IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received 8 October 2025, accepted 3 November 2025, date of publication 13 November 2025, date of current version 2 December 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3632654

RESEARCH ARTICLE

UnHIDE: A Novel Framework for Unsupervised Human-Interpretable Dialogue Exploration

PATRÍCIA FERREIRA [1], ANA ALVES [1,2], CATARINA SILVA [1], (Senior Member, IEEE), AND HUGO GONÇALO OLIVEIRA [1]

[1]Department of Informatics Engineering, CISUC/LASI–Centre for Informatics and Systems of the University of Coimbra, University of Coimbra, 3004-531 Coimbra, Portugal
[2]Coimbra Institute of Engineering (ISEC), Polytechnic Institute of Coimbra, 3045-093 Coimbra, Portugal

Corresponding author: Catarina Silva (catarina@dei.uc.pt)

This work is financed by the Portuguese Recovery and Resilience Plan (PRR) through project C645008882-00000055, Center for Responsible AI.

ABSTRACT Dialogue systems are increasingly central to applications in customer service, virtual assistance, and beyond, generating vast amounts of conversational data. While these systems have advanced with the exploitation of large language models (LLMs), they still face key limitations, some, in fact, strengthened by the black-box nature of such models, including the lack of feedback mechanisms and the absence of effective solutions for human-in-the-loop interaction and iterative improvement. As a result, understanding, refining, and debugging dialogue behavior remains a major challenge. To address this, we introduce UnHIDE, a novel, unsupervised framework for Human-Interpretable Dialogue Exploration. UnHIDE is designed to support human understanding of large collections of dialogues by surfacing interpretable structures and trends. It operates in three stages: 1) utterance clustering to group semantically similar dialogue turns, 2) flow discovery to build dialogue trajectories based on these clusters, and 3) the computation of interpretable metrics to analyze flow complexity, sentiment progression, and response times. We evaluate UnHIDE using a newly-created, automatically-generated, task-oriented dialogue dataset, where dialogue length, sentiment dynamics, and timing are systematically varied. Our results show that UnHIDE reliably captures these variations and provides actionable insights into dialogue structure and quality. By enabling transparent, human-interpretable analysis of dialogue without supervision, UnHIDE offers a powerful tool for diagnosing and improving dialogue systems. It not only fills a critical gap in feedback and interpretability, but also lays the groundwork for incorporating human-in-the-loop practices into future conversational Artificial Intelligence (AI) development.

INDEX TERMS Dialogue analysis, dialogue flows, flow discovery, human-interpretable AI, metrics, task-oriented dialogue, unsupervised learning.

I. INTRODUCTION

Natural language interaction capabilities have made dialogue systems a cornerstone of modern AI applications [1]. From customer service automation [2] to intelligent tutoring systems [3], these conversational agents are being deployed across a wide range of industries to provide timely assistance, answer questions, and support user needs. Their relevance has only intensified with the rapid rise of Large Language Models (LLMs), which have enabled more coherent, context-aware, and human-like responses [4], even if at the cost of lower transparency and control of internal processes. As a result, dialogue systems have become increasingly influential in shaping how people access information, receive support, and interact with digital services. Despite this progress, key challenges remain, particularly in ensuring that such systems can be effectively understood, evaluated, and improved through transparent, human-centered analysis [5].

The associate editor coordinating the review of this manuscript and approving it for publication was Maria Chiara Caschera.

However, a key challenge in the development and evaluation of dialogue systems lies in the lack of general-purpose, unsupervised methods that can adapt across diverse domains and tasks without relying on labeled data. This limitation hinders scalable analysis, especially in dynamic or low-resource settings where supervised training or annotation is impractical [6]. Furthermore, current approaches often fail to provide effective visual representations of the conversational structure, making it difficult for developers and analysts to explore or diagnose dialogue behaviors at scale [7]. Even when such tools exist, the insights they provide are typically not interpretable by humans, restricting their value in iterative system improvement and in integrating human feedback [8]. Addressing these limitations is essential for enabling more transparent, actionable, and scalable dialogue system development.

In an effort to bridge the gap between automated dialogue systems and human interpretability, a number of supervised approaches have been proposed that rely on extensive annotation and dialogue state tracking [9]. These methods often focus on labeling intents, slot values, or dialogue acts at the utterance level, enabling fine-grained, step-by-step tracking of conversation dynamics. Such techniques can support downstream tasks like dialogue policy learning and stateful response generation, and are frequently used to ensure that dialogue systems maintain consistency and goal orientation.

Nevertheless, these supervised methods face several limitations when it comes to scalable, interpretable dialogue analysis. First, the annotation processes involved are typically labor-intensive and domain-specific, making them difficult to generalize or apply in data-rich but label-scarce environments. Second, by focusing on dialogue state tracking at the turn level, these approaches often offer only a narrow, case-by-case view, rather than providing a global perspective on dialogue structure and flow across a corpus. Finally, their outputs tend to serve system-level optimization rather than human interpretability, leaving analysts and designers with limited insight into emergent patterns, structural complexity, or user sentiment trends across dialogues.

To address the limitations of existing approaches in dialogue analysis and interpretation, we introduce UnHIDE, an unsupervised framework for Human-Interpretable Dialogue Exploration. UnHIDE enables scalable exploration of large dialogue datasets by combining flow discovery with interpretable metrics, offering deep insights into conversational behavior without requiring labeled data or domain-specific assumptions.

UnHIDE is designed to uncover meaningful structure in large-scale dialogue data while enabling intuitive, human-centered insights. The framework operates in three key stages. First, it performs utterance clustering, grouping semantically similar utterances across dialogues to identify recurring conversational patterns. Next, it conducts flow discovery, constructing abstract representations of how conversations progress through these clusters. Finally, a set of interpretable metrics is computed, capturing aspects such as flow complexity, sentiment progression, response time, and distribution of dialogue lengths.

Unlike traditional visualization tools that offer only surface-level representations, UnHIDE combines structure and statistics to facilitate a deeper understanding of conversational dynamics. This makes it possible to identify frequent paths, detect anomalies such as spikes in negative sentiment or delayed responses, and validate the adherence to expected dialogue protocols.

We validate UnHIDE on MultiSynthiment, a novel, synthetically generated task-oriented dataset where dialogue length, sentiment, and response times are systematically varied. Experimental results confirm that these generation variables are reliably captured in both the discovered flows and the associated metrics, enabling rich, human-interpretable analysis.

Our contributions include:

- UnHIDE, a novel, fully unsupervised framework for large-scale, human-interpretable dialogue exploration that requires no annotated data or predefined structure beyond dialogues involving two participants communicating in turn-taking fashion.
- A three-stage analysis pipeline combining utterance clustering, flow discovery, and a set of interpretable metrics to uncover structural, temporal, and emotional patterns in dialogue.
- MultiSynthiment, a synthetic, task-oriented dialogue dataset with controlled variation in utterance count, sentiment, and response time, used to validate the effectiveness of the framework.
- Empirical results demonstrating that UnHIDE captures key generative variables and provides actionable insights into dialogue design, communication trends, and user experience.

The remainder of this paper is organized as follows. Section II reviews prior research on dialogue flow discovery and large-scale dialogue analysis. Section III presents the proposed UnHIDE framework, detailing its methodology for unsupervised flow discovery and the suite of interpretable metrics employed. Section IV outlines the experimental setup, including the synthetic generation of the MultiSynthiment task-oriented dialogue dataset. Section V reports and discusses the results, demonstrating the effectiveness of UnHIDE in capturing meaningful dialogue properties and validating key hypotheses. Finally, Section VI summarizes the main findings and outlines directions for future research.

II. RELATED WORK

A. DIALOGUE FLOW DISCOVERY

Dialogue flow discovery has been widely explored as a means to support the development, evaluation, and debugging of dialogue systems [10], [11], [12], [13]. Discovered flows provide structured abstractions of conversations, commonly modeled as directed graphs in which nodes represent clusters of semantically-similar utterances and edges correspond to

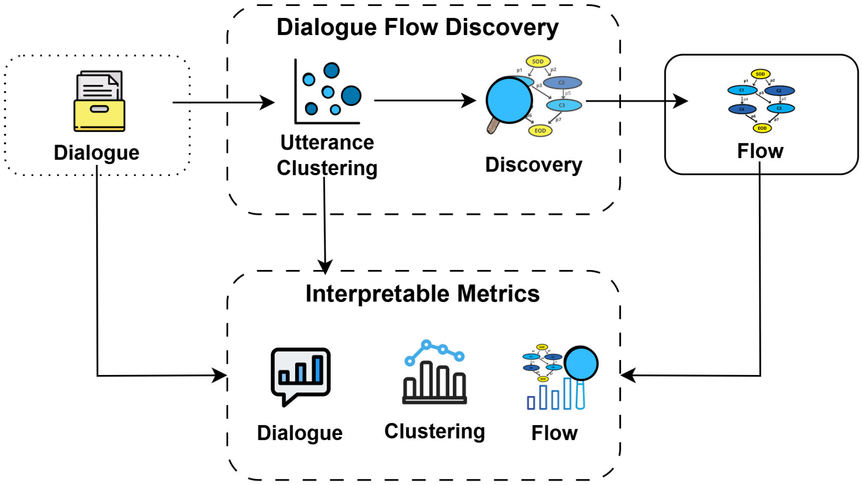

FIGURE 1. The UnHIDE framework: (1) utterance clustering identifies semantically similar dialogue units across a corpus, (2) flow discovery builds abstracted conversation trajectories based on cluster transitions, and (3) interpretable metrics quantify structural complexity, sentiment progression, and response times.

conversational transitions. This structure has proven useful in both task-oriented [13], [14] and open-domain dialogue contexts [1].

