ChatbotID: Identifying Chatbots with Granger Causality Test

Xiaoquan Yi[1], **Haozhao Wang**[1*], **Yining Qi**[1] **Wenchao Xu**[2],
Rui Zhang[1†], **Yuhua Li**[1], **Ruixuan Li**[1]
[1]School of Computer Science and Technology,
Huazhong University of Science and Technology, Wuhan, China
[2]School of Division of Integrative Systems and Design
Hong Kong University of Science and Technology, Hong Kong, China.
{yixiaoquan,hz_wang}@hust.edu.cn, rayteam@yeah.net

Abstract

With the increasing sophistication of Large Language Models (LLMs), it is crucial to develop reliable methods to accurately identify whether an interlocutor in real-time dialogue is human or chatbot. However, existing detection methods are primarily designed for analyzing full documents, not the unique dynamics and characteristics of dialogue. These approaches frequently overlook the nuances of interaction that are essential in conversational contexts. This work identifies two key patterns in dialogues: (1) Human-Human (H-H) interactions exhibit significant bidirectional sentiment influence, while (2) Human-Chatbot (H-C) interactions display a clear asymmetric pattern. We propose an innovative approach named ChatbotID, which applies the Granger Causality Test (GCT) to extract a novel set of interactional features that capture the evolving, predictive relationships between conversational attributes. By synergistically fusing these GCT-based interactional features with contextual embeddings and optimizing the model via a structured loss function, we significantly enhance the model's ability to capture asymmetric influence in H-C dialogues. Experimental results across multiple datasets and detection models demonstrate the effectiveness of our framework, with 15.92% improvements in accuracy for distinguishing between H-H and H-C dialogues.

1 Introduction

The rapid advancement and proliferation of Large Language Models (LLMs) have led to increasingly sophisticated conversational agents capable of generating remarkably human-like text [1, 2, 3]. By exploiting the sophisticated conversational abilities of LLMs, malicious actors can convincingly simulate human interactions [4, 5, 6], tricking unsuspecting individuals into believing they are communicating with a real person, thereby facilitating fraudulent activities such as scams and identity theft [7, 4, 8]. Considering this, it has become critically important to devise reliable methods for distinguishing between human and LLM-driven interactions.

Many state-of-the-art methods typically rely on identifying statistical anomalies in linguistic [9, 10, 11] or stylometric features [12, 13] for detecting LLMs-generated text. These approaches often require extensive manual feature engineering and may exhibit limited effectiveness against more advanced LLMs that are explicitly optimized to bypass such detection mechanisms. More recently, supervised learning approaches [14, 15, 16] have been developed to distinguish between human-written and LLM-generated text by analyzing both semantic content [17, 18] and high-level textual

*Haozhao Wang is corresponding authors.
†Homepage: https://www.ruizhang.info/

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

features [19, 20, 21]. Although these models are effective in certain contexts, they primarily analyze the static textual content and stylistic features of full documents. Consequently, they often overlook the fine-grained interactional nuances necessary for dialogue detection. These limitations underscore the urgent demand for innovative dialogue detection approaches capable of capturing fine-grained interaction dynamics and integrating them with semantic representations to enable more reliable identification of conversational participants.

In this work, we are developing a specialized detection framework to identify chatbot text in dialogues, focusing on unique linguistic features and interaction patterns. Particularly, we reveal two principal patterns of sentiment influence within dialogues: *Human-human (H-H) interactions are characterized by substantial bidirectional sentiment exchange, whereas Human-Chatbot (H-C) interactions demonstrate a distinct asymmetric influence*. Motivated by two principal patterns, we propose a novel approach named `ChatbotID` that integrates the deep contextual understanding capabilities of LLMs with a quantitative analysis of conversational interaction dynamics derived from Granger Causality tests [22] (GCT). To quantify these temporal dependencies, we first extract relevant time series features (e.g., sentiment scores per turn) from the dialogues. Subsequently, we apply GCT to these features to compute a feature vector, denoted as V_{GCT}. This vector is utilized to fine-tune LLMs specifically for the task of distinguishing dialogues generated by LLMs. By jointly modeling semantic content and interaction dynamics, the model becomes proficient at identifying chatbots. Experiments show our framework significantly improves accuracy in distinguishing H-H from H-C dialogues across multiple datasets. The main contributions of this work are:

- Grounded in Communication Accommodation Theory, this work is the first to systematically quantify and reveal two principal patterns of sentiment influence in dialogues: *H-H interactions are characterized by statistically significant bidirectional influence, whereas H-C interactions demonstrate a distinct asymmetric influence pattern*.
- *W*e propose a novel dialogue detection method based on GCT named `ChatbotID`. To our knowledge, *this work is the first to systematically address the detection of LLM-generated contributions specifically within conversational contexts.*
- Extensive experiments conducted on various datasets (DailyDialog, MultiWOZ, etc.) and advanced LLMs (Gemma, Qwen-2, Deepseek-R1, etc.), *demonstrate that our method outperforms state-of-the-art detection methods by up to 15.92%*.

2 Related Work

In this section, we discuss three critical dimensions of LLMs-generated text analysis, i.e., representative detection approaches, the Granger Causality Test [22] and Interaction Dynamics.

LLM-Generated Text Detection. Various approaches achieve differentiation between human and LLM-generated texts by capitalizing on the complex inner workings of LLMs [12, 13, 23], specifically examining aspects like intermediate layer outputs and model weights [24]. However, these methods reliant on internal model information also encounter notable limitations, such as their inapplicability to black-box proprietary models, and weaker generalization across diverse model architectures [25, 26]. Another category of detection approach shifts focus to the statistical properties of the text itself [9, 27, 28, 29]. They utilize statistical metrics (e.g. entropy, perplexity, frequency of specific words, sentence structure) to differentiate between LLMs-generated and human-written texts [30, 31, 32]. However, these detection performances can also be significantly affected by variations in text type, topic diversity, and specific linguistic characteristics, leading to insufficient stability and accuracy. Other researchers have adopted supervised learning methods [7, 14, 15, 16], training specialized classification models on large datasets of labeled texts. While these models demonstrate effectiveness in certain contexts, their primary analytical focus is on the static textual content and stylistic features of entire documents. They often fail to capture the interactional nuances for dialogue detection.

Granger Causality in Text Detection. GCT, an econometric concept by origin, is a standard statistical method used to determine if one time series improves the forecast of another [22, 33, 34]. GCT provides a valuable statistical framework for investigating directional predictive relationships between time-ordered data sequences, finding pertinent applications in detection tasks within Natural Language Processing [35, 36, 37, 38]. Existing methodologies for GCT are centered on its core concept of evaluating whether one-time series' past significantly improves the prediction of another's

future [39, 40, 41], beyond the information contained in the target series' history. These methods provide a comprehensive toolkit for identifying and characterizing directional predictive links in temporal data across various domains [42, 43, 44].

Linguistic Accommodation and Interaction Dynamics. A foundational concept for understanding dialogue is Communication Accommodation Theory, which posits that individuals adjust their communication strategies to signal social closeness, gain approval, or maintain social distance [45, 46]. This theory has motivated a significant body of work studying linguistic accommodation, where conversational partners tend to converge in their use of linguistic features, such as style, syntax, and sentiment [47, 48]. Many research have successfully leveraged metrics of accommodation to analyze social dynamics in various contexts [49, 50]. For instance, Danescu-Niculescu-Mizil et al. [51] demonstrated that power imbalances in conversations are reflected in asymmetric linguistic coordination patterns. Studies on online discussions have shown that interaction dynamics, including accommodation, are predictive of persuasion and argument outcomes, and can help in detecting disputes [52]. These works typically quantify accommodation using correlation-based metrics or measures of distributional similarity between speakers' features over a conversation.

3 Motivation: Asymmetric Influence in H-C Dialogues

By comparing H-H dialogues with H-C dialogues, we observe that *chatbots exert asymmetric conversational influence*. Human conversation is not merely a sequence of contextually relevant utterances [51, 53, 54]. It is a rich, dynamic process characterized by mutual influence, adaptation, and intricate feedback loops operating over time [55, 49]. This reciprocity shapes phenomena such as sentiment contagion, topic negotiation, and behavioral entrainment, reflecting an underlying dynamic coupling between participants [45, 50]. While an LLM might react with high sensitivity and predictability to user input (e.g., user sentiment strongly and immediately driving LLM sentiment), it may exert significantly less reciprocal influence in dynamically shaping the user's subsequent state or cognitive framing compared to a human partner [47, 56, 52].

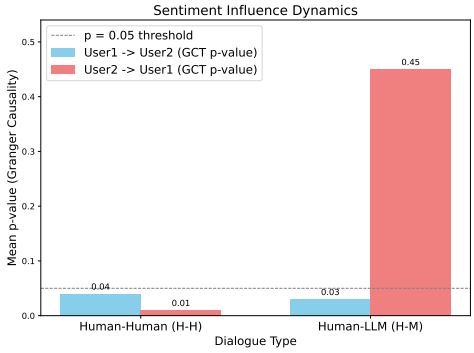

Figure 1: The figure demonstrates that H-H dialogues show significant bidirectional sentiment influences, whereas H-C dialogues feature a pronounced asymmetric pattern.

