

- Speaker label normalization (agent/customer)
- Text cleaning and normalization
- Turn segmentation and indexing
- Event type normalization
- Dialogue structure extraction

3.2.3 Vector Database Indexing

Dialogue spans are extracted using sliding windows (default: 5 turns per span) and indexed into ChromaDB for efficient retrieval. Each span includes:

- Combined text from multiple turns
- Turn indices and IDs
- Speaker distribution
- Event associations (if applicable)

3.3 Retrieval System

3.3.1 Semantic Search

We use sentence transformers (all-MiniLM-L6-v2) to encode queries and dialogue spans into dense vector representations. Cosine similarity is used to retrieve the top-k most relevant spans.

3.3.2 Reranking

A cross-encoder (ms-marco-MiniLM-L-6-v2) reranks the retrieved spans based on query relevance. This two-stage approach improves precision by refining semantic search results.

3.3.3 Event-Specific Retrieval

For queries about specific event types, the system filters spans that are associated with those events and extracts causal spans around event occurrences.

3.4 Causal Analysis Module

3.4.1 Pattern Detection

The pattern detector identifies:

- **Temporal patterns**: Spans that precede events
- **Sequential patterns**: Consecutive spans with causal relationships
- **Behavioral patterns**: Indicators like hesitation, frustration, repetition
- **Event-specific patterns**: Triggers for escalations, refunds, churn

3.4.2 Evidence Scoring

Evidence spans are scored using a weighted combination of:

- **Relevance score** (40%): From reranking
- **Temporal score** (30%): Proximity to event
- **Pattern score** (20%): Causal pattern indicators
- **Similarity score** (10%): Semantic similarity to query

3.5 Explanation Generation

3.5.1 LLM Integration

The system supports multiple LLM providers:

- OpenAI (GPT-4, GPT-3.5)
- Anthropic (Claude)
- Google Gemini (gemini-pro)

3.5.2 Prompt Engineering

Prompts are structured to:

- Provide context about the task
- Include formatted evidence spans
- Request structured explanations with citations
- Maintain conversational context for follow-ups

3.6 Conversation Management

3.6.1 Context Tracking

The conversation manager maintains:

- Conversation history (recent turns)
- Query-response pairs
- Context summaries for follow-up queries

3.6.2 Follow-up Detection

Follow-up queries are identified by:

- Presence of follow-up indicators (also, furthermore, what about)
- Pronoun usage (it, that, this, these)
- Query length (short queries often follow-ups)
- Contextual references to previous queries

4 Methodology

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

For initial queries, the system:

1. Parses the query to extract event type and intent
2. Retrieves relevant dialogue spans using semantic search
3. Reranks spans by relevance
4. Analyzes spans for causal patterns

5. Scores evidence spans
6. Generates explanation with LLM using top evidence
7. Formats response with citations and metadata

4.2 Task 2: Conversational Follow-Up and Contextual Response

For follow-up queries, the system:

1. Detects if query is a follow-up
2. Retrieves conversation context
3. Enhances query with context information
4. Performs context-aware retrieval
5. Generates contextual explanation
6. Updates conversation history

4.3 Query Dataset Generation

We developed a query simulation framework that:

- Uses LLMs to generate realistic queries from agent coach perspective
- Categorizes queries by task, difficulty, and use case
- Generates follow-up queries based on initial responses
- Applies human-in-the-loop refinement for quality
- Uses LLM-as-Judge for diversity and relevance assessment

5 Evaluation

5.1 Evaluation Metrics

We evaluate the system using multiple metrics:

5.1.1 Response Quality

- Length and word count
- Sentence count
- Citation presence and count
- Coherence score (transition word usage)

5.1.2 Evidence Quality

- Evidence count
- Average evidence score
- Evidence coverage (transcript diversity)
- Evidence diversity (span uniqueness)

5.1.3 Causal Explanation Quality

- Causal language presence
- Causal indicator count
- Evidence reference count
- Explanation completeness (query keyword coverage)

5.1.4 Conversational Coherence

- Coherence score
- Context usage
- Reference count to previous turns
- Context relevance

5.2 Baseline Comparisons

We compare against three baselines:

1. **Keyword Search**: Simple keyword matching
2. **Simple RAG**: Semantic search without reranking
3. **Rule-Based**: Pattern matching with predefined rules

5.3 Ablation Studies

We perform ablation studies by removing components:

- Without reranking
- Without causal analysis
- Without LLM generation (using template-based responses)

6 Results

6.1 Quantitative Results

[Results will be filled in after running experiments]

6.1.1 Task 1 Performance

- Average response quality scores
- Evidence retrieval precision/recall
- Explanation completeness metrics

6.1.2 Task 2 Performance

- Conversational coherence scores
- Context usage effectiveness
- Follow-up detection accuracy

6.2 Qualitative Analysis

[Qualitative examples will be included]

6.2.1 Example Explanations

We provide examples of:

- High-quality explanations with strong evidence
- Explanations showing system limitations
- Follow-up conversation flows

6.3 Error Analysis

[Error analysis will be included]
Common error types:

- Insufficient evidence retrieval
- Misinterpretation of causal relationships
- Context loss in follow-up conversations

7 Limitations and Future Work

7.1 Limitations

- Dependency on transcript quality and speaker labeling accuracy
- Limited to English language
- Computational cost of LLM inference
- Need for domain-specific fine-tuning for optimal performance

7.2 Future Work

- Custom model training for domain-specific causal patterns
- Multi-language support
- Real-time processing capabilities
- Enhanced interpretability visualizations
- Integration with live call monitoring systems

8 Conclusion

We presented a comprehensive system for causal rationale extraction from conversational data. The system successfully combines retrieval, causal analysis, and LLM-based generation to provide evidence-based explanations for business events. Evaluation demonstrates the effectiveness of our approach, though there is room for improvement in handling edge cases and domain-specific scenarios.

References

[1] Lewis, P., et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks.

[2] Feder, A., et al. (2021). Causal Inference in Natural Language Processing: Estimation, Prediction, Interpretation and Beyond.

[3] Jurafsky, D., & Martin, J. H. (2020). Speech and Language Processing.