Various clustering methods have been employed to discover dialogue states or turns, including co-clustering [10], DBSCAN [11], and k-means [15]. Transitions between clusters are typically inferred from the order of utterances in the dataset, and are quantified using either raw frequency [10], [11] or transition probabilities [15]. Thresholding is often applied to remove infrequent transitions, which improves the readability and interpretability of the resulting graphs [10], [15].

Although originally motivated by flow-based chatbot platforms such as Google DialogFlow [16] or Rasa [17], recent work has extended flow discovery to the post hoc analysis of human-human and human-machine conversations [11], [14], [18]. These efforts aim not just to support development, but also to uncover frequent behavioral patterns, identify failure cases, and guide design improvements.

B. INTERPRETABILITY AND LABELING

Beyond structural discovery, a significant body of work focuses on making dialogue flows interpretable to human analysts. A common strategy is to label clustered utterances, either through manual annotation, statistical heuristics, or automatic generation. Earlier efforts relied on labeling states with frequent words or n-grams [10], [15] or with manually assigned dialogue acts [19]. More recently, LLMs have been leveraged to generate more descriptive and context-aware labels for clusters [14], [20], and even to annotate entire dialogues in an unsupervised fashion [21].

Interpretability has also become a key component in synthetic dialogue generation, where dialogue flows are used to guide controllable generation using LLMs [22], [23], [24]. These works demonstrate that structured flows can serve not only for analysis, but also as scaffolding for consistent and coherent dialogue generation.

However, few approaches offer integrated tools that combine structure and interpretability in a way that directly supports human-in-the-loop diagnostics, scalable error analysis, or protocol validation across datasets.

C. METRICS FOR DIALOGUE STRUCTURE EVALUATION

A challenging area of research has focused on defining quantitative metrics to evaluate the structure and coverage of discovered dialogue flows. Some studies propose computing the proportion of transitions covered in unseen dialogue sets to assess generalizability [15]. Others evaluate graphs based on complexity (for example, number of states or branching factor) and path coverage to estimate flow completeness [25].

In addition to structural metrics, there is increasing interest in incorporating contextual or behavioral signals such as sentiment and response time. Sentiment analysis has been used both as an input signal for clustering [19], [21], [26] and as a feature for analyzing dialogue quality, satisfaction, or emotional progression [18], [26]. Similarly, response latency has been used to characterize user engagement and to distinguish between efficient and inefficient conversational paths [18].

While several metrics have been proposed individually, there is limited exploration of how they can be systematically combined to support a multifaceted, interpretable evaluation of dialogue behavior.

Summary: In summary, existing work has laid important foundations for discovering and analyzing dialogue flows, as well as for developing interpretable representations and evaluative metrics. However, most approaches focus on isolated components, either flow structure, labeling, or single metrics, and often require supervision or domain-specific tuning. In contrast, UnHIDE offers an unsupervised,

TABLE 1. Metrics computed in raw dialogues, discovered clusters, and flow. A dataset contains dialogues ($d \in D$), which contain utterances ($u \in U$), which contain tokens ($w \in W$). Utterances have an assigned sentiment (+, −, none) and timestamps. Flow states ($c \in C$) are connected by transitions ($e \in E$). *nc* and *nf* denote, respectively, the normalized complexity and the normalized average Fuzzy Dialogue-Graph Edit distance (FuDGE) score.

	Metrics	Description	Notation
Dialogue	Utterances in Dataset	For each dataset, total number of utterances	$\|U\|$
	Dialogues in Dataset	For each dataset, total number of dialogues	$\|D\|$
	Utterances per Dialogue	For each dataset, average number of utterances per dialogue	$\|U\|/\|D\| = \frac{\|U\|}{\|D\|}$
	Dialogue Duration	For each dataset, average time between the first and last utterance	$\Delta T/\|D\| = \frac{1}{\|D\|} \sum_{d \in D} T^d_{EOD} - T^d_{SOD})$
	Sentiment Distribution	For each dialogue, proportion of positive and negative sentiments	$S^+/\|D\| = \frac{1}{\|D\|} \sum_{d \in D} \frac{\|U^+_d\|}{\|U_d\|}, \quad S^-/\|D\| = \frac{1}{\|D\|} \sum_{d \in D} \frac{\|U^-_d\|}{\|U_d\|}$
Clustering	Number of States	For each dataset, total number of discovered dialogue states	$\|C\|$
	Silhouette Score	An indicator of cluster quality, considering cohesion and separation	$SS = \max\left(\frac{b(s)-a(s)}{\max(a(s),b(s))}\right)$
	Utterances/State	Average number of utterances per state	$\|U\|/\|C\| = \frac{\sum_{c \in C} \|U_c\|}{\|C\|}$
	Utterances in State	For each state, number of clustered utterances	$\|U_c\|$
	Tokens/utterance	For each state, average number of tokens per utterance (NLTK tokenizer)	$\|W\|/\|U\| = \frac{1}{\|C\|} \sum_{c=1}^{\|C\|} \left(\frac{\sum_{u \in U_c} \|W_u\|}{\|U_c\|}\right)$
	TimeSince(SOD)	For each state, average time since SOD	$\Delta T_{SOD} = \frac{\sum_{u \in U_c} T_{SOD_u}}{\|U_c\|}$
	TimeSince(Prev)	For each state, average time since the previous utterance	$\Delta T_{prev} = \frac{\sum_{i=1}^{\|U_c\|-1} (T_{i+1} - T_i)}{\|U_c\|-1}$
	Sentiment in State	For each state, average sentiment of utterances	$Sent_c = \frac{\sum_{u=1}^{\|U_c\|} Sent_u}{\|U_c\|}$
	Sentiment Cohesion	Average standard deviation of sentiment across all states.	$FSC = \frac{1}{\|C\|} \sum_{c=1}^{\|C\|} \sigma(Sent_c)$
Flow	Number of Transitions	Total number of transitions between dialogue states	$\|E\|$
	Initial Sentiment	Average sentiment of states connected to SOD	$Sent_{SOD} = \frac{\sum_{c \in \|C_{SOD}\|} Sent_c}{\|C_{SOD}\|}$
	Final Sentiment	Average sentiment of states connected to EOD	$Sent_{EOD} = \frac{\sum_{c \in \|C_{EOD}\|} Sent_c}{\|C_{EOD}\|}$
	Sentiment Variation	Sentiment variation between SOD and EOD	$\Delta Sent = Sent_{EOD} - Sent_{SOD}$
	Neg. Sentiment at End	Proportion of negative utterances in EOD states	$EOD_- = \frac{\|U^-_{EOD}\|}{\|U_{EOD}\|}$
	Flow Density	Graph density, an indicator of complexity	$FD = \frac{\|E\|}{\|C\|(\|C\|-1)}$
	Transition Coverage	Proportion of covered transitions in unseen dialogues	$EC = \frac{\|\forall e \in E_{Test} \cap e \in E_{Flow}\|}{\|E_{Test}\|}$
	Flow F1-Score	Harmonic mean of normalized complexity and FuDGE score	$FF1 = \frac{2(1-nc)(1-nf)}{(1-nc)+(1-nf)}$

integrated framework that combines flow discovery with a comprehensive set of interpretable metrics, enabling scalable, human-centered analysis of dialogue data across tasks and domains.

III. UnHIDE

UnHIDE, illustrated in Figure 1, is an unsupervised framework for analyzing large collections of dialogue. It centers on flow discovery and computes a diverse set of interpretable metrics using three data representations: the raw dialogues, the utterance clusters, and the flow graph itself. This section outlines the process of flow construction and describes the metrics used for quantitative interpretation.

A. DIALOGUE FLOW DISCOVERY

Given a collection of text-based dialogues between two participants (e.g., user and agent), the dialogue flow is discovered in two main stages, following the principles outlined by Ferreira et al. [15].

1) UTTERANCE CLUSTERING

Utterances are embedded and grouped based on semantic similarity, separately for each speaker. These clusters define the states of the dialogue flow. The number of clusters can be selected by optimizing the Silhouette Score [27], ensuring meaningful groupings. For enhanced interpretability, each cluster is assigned a human-readable label. This can be achieved using LLMs for summarizing representative utterances from each cluster [14], [20].

2) FLOW CONSTRUCTION

Transitions between discovered states are then extracted from the dialogue sequences, resulting in a directed graph $G(C, E)$, where nodes $c \in C$ represent clusters (hereafter, states), and edges $e(c_1, c_2, p_{12}) \in E$ indicate observed transitions between states, weighted by transition probability p_{12}, i.e., the proportion of utterances in c_1 followed by c_2. To ensure clarity, low-probability transitions can be pruned using a threshold θ. Each flow starts and ends with special states, namely: Start of Dialogue (SOD) and End of Dialogue (EOD).

To further enrich the flow, each state is associated with the average sentiment of the utterances it contains. This sentiment is reflected visually in the graph through the color of transitions: green for predominantly positive states, yellow for neutral, and red for negative.

As long as the dialogues are sequential and speaker turns are identifiable, UnHIDE can be applied to any dialogue corpus, regardless of language, domain, or structure.

B. INTERPRETABLE METRICS

To support human interpretation, UnHIDE computes a range of complementary metrics, categorized according to their input: raw dialogues, utterance clusters, or the discovered flow graph. A full summary of the computed metrics, their definitions, and formulae is provided in Table 1.

1) DIALOGUE METRICS

Dialogue metrics are computed directly from the original dialogues, prior to flow discovery. They provide high-level statistics on dialogue volume and structure:

- Total number of utterances (U) and average utterances per dialogue ($|U|/|D|$).
- Average dialogue duration ($\Delta T/|D|$), offering a proxy for response time.
- Sentiment composition, including the proportion of positive ($S^+/|D|$) and negative ($S^-/|D|$) utterances per dialogue.

2) CLUSTERING METRICS

Clustering metrics describe the discovered utterance clusters (i.e., flow states), regardless of flow structure:

- Number of states ($|C|$) and average utterances per state ($|U|/|C|$).
- Silhouette Score (SS) [27], indicating cluster separation and cohesion.
- Sentiment Cohesion (FSC) [26], measuring the consistency of sentiment within each cluster.

Additional per-cluster metrics include:

- Number of utterances assigned to the cluster (U_c).
- Average utterance length in tokens ($|W|/|U|$).
- Average time elapsed since dialogue start (ΔT_{SOD}) and since previous utterance (ΔT_{prev}).
- Average sentiment score ($Sent_c$).

3) FLOW METRICS

Flow metrics quantify the structural complexity of the resulting flow graph, which can impact its interpretability and overall quality:

- Number of transitions ($|E|$) and flow density (FD), indicating flow complexity.
- Sentiment progression, including average sentiment at the start ($Sent_{SOD}$) and end ($Sent_{EOD}$), proportion of negative sentiment at final states (EOD_-), and overall sentiment change ($\Delta Sent$).

Moreover, two quality metrics are computed using held-out dialogues, namely:

- Transition Coverage (EC) [15] evaluates the proportion of transitions in test dialogues that are represented in the flow.
- Flow F1-Score ($FF1$) [25] balances flow complexity and generalization, combining the number of states with the similarity of dialogue paths (measured via Levenshtein distance) between test dialogues and the flow.

C. EXPECTED INSIGHTS AND HYPOTHESES

UnHIDE is designed to surface interpretable patterns in dialogue datasets by combining flow discovery with a diverse set of structural and behavioral metrics. We aim to validate whether these metrics meaningfully reflect key dialogue properties, and to explore the type of insights that UnHIDE can provide to both researchers and practitioners. To this end, we formulate the following high-level hypotheses:

H1. **Dialogue volume and pacing**: Basic dialogue characteristics, such as the number of utterances and the pacing of responses, should be reflected in global metrics like total utterances ($|U|$) and average dialogue duration ($\Delta T/|D|$). These serve as foundational indicators of interaction length and intensity.

H2. **Flow complexity**: Variations in dialogue length and structure are expected to impact flow complexity. This should be observable through metrics such as the number of discovered states ($|C|$), transitions ($|E|$), and flow density (FD). Simpler dialogues may yield more linear flows, while richer conversations may produce more branched and diverse patterns.

H3. **Flow quality**: The coherence and generalizability of the discovered flow should be influenced by dialogue diversity and complexity. We anticipate this will be reflected in quality metrics such as the Silhouette Score (SS), Transition Coverage (EC), and Flow F1-Score ($FF1$), which together assess how well the flow captures conversational structure across instances.

H4. **Sentiment dynamics**: Sentiment should manifest both visually—through color-coded transitions in the flow—and quantitatively, via metrics capturing emotional trajectory (e.g., $Sent_{SOD}$, $Sent_{EOD}$, EOD_-, $\Delta Sent$). These indicators are expected to reveal trends such as sentiment improvement or escalation over time.

In real-world dialogues, response time may also correlate with sentiment, e.g., longer pauses could signal confusion or dissatisfaction. While this relationship is beyond the scope of the current setup, it represents a promising direction for applying UnHIDE to naturally occurring data.

IV. EXPERIMENTS

To evaluate the effectiveness of UnHIDE, we conducted experiments on a suite of synthetically generated dialogues designed with controlled variations in key conversational attributes. These variables include dialogue length, sentiment progression, and response timing factors that are later analyzed using UnHIDE's metrics and flow representations.

This section describes the experimental design, including the dialogue generation process, configuration of controlled variables, and the computational environment used.

A. EXPERIMENTAL SETTINGS

For flow discovery, all utterances were first embedded into 384-dimensional vectors using the sentence transformer model *all-MiniLM-L6-v2*.[1] Clustering was performed using the K-means algorithm from the scikit-learn library.[2]

[1] hf.co/sentence-transformers/all-MiniLM-L6-v2
[2] https://tinyurl.com/4ymet8ff

To ensure meaningful and interpretable flows, the number of clusters for each participant (agent and user) was optimized to maximize the Silhouette Score [27]. This was done using the Optuna hyperparameter optimization framework,[3] which explored cluster counts in the range of 3 to 10. This range was chosen to balance expressiveness and interpretability of the resulting dialogue flow. Transitions between clusters were computed from dialogue sequences, and a threshold $\theta = 0.1$ was applied across all experiments to remove low-probability transitions. This pruning step improves the readability of the flow graphs by filtering out infrequent paths.

Cluster labels were generated using a quantized Llama3–8B model running in Ollama.[4] For each cluster, the model was prompted to summarize the dominant intent or action from a random sample of 30 utterances. The labeling prompt template is shown in Figure 2.

```
Please provide a label that captures the main
actions of the following sentences in English:
{input}. Answer with the label only in English,
using the format label: {your_label}, and with a
maximum of three words.
```

FIGURE 2. Prompt template for generating labels for dialogue states, where the `{input}` variable is to be replaced by a random sample of 30 clustered utterances.

The same model was used to classify the sentiment polarity of each utterance as positive, negative, or neutral. The sentiment classification prompt template is shown in Figure 3. For both prompting tasks, the temperature was set to 0.1 to promote consistency and minimize variance in model responses. All experiments were conducted on a system equipped with an NVIDIA RTX A6000 GPU (48GB VRAM).

```
You are an assistant specialized in analyzing
sentiment in customer service dialogues.
Classify the sentiment of the utterance below
according to the following definitions:
Negative: the utterance expresses complaints,
frustration, dissatisfaction, or something with a
negative context.
Neutral: the utterance is neutral, factual, or
does not convey a clear positive or negative
sentiment.
Positive: the utterance conveys satisfaction,
gratitude, or something positive.
{utterance}
Only respond with the corresponding word
(Negative, Neutral, or Positive).
```

FIGURE 3. Prompt template for sentiment classification of each dialogue utterance, where `{utterance}` is to be replaced by the utterance to classify.

B. SYNTHETIC DIALOGUE GENERATION

To meaningfully evaluate the interpretability and diagnostic power of UnHIDE, we require dialogue data where specific conversational attributes are both systematically varied and precisely known.

```
Generate a new dialogue based on the structure and
style of the original dialogue provided below. The
new dialogue should meet the following criteria:
- The sentiment of the dialogue {sentiment}.
- The dialogue must have {number_utterances}.
Original dialogue: {original_dialogue}.
Each utterance of the dialogue must start with
'USER:' or 'SYSTEM:'.
Just put the dialogue without any text before or
after the dialogue.
```

FIGURE 4. Prompt template for dialogue generation, with variables: `{sentiment}`, with possible values *positive*, *negative*, *positive to negative*, *negative to positive*; `{number_utterances}`, *at most 10 utterances*, *at least 18 utterances*; `{original_dialogue}`, obtained from MultiWOZ.

To this end, we introduce MultiSynthiment, a synthetically generated, task-oriented dialogue dataset designed with specific subsets that target key variables analyzed by UnHIDE, namely: (i) sentiment progression (i.e., how sentiment evolves throughout the dialogue), (ii) number of utterances, and (iii) response time. Rather than generating dialogues entirely from scratch, MultiSynthiment was built by aligning synthetic content with MultiWOZ 2.2 [28], a widely used benchmark of task-oriented dialogues without sentiment or temporal dynamics. For each MultiWOZ dialogue, LLaMA3 was prompted to generate a new dialogue based on the original (see Figure 4 for the prompt template), while incorporating a predefined sentiment progression and a target number of utterances. The dialogue subsets can be reliably recreated using the same prompt template and controlled variable settings, ensuring reproducibility and enabling comparable analyses across subsets, despite slight variations in individual dialogues due to the generative nature of LLMs.

To ensure both diversity and fidelity to the underlying dialogue structure, the generation temperature was set to 0.7. A summary of the 16 dialogue subsets with their respective generation variables is in Table 2.

TABLE 2. Synthetic dialogue subsets (S1–S16), defined by combinations of generation variables: sentiment progression (static positive or negative, or a transition between the two), response time (fast or slow), and utterance count (few or many).

Subset	Sentiment	Resp. Time	Utterances
S1	Pos	Fast	Few
S2	Neg	Fast	Few
S3	Pos→Neg	Fast	Few
S4	Neg→Pos	Fast	Few
S5	Pos	Slow	Few
S6	Neg	Slow	Few
S7	Pos→Neg	Slow	Few
S8	Neg→Pos	Slow	Few
S9	Pos	Fast	Many
S10	Neg	Fast	Many
S11	Pos→Neg	Fast	Many
S12	Neg→Pos	Fast	Many
S13	Pos	Slow	Many
S14	Neg	Slow	Many
S15	Pos→Neg	Slow	Many
S16	Neg→Pos	Slow	Many

To construct a diverse and balanced dataset, we systematically combined: four types of sentiment progression, two

[3]https://optuna.org
[4]https://ollama.com

levels of utterance count, and two response time conditions, resulting in a total of 16 dialogue subsets ($4 \times 2 \times 2$). Since the original MultiWOZ dataset contains 8,436 training and 1,000 test dialogues, quantities not evenly divisible by 16, we added 12 dialogues to the training set and 8 to the test set to ensure balanced subset sizes. This results in 528 dialogues per subset in the training split and 63 dialogues per subset in the test split. The final distribution of dialogues across subsets for both the original and synthetic datasets is shown in Table 3.

TABLE 3. Distribution of dialogues ($|D|$) and utterances ($|U|$) across the training and test splits of the MultiWOZ 2.2 and MultiSynthiment.

Dataset	Split	|U|	|D|
MultiWOZ 2.2	Train	113,552	8,436
	Test	14,748	1,000
MultiSynthiment	Train	140,162	8,448
	Test	16,098	1,008

To define dialogue length categories, we analyzed the distribution of utterance counts in the original MultiWOZ 2.2 dataset. Dialogues in the first quartile (with at most 10 utterances) served as references for generating short dialogues, while those in the third quartile (with at least 18 utterances) guided the generation of long dialogues. This approach enabled a clear separation between the "few" and "many" utterance categories, while maintaining alignment with the original train-test split of MultiWOZ.

After generation, timestamps were assigned to each utterance to simulate either fast or slow response times, reflecting different interaction paces between the user and the system. Response intervals were randomly sampled from normal distributions: (i) **Fast responses** were drawn from a distribution with a mean of 10 seconds and a standard deviation of 5 seconds; (ii) **Slow responses** used a mean of 120 seconds with a standard deviation of 60 seconds.

This setup introduces realistic temporal variability, enabling analysis of how dialogue pace influences flow dynamics and metrics. To illustrate the diversity of the generated dialogues, Table 4 shows examples from subsets S5 and S6, while Table 5 presents an example from subset S12. In each case, the original dialogue from the MultiWOZ dataset and its corresponding generated version from MultiSynthiment are included, highlighting controlled variations in sentiment, pacing, and number of utterances. These examples showcase different patterns: (i) a dialogue with predominantly positive sentiment, slower response times, and few utterances; (ii) a dialogue with negative sentiment, also with slow response times and few utterances; and (iii) a dialogue that shifts from negative to positive sentiment, with faster response times and a larger number of utterances.

V. RESULTS AND ANALYSIS

The validation of UnHIDE is conducted through its application to the MultiSynthiment, focusing on how effectively the generated variables are reflected in the computed metrics.

A. DIALOGUE METRICS

Table 6 presents the computed dialogue-level metrics for each subset. As expected, there is a clear distinction in the number of utterances ($|U|$), with subsets S1–S8 containing fewer utterances and S9–S16 containing more. This variation is also reflected in the average dialogue duration ($\Delta T/|D|$), which shows an inverse correlation with response time, confirming that faster responses result in shorter dialogues, as hypothesized in H1.

Table 7 reports the global clustering and flow-level metrics for each dialogue subset. While certain trends can be observed directly from the table, a more comprehensive understanding is gained by examining broader patterns. To this end, we analyze the correlation heatmap between generation variables and computed metrics (Figure 5) and visualize selected metrics across all subsets to highlight key differences (Figure 6).

This suggests that dialogue length does not influence flow complexity, contrary to our expectation in Hypothesis H2. Instead, its primary effect is on the average number of utterances per state ($|U|/|C|$), indicating denser states in longer dialogues. A potential explanation lies in the clustering constraints imposed during flow discovery. Specifically, the number of clusters was capped at 10 per speaker during Silhouette-based optimization. Although the theoretical maximum number of states (22, including SOD and EOD) is not reached in practice, this upper bound may limit flow complexity. Furthermore, the transition pruning threshold ($\theta = 0.1$) may suppress additional structure by removing low-probability edges, effectively eliminating transitions to states that were rarely or never reached.

Despite the clustering optimization process, the Silhouette Score (SS) remains consistently low across subsets. This may be attributed to the low-dimensional sentence embeddings (384 dimensions), which may not fully capture the nuances required for precise utterance grouping. Another likely factor is the upper bound placed on the number of clusters during optimization. While intended to preserve interpretability, this constraint may have limited the model's ability to separate diverse interaction types. In many cases, utterances with distinct semantics may have been grouped into the same cluster, leading to larger, less coherent states and ultimately weaker separation—reflected in the persistently low SS values.

In contrast, the quality of the discovered transitions is noticeably stronger. Although Transition Coverage (EC) appears relatively insensitive to the generation variables, we observe a strong negative correlation between utterance count and Flow F1-Score ($FF1$), supporting Hypothesis H3. As dialogues grow longer and more structurally complex, it becomes increasingly difficult for the flow to capture all observed paths. This results in reduced path overlap between the test dialogues and the generated flow and increasing the trade-off penalty between complexity and coverage embedded in the $FF1$ metric. Although most transitions are included (high EC), rare sequences of utterances in complex

TABLE 4. MultiWOZ S5 and S6 dialogues and their MultiSynthiment generated counterparts, showcasing controlled variation in sentiment, pacing, and length.

Speaker	Dialogue	Sent.	ΔT_{SOD}
Original dialogue example from subset S5			
User	Can you help me find a train going to Cambridge leaving on Tuesday?	-	-
System	Where are you departing from and what is your departure time?	-	-
User	I'll be departing Bishops Stortford. My departure time is flexible, but I need to arrive by 16:30.	-	-
System	The Bishops Stortford to Cambridge at 15:29 is scheduled to arrive at 16:07. The one before that is two hours earlier.	-	-
User	Could you give me the price and train ID for that one please?	-	-
System	TR1392. 10.10 pounds. Is there anything else?	-	-
User	I'm looking for places to go in town centre. I'm interested in entertainment. Can you suggest a place to go?	-	-
System	There are many attractions do you want a church,or park or water slide ?	-	-
Generated dialogue with positive sentiment, slow response time, and a few number of utterances (S5)			
User	Hi! I'm excited to find a train to Cambridge for Tuesday. Can you help me?	Pos	0.00
System	Absolutely, I'd love to help! Where will you be departing from and what time works best for you?	Pos	180
User	I'm leaving from Bishops Stortford. I'm flexible but want to get there by 16:30.	Pos	300.88
System	Perfect! There's a train at 17:30 that arrives at 18:29.	Pos	391.04
User	That sounds ideal! Can you tell me the price and train ID?	Pos	557.57
System	Sure! It's train TR3213, and the ticket costs £10.10.	Pos	729.94
User	Great! Also, could you recommend some fun spots to check out in Cambridge town center?	Pos	909.95
System	Absolutely! You'll find lovely parks, charming churches, and great entertainment venues all nearby.	Pos	1089.94
Original dialogue example from subset S6			
User	I want to book a train. I would like to depart from cambridge and arrive by 16:45	-	-
System	What is your destination?	-	-
User	The train should depart from cambridge and should arrive by 16:45. Going to Ely	-	-
System	What day will you be travelling?	-	-
User	I would like to go on Saturday.	-	-
System	How about TR8860?	-	-
User	Sure, I would like to book that train for 4 passengers.	-	-
System	Booking was successful, the total fee is 14.08 GBP payable at the station. Reference number is : VX48ZGH0.	-	-
User	That's wonderful. I am all set. Thank you. Goodbye.	-	-
System	You're very welcome! Goodbye.	-	-
Generated dialogue with negative sentiment, slow response time, and a few number of utterances (S6)			
User	I need to cancel my train ticket.	Neg	0.00
System	Sorry to hear that, can you please provide your reference number?	Neg	139.10
User	It's... um... VX48ZGH0. Yeah, that's it.	Neut	280.09
System	Okay, I've located your booking. What seems to be the issue?	Neut	460.10
User	The train is no longer suitable for us. We need a different one.	Neg	580.81
System	Alright, let me see what else is available. Unfortunately, there are no trains departing from Cambridge on Saturday.	Neg	718.62
User	That's frustrating. Can you suggest an alternative route or schedule?	Neg	807.63
System	Not really. You might want to consider driving instead.	Neut	987.63
User	Ugh, that's not ideal. Is there anything else we can do?	Neg	1129.69
System	Not at the moment. Sorry for the inconvenience.	Neg	1309.70

dialogues are not fully captured, thus reflecting lower generalizability. Notably, this is quantified by a correlation of -0.89 between utterance count and $FF1$, indicating a sharp decline in flow alignment quality as dialogue length increases.

We also observe a positive correlation between flow density (FD) and transition coverage (EC). This relationship arises because denser flow graphs contain more transitions per state, thereby increasing the likelihood that transitions observed in unseen test dialogues are already represented in the flow. As a result, higher density contributes to broader structural coverage, even if it does not necessarily imply better clustering quality.

Regarding sentiment, results are consistent with expectations outlined in Hypothesis H4. Dialogues labeled as positive or negative exhibit a corresponding dominance of positive

TABLE 5. MultiWOZ S12 dialogue and its MultiSynthiment generated counterpart, showcasing controlled variation in sentiment, pacing, and length.

Speaker	Dialogue	Sent.	ΔT_{SOD}
Original dialogue example from subset S12			
User	I'm in the mood for some modern eclectic food so please find me a cheap restaurant that sells some	-	-
System	I'm sorry, I wasn't able to find any of those.	-	-
User	How about Indian, instead?	-	-
System	There are four cheap Indian restaurants in Cambridge. What area of town were you interested in?	-	-
User	I want one on the west side of town. If there's not one there then the centre of town will be fine.	-	-
System	We don't have any on the westside, but three in the city centre. Might I suggest the Kohinoor?	-	-
User	That sounds great! Could you book a table there for 12:00 on Saturday? Just me.	-	-
System	Booking was successful. The table will be reserved for 15 minutes. Reference number is : VSJ2W6N3 . Can I help you with anything else?	-	-
User	I am also looking for a place to stay.	-	-
System	What area would you like it to be in?	-	-
User	The area doesn't matter.	-	-
System	Well do you have a price preference?	-	-
User	I would prefer a guesthouse and free wifi and parking.	-	-
System	Do you have a price range or star rating in mind?	-	-
User	I don't care, book it for the same group of people and 2 nights starting from the same day.	-	-
System	Your booking number is GC8V34OD for the Acorn guest house at 154 chesterton road for 1 person on Saturday for 2 days is set.	-	-
User	Thank you, that is all that I need.	-	-
System	Thank you for using our services. Have a great stay!	-	-
Generated dialogue with negative to positive sentiment, fast response time, and many utterances (S12)			
User	I'm desperate to find a cheap restaurant that serves modern food, but I can't find anything.	Neg	0.00
System	Sorry, I couldn't find any restaurants like that right now.	Neg	8.26
User	This is frustrating... What about Indian food instead?	Neg	10.75
System	There are four affordable Indian restaurants in Cambridge. Which area do you prefer?	Neut	25.75
User	I'd like one on the west side, but if not, the city center will do.	Neut	39.02
System	We don't have any on the west side, but there are three in the city center. How about "Spice House"?	Neut	45.65
User	That sounds good! Can you book a table for Saturday at 7 PM, just for me?	Pos	53.59
System	Booking confirmed! The table will be held for 15 minutes. Reference code: SHB12345. Need anything else?	Pos	56.15
User	I'm also trying to book a place to stay.	Neut	66.20
System	Do you have a preferred area?	Neut	73.03
User	Location doesn't matter much.	Neut	88.03
System	Any preferences on price or amenities?	Neut	96.93
User	I'd like a guesthouse with free Wi-Fi and parking.	Neut	101.31
System	Any price range or star rating in mind?	Neut	104.88
User	Not really, just booking for me and two friends, for two nights starting Saturday. It's going to be a lot of fun!	Pos	119.88
System	Reservation confirmed at Greenleaf Guesthouse, 78 Baker Street, for 3 people, two nights from Saturday. Booking code: GG5678.	Pos	122.95
User	Great! Thanks so much for your help.	Pos	134.56
System	You're welcome! Hope you have a wonderful stay. Let me know if you need anything else!	Pos	136.50

TABLE 6. Dialogue metrics computed for each subset, including: number of utterances ($|U|$), average number of utterances per dialogue ($|U|/|D|$), average dialogue duration ($\Delta T/|D|$), proportion of utterances with positive ($S^+/|D|$) and negative ($S^-/|D|$) sentiment per dialogue, and average utterances per discovered state ($|U|/|C|$).

Subset	|U|	$\|U\|/\|D\|$	$\Delta T/\|D\|$	$S^+/\|D\|$	$S^-/\|D\|$	$\|U\|/\|C\|$
S1	4,909	9.32 ± 1.2	79.21 ± 17.2	0.54	0.00	272.72 ± 70.2
S2	5,145	9.81 ± 0.6	83.70 ± 13.6	0.13	0.30	285.83 ± 113.3
S3	5,076	9.67 ± 0.7	83.37 ± 14.3	0.49	0.32	282.00 ± 103.6
S4	5,000	9.72 ± 0.7	82.63 ± 17.5	0.44	0.34	333.33 ± 144.9
S5	4,998	9.47 ± 1.0	977.70 ± 191.6	0.52	0.00	294.00 ± 183.3
S6	5,085	9.76 ± 0.7	$1{,}005.15 \pm 185.6$	0.13	0.29	635.00 ± 280.9
S7	5,022	9.68 ± 0.7	986.05 ± 191.3	0.48	0.33	456.55 ± 199.3
S8	4,977	9.70 ± 0.7	977.19 ± 220.6	0.44	0.34	355.52 ± 189.7
S9	11,727	22.83 ± 5.8	206.14 ± 48.4	0.53	0.00	781.80 ± 414.3
S10	12,875	25.65 ± 7.1	226.88 ± 49.6	0.16	0.24	858.33 ± 459.3
S11	13,976	27.70 ± 7.4	247.69 ± 54.2	0.44	0.17	735.58 ± 269.7
S12	11,138	22.92 ± 6.8	194.95 ± 42.8	0.40	0.21	795.57 ± 242.2
S13	11,699	23.39 ± 6.1	$2{,}458.81 \pm 534.1$	0.53	0.00	835.64 ± 315.2
S14	13,408	26.92 ± 7.6	$2{,}830.63 \pm 688.1$	0.18	0.27	957.71 ± 370.8
S15	13,711	27.66 ± 7.9	$2{,}918.13 \pm 621.9$	0.42	0.20	761.72 ± 248.9
S16	11,416	23.35 ± 6.5	$2{,}377.37 \pm 467.8$	0.41	0.16	761.07 ± 243.0

($S^+/|D|$) or negative ($S^-/|D|$) utterances, respectively. Interestingly, dialogues with sentiment transitions tend to show a higher proportion of positive utterances overall. This may reflect an underlying bias in the MultiWOZ dataset, where

TABLE 7. Clustering and flow metrics per subset. Clustering metrics include: number of dialogue states ($|C|$), Silhouette Score (SS_{usr}, SS_{sys}) and sentiment cohesion (FSC_{usr}, FSC_{sys}), both for user and system utterances. Flow metrics include: number of transitions between states ($|E|$), average sentiment of SOD states ($Sent_{SOD}$), average sentiment of EOD ($Sent_{EOD}$), sentiment variation across dialogue($\Delta Sent$), proportion of negative sentiment in EOD states ($EOD\_$), flow density (FD), transition coverage (EC), and flow F1-score ($FF1$).

, proportion of negative sentiment in EOD states ($EOD\_$), flow density (FD), transition coverage (EC), and flow F1-score ($FF1$).

Subset	Clustering Metrics					Flow Metrics							
	$\|C\|$	SS_{usr}	SS_{sys}	FSC_{usr}	FSC_{sys}	$\|E\|$	$Sent_{SOD}$	$Sent_{EOD}$	$\Delta Sent$	$EOD\_$	FD	EC	FF1
S1	18	0.06	0.06	0.22	0.23	54	0.75	0.75	0.00	0.00	0.14	0.70	0.46
S2	18	0.04	0.05	0.29	0.31	67	0.41	0.43	0.02	0.42	0.18	0.80	0.57
S3	18	0.05	0.05	0.37	0.40	52	0.84	0.42	-0.42	0.78	0.14	0.69	0.58
S4	15	0.06	0.07	0.38	0.33	51	0.29	0.86	0.57	0.00	0.19	0.83	0.57
S5	17	0.06	0.06	0.24	0.23	62	0.76	0.76	0.00	0.00	0.18	0.74	0.66
S6	8	0.05	0.06	0.28	0.32	27	0.43	0.38	-0.05	0.15	0.30	0.93	0.58
S7	11	0.05	0.05	0.38	0.40	37	0.82	0.40	-0.42	0.58	0.24	0.85	0.58
S8	14	0.06	0.07	0.34	0.32	41	0.12	0.78	0.66	0.00	0.17	0.82	0.58
S9	14	0.06	0.07	0.25	0.23	39	0.77	0.96	0.19	0.00	0.16	0.75	0.24
S10	14	0.05	0.06	0.29	0.30	45	0.42	0.62	0.20	0.68	0.19	0.74	0.43
S11	19	0.05	0.05	0.31	0.34	49	0.79	0.40	-0.39	0.37	0.12	0.65	0.40
S12	14	0.06	0.08	0.34	0.30	44	0.39	0.98	0.59	0.00	0.18	0.81	0.38
S13	14	0.05	0.05	0.24	0.22	50	0.78	0.81	0.03	0.00	0.21	0.83	0.27
S14	14	0.05	0.06	0.30	0.32	51	0.43	0.68	0.25	0.00	0.21	0.85	0.41
S15	18	0.05	0.05	0.33	0.35	51	0.76	0.40	-0.36	0.44	0.13	0.64	0.29
S16	15	0.05	0.08	0.32	0.27	50	0.42	0.99	0.57	0.00	0.18	0.81	0.41

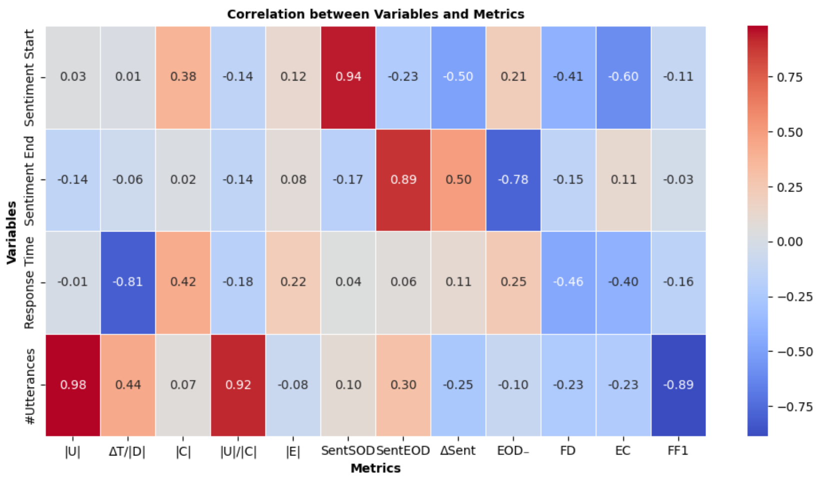

FIGURE 5. Heatmap with correlations between variables and UnHIDE metrics.

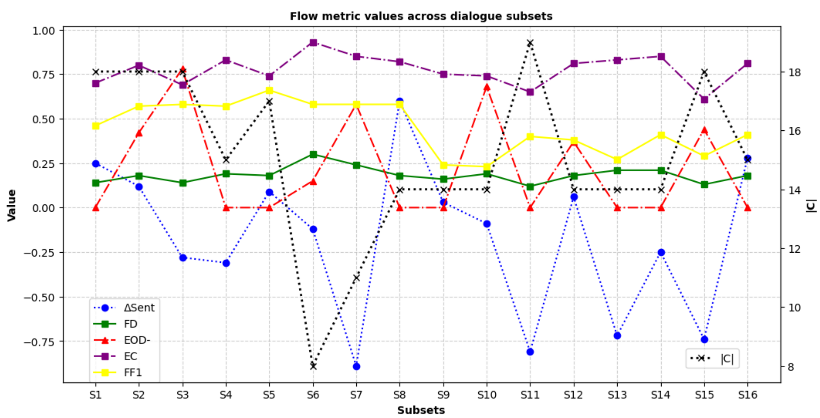

FIGURE 6. Flow metric values across dialogue subsets. All metrics are plotted on the left axis, except for the number of discovered states ($|C|$), which is shown on the right axis for clarity.

negative utterances are comparatively scarce. EmoWOZ [29], built on MultiWOZ with emotion annotations, shows a similar scarcity of negative utterances due to its task-oriented nature. To systematically study sentiment and other dialogue

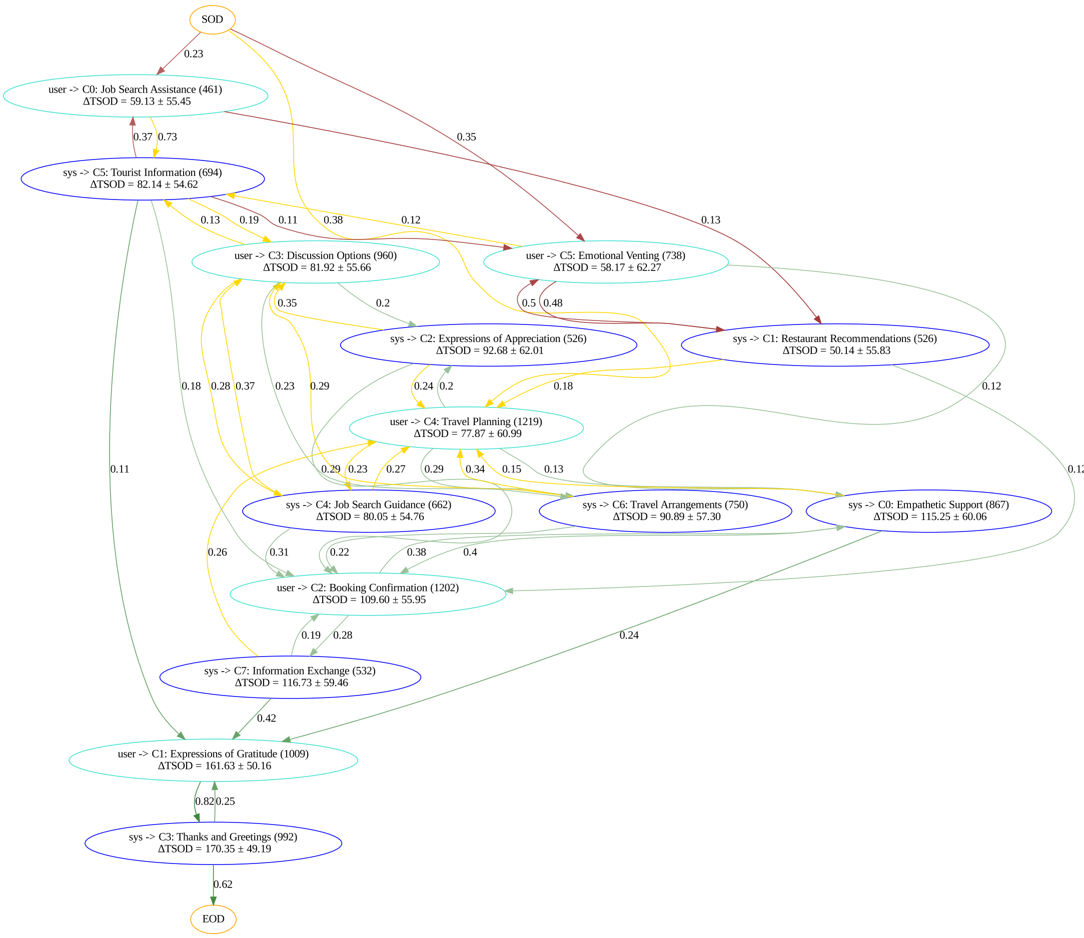

FIGURE 7. Dialogue flow discovered for Subset 12.

TABLE 8. Clustering metrics for Subset 12: number of utterances (U_c), tokens per utterance ($|W|/|U|$), time since dialogue start (ΔT_{SOD}), time since previous utterance (ΔT_{prev}), and average sentiment ($Sent_c$).

	U_c	$\|W\|/\|U\|$	ΔT_{SOD}	ΔT_{prev}	$Sent_c$
User					
C0: Job Search Assistance	461	16.19 ± 5.0	59.13 ± 55.5	7.19 ± 5.6	0.37 ± 0.4
C1: Expressions of Gratitude	1009	9.74 ± 3.9	161.63 ± 50.2	9.90 ± 4.2	0.90 ± 0.2
C2: Booking Confirmation	1202	11.62 ± 5.4	109.60 ± 55.9	9.81 ± 4.2	0.63 ± 0.3
C3: Discussion Options	960	12.65 ± 4.3	81.92 ± 55.7	9.75 ± 4.2	0.53 ± 0.4
C4: Travel Planning	1219	15.42 ± 4.7	77.87 ± 60.9	8.03 ± 5.2	0.52 ± 0.4
C5: Emotional Venting	738	14.31 ± 4.9	58.17 ± 62.3	6.98 ± 5.4	0.28 ± 0.4
System					
C0: Empathetic Support	867	20.07 ± 8.0	115.25 ± 60.1	9.67 ± 4.0	0.64 ± 0.3
C1: Restaurant Recommendations	526	18.88 ± 7.9	50.14 ± 55.8	9.31 ± 4.2	0.32 ± 0.2
C2: Expressions of Appreciation	526	21.97 ± 7.0	92.68 ± 62.0	9.56 ± 4.1	0.63 ± 0.4
C3: Thanks and Greetings	992	16.98 ± 7.5	170.35 ± 49.2	9.66 ± 4.1	0.98 ± 0.1
C4: Job Search Guidance	662	22.11 ± 8.0	80.05 ± 54.8	9.77 ± 4.1	0.49 ± 0.3
C5: Tourist Information	694	23.36 ± 8.1	82.14 ± 54.6	9.75 ± 4.2	0.55 ± 0.4
C6: Travel Arrangements	750	22.23 ± 7.3	90.89 ± 57.3	9.68 ± 4.1	0.62 ± 0.3
C7: Information Exchange	532	19.43 ± 7.1	116.73 ± 59.5	9.68 ± 4.0	0.60 ± 0.3

variables, synthetic datasets were used, allowing controlled analysis of their impact on the metrics.

To facilitate clearer analysis, Figure 5 splits sentiment progression into two independent variables: Sentiment Start and Sentiment End, computed respectively as the average sentiment of utterances in the initial and final dialogue clusters. As expected, Sentiment Start shows a strong correlation with the initial sentiment metric ($Sent_{SOD}$), while Sentiment End correlates with both the final sentiment ($Sent_{EOD}$) and the proportion of negative sentiment in final states ($EOD\_$). Both start and end sentiment values exhibit moderate correlation with overall sentiment change ($\Delta Sent$), further validating the interpretability of the derived metrics.

In addition, changes in sentiment across dialogues are reflected in flow sentiment cohesion (FSC), which captures the internal consistency of sentiment within discovered dialogue states.

B. CLUSTERING AND FLOW METRICS

To illustrate state-level clustering characteristics, we analyze the flow discovered for subset S12, a group of dialogues defined by a negative-to-positive sentiment shift, fast response time, and many utterances. The corresponding flow graph is shown in Figure 7, and detailed per-state metrics are reported in Table 8. Sentiment values range from 0 (indicating more negative sentiment) to 1 (indicating more positive sentiment). In subset S12, as expected, sentiment progresses from negative to positive—with an initial sentiment of $Sent_{SOD} = 0.39$ and a final sentiment of $Sent_{EOD} = 0.98$. This progression is also visually evident in the flow diagram (Figure7), where transition colors shift from red to green, reflecting the underlying change in emotional tone throughout the dialogue.

Approximately a quarter of the dialogues begin with the user expressing dissatisfaction with their current job, a situation typically labeled as *Job Search Assistance*. In about a third of the cases (35%), the dialogues open with more general expressions of negative sentiment, often reflecting frustration or emotional distress—categorized as *Emotional Venting*.

This pattern is illustrated in the second generated dialogue of Table 5, where the user begins by expressing frustration about being unable to find suitable restaurants. In 48% of similar cases, the system responds with a *Restaurant Recommendation*, aiming to improve sentiment by offering a practical solution. As the dialogue progresses, a typical closing sequence emerges: the user proceeds with a *Booking Confirmation*, the system provides *Empathetic Support*, the user offers *Expressions of Gratitude*, and the system concludes with *Thanks and Greetings*. This complete trajectory, from initial frustration to resolution and mutual appreciation, is a recurring flow observed in multiple dialogues, including the example in Table 5.

The user states C2 and C4 exhibit a higher number of utterances ($|U_c|$), as they correspond to frequently occurring actions in the source data, specifically, booking confirmations and travel planning. In contrast, user state C1 and system state C3 contain fewer tokens per utterance ($|W|/|U|$), as they represent brief interactions such as expressions of gratitude, which typically occur at the end of the dialogue and are marked by high values of ΔT_{SOD}. Notably, these final states also have the highest sentiment scores, reflecting the consistently positive tone of concluding interactions.

VI. CONCLUSION AND FUTURE WORK

UnHIDE is an unsupervised framework for dialogue analysis that facilitates human interpretation of large-scale conversational data. It operates in two stages: first, by discovering dialogue flows through utterance clustering and graph-based flow construction; and second, by computing a comprehensive set of interpretable metrics that reflect key properties of the dialogues at different levels, utterance, transition, state, and graph. This dual focus allows UnHIDE to reveal structural and emotional patterns in dialogue data without the need for supervision or manual labeling.

To evaluate the framework's capabilities, we applied it to MultiSynthiment,[5] a novel, synthetically generated dataset inspired by MultiWOZ. Dialogues were systematically varied on three dimensions: sentiment progression, response time, and number of utterances. These variables served as controlled factors for evaluating whether and how UnHIDE captures relevant dialogue properties through flow structures and metric outputs.

Our evaluation was guided by four hypotheses. First, we expected that basic dialogue characteristics, such as total utterances and response pacing, would be reflected in high-level metrics like $|U|$ and $\Delta T/|D|$, providing foundational insight into dialogue length and intensity. Second, we hypothesized that dialogue structure would impact flow complexity, with simpler dialogues yielding more linear flows and longer ones resulting in richer, more branched graphs, observable through metrics like the number of states ($|C|$), transitions ($|E|$), and graph density (FD). Third, we anticipated that the quality and coherence of the discovered flows would vary with dialogue diversity, and that metrics such as Silhouette Score (SS), Transition Coverage (EC), and Flow F1-Score ($FF1$) would reflect this relationship. Finally, we expected that sentiment dynamics would be evident both visually, through color-coded transitions, and quantitatively, through trajectory metrics such as $Sent_{SOD}$, $Sent_{EOD}$, $EOD\_$, and $\Delta Sent$, helping to characterize emotional progression within conversations.

The results largely confirmed these hypotheses. UnHIDE successfully surfaced the variation in dialogue length, pacing, and sentiment progression through both its flow representations and computed metrics. However, some limitations emerged. Flow complexity did not always scale with dialogue length, mainly due to constraints on the maximum number of clusters and the pruning of low-probability transitions. Without this pruning, the resulting flows became cluttered and difficult to interpret. This trade-off, while improving interpretability and ensuring comparability across subsets, also introduced limits on flow expressiveness. Similarly, while transition-level quality metrics such as EC and $FF1$ were informative, their sensitivity to dialogue length highlighted challenges in generalizing across more complex

[5]The source code and synthetic dataset are available at: https://github.com/NLP-CISUC/FlowDisco/tree/main

interactions. Sentiment trends were robustly captured and closely aligned with the controlled generation settings, further supporting the framework's analytical value.

Despite its reliance on synthetic data, this study provides strong evidence that UnHIDE can uncover meaningful insights from dialogue flows and metrics. The framework's components, clustering, flow construction, and metric computation, are computationally efficient and can be run locally. Moreover, LLM-based steps, like labeling and sentiment classification, can be substituted with lighter alternatives where necessary.

In addition to integrating UnHIDE in the analysis of real-world dialogues, in the future, we aim to simulate more diverse conversational scenarios, including diverse user personas and LLM temperature adjustments, to increase generative diversity. We also plan to expand the metric suite to capture other dialogue qualities such as engagement or politeness. Finally, we envision applying UnHIDE not only to analyze dialogue, but also to support the interpretability of LLMs themselves, offering a broader contribution to model transparency and responsible AI deployment.

LIMITATIONS

The main limitations of this work arise from the use of synthetic data. Although synthetic data provided an efficient and controlled environment for validating UnHIDE, it may not fully reflect the variability and nuance of real-world conversations. This choice was partially motivated by the fact that the real dialogue data available to us is proprietary and cannot be publicly released. As an alternative, generation via prompting allowed us to vary specific dialogue attributes systematically, such as sentiment, utterance count, and response time, offering valuable ground truth for evaluating the framework.

Additionally, while the analysis includes temporal features, that is, response times, these were also synthetically assigned and are not grounded in actual human timing behaviors. As such, their representativeness in practical settings may be limited. Nevertheless, the structure-oriented evaluation based on clustering and flow metrics still offers a meaningful lens into the framework's interpretability and performance.

Some limitations are also related to specific implementation choices. Utterances were embedded using generic sentence transformers, which may not be optimal for capturing dialogue-specific nuances, like turns, context, or intentions. The number of clusters was selected using the Silhouette Score, with an upper limit imposed to maintain interpretability, potentially constraining flow complexity. Furthermore, LLMs (Llama3-8B) were used in a zero-shot setting to assign cluster labels and classify sentiment. These models are known to be susceptible to inconsistency and hallucination, and their outputs may not always align with human judgment. Still, we found that dominant sentiment trends were consistently captured and that the resulting flows remained interpretable, which we view as the central success of this validation.

Lastly, we emphasize that the core components of UnHIDE, including utterance clustering, flow discovery, and metric computation, are fully self-contained and can be run locally without specialized hardware. Additionally, the reliance on LLMs for cluster labeling and sentiment analysis is optional: these components can be replaced by simpler, more traditional methods such as keyword-based approaches or lightweight classifiers, depending on the application context.

ACKNOWLEDGMENT

This work was also supported by national funds through FCT - Fundação para a Ciência e a Tecnologia, I.P., within the projects UIDB/00326/2025 and UIDP/00326/2025. Parts of this paper were reviewed with the assistance of tools that make use of LLMs, namely ChatGPT and Grammarly, but only for language corrections. The overall content and ideas remain solely the responsibility of the authors.

REFERENCES

[1] Y. Kim, Y. Hwang, H. Bae, T. Kang, and K. Jung, "Flowlogue: A novel framework for synthetic dialogue generation with structured flow from text passages," *IEEE Access*, vol. 12, pp. 151920–151929, 2024.

[2] L. Vanel, "Natural language generation for socially competent task-oriented agent," in *Proc. 11th Int. Conf. Affect. Comput. Intell. Interact. Workshops Demos (ACIIW)*, Sep. 2023, pp. 1–5.

[3] J. Paladines and J. Ramirez, "A systematic literature review of intelligent tutoring systems with dialogue in natural language," *IEEE Access*, vol. 8, pp. 164246–164267, 2020.

[4] L. Wang, C. Ma, X. Feng, Z. Zhang, H. Yang, J. Zhang, Z. Chen, J. Tang, X. Chen, Y. Lin, W. X. Zhao, Z. Wei, and J. Wen, "A survey on large language model based autonomous agents," *Frontiers Comput. Sci.*, vol. 18, no. 6, pp. 1–26, Dec. 2024.

[5] N. Chondamrongkul, G. Hristov, and P. Temdee, "Addressing technical challenges in large language model-driven educational software system," *IEEE Access*, vol. 13, pp. 12846–12858, 2025.

[6] L. Ma, M. Li, W.-N. Zhang, J. Li, and T. Liu, "Unstructured text enhanced open-domain dialogue system: A systematic survey," *ACM Trans. Inf. Syst.*, vol. 40, no. 1, pp. 1–44, Sep. 2021, doi: 10.1145/3464377.

[7] Y. Guo, "Design of improved artificial intelligence generative dialogue algorithm and dialogue system model based on knowledge graph," *IEEE Access*, vol. 12, pp. 102637–102648, 2024.

[8] S. Lei, X. Wang, and C. Yuan, "Word-based POMDP dialog management via hybrid learning," *IEEE Access*, vol. 7, pp. 39236–39243, 2019.

[9] H. Brabra, M. Báez, B. Benatallah, W. Gaaloul, S. Bouguelia, and S. Zamanirad, "Dialogue management in conversational systems: A review of approaches, challenges, and opportunities," *IEEE Trans. Cognit. Develop. Syst.*, vol. 14, no. 3, pp. 783–798, Sep. 2022.

[10] J.-L. Bouraoui, S. Le Meitour, R. Carbou, L. M. R. Barahona, and V. Lemaire, "Graph2Bots, unsupervised assistance for designing chatbots," in *Proc. 20th Annu. SIGdial Meeting Discourse Dialogue*, 2019, pp. 114–117.

[11] J. M. Sastre Martinez and A. Nugent, "Inferring ranked dialog flows from human-to-human conversations," in *Proc. 23rd Annu. Meeting Special Interest Group Discourse Dialogue*, Sep. 2022, pp. 312–324.

[12] X. Xi, C. Lv, Y. Hua, W. Ye, C. Sun, S. Liu, F. Yang, and G. Wan, "A low-cost, controllable and interpretable task-oriented chatbot: With real-world after-sale services as example," in *Proc. 45th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2022, pp. 3398–3402.

[13] S. Sohn, Y. Lyu, A. Liu, L. Logeswaran, D.-K. Kim, D. Shim, and H. Lee, "TOD-flow: Modeling the structure of task-oriented dialogues," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Singapore, Dec. 2023, pp. 3355–3371.

[14] S. Agrawal, P. Pillai, N. Uppuluri, R. Gangi Reddy, S. Li, G. Tur, D. Hakkani-Tur, and H. Ji, "Dialog flow induction for constrainable LLM-based chatbots," in *Proc. 25th Annu. Meeting Special Interest Group Discourse Dialogue*, Kyoto, Japan, Sep. 2024, pp. 66–77.

[15] P. Ferreira, D. Martins, A. Alves, C. Silva, and H. Gonçalo Oliveira, ''Unsupervised flow discovery from task-oriented dialogues,'' 2024, *arXiv:2405.01403*.
[16] N. Sabharwal and A. Agrawal, ''Introduction to Google dialogflow,'' in *Cognitive Virtual Assistants Using Google Dialogflow: Develop Complex Cognitive Bots Using the Google Dialogflow Platform*. Berkeley, CA, USA: Apress, 2020, pp. 13–54.
[17] X. Kong, G. Wang, and A. Nichol, *Conversational AI With Rasa: Build, Test, and Deploy AI-Powered, Enterprise-Grade Virtual Assistants and Chatbots*. Birmingham, U.K.: Packt Publishing Ltd, 2021.
[18] I. Carvalho, P. Ferreira, A. Alves, C. Silva, and H. G. Oliveira, ''Analysing customer-support trends in social networks through dialogue flow discovery,'' in *Proc. IberSPEECH*, Nov. 2024, pp. 56–60.
[19] A. Ritter, C. Cherry, and B. Dolan, ''Unsupervised modeling of Twitter conversations,'' in *Proc. Human Lang. Technol., Annu. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2010, pp. 172–180.
[20] D. Ressurreição, P. Ferreira, A. Alves, and H. G. Oliveira, ''Optimising dialogue flow discovery,'' in *Proc. EPIA Conf. Artif. Intell.* Cham, Switzerland: Springer, 2024, pp. 163–175.
[21] B. King and J. Flanigan, ''Unsupervised end-to-end task-oriented dialogue with LLMs: The power of the noisy channel,'' in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Nov. 2024, pp. 8283–8300.
[22] Z. Li, W. Chen, S. Li, H. Wang, J. Qian, and X. Yan, ''Controllable dialogue simulation with in-context learning,'' in *Proc. Findings Assoc. Comput. Linguistics, EMNLP*, Abu Dhabi, United Arab Emirates, Dec. 2022, pp. 4330–4347.
[23] S. Terragni, M. Filipavicius, N. Khau, B. Guedes, A. Manso, and R. Mathis, ''In-context learning user simulators for task-oriented dialog systems,'' 2023, *arXiv:2306.00774*.
[24] S. K. Suresh, W. Mengjun, T. Pranav, and E. S. Chng, ''Diasynth: Synthetic dialogue generation framework for low resource dialogue applications,'' 2024, *arXiv:2409.19020*.
[25] M. Mirtaheri, N. Varghese, C. Khatri, and A. Kelkar, ''Towards automatic evaluation of task-oriented dialogue flows,'' 2024, *arXiv:2411.10416*.
[26] P. Ferreira, I. Carvalho, A. Alves, C. Silva, and H. G. Oliveira, ''Sentiment-aware dialogue flow discovery for interpreting communication trends,'' in *Proc. 25th Annu. Meeting Special Interest Group Discourse Dialogue*, Kyoto, Japan, Sep. 2024, pp. 274–288.
[27] P. J. Rousseeuw, ''Silhouettes: A graphical aid to the interpretation and validation of cluster analysis,'' *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, Nov. 1987.
[28] X. Zang, A. Rastogi, and J. Chen, ''MultiWOZ 2.2 : A dialogue dataset with additional annotation corrections and state tracking baselines,'' in *Proc. 2nd Workshop Natural Lang. Process. Conversational AI*, Jul. 2020, pp. 109–117.
[29] S. Feng, N. Lubis, C. Geishauser, H.-C. Lin, M. Heck, C. van Niekerk, and M. Gašić, ''Emowoz: A large-scale corpus and labelling scheme for emotion recognition in task-oriented dialogue systems,'' 2021, *arXiv:2109.04919*.

PATRÍCIA FERREIRA received the B.S. and M.S. degrees in computer engineering from the Instituto Superior de Engenharia de Coimbra, Portugal. She is currently pursuing the Ph.D. degree with the Department of Computer Engineering, Faculty of Sciences and Technology, University of Coimbra. Her research interests include natural language processing, artificial intelligence, and dialogue systems.

ANA ALVES is currently an Associate Professor with Coimbra Institute of Engineering, Polytechnic University of Coimbra, and a Researcher with CISUC. She has participated in nationally funded projects in natural language processing (NLP), responsible artificial intelligence (AI), and ambient intelligence (AmI). In NLP, AI, and AmI, she has also authored and co-authored multiple publications and served as a member of the scientific committees of several peer-reviewed journals and conferences. She is a member of ACL and APPIA and is on the Editorial Board of *Journal of Ambient Intelligence and Smart Environments*.

CATARINA SILVA (Senior Member, IEEE) received the Ph.D. degree in computer engineering. She is currently an Associate Professor with the Department of Informatics Engineering, University of Coimbra, Portugal. With 25 years of teaching experience in computer engineering the B.Sc. and M.Sc., while also supervising the M.Sc. and Ph.D. students. She is also a Senior Researcher with the Adaptive Computation Group, Centre for Informatics and Systems of the University of Coimbra (CISUC), with machine learning and pattern recognition as main areas of research. More specific recent research projects on Transparency in AI in Finance (COST); