TECHNICAL REPORT

1. Introduction

Large-scale conversational platforms produce vast amounts of multi-turn dialogue data involving customers and service agents. Many of these conversations result in operationally significant outcomes such as refunds, escalations, or service delays. However, conventional systems typically capture only the final outcome, without revealing which aspects of the conversation contributed to that result.

The aim of this project is to go beyond basic outcome prediction and develop a system that:

- Identifies the causal conversational factors that lead to a specific outcome,
- Extracts individual dialogue turns that serve as supporting evidence, and
- Enables multi-turn, context-aware analytical queries over conversations.

This work introduces an end-to-end machine learning framework that integrates outcome prediction, causal evidence extraction, and interactive reasoning over conversational transcripts.

Problem Statement

Given a dataset of conversational transcripts composed of structured dialogue turns and their corresponding outcome events, the objective is to:

- Generate causal explanations that connect conversational behaviours to observed outcomes,
- Extract explicit dialogue-level evidence to support each explanation, and
- Enable multi-turn analytical queries while preserving contextual consistency across interactions.

The system must prioritize faithfulness and interpretability, ensuring that all generated explanations are directly grounded in the underlying conversational data, rather than relying on spurious correlations or hallucinated patterns.

Dataset Description

The dataset comprises structured conversational transcripts, where each conversation is represented as a sequence of dialogue turns. The available fields include:

- transcript: A unique identifier for each conversation
- turn_id: The sequential index of a dialogue turns within a conversation
- speaker: The role of the speaker (Agent or Customer)
- text: The textual content of the utterance
- intent: The high-level issue category (e.g., delivery, appointment)
- domain: The associated business domain
- reason: The labelled outcome explanation
- time: The timestamp of the dialogue turn

Each conversation consists of multiple dialogue turns, while the outcome label is assigned at the conversation level rather than the individual turn level.

To ensure reliable modelling, only spoken dialogue turns from Agents and Customers were retained. System-generated summaries and metadata entries were excluded to avoid introducing noise into the analysis.

System Architecture Overview

The proposed system follows a modular, three-stage architecture designed to support prediction, explanation, and interactive analysis:

Stage 1 – Outcome Prediction

A Transformer-based classification model is used to predict the final outcome of a conversation.

Stage 2 – Turn-Level Causal Attribution

A leave-one-out attribution strategy is applied to determine which dialogue turns contribute most significantly to the predicted outcome.

Stage 3 – Query-Driven Interactive Reasoning

A deterministic query-processing pipeline enables analytical and follow-up queries while maintaining contextual memory across interactions.

Stage 1: Outcome Prediction Model Label Consolidation

The original labels were highly granular, leading to hundreds of low-frequency classes.
To enable reliable training and evaluation, outcome labels were consolidated into a small set of high-level categories:

- RESOLVED
- REFUND
- ESCALATION
- DELAY
- CANCELLATION

This abstraction improves generalization and interpretability without losing semantic meaning.

Input Representation

Each conversation was constructed by concatenating dialogue turns in chronological order using the format:
Speaker: Utterance [SEP] Speaker: Utterance ...
This representation preserves:

- The temporal ordering of dialogue turns,
- Explicit speaker roles, and
- The overall conversational flow.

Model Architecture

The outcome prediction model is based on the BERT (Bert-base-uncased) architecture and is formulated as a multi-class sequence classification task.

- Base model: BERT (Bert-base-uncased)
- Objective: Multi-class classification
- Loss function: Cross-entropy loss

The model processes the full conversational context and outputs a single outcome prediction for each transcript.

Training and Evaluation.

The dataset was split into training and validation sets using an 80/20 stratified split to preserve label distribution.

- Optimizer: Adam
- Learning rate: 2×10^{-5}
- Epochs: 130
- Evaluation metrics: Accuracy, Macro F1-score

Results:

- Validation Accuracy: 0.88
- Macro F1-score: 0.80

These results demonstrate a strong predictive foundation, which is subsequently leveraged for causal attribution and reasoning.

Stage 2: Turn-Level Causal Attribution

Although outcome prediction provides useful insights, it does not explain the underlying reasons for a given outcome. Stage 2 addresses this limitation by identifying which specific dialogue turns are responsible for the model's prediction.

Leave-One-Out Attribution Method

For a given conversation, causal importance is estimated using a leave-one-out attribution approach:

1. Compute the model's confidence score for the predicted outcome.
2. Remove a single dialogue turn from the conversation.
3. Recompute the model's prediction confidence.
4. Measure the change in confidence resulting from the removal.

The importance of a dialogue turn is defined as the decrease in prediction confidence caused by its exclusion.

Importance = Confidence (full conversation) – Confidence (with turn removed)

Evidence Extraction

Dialogue turns are ranked by importance score.

Top-ranked turns are returned as **causal evidence**, including:

- Turn text,
- Speaker role,
- Importance value.

Low-impact turns (e.g., greetings) naturally receive near-zero importance, demonstrating faithfulness.

Stage 3: Query-Driven and Multi-Turn Reasoning

Query Handling

User queries are processed through a deterministic pipeline that combines:

- Lightweight, keyword-based intent parsing,
- Conversation filtering using available metadata, and
- Causal evidence extraction produced in Stage 2.

This approach avoids free-form generative reasoning, thereby reducing hallucination risks while maintaining interpretability and traceability.

Context Memory

A structured context memory component is maintained to store:

- The previous user query,
- The set of selected conversations, and
- The extracted causal evidence.

This design enables coherent follow-up queries such as:

- *"Which dialogue turn had the greatest impact?"*
- *"Was customer frustration the primary contributing factor?"*

By preserving contextual state across interactions, the system ensures consistency in responses and satisfies the requirements for multi-turn reasoning.

Faithfulness and Interpretability

The system enforces faithfulness through the following design principles:

- Model-based causal attribution,
- Explicit extraction of dialogue-level evidence, and
- The absence of unconstrained natural language generation.

As a result, every explanation produced by the system can be directly traced to:

- Specific conversational transcripts,
- Individual dialogue turns, and
- Quantitative importance scores derived from the attribution process.

Evaluation Criteria Alignment

Judging Criterion	System Alignment
ID Recall	Exact extraction of dialogue turns identifiers

Judging Criterion	System Alignment
Faithfulness	Leave-one-out confidence-based attribution
Relevancy	Query-conditioned causal analysis
Context Awareness	Deterministic context memory across turns

Conclusion

This project presents a complete end-to-end framework for causal analysis and interactive reasoning over conversational data. By integrating Transformer-based outcome prediction with faithful turn-level causal attribution and context-aware query processing, the system advances beyond correlation-based analysis toward interpretable, evidence-backed explanations.

The modular architecture promotes scalability, transparency, and strong alignment with the defined problem objectives, making the approach suitable for real-world conversational analytics and decision-support applications.

Future Work

Several extensions can further enhance the system, including:

- Hierarchical Transformer models for turn-level reasoning,
- Temporal modelling that explicitly incorporates timestamps,
- Cross-conversation causal pattern discovery, and
- Interactive visualization dashboards for inspecting causal evidence.