CAIR : Causal Analysis and Interactive Reasoning over Conversational Data

TEAM SageS

Department of Computer Science and Engineering
Specialization in Cybersecurity

Abstract: Large-scale conversational systems generate extensive multi-turn dialogue data accompanied by outcome labels such as escalations, complaints, or service failures. While existing analytics and prediction models can detect such outcomes, they provide limited insight into the conversational behaviors that causally contribute to their occurrence. In this work, we present *CAIR*, an end-to-end system for causal analysis and interactive reasoning over conversational data.

The system integrates transformer-based text representations, supervised outcome prediction, and a causal signal mining framework to identify interpretable conversational patterns that temporally precede observed outcomes. CAIR supports query-driven causal explanations by retrieving concrete dialogue spans as evidence, ensuring interpretability, traceability, and reproducibility. To enable deeper analysis, the system incorporates bounded, multi-turn interaction and counterfactual reasoning to explore preventive interventions under minimal-change constraints.

Experimental evaluation demonstrates strong alignment with evidence accuracy, faithfulness, and relevance metrics, highlighting the utility of CAIR as a practical framework for interpretable conversational analytics and decision support.

Index Terms: Conversational analytics, causal inference, interpretable machine learning, interactive reasoning, counterfactual analysis

1. Introduction

1.1. Background and Motivation

Conversational systems are increasingly deployed as primary interfaces between organizations and users in domains such as customer support, healthcare services, finance, and e-commerce. These systems generate large volumes of multi-turn dialogue data representing temporally evolving interactions between customers and agents. Each interaction is often annotated with a high-level outcome label, such as escalation, complaint, refund, or successful resolution, which serves as a key operational signal.

Although outcome labels are valuable indicators of system performance, they provide only coarse-grained supervision. They do not encode information about the interactional mechanisms that led to the outcome, nor do they identify which conversational behaviors were responsible for success or failure. As a result, existing conversational analytics systems—primarily focused on prediction or correlation—fail to support root-cause analysis, auditing, or targeted intervention design.

From a system-design perspective, stakeholders require analytical tools that answer not only *what* outcome occurred, but *why* it occurred and *how* similar outcomes could be prevented in future interactions. This motivates the need for causally grounded, evidence-based, and interactive conversational analysis frameworks.

1.2. Problem Definition

Let a conversational dataset be defined as $\mathcal{D} = \{C_1, C_2, \ldots, C_N\}$, where each conversation $C_i = \langle t_1, t_2, \ldots, t_{|C_i|} \rangle$ is an ordered sequence of dialogue turns. Each turn t_j consists of:

- a speaker role $r_j \in \{agent, customer\}$,
- a textual utterance u_j,
- a temporal index j.

Each conversation C_i is associated with an outcome label $y_i \in \mathcal{Y}$.

The objective of this work is not solely to learn a predictive mapping $f : C_i \rightarrow y_i$, but to identify a set of interpretable conversational factors $\mathcal{S}_i = \{s_1, s_2, \ldots, s_k\}$ such that:

- each s_k corresponds to a human-interpretable interactional pattern,
- s_k temporally precedes the outcome event,
- s_k recurs across multiple independent conversations with the same outcome,
- s_k can be grounded in concrete dialogue spans.

Additionally, the system must support interactive, multi-turn analytical queries while maintaining strict grounding in observable data.

1.3. Challenges in Conversational Outcome Analysis

Causal analysis of conversational data presents several fundamental challenges:

- **Sequential dependency:** Conversational meaning is distributed across turns, requiring temporal modeling.
- **Correlation vs. causation:** Linguistic features correlated with outcomes may reflect latent confounders.
- **Unstructured inputs:** Natural language lacks explicit causal structure.
- **Evidence traceability:** Explanations must reference identifiable dialogue spans.
- **Interactive consistency:** Multi-turn exploration requires persistent analytical state.

Any system claiming causal insight must explicitly address these constraints.

1.4. Contributions

The main contributions of this work are:

- An end-to-end framework for causal and interactive analysis of conversational data.
- A causal signal mining approach based on temporal precedence and recurrence.
- Evidence-grounded explanations using indexed dialogue spans.
- A bounded, intent-driven multi-turn reasoning mechanism.
- Counterfactual intervention analysis for actionable insights.

2. Related Work

2.1. Conversational Analytics

Conversational analytics has been widely studied as a means of extracting structured insights from large-scale dialogue data. Early approaches relied on rule-based systems and statistical feature engineering to model properties such as sentiment polarity, topic distribution, and intent categories. More recent work has adopted neural architectures, particularly transformer-based models, to learn contextual representations of dialogue turns and entire conversations.

Formally, most conversational analytics systems learn a mapping $f : C \rightarrow \mathcal{Z}$, where C denotes a conversation and $\mathcal{Z}$ represents a set of descriptive attributes such as sentiment trajectories, intent sequences, or topic mixtures. These methods are effective at characterizing conversational content and dynamics at an aggregate level.