GCT offers a powerful statistical framework for examining the predictive relationships between time series derived from dialogue features (e.g., sentiment scores and utterance lengths). It allows us to test whether the history of one participant's conversational features (e.g., the User's sentiment time series) significantly improves the prediction of the other participant's future features. As shown in Figure 1, we apply GCT to examine sentiment dynamics within 200 dialogues from the DailyDialog dataset. In H-H interactions, there is statistically significant mutual influence. User1 significantly affects User2 (mean p-value = 0.04), and User2 reciprocates with a stronger influence on User1 (mean p-value = 0.01). In contrast, H-C dialogues exhibit a clear asymmetric pattern. GCT p-values indicate that only the human user exerts a statistically significant influence on the LLM's sentiment (mean p-value = 0.03), whereas the LLM's sentiment does not significantly influence the user (mean p-value = 0.45). These distinct sentiment causality structures highlight the contrast between statistically significant bidirectional influence in H-H dialogues and the unidirectional, user-driven dominance observed in H-C interactions.

4 Methodology

In this section, we detail a method named `ChatbotID` for human versus LLM dialogue classification by fine-tuning an LLM using Multi-Task Learning. Our approach combines the LLM's semantic representations with GCT metrics derived from interaction dynamics. `ChatbotID` leverages interaction dynamics and semantic-focused attribution to enhance classification accuracy.

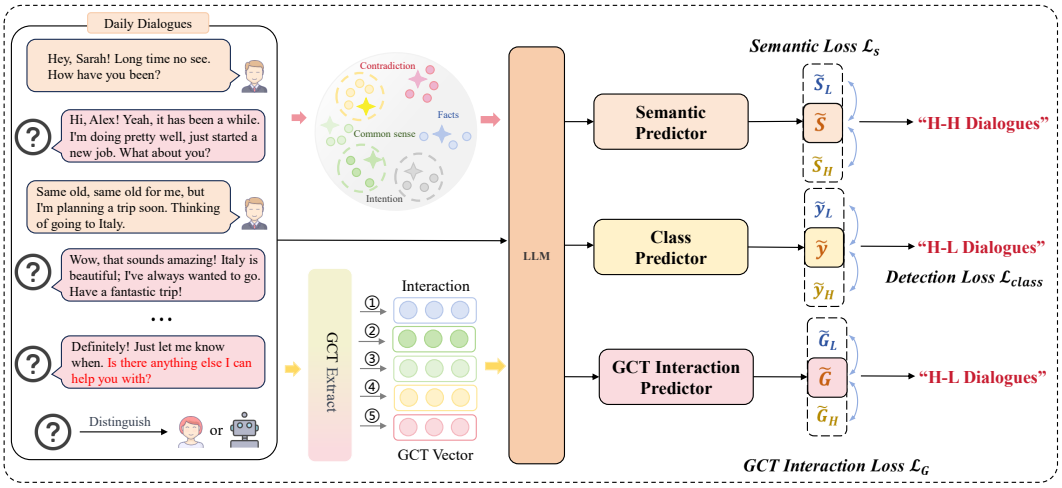

Figure 2: `ChatbotID` is purpose-built for the detection and analysis of dialogue scenarios. The Semantic Loss L_s associated with the Semantic Predictor focuses on capturing the deep semantic understanding of dialogues. The Detection Loss (L_{class}) drives the model to perform classification tasks, categorizing dialogues into predefined classes. The GCT Interaction Loss (L_G) ensures the model learns and leverages the interactional features and patterns extracted by the GCT module.

4.1 Problem Formulation

Let $\mathcal{D} = \{(C^{(i)}, y^{(i)})\}_{i=1}^{N}$ be a dataset comprising N dialogues. Each dialogue $C^{(i)}$ consists of a sequence of user (U) and agent (A) utterances, paired with a label $y^{(i)} \in \{0, 1\}$ denoting the agent's type (0: Human, 1: Chatbot). The core problem is to learn a parameterized function $f_\theta : C \mapsto [0, 1]$, where θ represents the model parameters. This function aims to estimate the posterior probability $P(y = 1|C)$ for any given dialogue C. The parameter θ is optimized to minimize loss function reflecting the classification error on the dataset $\mathcal{D}$.

4.2 Feature Engineering

Dialogue Time Series Extraction. For each dialogue $C^{(i)}$ comprising T_i interaction points (e.g., turns, utterances per participant), we extract features for both the user and agent at each point $t \in \{1, \ldots, T_i\}$. The pre-defined feature set Q includes basic metrics (e.g., utterance length, topic embedding components) along with semantic features(e.g., sentiment scores, topic embedding components), extracted using LLMs. Let the feature extraction function be $\mathcal{E}$. This yields paired numerical time series:

$$\mathcal{E} : C^{(i)} \mapsto \left(\{X_{U,t}^{(f)}\}_{t=1}^{T_i}, \{X_{A,t}^{(f)}\}_{t=1}^{T_i} \right)_{f \in Q} \tag{1}$$

where $X_{U,t}^{(f)} \in \mathbb{R}$ and $X_{A,t}^{(f)} \in \mathbb{R}$ are the values for feature f at time t.

Granger Causality Feature Vector Calculation. The GCT is employed to assess predictive causality between selected pairs of user and agent feature time series. We perform a specific test to determine whether time series X_t Granger-causes time series Y_t, employing a lag order of p. X_t and Y_t represent specific feature sequences, e.g. $X_U^{(f_1)}(t)$ or $X_A^{(f_1)}(t)$ for X_t, and $X_U^{(f_2)}(t)$ or $X_A^{(f_2)}(t)$ for Y_t. To conduct this test, two linear autoregressive models are estimated using Ordinary Least Squares over the effective sample period $t = p + 1, \ldots, T_i$. This period corresponds to an effective sample size of $n = T_i - p$.

The baseline is the restricted model, where Y_t is modeled solely based on its own p past values:

$$Y_t = \alpha_0 + \sum_{k=1}^{p} \alpha_k Y_{t-k} + \epsilon_{R,t} \tag{2}$$

The fit of this model is measured by its Sum of Squared Residuals, $SSR_R = \sum_{t=p+1}^{T_i} \hat{\epsilon}_{R,t}^2$. This is contrasted with the unrestricted model, which incorporates p lagged values of X_t as potential predictors for Y_t:

$$Y_t = \beta_0 + \sum_{k=1}^{p} \beta_k Y_{t-k} + \sum_{j=1}^{p} \gamma_j X_{t-j} + \epsilon_{UR,t} \tag{3}$$

The corresponding Sum of Squared Residuals for this model is $SSR_{UR} = \sum_{t=p+1}^{T_i} \hat{\epsilon}_{UR,t}^2$. A statistically significant reduction from SSR_R to SSR_{UR} indicates that X_t Granger-causes Y_t. Such predictive relationships within the dialogue's interaction dynamics, captured by comparing these models, offer valuable signals for distinguishing between human and chatbot.

The null hypothesis $H_0 : \gamma_1 = \gamma_2 = \cdots = \gamma_p = 0$ posits that X does not Granger-cause Y. $k_{UR} = 2p + 1$ is the number of parameters in the unrestricted model. This hypothesis is tested using the F-statistic:

$$F = \frac{(SSR_R - SSR_{UR})/p}{(SSR_{UR})/(n - k_{UR})} = \frac{(SSR_R - SSR_{UR})/p}{(SSR_{UR})/(n - 2p - 1)} \tag{4}$$

Under H_0, the statistic follows an F-distribution, $F \sim F(p, n - 2p - 1)$. The p-value is computed as:

$$p\text{-value} = P(F_{p,n-2p-1} \geq F | H_0) \tag{5}$$

We specify the pairs of conversational features to compare (e.g., user sentiment vs. agent reply length), the direction of potential causality being tested (User-to-Agent or Agent-to-User), and the relevant time lags p. Subsequently, for each dialogue $C^{(i)}$ in our dataset, this entire suite of pre-defined tests is performed; let d_{GCT} be the total number of such tests. Each test k ($k = 1, \ldots, d_{GCT}$) yields a statistical outcome $g_k^{(i)}$ for dialogue $C^{(i)}$, typically a p-value reflecting the significance of that specific predictive relationship. Finally, all these d_{GCT} outcomes about dialogue $C^{(i)}$ are gathered and concatenated into a single numerical list, forming the dialogue's GCT feature vector:

$$V_{GCT}^{(i)} = [g_1^{(i)}, g_2^{(i)}, \ldots, g_{d_{GCT}}^{(i)}] \tag{6}$$

4.3 Classification Loss

The classification loss is employed to enable the model to distinguish between human users and LLMs within a conversational environment. For a given dialogue input p, associated with its true class label y and predicted probability distribution $\hat{y}$, $K = 2$, corresponds to the two possible sources of the dialogue: human and LLMs-generated, and the cross-entropy loss is formulated as:

$$L_{\text{class}} = -\sum_{k=1}^{K} y_k \log \hat{y}_k \tag{7}$$

4.4 Sesmantic-Focused Attribution Supervision

To enhance the model's understanding of semantic differences between H-H and H-C dialogues, particularly the quality of LLMs contributions in H-C contexts, we introduce a supervision mechanism based on semantic-focused attributions. These attributions are identified by querying an LLM to detect specific undesirable characteristics or failures within a given dialogue. The LLM is prompted using the following method to generate these semantic attributions:

Let $C = \{c_{\text{goal}}, c_{\text{fact}}, c_{\text{common}}, c_{\text{logic}}\}$ be the predefined set of pragmatic deficiencies. For each dialogue D_j in our training set, the LLM's output is parsed to generate a binary deficiency attribution vector $\mathbf{a}_j = [a_{j,\text{goal}}, a_{j,\text{fact}}, a_{j,\text{common}}, a_{j,\text{logic}}]$. Each element $a_{j,k}$ (where k corresponds to a deficiency in C) is defined as:

$$a_{j,k} = \begin{cases} 1, & \text{if dialogue } D_j \text{ is identified as exhibiting deficiency } c_k, \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

Input Dialogue: [Dialogue Text]
Contextual Focus (if identifiable as potential H-C): Contributions from the suspected chatbot.
Question: Which of the following pragmatic semantic deficiencies does this dialogue exhibit, particularly concerning the contextual focus if applicable?

1. **Goal Obfuscation/Failure (c_{goal})**: The primary user's goals seem unmet, poorly addressed, or significantly side-tracked.
2. **Factual Inconsistency (c_{fact})**: The dialogue contains statements that are demonstrably false, misleading, or internally inconsistent with established facts.
3. **Commonsense Violation (c_{common})**: The dialogue includes statements, reasoning, or assumptions that clearly contradict basic, everyday commonsense.
4. **Logical Incoherence (c_{logic})**: The dialogue displays internal contradictions in reasoning, significant logical fallacies, or a breakdown in coherent argumentation.

If multiple deficiencies are applicable, provide a comma-separated list of the corresponding labels (e.g., "$c_{\text{goal}}, c_{\text{common}}$"). Answer "None" if none of the options apply.

To guide the main classification model using these semantic deficiency attributions, we train it to jointly predict these attributes. This is achieved by defining an auxiliary semantic deficiency Attribution loss (L_{S}). Assuming the model produces a corresponding vector of predicted probabilities $\hat{\mathbf{a}}_j = [\hat{a}_{j,\text{goal}}, \hat{a}_{j,\text{fact}}, \hat{a}_{j,\text{common}}, \hat{a}_{j,\text{logic}}]$ for each deficiency type for dialogue D_j, the loss is:

$$L_{\text{S}} = \frac{1}{N_D} \sum_{j=1}^{N_D} \sum_{k \in C} \text{BCE}(a_{j,k}, \hat{a}_{j,k}) \tag{9}$$

where N_D is the total number of dialogues in the training batch, k iterates over the set of deficiencies C, and BCE denotes the binary cross-entropy loss.

4.5 Causal Interaction Dynamics Supervision using GCT

In our motivation phase, we discover a pattern indicating that the presence or absence of statistically significant causal links, as reflected by the corresponding p-values, serves as a distinguishing feature between H-H and H-C interactions. To leverage this, we transform the GCT p-values into binary indicators of significant causal effects. Let α_{sig} be a pre-defined significance level (e.g., 0.05). For each dialogue $C^{(i)}$ and each GCT test outcome $g_k^{(i)}$, we define a binary causality indicator $b_k^{(i)}$:

$$b_k^{(i)} = \begin{cases} 1, & \text{if } g_k^{(i)} < \alpha_{\text{sig}} \text{ (indicating a significant causal link)}, \\ 0, & \text{otherwise (no significant causal link detected)}. \end{cases} \tag{10}$$

This process yields a binary GCT vector $\mathbf{b}^{(i)} = [b_1^{(i)}, b_2^{(i)}, \ldots, b_{d_{GCT}}^{(i)}]$ for each dialogue $C^{(i)}$.

Our main model is then tasked with jointly predicting these binary significance indicators. This is achieved by incorporating an auxiliary GCT Significance Loss (L_{G}). Assuming the model produces a corresponding vector of predicted probabilities $\hat{\mathbf{b}}^{(i)} = [\hat{b}_1^{(i)}, \hat{b}_2^{(i)}, \ldots, \hat{b}_{d_{GCT}}^{(i)}]$ for dialogue $C^{(i)}$, the loss is defined as:

$$L_{\text{G}} = \frac{1}{N_D} \sum_{i=1}^{N_D} \sum_{k=1}^{d_{GCT}} \text{BCE}(b_k^{(i)}, \hat{b}_k^{(i)}) \tag{11}$$

where d_{GCT} is the number of GCT tests performed, and BCE is the binary cross-entropy loss.

4.6 Final Objective Function

This L_{G} is added to the overall loss function, alongside the primary classification loss L_{class} and semantic-focused attribution L_{S}:

$$L = L_{\text{class}} + L_{\text{S}} + L_{\text{G}} \tag{12}$$

5 Experiments

In this section, we present the experimental setup, detailing the datasets used and the implementation of our methods. We evaluate the performance across multiple datasets using metrics (e.g., accuracy, F1-score) with state-of-the-art methods.

5.1 Experimental Setup

Datasets: To evaluate our proposed methodology across different conversational settings, we utilize four prominent English-language dialogue datasets: two focused on open-domain chit-chat (e.g., DailyDialog [57], PersonaChat [58]) and two on task-oriented interactions (e.g., MultiWOZ [59], Taskmaster-1 [60]). These datasets serve as the foundation for constructing both our H-H and H-C dialogue corpora, ensuring comparability in style and domain. **The H-H corpus** used in our experiments is formed by selecting dialogues directly from the aforementioned datasets. Dialogues below a pre-defined length are filtered out to ensure suitability for Granger Causality analysis. **The H-C corpus** is a semi-synthetic dataset derived from the H-H corpus to ensure high comparability in user input and conversational context. The construction involves selecting H-H dialogues, each comprising user turns and original human agent turns. For every dialogue, we identify the human agent's utterances and then prompt an LLM (e.g., Llama-2-Chat 70B or GPT-4) to generate new responses for those specific turns.

Metrics. To rigorously evaluate our method's ability to distinguish between H-H and H-C dialogues. We employ three primary performance metrics on a held-out test set: Accuracy (ACC), Area Under the Receiver Operating Characteristic Curve (AUROC), and the F1-score (F1).

Baselines. We conduct a rigorous comparative analysis of our proposed methodology against several state-of-the-art methods for detecting LLM-generated text. DetectGPT [12] identifies synthetic text by scrutinizing the local curvature of a source language model's probability function around a given text passage. Fast-DetectGPT [13] detects LLMs-generated text by evaluating the conditional probability curvatures of sampled token alternatives. T5-Sentinel [61] introduces a supervised learning approach that reframes LLM-generated text detection as a token prediction task, using labeled data to fine-tune T5 models to directly predict text sources. COCO [62] employs contrastive learning to enhance detection by learning discriminative representations that separate LLM-generated from human-authored texts in the embedding space. RoBERTa-MPU [63] is a standard RoBERTa model fine-tuned specifically for LLM-generated text detection. OUTFOX [7] enhances the robustness of detecting LLMs-generated texts by implementing iterative in-context learning between the detector and an attacker that generates adversarial examples. LLMDet [64] employs surrogate perplexity calculations specifically tailored to individual LLMs. Shifting to a structural representation. SeqXGPT [11] transforms sentences into waveforms, utilizing convolutional networks and self-attention mechanisms for detection at the sentence level. GECScore [65] provides a robust metric for discerning LLM origins by evaluating text similarity through the lens of a grammar error correction model.

General-purpose LLMs. In our study, we employ a selection of representative general-purpose LLMs as analytical benchmarks, leveraging their inherent capability for zero-shot veracity prediction. This approach facilitates the direct assessment of truthfulness without necessitating specialized

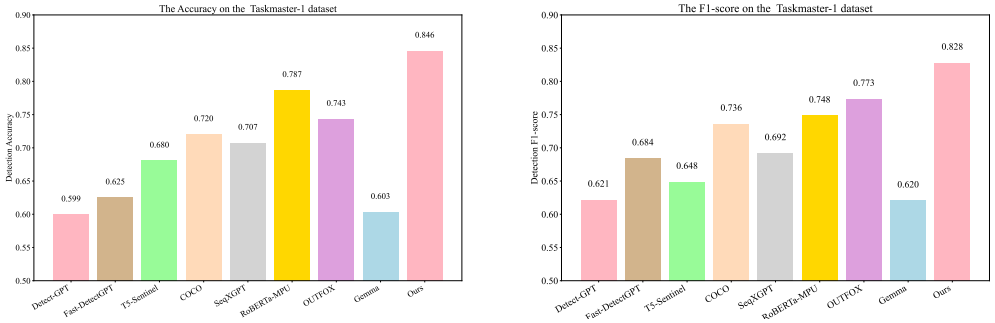

Figure 3: The visual data presented in the graphs clearly indicates that our methodology excels in detection accuracy and F1-score on the Taskmaster-1 dataset.

fine-tuning procedures. The LLMs selected for analysis include LLaMA-7B [66], LLaMA-13B, GPT-3.5-turbo, GPT-4, Gemma [67], Qwen-2 [68], and Deepseek-R1 [69]. These models are utilized as standards to systematically evaluate both the capabilities and limitations of LLMs in performing zero-shot detection of content-generated tasks.

Implementation Details. We implement our proposed methodology using the Hugging Face Transformers library. The model is fine-tuned on the constructed H-H and H-M datasets, with a batch size of 16 and a learning rate of 2e-5. The GCT analysis is performed using the statsmodels library, with a maximum lag of 5 for Granger causality tests. The model is trained for 20 epochs, with early stopping based on validation loss. The model is evaluated on a separate test set, and the results were averaged over 5 runs to account for variability in training. The model is trained to utilize the AdamW optimizer, incorporating weight decay to enhance regularization.

5.2 Performance evaluation

Table 1: Experimental results on the DailyDialog, PersonaChat, and MultiWOZ datasets. The best number is highlighted in bold, while the second best one is underlined. Our approach consistently outperforms other methods, achieving the highest accuracy in each dataset.

Method	DailyDialog		PersonaChat		MultiWOZ	
	ACC	F1	ACC	F1	ACC	F1
DetectGPT (ICML 2023)	60.36 ± 1.37	66.17 ± 0.42	65.42 ± 2.20	65.24 ± 1.96	64.59 ± 5.19	62.36 ± 1.23
COCO (EMNLP 2023)	77.36 ± 0.81	77.30 ± 1.34	78.56 ± 1.74	77.64 ± 2.13	78.86 ± 0.85	75.41 ± 3.54
LLMDet (EMNLP 2023)	64.77 ± 2.25	70.28 ± 0.18	67.22 ± 0.56	66.00 ± 0.61	67.24 ± 1.88	67.27 ± 1.97
SeqXGPT (EMNLP 2023)	65.63 ± 2.57	66.82 ± 3.73	69.54 ± 1.30	65.02 ± 2.06	67.14 ± 1.56	67.05 ± 0.19
Fast-DetectGPT (ICLR 2024)	62.71 ± 2.65	62.28 ± 1.32	63.43 ± 3.02	64.17 ± 1.71	59.86 ± 0.37	62.46 ± 1.79
T5-Sentinel (EMNLP 2024)	76.68 ± 2.39	74.63 ± 3.15	72.84 ± 2.05	73.77 ± 0.61	84.52 ± 2.11	77.47 ± 1.08
RoBERTa-MPU (ACL 2024)	78.62 ± 0.38	81.15 ± 0.96	<u>82.05</u> ± 2.75	83.35 ± 1.40	<u>83.24</u> ± 1.53	<u>82.04</u> ± 0.24
DeTeCtive (NeurIPS 2024)	72.31 ± 0.53	74.83 ± 0.56	75.95 ± 0.38	73.99 ± 1.25	80.12 ± 1.27	77.74 ± 0.31
OUTFOX (AAAI 2024)	<u>78.80</u> ± 0.90	<u>83.46</u> ± 1.13	80.06 ± 0.31	<u>84.08</u> ± 1.06	82.77 ± 2.04	81.11 ± 0.22
GECScore (ACL 2025)	69.05 ± 1.69	72.81 ± 1.06	75.36 ± 4.12	73.60 ± 2.33	75.54 ± 3.32	67.55 ± 0.16
GPT-3.5-turbo (2023)	57.42 ± 1.68	59.11 ± 2.81	61.00 ± 1.07	66.84 ± 0.63	60.37 ± 0.54	59.62 ± 1.67
LLaMA-7B (2024)	58.17 ± 1.01	60.52 ± 2.40	61.61 ± 3.06	58.96 ± 1.84	65.47 ± 1.82	59.54 ± 2.32
LLaMA-13B (2024)	60.94 ± 3.30	62.81 ± 2.27	63.32 ± 0.26	63.53 ± 2.85	65.16 ± 0.30	62.35 ± 1.13
GPT-4 (2024)	62.61 ± 3.93	64.56 ± 2.96	65.28 ± 0.87	62.09 ± 2.56	62.74 ± 0.87	58.94 ± 2.59
Gemma (2025)	63.67 ± 2.76	65.19 ± 0.77	66.31 ± 1.86	68.35 ± 0.48	66.99 ± 4.72	64.91 ± 1.40
Qwen-2 (2025)	61.92 ± 1.27	66.35 ± 1.03	65.50 ± 0.71	64.95 ± 0.49	65.43 ± 2.84	63.97 ± 4.66
Deepseek-R1 (2025)	65.68 ± 0.20	67.96 ± 1.47	66.27 ± 2.60	62.25 ± 0.89	68.03 ± 4.55	69.44 ± 1.24
`ChatbotID` **(Ours)**	**82.77** ± 0.56	**84.74** ± 2.72	**82.23** ± 1.29	**87.01** ± 2.92	**87.38** ± 4.18	**84.38** ± 0.91

Accuracy. As shown in Table 1, `ChatbotID` model consistently achieves the highest accuracy across multiple diverse dialogue datasets. For instance, on the MultiWOZ dataset, `ChatbotID`'s accuracy reaches 87.38%, which is notably higher than other leading specialized detectors such as RoBERTa-MPU (83.24%) and T5-Sentinel (84.52%). In contrast to detection methods that primarily rely on static text features or stylistic analysis (e.g., DetectGPT, COCO, LLMDet), `ChatbotID` gains its performance edge by analyzing the dynamic interactive features within a dialogue, particularly by employing the GCT to capture predictive relationships in attributes like sentiment. The GCT provides a statistically grounded way to quantify influence and predictive causality within a dialogue. This is a more targeted approach than relying solely on learned representations from LLMs, which might not inherently focus on these subtle interactional cues crucial for distinguishing nuanced LLM behavior from human behavior.

F1-Score. As shown in Figure 3, `ChatbotID` records an F1-score of 0.828. This is considerably higher than the other methods, including OUTFOX (0.773), RoBERTa-MPU (0.748), and COCO (0.736). The general-purpose LLM, Gemma, shows a much lower F1-score of 0.620. `ChatbotID` provides a statistically grounded way to quantify influence and predictive causality within a dialogue. This is a more targeted approach than relying solely on learned representations from large pre-trained models, which might not inherently focus on these subtle interactional cues crucial for distinguishing nuanced LLM behavior from human behavior.

AUROC. As shown in Figure 4, across four distinct dialogue datasets, `ChatbotID` consistently achieves the highest AUROC scores when compared against seven other text detection models.

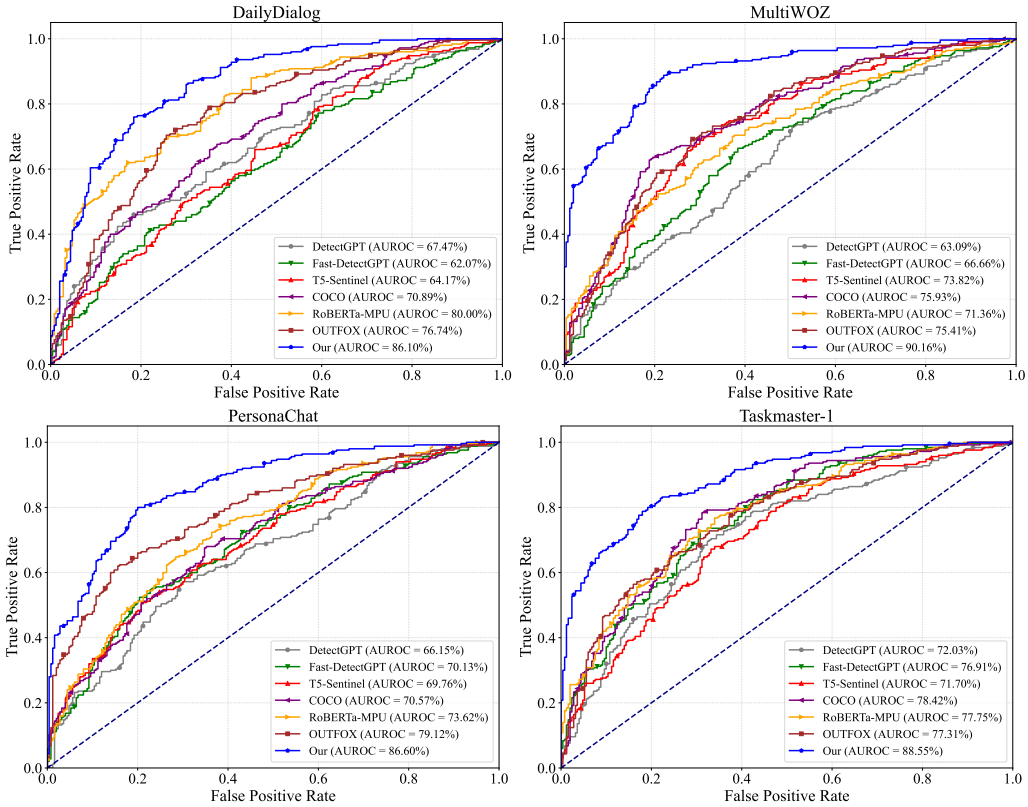

Figure 4: The figure displays ROC curves illustrating the comparative performance of seven different text detection models across various datasets.

Specifically, `ChatbotID` attains an AUROC of 86.10% on DailyDialog, a remarkable 90.16% on MultiWOZ, 86.69% on PersonaChat, and 88.55% on Taskmaster-1. This consistent outperformance across varies conversational contexts, from open-ended chit-chat to structured task completion, underscores a key advantage of our approach. Unlike methods that rely predominantly on static textual features or stylistic anomalies, `ChatbotID` incorporates an analysis of interaction dynamics. Employing Granger Causality tests quantifies the predictive influence between conversational attributes of the user and the agent over time.

Table 2: On various LLM backbones, `ChatbotID` demonstrates consistent improvements in accuracy.

Method	**DailyDialog**	**PersonaChat**	**MultiWOZ**	**Taskmaster-1**
LLaMA-7B	61.74 ± 1.38	59.05 ± 3.17	58.32 ± 3.34	55.44 ± 1.65
`ChatbotID`-LLaMA-7B	**70.26** ± 0.64	**72.87** ± 2.62	**72.47** ± 1.28	**66.48** ± 1.46
Gemma	63.20 ± 4.94	57.62 ± 1.08	63.59 ± 3.21	59.17 ± 3.15
`ChatbotID`-Gemma	**69.12** ± 0.69	**71.91** ± 1.34	**74.75** ± 4.64	**76.25** ± 1.43
Qwen-2	60.80 ± 2.11	63.59 ± 0.66	61.70 ± 0.01	63.04 ± 2.68
`ChatbotID`-Qwen-2	**79.99** ± 3.50	**82.56** ± 1.59	89.63 ± 0.01	**85.34** ± 2.48
Deepseek-R1	60.37 ± 2.16	65.29 ± 1.24	61.65 ± 2.48	63.50 ± 2.21
`ChatbotID`-Deepseek-R1	**82.07** ± 0.66	**84.35** ± 1.25	**85.77** ± 1.58	**84.15** ± 1.28

Different LLMs backbones. Table 2 systematically demonstrates that integrating the `ChatbotID` framework leads to substantial and consistent accuracy improvements when applied to a variety of LLM backbones for the task of distinguishing human-LLM dialogues. For LLaMA-7B, the introduction of `ChatbotID` elevates accuracy from a baseline of 61.74% to 70.26% on DailyDialog, from 59.05% to 72.87% on PersonaChat. When applied to Qwen-2, `ChatbotID` shows particularly striking gains, boosting accuracy on Taskmaster-1 from 63.04% to 85.34%. The magnitude of these improvements, often exceeding 10-20 percentage points (e.g., Qwen-2 on MultiWOZ shows an

increase of nearly 28 percentage points), highlights the significant value added by ChatbotID. This advantage stems from ChatbotID's use of GCT to extract and integrate distinctive interactional dynamics, combined with LLM-based semantic understanding through a structured multi-task learning framework, enabling more nuanced detection.

Table 3: Ablation study: This table illustrates the individual and combined contributions of the ChatbotID's distinct loss components to its overall accuracy in distinguishing H-C dialogues

Method	DailyDialog	PersonaChat	MultiWOZ	Taskmaster-1
L_{class}	63.19 ± 1.63	67.86 ± 1.97	69.16 ± 0.63	66.18 ± 4.93
L_{class} + L_S	67.58 ± 2.45	72.73 ± 0.79	74.71 ± 1.08	74.35 ± 1.59
L_{class} + L_G	70.99 ± 3.88	78.70 ± 8.55	77.70 ± 2.49	80.38 ± 1.29
L_{class} + L_G + L_S	80.23 ± 3.82	83.63 ± 1.74	87.01 ± 0.64	83.37 ± 2.70

Ablation study. As shown in Table 3, the baseline model, relying solely on the classification loss L_{class}, establishes a foundational level of performance across datasets. The introduction of the semantic-focused attribution supervision L_S yields consistent accuracy improvements (e.g., from 63.19% to 67.58% on DailyDialog) demonstrating the value of guiding the model to recognize specific semantic deficiencies often present in chatbots. More profoundly, the integration of the causal interaction dynamics supervision using L_G provides a more substantial boost in accuracy (e.g., from 63.19% to 70.99% on DailyDialog when combined with L_{class}. The semantic deficiency attribution loss L_S helps the model identify common LLM pitfalls, further refining its classification and potentially reducing false negatives where an LLM produces superficially coherent but pragmatically flawed dialogue.

Table 4: Performance across dialogue turn ranges on DailyDialog, PersonaChat, and MultiWOZ.

Turns	DailyDialog		PersonaChat		MultiWOZ	
	ACC	F1	ACC	F1	ACC	F1
1–5	60.07 ± 0.38	62.39 ± 0.98	61.29 ± 1.97	61.86 ± 0.51	60.89 ± 0.81	63.26 ± 1.11
6–10	75.88 ± 0.60	78.12 ± 1.48	76.90 ± 3.36	75.96 ± 0.27	78.96 ± 1.50	78.34 ± 0.48
10–15	83.41 ± 0.66	82.83 ± 1.93	84.96 ± 0.54	82.44 ± 0.83	84.86 ± 3.08	83.72 ± 2.53
15+	86.82 ± 0.33	85.49 ± 1.45	86.95 ± 1.24	83.91 ± 1.95	89.17 ± 2.75	84.77 ± 0.29

Dialogue Turns. In the initial stages of the dialogues, from 1-5 turns up to 10-15 turns, the model exhibits a dramatic and consistent improvement in both accuracy and F1-score across all three datasets. For instance, on the MultiWOZ dataset, accuracy skyrockets from 60.89% in the 1-5 turn bucket to 84.86% in the 10-15 turn bucket. In very short dialogues, there is insufficient interaction history to establish a stable pattern of influence. As turns accumulate, the cause-and-effect chain between speakers becomes more robust, allowing ChatbotID to distinguish H-H interaction and H-C interaction more reliably.

6 Conclusion

This work introduces a novel framework named ChatbotID that effectively distinguishes between H-H and H-C dialogues by analysing interactional dynamics, particularly sentiment influence, using the GCT. ChatbotID demonstrates superior performance over existing methods across various datasets and LLM backbones.

Acknowledgments and Disclosure of Funding

This work is supported by the National Key Research and Development Program of China under grant 2024YFC3307900; the National Natural Science Foundation of China under grants 62376103, 62302184, 62436003 and 62206102; Major Science and Technology Project of Hubei Province under grant 2024BAA008; Hubei Science and Technology Talent Service Project under grant 2024DJC078; Ant Group through CCF-Ant Research Fund. The computation is completed in the HPC Platform of Huazhong University of Science and Technology.

References

[1] Xinru Wang, Hannah Kim, Sajjadur Rahman, Kushan Mitra, and Zhengjie Miao. Human-llm collaborative annotation through effective verification of llm labels. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–21, 2024.

[2] Subbarao Kambhampati, Karthik Valmeekam, Lin Guan, Mudit Verma, Kaya Stechly, Siddhant Bhambri, Lucas Paul Saldyt, and Anil B Murthy. Position: Llms can't plan, but can help planning in llm-modulo frameworks. In *Forty-first International Conference on Machine Learning*, 2024.

[3] Zijie Zeng, Shiqi Liu, Lele Sha, Zhuang Li, Kaixun Yang, Sannyuya Liu, Dragan Gašević, and Guanliang Chen. Detecting ai-generated sentences in human-ai collaborative hybrid texts: Challenges, strategies, and insights. *arXiv preprint arXiv:2403.03506*, 2024.

[4] Jiaying Wu, Jiafeng Guo, and Bryan Hooi. Fake news in sheep's clothing: Robust fake news detection against llm-empowered style attacks. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pages 3367–3378, 2024.

[5] Shiwei Li, Yingyi Cheng, Haozhao Wang, Xing Tang, Shijie Xu, Weihong Luo, Yuhua Li, Dugang Liu, Xiuqiang He, and Ruixuan Li. Masked random noise for communication-efficient federated learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 3686–3694, 2024.

[6] Shiquan Yang, Rui Zhang, Sarah Erfani, and Jey Han Lau. An interpretable neuro-symbolic reasoning framework for task-oriented dialogue generation. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4918–4935, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[7] Ryuto Koike, Masahiro Kaneko, and Naoaki Okazaki. Outfox: Llm-generated essay detection through in-context learning with adversarially generated examples. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 21258–21266, 2024.

[8] Shiquan Yang, Rui Zhang, and Sarah Erfani. GraphDialog: Integrating graph knowledge into end-to-end task-oriented dialogue systems. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1878–1888, Online, November 2020. Association for Computational Linguistics.

[9] Sebastian Gehrmann, Hendrik Strobelt, and Alexander Rush. GLTR: Statistical detection and visualization of generated text. In Marta R. Costa-jussà and Enrique Alfonseca, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 111–116, Florence, Italy, July 2019. Association for Computational Linguistics.

[10] Kangxi Wu, Liang Pang, Huawei Shen, Xueqi Cheng, and Tat-Seng Chua. LLMDet: A third party large language models generated text detection tool. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2113–2133, Singapore, December 2023. Association for Computational Linguistics.

[11] Pengyu Wang, Linyang Li, Ke Ren, Botian Jiang, Dong Zhang, and Xipeng Qiu. SeqXGPT: Sentence-level AI-generated text detection. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1144–1156, Singapore, December 2023. Association for Computational Linguistics.

[12] Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. Detectgpt: zero-shot machine-generated text detection using probability curvature. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

[13] Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. Fast-detectgpt: Efficient zero-shot detection of machine-generated text via conditional probability curvature. In *The Twelfth International Conference on Learning Representations*, 2023.

[14] Xun Guo, Yongxin He, Shan Zhang, Ting Zhang, Wanquan Feng, Haibin Huang, and Chongyang Ma. Detective: Detecting ai-generated text via multi-level contrastive learning. *Advances in Neural Information Processing Systems*, 37:88320–88347, 2024.

[15] Lucio La Cava, Davide Costa, and Andrea Tagarelli. Is contrasting all you need? contrastive learning for the detection and attribution of ai-generated text. In *ECAI 2024*, pages 3179–3186. IOS Press, 2024.

[16] Annepaka Yadagiri, Reddi Mohana Krishna, and Partha Pakray. Cnlp-nits-pp at genai detection task 2: Leveraging distilbert and xlm-roberta for multilingual ai-generated text detection. In *Proceedings of the 1stWorkshop on GenAI Content Detection (GenAIDetect)*, pages 307–311, 2025.

[17] Zhixin Lai, Xuesheng Zhang, and Suiyao Chen. Adaptive ensembles of fine-tuned transformers for llm-generated text detection. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2024.

[18] Yichen Li, Haozhao Wang, Wenchao Xu, Tianzhe Xiao, Hong Liu, Minzhu Tu, Yuying Wang, Xin Yang, Rui Zhang, Shui Yu, Song Guo, and Ruixuan Li. Unleashing the power of continual learning on non-centralized devices: A survey, 2024.

[19] Hao Wang, Jianwei Li, and Zhengyu Li. Ai-generated text detection and classification based on bert deep learning algorithm. *arXiv preprint arXiv:2405.16422*, 2024.

[20] Yichen Li, Wenchao Xu, Yining Qi, Haozhao Wang, Ruixuan Li, and Song Guo. Sr-fdil: Synergistic replay for federated domain-incremental learning. *IEEE Transactions on Parallel and Distributed Systems*, 35(11):1879–1890, 2024.

[21] Ruohong Huan, Guowei Zhong, Peng Chen, and Ronghua Liang. Unimf: A unified multimodal framework for multimodal sentiment analysis in missing modalities and unaligned multimodal sequences. *IEEE Transactions on Multimedia*, 26:5753–5768, 2024.

[22] Craig Hiemstra and Jonathan D Jones. Testing for linear and nonlinear granger causality in the stock price-volume relation. *The Journal of Finance*, 49(5):1639–1664, 1994.

[23] Kaito Taguchi, Yujie Gu, and Kouichi Sakurai. The impact of prompts on zero-shot detection of ai-generated text. *arXiv preprint arXiv:2403.20127*, 2024.

[24] Xiao Pu, Jingyu Zhang, Xiaochuang Han, Yulia Tsvetkov, and Tianxing He. On the zero-shot generalization of machine-generated text detectors. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4799–4808, Singapore, December 2023. Association for Computational Linguistics.

[25] Yichen Li, Qunwei Li, Haozhao Wang, Ruixuan Li, Wenliang Zhong, and Guannan Zhang. Towards efficient replay in federated incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12820–12829, June 2024.

[26] Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Spotting llms with binoculars: zero-shot detection of machine-generated text. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

[27] Christoforos Vasilatos, Manaar Alam, Talal Rahwan, Yasir Zaki, and Michail Maniatakos. Howkgpt: Investigating the detection of chatgpt-generated university student homework through context-aware perplexity analysis. *arXiv preprint arXiv:2305.18226*, 2023.

[28] Ryo Kamoi, Sarkar Snigdha Sarathi Das, Renze Lou, Jihyun Janice Ahn, Yilun Zhao, Xiaoxin Lu, Nan Zhang, Yusen Zhang, Ranran Haoran Zhang, Sujeeth Reddy Vummanthala, et al. Evaluating llms at detecting errors in llm responses. *arXiv preprint arXiv:2404.03602*, 2024.

[29] Irina Tolstykh, Aleksandra Tsybina, Sergey Yakubson, Aleksandr Gordeev, Vladimir Dokholyan, and Maksim Kuprashevich. Gigacheck: Detecting llm-generated content. *arXiv preprint arXiv:2410.23728*, 2024.

[30] Jiazhou Ji, Jie Guo, Weidong Qiu, Zheng Huang, Yang Xu, Xinru Lu, Xiaoyu Jiang, Ruizhe Li, and Shujun Li. " i know myself better, but not really greatly": Using llms to detect and explain llm-generated texts. *arXiv preprint arXiv:2502.12743*, 2025.

[31] Luyang Lin, Lingzhi Wang, Jinsong Guo, and Kam-Fai Wong. Investigating bias in llm-based bias detection: Disparities between llms and human perception. *arXiv preprint arXiv:2403.14896*, 2024.

[32] Yichen Li, Yuying Wang, Haozhao Wang, Yining Qi, Tianzhe Xiao, and Ruixuan Li. Fedssi: Rehearsal-free continual federated learning with synergistic synaptic intelligence. In *Forty-second International Conference on Machine Learning*.

[33] Steven L Bressler and Anil K Seth. Wiener–granger causality: a well established methodology. *Neuroimage*, 58(2):323–329, 2011.

[34] Alex Tank, Ian Covert, Nicholas Foti, Ali Shojaie, and Emily B Fox. Neural granger causality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):4267–4279, 2021.

[35] Huida Qiu, Yan Liu, Niranjan A Subrahmanya, and Weichang Li. Granger causality for time-series anomaly detection. In *2012 IEEE 12th international conference on data mining*, pages 1074–1079. IEEE, 2012.

[36] Zhewen Zhang and Lifeng Wu. Graph neural network-based bearing fault diagnosis using granger causality test. *Expert Systems with Applications*, 242:122827, 2024.

[37] Zehao Liu, Mengzhou Gao, and Pengfei Jiao. Gcad: Anomaly detection in multivariate time series from the perspective of granger causality. *arXiv preprint arXiv:2501.13493*, 2025.

[38] Cheng-Ming Lin, Ching Chang, Wei-Yao Wang, Kuang-Da Wang, and Wen-Chih Peng. Root cause analysis in microservice using neural granger causal discovery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 206–213, 2024.

[39] Jian-Guo Wang, Rui Chen, Xiang-Yun Ye, Zhong-Tao Xie, Yuan Yao, and Li-Lan Liu. A hierarchical granger causality analysis framework based on information of redundancy for root cause diagnosis of process disturbances. *Computers & Chemical Engineering*, 182:108589, 2024.

[40] Sipan Aslan and Hernando Ombao. Granger causality in high-dimensional networks of time series. *arXiv preprint arXiv:2406.02360*, 2024.

[41] Ziyi Zhang, Shaogang Ren, Xiaoning Qian, and Nick Duffield. Learning flexible time-windowed granger causality integrating heterogeneous interventional time series data. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4408–4418, 2024.

[42] Victor Troster. Testing for granger-causality in quantiles. *Econometric Reviews*, 37(8):850–866, 2018.

[43] Shiwei Li, Huifeng Guo, Xing Tang, Ruiming Tang, Lu Hou, Ruixuan Li, and Rui Zhang. Embedding compression in recommender systems: A survey. *ACM Computing Surveys*, 56(5):1–21, 2024.

[44] Ming Ke, Yaru Hou, Li Zhang, and Guangyao Liu. Brain functional network changes in patients with juvenile myoclonic epilepsy: a study based on graph theory and granger causality analysis. *Frontiers in Neuroscience*, 18:1363255, 2024.

[45] Howard Giles, Nikolas Coupland, and Justine Coupland. Accommodation theory: Communication, context, and consequence. *Contexts of accommodation: Developments in applied sociolinguistics*, 1:1–68, 1991.

[46] Yichen Li, Yijing Shan, Yi Liu, Haozhao Wang, Wei Wang, Yi Wang, and Ruixuan Li. Personalized federated recommendation for cold-start users via adaptive knowledge fusion. In *Proceedings of the ACM on Web Conference 2025*, WWW '25, page 2700–2709, New York, NY, USA, 2025. Association for Computing Machinery.

[47] Paul Schrodt, Kristina M Scharp, and Dawn O Braithwaite. *Engaging theories in interpersonal communication: Multiple perspectives*. Routledge, 2021.

[48] Shiwei Li, Wenchao Xu, Haozhao Wang, Xing Tang, Yining Qi, Shijie Xu, Weihong Luo, Yuhua Li, Xiuqiang He, and Ruixuan Li. Fedbat: communication-efficient federated learning via learnable binarization. In *Proceedings of the 41st International Conference on Machine Learning*, pages 29074–29095, 2024.

[49] Yi Mao and Guy Lebanon. Isotonic conditional random fields and local sentiment flow. *Advances in neural information processing systems*, 19, 2006.

[50] Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th international conference on world wide web*, pages 613–624, 2016.

[51] Cristian Danescu-Niculescu-Mizil, Michael Gamon, and Susan Dumais. Mark my words! linguistic style accommodation in social media. In *Proceedings of the 20th international conference on World wide web*, pages 745–754, 2011.

[52] Lu Wang and Claire Cardie. A piece of my mind: A sentiment analysis approach for online dispute detection. In Kristina Toutanova and Hua Wu, editors, *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 693–699, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

[53] Haozhao Wang, Shengyu Wang, Jiaming Li, Hao Ren, Xingshuo Han, Wenchao Xu, Shangwei Guo, Tianwei Zhang, and Ruixuan Li. Bsemifl: Semi-supervised federated learning via a bayesian approach. In *Forty-second International Conference on Machine Learning*.

[54] Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. Echoes of power: Language effects and power differences in social interaction. In *Proceedings of the 21st international conference on World Wide Web*, pages 699–708, 2012.

[55] Robin IM Dunbar, Anna Marriott, and Neil DC Duncan. Human conversational behavior. *Human nature*, 8:231–246, 1997.

[56] Haozhao Wang, Peirong Zheng, Xingshuo Han, Wenchao Xu, Ruixuan Li, and Tianwei Zhang. Fednlr: Federated learning with neuron-wise learning rates. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3069–3080, 2024.

[57] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*, 2017.

[58] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*, 2018.

[59] Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. Multiwoz–a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. *arXiv preprint arXiv:1810.00278*, 2018.

[60] Bill Byrne, Karthik Krishnamoorthi, Chinnadhurai Sankar, Arvind Neelakantan, Daniel Duckworth, Semih Yavuz, Ben Goodrich, Amit Dubey, Andy Cedilnik, and Kyu-Young Kim. Taskmaster-1: Toward a realistic and diverse dialog dataset. *arXiv preprint arXiv:1909.05358*, 2019.

[61] Yutian Chen, Hao Kang, Vivian Zhai, Liangze Li, Rita Singh, and Bhiksha Raj. Token prediction as implicit classification to identify llm-generated text. *arXiv preprint arXiv:2311.08723*, 2023.

[62] Xiaoming Liu, Zhaohan Zhang, Yichen Wang, Hang Pu, Yu Lan, and Chao Shen. CoCo: Coherence-Enhanced Machine-Generated Text Detection Under Data Limitation With Contrastive Learning, October 2023. arXiv:2212.10341 [cs].

[63] Rongsheng Wang, Haoming Chen, Ruizhe Zhou, Han Ma, Yaofei Duan, Yanlan Kang, Songhua Yang, Baoyu Fan, and Tao Tan. Llm-detector: Improving ai-generated chinese text detection with open-source llm instruction tuning. *arXiv preprint arXiv:2402.01158*, 2024.

[64] Kangxi Wu, Liang Pang, Huawei Shen, Xueqi Cheng, and Tat-Seng Chua. Llmdet: A third party large language models generated text detection tool. *arXiv preprint arXiv:2305.15004*, 2023.

[65] Junchao Wu, Runzhe Zhan, Derek F Wong, Shu Yang, Xuebo Liu, Lidia S Chao, and Min Zhang. Who wrote this? the key to zero-shot llm-generated text detection is gecscore. *arXiv preprint arXiv:2405.04286*, 2024.

[66] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[67] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

[68] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[69] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the claims made, including the contributions made in the paper and important assumptions and limitations.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations are thoroughly discussed in Appendix A.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The dataset and code are in `https://anonymous.4open.science/r/Distinguishing-LLMs-by-Analyzing-Dialogue-Dynamics-with-Granger-Causality-56E4/`. This direct provision of code and the newly constructed dataset is the strongest factor supporting reproducibility. Other researchers can directly access these resources to replicate the experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The dataset and code are in https://anonymous.4open.science/r/Distinguishing-LLMs-by-Analyzing-Dialogue-Dynamics-with-Granger-Causality-56E4/. This direct provision of code and the newly constructed dataset is the strongest factor supporting reproducibility. Other researchers can directly access these resources to replicate the experiments.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The dataset and code are in https://anonymous.4open.science/r/Distinguishing-LLMs-by-Analyzing-Dialogue-Dynamics-with-Granger-Causality-56E4/. All the experimental settings can be found in the code. What's more, other settings are thoroughly discussed in Appendix B.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: All experiments reported in the paper include properly defined error bars.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Experiments compute resources are thoroughly discussed in Appendix B.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [NA]

   Justification: The paper poses no such risks.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Impacts are thoroughly discussed in Appendix C.

    Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: This work demonstrates a commitment to properly crediting existing assets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We provide detailed code and datasets in https://anonymous.4open.science/r/Distinguishing-LLMs-by-Analyzing-Dialogue-Dynamics-with-Granger-Causality-56E4/.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [No]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Limitations

Noise-free Setting: The experiments are conducted on clean, curated datasets without considering real-world noise. These factors may significantly impact the robustness of detection models when deployed in practical settings.

Well-specified Model: We assume that pre-trained language models used in our benchmarking are well-suited for the detection task. However, these models were originally trained for language generation rather than detection, and suboptimal fine-tuning or domain mismatch may limit their effectiveness in distinguishing H-H and H-C dialogue.

Asymptotic Approximations: Some of the statistical analysis techniques employed rely on asymptotic assumptions that require large sample sizes to achieve accurate estimation. In practice, especially with limited or imbalanced datasets, these approximations may not hold, potentially affecting the validity of the results.

Only Applicability to Two-Party Dialogues: Our current methodology and experimental validation are exclusively focused on two-party dialogues. While applying Granger Causality to multi-party $(N > 2)$ interactions is theoretically feasible, it introduces significant complexity. Specifically, the number of potential causal relationships to analyze grows quadratically from 2 to $N * (N - 1)$, making the current approach computationally challenging. Future work is required to extend our framework to model the more complex dynamics inherent in multi-party conversations, such as coalitions or mediation effects.

B Implementation Details

B.1 Hardware devices

All our experiments were meticulously conducted on a high-performance computing platform running Ubuntu. The platform is powered by an Intel(R) Xeon(R) Platinum 8176 CPU @ 2.10GHz, delivering robust computational capabilities. The system is equipped with a substantial 503 GB of memory, ensuring efficient data processing and storage. Additionally, to further enhance computational power, we utilized four NVIDIA Corporation GA102GL RTX A6000 GPUs. These GPUs provided the necessary parallel processing power to handle the intensive computational tasks associated with our research. The stability and broad support of the Ubuntu operating system allowed us to fully leverage the hardware's performance, ensuring the smooth execution of experiments and the reliability of our results.

B.2 Datasets

- **DailyDialog**: This dataset contains high-quality, multi-turn dialogues reflecting everyday human communication. The conversations cover various topics and exhibit natural language usage. We utilized dialogues directly from this corpus as part of our H-H chit-chat data, selecting conversations exceeding a minimum turn length threshold suitable for GCT analysis.
- **PersonaChat**: This dataset consists of chit-chat dialogues where participants are assigned specific persona profiles that they are expected to condition their conversation on. It encourages engaging and consistent dialogue. Similar to DailyDialog, naturally occurring dialogues between human participants in this dataset were included in our H-H chit-chat corpus.
- **MultiWOZ** : A large-scale, multi-domain dataset for task-oriented dialogues, covering domains like restaurants, hotels, transportation, etc. It is a standard benchmark in dialogue state tracking and end-to-end dialogue systems. We used the human-human Wizard-of-Oz collected dialogues within this dataset, where one human plays the user and another simulates a constrained system based on database information, as representative examples for our H-H task-oriented corpus.
- **Taskmaster-1**: This dataset contains goal-oriented dialogues, covering tasks such as ordering pizza, creating auto repair appointments, and booking flights. It includes both spoken and written conversations collected via a Wizard-of-Oz setup. We specifically used the written

dialogues where both the 'user' and the 'wizard' (simulating the system) were human participants to form part of our H-H task-oriented corpus.

H-H Corpus: The Human-Human (H-H) corpus used in our experiments was formed by selecting dialogues directly from the aforementioned datasets (DailyDialog, PersonaChat, the human-controlled segments of MultiWOZ, and Taskmaster-1 WoZ data). Dialogues below a pre-defined length (e.g., 20 turns) were filtered out to ensure suitability for Granger Causality analysis.

H-M Corpus Construction: The Human-LLM (H-M) corpus was constructed semi-synthetically, derived directly from the dialogues selected for the H-H corpus to ensure maximal comparability of user input and conversational context. For each selected H-H dialogue $C^{(i)} = \{(U_1, A_1), (U_2, A_2), \dots\}$ (where U_t denotes a user utterance and A_t denotes the original human agent's utterance at turn t), we identified all turns originally spoken by the human agent A. We then employed a specific pre-trained Large Language Model (LLM-X, e.g., specify model like Llama-2-Chat 70B or GPT-4) to generate alternative responses for these turns.

Specifically, for each agent turn A_t, the dialogue history preceding it, typically ending with the user's utterance U_t, was provided as context to LLM-X. Let the history be $H_t = (U_1, A'_1, U_2, A'_2, \dots, A'_{t-1}, U_t)$ where A'_k are the previously generated LLM responses (or original A_k for $k = 1$ if the agent starts). The LLM was prompted to generate a suitable response A'_t given this history:

$$A'_t = \text{LLM-X}(H_t) \tag{13}$$

This generated response A'_t then replaced the original human response A_t in the dialogue sequence. This process was repeated for all agent turns in the dialogue, resulting in a new H-M dialogue $C'^{(i)} = \{(U_1, A'_1), (U_2, A'_2), \dots\}$. Note that the user utterances U_t remain identical to those in the original H-H dialogue $C^{(i)}$.

For dialogues derived from PersonaChat, the corresponding persona information was included in the prompt for LLM-X to encourage consistent persona adoption. For task-oriented dialogues derived from MultiWOZ and Taskmaster-1, relevant task goals or simulated dialogue states (if available and applicable) were potentially included in the prompt history H_t to guide the LLM towards task completion, mimicking the information available to the original human agent/wizard. This construction method yields an H-M corpus where the user's side of the conversation is natural human language drawn from established datasets, while the agent's side is generated by the target LLM conditioned on that human input, allowing for a controlled comparison of response patterns and interaction dynamics against the original H-H dialogues. Similar length filtering was applied to the resulting H-M dialogues.

B.3 Metrics

To ensure the accuracy and reliability of the results, each experiment was conducted in triplicate, and the standard deviations were calculated. This approach effectively assesses the stability and consistency of the data, thereby enhancing the credibility of our conclusions. To assess the detector's capability to differentiate between texts generated by large language models (LLMs) and those written by humans, we utilize Accuracy (A) and the Area Under the Receiver Operating Characteristic Curve (AUROC) as primary performance metrics. Additionally, we consider other metrics, such as F1 scores (F1) and Recall (R), to provide a more comprehensive evaluation.

$$A = \frac{TP + TN}{TP + TN + FP + FN} \tag{14}$$

$$R = \frac{TP}{TP + FN}; \quad F1 = 2 \times \frac{P \times R}{P + R} \tag{15}$$

True Positives (TP) refer to H-H dialogue correctly identified by the model. True Negatives (TN) represent H-C dialogue accurately classified as H-c dialogue. False Positives (FP) denote H-C dialogue incorrectly labeled H-H dialogue, while False Negatives (FN) correspond to H-H dialogue the model fails to identify correctly.

C Performance evaluation

C.1 Cross-Domain Evaluation

Table 5: Cross-domain evaluation results. All methods were trained exclusively on the DailyDialog dataset and evaluated on the entirely unseen Taskmaster-1, PersonaChat, and MultiWOZ test sets. The best number is highlighted in bold, while the second best one is underlined. Our approach consistently outperforms other methods.

Method	Taskmaster-1		PersonaChat		MultiWOZ	
	ACC	F1	ACC	F1	ACC	F1
DetectGPT (ICML 2023)	59.12 ± 0.82	55.66 ± 1.54	60.70 ± 0.98	59.39 ± 1.31	61.30 ± 3.69	61.78 ± 2.75
COCO (EMNLP 2023)	66.80 ± 0.70	68.29 ± 2.16	69.56 ± 0.41	65.58 ± 0.93	65.67 ± 1.36	68.88 ± 2.46
LLMDet (EMNLP 2023)	60.13 ± 1.14	64.83 ± 2.37	62.13 ± 1.33	61.81 ± 0.57	63.43 ± 1.20	62.96 ± 1.46
SeqXGPT (EMNLP 2023)	59.75 ± 0.91	64.66 ± 0.99	58.48 ± 1.56	60.58 ± 1.77	61.00 ± 1.22	61.18 ± 0.77
Fast-DetectGPT (ICLR 2024)	60.93 ± 1.96	60.66 ± 1.64	63.02 ± 2.64	62.05 ± 1.10	64.01 ± 0.68	61.28 ± 0.59
T5-Sentinel (EMNLP 2024)	69.24 ± 1.31	70.07 ± 0.16	<u>74.95</u> ± 0.42	72.41 ± 1.45	74.90 ± 0.94	71.67 ± 2.57
RoBERTa-MPU (ACL 2024)	69.35 ± 0.93	74.15 ± 0.56	73.51 ± 1.29	72.15 ± 2.30	70.99 ± 4.82	<u>74.79</u> ± 0.96
DeTeCtive (NeurIPS 2024)	67.85 ± 0.45	70.69 ± 1.58	71.46 ± 0.55	71.98 ± 1.05	71.87 ± 1.30	73.40 ± 4.89
OUTFOX (AAAI 2024)	<u>76.39</u> ± 1.88	<u>78.88</u> ± 0.72	72.29 ± 1.01	<u>78.10</u> ± 3.07	<u>78.64</u> ± 0.87	76.70 ± 1.73
GFCScore (ACL 2025)	67.73 ± 1.97	71.12 ± 0.38	73.85 ± 1.99	73.98 ± 0.22	71.82 ± 3.20	65.30 ± 1.20
GPT-3.5-turbo (2023)	60.72 ± 1.73	59.28 ± 0.95	59.38 ± 0.14	61.14 ± 1.09	63.07 ± 1.20	60.18 ± 1.88
LLaMA-7B (2024)	59.74 ± 1.23	58.85 ± 1.17	60.54 ± 1.76	60.36 ± 1.56	61.76 ± 0.78	58.05 ± 2.14
LLaMA-13B (2024)	60.94 ± 3.30	62.81 ± 2.27	63.32 ± 0.26	63.53 ± 2.85	65.16 ± 0.30	62.35 ± 1.13
GPT-4 (2024)	58.90 ± 2.31	64.61 ± 1.64	64.23 ± 1.25	59.87 ± 0.60	67.70 ± 0.43	63.29 ± 0.98
Gemma (2025)	62.26 ± 0.14	66.68 ± 1.01	66.87 ± 1.00	62.98 ± 3.70	67.98 ± 0.76	62.12 ± 1.51
Qwen-2 (2025)	59.39 ± 0.71	61.67 ± 2.43	64.81 ± 0.05	61.78 ± 3.62	64.98 ± 0.90	62.67 ± 1.05
Deepseek-R1 (2025)	63.29 ± 1.99	64.73 ± 1.69	65.22 ± 2.23	66.46 ± 1.84	66.27 ± 0.12	64.62 ± 1.06
ChatbotID (Ours)	**80.59** ± 1.18	**81.96** ± 2.30	**83.84** ± 1.38	**84.72** ± 2.19	**82.28** ± 0.62	**83.40** ± 0.45

To test for robustness against domain shift, we performed a cross-domain evaluation. We trained ChatbotID and all baseline models exclusively on the DailyDialog dataset and evaluated their performance on the entirely unseen Taskmaster-1, PersonaChat, and MultiWOZ test sets. General-purpose LLMs (e.g. GPT-3.5-turbo, LLaMA-7B, LLaMA-7B, LLaMA-13B, Gemma, etc.) adopt a zero-shot detection approach. The results of our cross-domain evaluation demonstrate the robustness of our approach. While all methods were trained exclusively on DailyDialog, ChatbotID maintains a high F1-score of over 83% across all datasets. This performance represents a substantial margin of 5-20% F1 points over all baseline methods.

C.2 Zero-Shot Evaluation on WildChat dataset

Table 6: Experimental results on the WildChat dataset.

Method	WildChat ACC
DetcctGPT (ICML 2023)	58.04 ± 2.62
COCO (EMNLP 2023)	64.76 ± 0.37
LLMDet (EMNLP 2023)	69.67 ± 0.90
SeqXGPT (EMNLP 2023)	60.19 ± 1.05
Fast-DetectGPT (ICLR 2024)	62.19 ± 1.05
T5-Sentinel (EMNLP 2024)	68.03 ± 1.24
RoBERTa-MPU (ACL 2024)	67.68 ± 2.97
DeTeCtive (NeurIPS 2024)	67.85 ± 0.45
OUTFOX (AAAI 2024)	78.44 ± 0.60
GECScore (ACL 2025)	68.44 ± 0.60
ChatbotID (Ours)	**79.12 ± 1.53**

To evaluate our model's performance on naturally occurring H-C dialogues, we tested our model in a zero-shot setting on the WildChat dataset. ChatbotID achieves the highest accuracy (79.12%) among all methods, outperforming even the most competitive baselines like OUTFOX. This result

demonstrates that the unidirectional influence signal captured by ChatbotID is not merely an artifact of our semi-synthetic data generation process. Instead, it is a genuine and detectable characteristic present in real-world human-LLM interactions.

C.3 Inference Complexity Comparison

Table 7: Inference Complexity Comparison.

Method	Inference Complexity
DetcctGPT (ICML 2023)	$O(n)$
COCO (EMNLP 2023)	$O(1)$
LLMDet (EMNLP 2023)	$O(n)$
SeqXGPT (EMNLP 2023)	$O(1)$
Fast-DetectGPT (ICLR 2024)	$O(n)$
T5-Sentinel (EMNLP 2024)	$O(n)$
RoBERTa-MPU (ACL 2024)	$O(1)$
DeTeCtive (NeurIPS 2024)	$O(1)$
OUTFOX (AAAI 2024)	$O(n)$
GECScore (ACL 2025)	$O(n)$
ChatbotID (Ours)	$O(1)$

In the training phase, we acknowledge that our proposed method, ChatbotID, has a higher computational overhead compared to some lightweight detection approaches. The computational complexity is primarily concentrated in two stages. The calculation of GCT features requires additional processing time. The fine-tuning process, which incorporates auxiliary losses, is slightly more complex than a standard single-task classification setup.

However, the primary advantage of `ChatbotID` lies in its inference efficiency. Once trained, making a prediction is extremely fast, achieving an inference complexity of $O(1)$. This is because it only requires a single forward pass through the model to make a prediction.

This stands in contrast to perturbation-based approaches, such as **DetectGPT**, **LLMDet**, and **OUTFOX**, which are computationally heavy at inference time. For every single dialogue they need to evaluate, they must perform multiple forward passes through a large language model (LLM) to generate perturbations and calculate scores. This characteristic makes them prohibitively slow and expensive for any real-time or large-scale application.

D Potential Positive Societal Impacts

Enhanced Dialogue Understanding and Interaction: By leveraging interaction dynamics and semantic-focused attribution, this research aims to improve dialogue understanding and classification accuracy beyond purely semantic analysis. This could lead to more effective communication tools, such as chatbots and virtual assistants, enhancing user experience and satisfaction across various applications.

Improved Detection of AI-Generated Text: The development of sophisticated models for detecting machine-generated text can play a crucial role in combating misinformation and ensuring content authenticity. In an era where LLM-generated is becoming increasingly prevalent, having reliable methods to distinguish between human and AI-generated texts is vital for maintaining trust in digital communications.

Promotion of Ethical Use of AI: Through advancements in identifying LLM-generated, this research supports the ethical use of technology by helping prevent misuse and manipulation. It contributes to the broader conversation on AI ethics and responsibility, encouraging transparency and accountability in how AI technologies are deployed and managed.