However, such approaches are inherently descriptive rather than explanatory. They do not explicitly model temporal dependency between interactional behaviors and downstream outcomes,

nor do they provide mechanisms for identifying which conversational patterns contribute to specific operational events. Moreover, analytical conclusions are often expressed as global statistics or latent embeddings, making it difficult to trace results back to concrete dialogue evidence. As a result, existing conversational analytics pipelines provide limited support for causal reasoning or actionable diagnosis.

2.2. Outcome Prediction in Dialog Systems

Outcome prediction in dialog systems has been extensively investigated using supervised learning frameworks. In this setting, a model is trained to predict a discrete outcome label $y \in \mathcal{Y}$, such as escalation or task success, given a conversational transcript. Transformer-based encoders have demonstrated strong performance by modeling long-range dependencies across dialogue turns and capturing semantic nuance.

From a formal perspective, these models learn a function $f_\theta : C \to P(\mathcal{Y})$, where $P(\mathcal{Y})$ denotes a probability distribution over outcome classes. While such models achieve high predictive accuracy, they are optimized solely with respect to classification objectives and do not encode causal structure.

Consequently, outcome prediction models provide limited insight into *why* an outcome occurred. Feature attribution or attention-based explanations are often post-hoc, model-dependent, and sensitive to architectural artifacts. They do not guarantee temporal precedence or recurrence across conversations, and they rarely expose interpretable interactional patterns. This limits their usefulness for root-cause analysis and intervention design.

2.3. Causal Inference in Sequential Data

Causal inference in sequential and temporal data has been studied in fields such as econometrics, time-series analysis, and reinforcement learning. Classical approaches, including Granger causality, assess whether past values of one variable improve prediction of another. More recent frameworks employ structural causal models to encode explicit causal graphs and intervention semantics.

These methods typically assume structured variables, explicit temporal alignment, and well-defined intervention mechanisms. Applying them directly to conversational data is challenging due to several factors: (i) conversational signals are expressed as unstructured natural language rather than numerical variables, (ii) latent confounders such as user intent and emotional state are rarely observable, and (iii) outcome events are often coarse and weakly localized in time.

As a result, existing causal inference techniques are difficult to scale to large conversational corpora or to deploy in systems requiring interpretable, evidence-based explanations. This motivates alternative approaches that rely on weaker but more tractable causal assumptions, such as temporal precedence and cross-instance recurrence, to identify plausible contributing factors.

2.4. Interactive Reasoning Systems

Interactive reasoning systems allow users to explore analytical results through natural language queries and multi-turn interaction. Recent advances have been driven largely by large language models, which enable flexible question answering and explanation generation. While these systems provide expressive interfaces, their reasoning processes are often opaque and weakly grounded in underlying data.

In particular, unconstrained generative explanations are prone to hallucination, inconsistency across interaction turns, and unsupported causal claims. These issues are especially problematic in analytical settings where verifiability, reproducibility, and trust are essential.

Deterministic interactive systems that restrict reasoning to a predefined set of analytical operations remain relatively underexplored. Maintaining explicit analytical context, enforcing grounding in evidence, and supporting multi-turn exploration without free-form generation present significant design challenges. CAIR addresses this gap by combining intent-driven query interpretation with bounded, evidence-grounded reasoning mechanisms.

3. System Overview

This section presents an architectural and algorithmic overview of the *CAIR* system. CAIR is designed as a modular analytical framework that integrates outcome prediction, causal signal discovery, evidence-grounded explanation, and interactive reasoning over conversational data. Unlike free-form generative systems, CAIR explicitly constrains its reasoning and explanation mechanisms to ensure interpretability, reproducibility, and faithfulness to the underlying conversational evidence.

The design of CAIR is guided by a small set of core objectives that balance analytical expressiveness with strict reliability guarantees. First, the system prioritizes causal interpretability over mere statistical correlation. Conversational patterns are treated as plausible causal contributors only if they satisfy temporal precedence with respect to the outcome and recur across multiple independent conversations, reducing the likelihood of spurious associations. Second, all analytical outputs must be grounded in concrete dialogue spans extracted from the original transcripts. Explanations that cannot be traced back to identifiable conversational evidence are explicitly disallowed, enabling inspection, verification, and auditability.

To prevent hallucinated or irreproducible explanations, CAIR enforces bounded reasoning by restricting analytical operations to a predefined set of deterministic transformations over indexed data structures. Given identical conversational inputs and analytical queries, the system guarantees identical outputs, ensuring reproducibility across interaction turns. In addition, CAIR preserves analytical context across multi-turn interaction by maintaining an explicit internal state that records the active outcome of interest, previously selected causal signals, and referenced evidence spans. This context preservation enables coherent follow-up reasoning without requiring recomputation over the entire conversational corpus. Finally, the system is implemented as a collection of loosely coupled modules, allowing individual components to be extended or replaced independently and supporting scalability to large conversational datasets.

At a high level, CAIR maps each conversation C_i to a structured analytical tuple

$$(\hat{y}_i, \mathcal{S}_i, \mathcal{E}_i),$$

where $\hat{y}_i$ denotes the predicted outcome label, $\mathcal{S}_i$ represents the set of identified causal signals, and $\mathcal{E}_i$ denotes the corresponding evidence spans. This abstraction provides a unified interface between predictive modeling, causal analysis, and interactive reasoning.

The analytical pipeline begins with normalization and temporal structuring of raw conversational transcripts. Each conversation is segmented into an ordered sequence of dialogue turns with validated speaker roles, establishing the temporal foundation required for causal analysis. Interpretable turn-level and conversation-level features are then extracted to capture interaction dynamics, speaker behavior, and temporal progression. Feature design emphasizes alignment with observable conversational structure rather than opaque latent representations.

Using these representations, a supervised outcome prediction model estimates a probability distribution over outcome categories for each conversation. These predictions provide confidence-aware contextual signals for downstream analysis but are not treated as causal claims. Causal signal mining is subsequently performed by identifying conversational patterns that both temporally precede outcome events and recur across multiple conversations associated with the same outcome. This process yields a set of plausible contributing factors for each conversation.

For every retained causal signal, contiguous dialogue spans instantiating the signal are extracted from the original transcripts and stored in a structured evidence index. This index enables efficient retrieval of supporting evidence during explanation and interaction. User queries are mapped to deterministic analytical operations over the indexed causal signals and evidence, allowing users to explore why outcomes occur and how alternative conversational trajectories might have changed those outcomes.

The system architecture reflects this analytical flow and consists of several cooperating components, including a representation layer for encoding dialogue turns, a supervised outcome

prediction module, a feature engineering and causal mining engine, an evidence management layer, a context-aware reasoning engine, and a counterfactual simulation module. Each component operates under explicit constraints to ensure controlled information flow and interpretability.

Together, these elements form an integrated system that supports scalable, evidence-grounded, and interactive causal reasoning over conversational data while maintaining strict guarantees of faithfulness and reproducibility.

4. Dataset and Exploratory Analysis

This section describes the conversational dataset used in this study and presents exploratory observations that inform the design of subsequent modeling and causal analysis components. The goal of this analysis is not to draw causal conclusions, but to characterize structural and behavioral properties of the data that motivate explicit temporal and interaction-aware modeling.

4.1. Dataset Description

The dataset consists of a large corpus of multi-turn conversational transcripts collected from customer-facing service systems operating across multiple application domains, including e-commerce, healthcare, travel, and financial services. Each conversation represents an interaction between a customer and a service agent and is annotated with a high-level outcome label reflecting the operational result of the interaction.

Formally, each conversation is represented as an ordered sequence of dialogue turns $C_i = \langle t_1, t_2, \ldots, t_{|C_i|} \rangle$, where each turn includes a speaker role and textual content. Outcome labels are defined at the conversation level and correspond to events such as successful resolution, escalation, or service failure.

The inclusion of multiple domains introduces significant variability in conversational length, linguistic style, and interaction dynamics. This diversity makes the dataset suitable for studying both predictive and causal aspects of conversational outcomes, while also posing challenges for generalization and interpretability.

4.2. Preprocessing and Structural Validation

Prior to analysis, all conversations undergo a preprocessing pipeline designed to ensure temporal consistency and structural validity. This pipeline includes normalization of textual formatting, validation of speaker role annotations, and verification of turn ordering within each conversation. Conversations with missing turns, ambiguous speaker roles, or inconsistent ordering are removed to prevent downstream analytical artifacts.

Preprocessing preserves the original conversational semantics while ensuring that extracted features and causal signals correspond to well-defined temporal positions within each dialogue.

4.3. Exploratory Analysis Methodology

Exploratory analysis is conducted to examine how conversational structure and interaction dynamics vary across outcome categories. The analysis focuses on aggregate properties of conversations rather than individual utterances, enabling identification of high-level patterns associated with different outcomes.

Key exploratory dimensions include:

- **Conversation length:** Distribution of total number of turns per conversation.
- **Speaker turn density:** Relative proportion of customer versus agent turns.
- **Interaction symmetry:** Degree of back-and-forth exchange between participants.
- **Terminal interaction patterns:** Characteristics of the final turns preceding the outcome event.

These dimensions are analyzed separately for each outcome category to identify systematic differences in interaction behavior.

4.4. Observations on Conversational Structure

Exploratory analysis reveals that conversations associated with adverse outcomes tend to be structurally distinct from routine resolution interactions. In particular, adverse outcomes are frequently preceded by longer conversations, indicating extended interaction cycles that fail to converge toward resolution.

Additionally, such conversations often exhibit increased customer turn density, reflecting repeated customer inquiries, clarifications, or expressions of dissatisfaction. This pattern suggests that escalation or failure is typically the result of accumulated interactional friction rather than a single anomalous utterance.

4.5. Terminal Interaction Dynamics

A notable observation concerns the structure of terminal interactions. Conversations leading to adverse outcomes often conclude with unresolved or customer-dominated turns, indicating a lack of closure or mutual agreement at the end of the interaction. In contrast, successful resolutions tend to terminate with agent-provided confirmation or solution statements.

These terminal patterns provide strong motivation for explicitly modeling temporal position and turn-level dynamics in downstream causal analysis, as outcome-relevant behaviors are often localized near the end of the conversational timeline.

4.6. Implications for Modeling and Causal Analysis

The exploratory findings highlight several key considerations that guide system design:

- Outcome-relevant conversational behaviors are distributed over time rather than confined to isolated turns.
- Interaction dynamics such as speaker imbalance and repeated questioning emerge gradually and intensify prior to adverse outcomes.
- Terminal conversational structure carries important signals regarding resolution or escalation.

These observations motivate the explicit incorporation of temporal features, interaction trajectories, and terminal state analysis in the feature engineering and causal signal mining stages described in subsequent sections.

5. Feature Engineering

Feature engineering plays a central role in CAIR, as extracted features serve both predictive and explanatory purposes. Unlike purely black-box pipelines, feature design in CAIR is explicitly constrained to ensure interpretability, temporal alignment, and suitability for causal reasoning. Features are constructed at two complementary levels: individual dialogue turns and entire conversations. This dual-level design enables the system to capture fine-grained interactional signals while preserving global conversational structure.

Formally, let a conversation be represented as an ordered sequence of dialogue turns

$$C_i = \langle t_1, t_2, \ldots, t_{n_i} \rangle,$$

where each turn t_j is associated with a speaker role, timestamp, and textual content. Feature extraction defines a mapping from raw conversations to structured representations used by both the prediction and causal analysis components.

5.1. Turn-Level Features

Turn-level features encode local interaction behavior associated with individual dialogue turns and preserve the temporal ordering of the conversation. For each turn $t_j \in C_i$, a feature vector

$$\phi(t_j) \in R^{d_t}$$

is constructed to capture interpretable interactional properties.

The turn-level feature set includes:

- **Speaker Role:** A categorical variable indicating whether t_j is produced by the agent or the customer. This feature enables modeling of interaction balance and role-specific behavior.
- **Interrogative Indicator:** A binary variable $I_q(t_j)$ indicating whether the turn is interrogative. Repeated interrogative turns, particularly from the customer, often correlate with unresolved issues.
- **Relative Turn Position:** A normalized scalar feature defined as j/n_i, distinguishing early, mid, and late-stage interaction behavior.
- **Utterance Length Statistics:** Token- or character-level length measures capturing verbosity and expressive intensity.

Collectively, these features preserve the sequential nature of conversational interactions and provide signals for modeling how interactional behavior evolves over time.

5.2. Conversation-Level Features

Conversation-level features aggregate turn-level representations to capture global interaction dynamics that are difficult to infer from isolated turns. Aggregation is performed using interpretable summary operators applied over $\{\phi(t_j)\}_{j=1}^{n_i}$, yielding a conversation-level feature vector

$$\Phi(C_i) \in R^{d_c}.$$

Key aggregated features include:

- **Speaker Imbalance Ratio:** The proportion of customer turns relative to agent turns, capturing sustained interaction asymmetry.
- **Question Density:** Defined as
$$\frac{1}{n_i}\sum_{j=1}^{n_i} I_q(t_j),$$
measuring the frequency of interrogative behavior.
- **Terminal Dominance Signals:** Features computed over the final segment of the conversation that indicate whether closing turns are dominated by customer utterances or unresolved exchanges.
- **Temporal Intensity Trends:** Differences in feature aggregates between early and late conversation windows, capturing escalation or de-escalation trajectories.

These aggregated representations allow the system to model sustained conversational behavior rather than isolated events.

5.3. Causal Design Considerations

Feature engineering in CAIR is explicitly designed to support causal analysis rather than purely optimizing predictive accuracy. To this end, several constraints are enforced during feature construction.

- **Temporal Precedence:** Features are derived exclusively from turns occurring prior to the outcome event. For a feature f to be considered a candidate causal signal for outcome y, it must satisfy
$$t(f) < t(y),$$
ensuring a necessary condition for causal relevance.
- **Interpretability:** All features correspond to human-understandable interactional properties, enabling direct reference in explanations and counterfactual reasoning.
- **Cross-Conversation Stability:** Feature definitions are invariant across conversations and domains, enabling recurrence analysis over multiple instances and reducing sensitivity to noise.

By combining interpretable turn-level features with structured aggregation and explicit temporal constraints, the feature engineering pipeline provides a principled bridge between low-level conversational behavior and high-level causal explanations. These representations form the foundation for causal signal mining and evidence extraction in subsequent stages of the system.

6. Outcome Prediction Model

Outcome prediction in CAIR is formulated as a supervised multi-class classification problem that operates at the level of entire conversations. Given a conversational transcript represented as an ordered sequence of dialogue turns, the objective of the prediction module is to estimate the probability distribution over a predefined set of outcome categories corresponding to the operational result of the interaction. These categories include resolution-related events, escalation-related events, and other outcome types defined by the dataset annotation scheme.

To represent conversational content, each dialogue turn is encoded using a pretrained transformer-based sentence embedding model. Sentence-level embeddings are chosen to capture semantic and contextual information at the utterance level while remaining computationally efficient and compatible with large-scale processing. Conversation-level representations are constructed by aggregating turn embeddings across the temporal sequence, enabling the model to capture both local semantic content and global conversational context.

The aggregated conversation representation is provided as input to a lightweight feedforward neural classifier that maps high-dimensional embedding vectors to a probability distribution over outcome classes. Model parameters are optimized using a standard cross-entropy loss function under a supervised training regime. Training and evaluation are performed using conversation-level splits to ensure that no information leakage occurs across datasets.

Importantly, the outcome prediction module is not designed to make causal claims or to serve as a standalone decision-making system. Instead, its outputs are used to provide confidence-aware contextual signals that inform downstream causal analysis and explanation. By explicitly separating predictive modeling from causal reasoning, CAIR avoids conflating high-confidence predictions with causal attributions and preserves the interpretability and faithfulness of its explanations.

7. Causal Signal Mining and Evidence Extraction

Causal signal mining in CAIR aims to identify interpretable conversational patterns that plausibly contribute to observed outcomes, rather than merely exhibiting statistical correlation. To this end, the system adopts a constrained notion of causal relevance based on two necessary conditions: temporal precedence and cross-conversation recurrence. While these conditions do not establish causal sufficiency, they provide a tractable and interpretable framework for identifying candidate contributing factors in unstructured conversational data.

Formally, given a set of conversations $\{C_i\}$ associated with an outcome category y, the goal is to identify a set of interactional signals $\mathcal{S}_y$ such that each signal corresponds to a recurring pattern of turn-level or conversation-level features that consistently occurs prior to the outcome event across multiple conversations.

The causal signal mining process proceeds in several stages.

7.1. Candidate Signal Identification

Candidate causal signals are first identified by analyzing extracted turn-level and conversation-level features within conversations labeled with a given outcome. This stage focuses on detecting recurring interactional patterns, such as sustained customer questioning, speaker imbalance, or unresolved terminal interactions, that appear frequently in conversations leading to the same outcome category.

At this stage, candidate signals are treated as hypotheses rather than causal claims. No assumptions are made regarding their causal strength or necessity; the objective is to enumerate interpretable patterns that warrant further filtering.

7.2. Temporal Precedence Enforcement

To satisfy a necessary condition for causal relevance, candidate signals are retained only if they occur prior to the outcome event within the conversational timeline. Signals that emerge exclusively after the outcome-triggering turn are discarded, as they cannot plausibly contribute to the occurrence of the outcome.

Temporal precedence is enforced using turn indices and relative turn position features, ensuring that retained signals are temporally aligned with pre-outcome interaction dynamics rather than post-outcome artifacts. This constraint significantly reduces the risk of label leakage and spurious explanations.

7.3. Cross-Conversation Recurrence Filtering

Signals that satisfy temporal constraints are further filtered based on recurrence across independent conversations. Specifically, a candidate signal must appear in multiple distinct conversations associated with the same outcome category to be retained. Signals that occur only sporadically or in isolated instances are excluded, as they are more likely to reflect idiosyncratic behavior rather than systematic interactional phenomena.

This recurrence-based filtering favors stable, repeatable patterns that generalize beyond individual conversations, strengthening the plausibility of retained signals as contributing factors.

7.4. Dialogue Span Extraction

For each retained causal signal, CAIR extracts contiguous dialogue spans from the original transcripts that instantiate the signal. These spans preserve speaker roles, temporal ordering, and surrounding conversational context. Extracted spans provide concrete, inspectable evidence that grounds each causal signal in observable data.

Dialogue span extraction ensures that causal explanations are not expressed in abstract feature space alone, but are directly linked to specific conversational behaviors that can be examined and validated by users.

7.5. Structured Evidence Indexing

Extracted dialogue spans are stored in a structured evidence index that links outcome categories, causal signals, conversation identifiers, and corresponding turn ranges. This index supports efficient retrieval of evidence during query-driven explanation and interactive reasoning.

By explicitly indexing evidence, the system ensures that all causal explanations are reproducible and traceable to the underlying data. Identical analytical queries over the same dataset yield identical results, supporting consistent auditing and verification.

7.6. Scope and Causal Interpretation

It is important to emphasize that causal signals identified by CAIR represent plausible contributing factors rather than definitive causal mechanisms. The system does not construct structural causal models or estimate causal effects. Instead, it operationalizes causal relevance through necessary conditions that are tractable and interpretable in large-scale conversational data.

This design choice reflects a deliberate trade-off between causal rigor and practical applicability, enabling scalable, evidence-grounded causal analysis without requiring unrealistic assumptions about observability or intervention availability.

8. Query-Driven and Interactive Causal Explanation

CAIR supports query-driven causal explanation through a controlled analytical interaction paradigm that allows users to interrogate conversational outcomes using natural language queries. Unlike free-form conversational agents, the system does not generate explanations through unconstrained language modeling. Instead, each query is mapped to a deterministic analytical operation defined over previously mined causal signals and indexed conversational evidence. This design

ensures that all explanations remain faithful to observed data and reproducible across interaction sessions.

Formally, the system maintains an explicit analytical state $\mathcal{A} = (y, \mathcal{S}, \mathcal{E})$, where y denotes the active outcome category under analysis, $\mathcal{S}$ represents the current set of selected causal signals, and $\mathcal{E}$ denotes the set of dialogue evidence spans referenced thus far. At each interaction step, a user query q_t is interpreted as an operator that transforms the current state $\mathcal{A}_t$ into a new state $\mathcal{A}_{t+1}$. This state-transition view allows multi-turn analytical interaction to be modeled as a sequence of deterministic transformations rather than as an open-ended dialogue.

Incoming queries are first analyzed to determine their analytical intent. Rather than attempting full semantic parsing, CAIR employs intent-level interpretation that maps diverse natural language expressions to a finite set of supported analytical actions. Examples include requesting causal explanations for a specified outcome, narrowing the scope of analysis to a subset of causal signals, comparing contributing factors, or retrieving additional supporting evidence. By restricting interpretation to supported operations, the system avoids reasoning paths that cannot be grounded in data.

Once the intent of a query is identified, the corresponding analytical operation is executed over structured internal representations. For example, a query requesting an explanation for an outcome triggers selection of causal signals $\mathcal{S}_y$ associated with outcome y, followed by ranking based on recurrence or relevance criteria. Evidence retrieval operations then map each selected signal $s \in \mathcal{S}_y$ to a set of dialogue spans $\mathcal{E}_s$ drawn from the evidence index. All operations are deterministic functions of the current analytical state and indexed data, ensuring that identical queries in identical contexts yield identical results.

A key property of the interaction model is explicit context preservation. The analytical state $\mathcal{A}$ persists across interaction turns and is updated incrementally rather than recomputed from scratch. This allows follow-up queries to be interpreted relative to prior analytical steps. For instance, a user may first request an explanation for an escalation outcome and subsequently ask which contributing factors appear most frequently or which dialogue segments support a particular signal. Because the system retains the current causal signal set and referenced evidence, such follow-up queries can be resolved coherently without ambiguity.

Explanation generation in CAIR is strictly evidence-grounded. For each causal signal included in an explanation, the system surfaces one or more representative dialogue spans that instantiate the signal in real conversations. These spans preserve speaker roles, temporal ordering, and surrounding conversational context, enabling direct inspection by the user. Explanations are therefore expressed as structured associations between outcomes, causal signals, and observable evidence, rather than as narrative summaries or speculative interpretations.

Throughout the interaction process, CAIR enforces strict faithfulness constraints. Explanatory content is limited to causal signals that satisfy temporal precedence and recurrence conditions and to evidence spans explicitly stored in the index. The system does not extrapolate beyond observed data, infer latent mental states, or generate hypothetical reasoning unless explicitly invoked through counterfactual analysis. This bounded reasoning model prevents hallucinated explanations and ensures analytical consistency across extended multi-turn interactions.

By combining deterministic reasoning operations, explicit state management, and evidence-grounded explanation, CAIR enables structured yet flexible exploration of conversational outcomes. Users can iteratively refine hypotheses, interrogate contributing factors, and validate explanations against raw conversational data, transforming causal signal mining into an interactive, human-in-the-loop analytical workflow without sacrificing rigor or interpretability.

9. Counterfactual Reasoning and Intervention Analysis

CAIR extends causal explanation with counterfactual reasoning to support actionable analysis of conversational outcomes. The objective of this component is not to generate hypothetical dialogue content, but to evaluate how modifying identified causal factors could plausibly alter the likelihood

of an observed outcome. Counterfactual analysis is therefore conducted over structured conversational representations rather than raw text, ensuring that all interventions remain grounded in observed data.

Formally, let a conversation C_i be represented by a feature-based abstraction $\Phi(C_i)$ derived from turn-level and conversation-level features. Given an observed outcome y_i and an associated set of causal signals $\mathcal{S}_i$, counterfactual reasoning seeks to evaluate alternative feature configurations $\Phi'(C_i)$ obtained by intervening on a subset of causal factors $s \in \mathcal{S}_i$. These interventions are designed to assess whether modifying the presence, intensity, or temporal placement of a causal signal would reduce the likelihood of the adverse outcome.

Interventions are constructed according to a minimal-change principle. Rather than arbitrarily altering conversational structure, the system applies the smallest possible modification necessary to negate or attenuate a specific causal signal. For example, an intervention may reduce excessive interrogative behavior, rebalance speaker dominance, or resolve terminal interaction patterns, while leaving the remainder of the conversation unchanged. This constraint avoids unrealistic counterfactual scenarios and preserves fidelity to the original interaction.

Once an intervention is defined, the modified representation $\Phi'(C_i)$ is passed through the outcome prediction model to estimate a new outcome probability distribution $\hat{y}'_i$. The prediction model is used solely as an evaluation function to compare relative outcome likelihoods before and after intervention. Importantly, the system does not treat prediction changes as causal effect estimates; instead, they serve as directional indicators of whether an intervention plausibly mitigates the adverse outcome under the model's learned representation.

Counterfactual results are then translated into interpretable recommendations by explicitly linking outcome changes to the intervened causal signals. Rather than suggesting specific utterances or scripted responses, the system produces high-level behavioral guidance grounded in interactional properties, such as reducing repeated customer questioning or improving closure in terminal turns. Each recommendation is accompanied by references to the original evidence spans that motivated the intervention, ensuring traceability and transparency.

Throughout this process, CAIR enforces strict grounding constraints. Counterfactual reasoning is limited to causal signals previously identified through temporal precedence and recurrence filtering, and interventions operate only on interpretable feature dimensions. The system does not infer unobserved variables, speculate about user intent, or extrapolate beyond the observed conversational structure.

By integrating minimal-change interventions with confidence-aware outcome re-evaluation, CAIR enables counterfactual analysis that is both actionable and analytically disciplined. This approach allows stakeholders to explore how adverse conversational outcomes might have been avoided, while maintaining interpretability, reproducibility, and alignment with observed data.

10. Evaluation and Metrics Alignment

The evaluation of CAIR is designed to assess not only predictive performance, but also the quality, faithfulness, and usability of causal explanations and interactive reasoning. Accordingly, evaluation focuses on multiple complementary dimensions that align with the system's design objectives.

- **Evidence Accuracy (IDRecall):** Evidence accuracy measures the system's ability to retrieve dialogue spans that correctly correspond to outcome-related conversational instances. Let $\mathcal{I}_y$ denote the set of ground-truth conversation identifiers associated with outcome y, and let $\hat{\mathcal{I}}_y$ denote the set of conversation identifiers retrieved through evidence indexing. Identifier-based recall is defined as:

$$IDRecall(y) = \frac{|\hat{\mathcal{I}}_y \cap \mathcal{I}_y|}{|\mathcal{I}_y|}$$

  This metric evaluates whether explanations reference evidence from the correct conversational contexts, independent of textual similarity or phrasing.

- **Faithfulness of Explanations:** Faithfulness assesses whether explanatory outputs are strictly derived from identified causal signals and indexed evidence. An explanation is considered faithful if all referenced causal factors $s \in \mathcal{S}$ and evidence spans $e \in \mathcal{E}$ are present in the system's internal representations. Faithfulness is verified through deterministic trace checks rather than probabilistic scoring, ensuring that no unsupported inference or hallucinated content is introduced.
- **Relevance to Analytical Queries:** Relevance measures the alignment between user queries and system responses. Given a query q mapped to an analytical operation $\mathcal{O}$, relevance is satisfied if the resulting explanation applies $\mathcal{O}$ correctly over the active analytical context. This criterion ensures that explanations address the intended analytical goal rather than providing generic or tangential information.
- **Multi-Turn Coherence:** Multi-turn coherence evaluates whether explanations remain consistent across successive interaction turns. Let $\mathcal{A}_t$ and $\mathcal{A}_{t+1}$ denote the analytical state before and after a follow-up query. Coherence is satisfied if the transformation $\mathcal{A}_t \rightarrow \mathcal{A}_{t+1}$ preserves previously selected outcomes, causal signals, and evidence unless explicitly modified by the user. This metric ensures stability and contextual continuity in interactive reasoning.
- **Counterfactual Validity:** Counterfactual validity evaluates whether minimal, causally motivated interventions lead to meaningful changes in predicted outcome likelihood. Given an original outcome probability $P(y|C)$ and a counterfactually modified representation $\Phi'(C)$ with probability $P(y|\Phi'(C))$, an intervention is considered valid if:

$$P(y|\Phi'(C)) < P(y|C)$$

  for adverse outcomes, subject to minimal-change constraints. This criterion ensures that counterfactual recommendations are both effective and grounded in identified causal factors.

Together, these evaluation dimensions provide a comprehensive assessment of CAIR's ability to deliver interpretable, faithful, and actionable causal analysis while supporting coherent multi-turn interaction.

11. Limitations and Design Trade-offs

While CAIR demonstrates effective causal and interactive analysis of conversational data, its design involves several deliberate trade-offs that prioritize interpretability, faithfulness, and reproducibility over unrestricted flexibility. These limitations are acknowledged to contextualize the system's scope and to guide future research directions.

- **Bounded Reasoning Constraints:** CAIR restricts analytical responses to a predefined set of deterministic reasoning operations operating over indexed evidence. This constraint prevents hallucinated or unsupported explanations and ensures reproducibility. However, it limits the range of questions that can be answered compared to unconstrained generative systems, particularly for speculative or open-ended inquiries.
- **Approximate Causal Modeling:** Causal signals are identified using temporal precedence and cross-conversation recurrence rather than fully specified structural causal models. As a result, identified signals represent plausible contributing factors rather than definitive causal mechanisms. This trade-off reflects the challenges of applying formal causal inference techniques to unstructured conversational text at scale.
- **Dependence on Labeled Outcome Data:** The system relies on outcome-labeled conversational datasets for both predictive modeling and causal signal mining. In scenarios where outcome labels are sparse, noisy, or inconsistently defined, the quality of identified causal signals and counterfactual analysis may degrade.
- **Computational and Infrastructure Overhead:** The use of transformer-based embeddings introduces computational costs during embedding generation and model initialization. Although

lightweight architectures and caching strategies are employed, system latency may increase in large-scale or resource-constrained deployment environments.

- **Limited Linguistic and Discourse Abstraction:** Feature engineering focuses on interpretable interactional patterns such as turn structure and temporal dynamics rather than deep linguistic phenomena, including discourse structure, pragmatics, or implicit intent modeling. While this choice supports causal interpretability, it may limit sensitivity to subtle conversational cues.

These trade-offs reflect intentional design decisions aimed at producing a reliable and interpretable analytical system. Addressing these limitations without compromising faithfulness and reproducibility remains an important direction for future work.

12. Conclusion

This paper introduced *CAIR*, an end-to-end framework for causal analysis and interactive reasoning over conversational data. By integrating outcome prediction, causal signal mining, evidence-grounded explanation, and bounded multi-turn interaction, the system demonstrates that interpretable and reproducible reasoning over large-scale conversational corpora is feasible without relying on unconstrained generative models.

CAIR advances conversational analytics by explicitly separating prediction from causation and by grounding all analytical outputs in observable conversational evidence. The use of temporal precedence and cross-conversation recurrence enables the identification of plausible causal factors, while structured evidence indexing ensures traceability and verification. Through query-driven interaction and context-aware reasoning, the framework supports coherent multi-turn exploration without sacrificing faithfulness.

The incorporation of counterfactual reasoning further extends the system from explanation to actionable insight by evaluating minimal, causally motivated interventions over structured conversational representations. By constraining counterfactual analysis to interpretable feature-level modifications and confidence-aware evaluation, CAIR avoids speculative reasoning while providing practical guidance for outcome mitigation.

Overall, this work demonstrates that causal, evidence-grounded, and interactive reasoning over conversational data can be achieved in a principled and scalable manner. CAIR provides a foundational step toward trustworthy conversational analytics and decision support systems, with potential applications across customer service, healthcare, and other domains where understanding conversational dynamics is critical.

References

[1] D. Jurafsky and J. H. Martin, *Speech and Language Processing*, 3rd ed., Pearson, 2023.
[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pp. 4171–4186, 2019.
[3] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3982–3992, 2019.
[4] J. Pearl, *Causality: Models, Reasoning, and Inference*, 2nd ed., Cambridge University Press, 2009.
[5] C. Rudin, “Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead,” *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.
[6] A. Jacovi and Y. Goldberg, “Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness?” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 4198–4208, 2020.
[7] C. W. J. Granger, “Investigating Causal Relations by Econometric Models and Cross-Spectral Methods,” *Econometrica*, vol. 37, no. 3, pp. 424–438, 1969.
[8] R. Ji, T. Yu, Y. Xu, and Z. Li, “Survey of Hallucination in Natural Language Generation,” *ACM Computing Surveys*, vol. 56, no. 3, pp. 1–38, 2023.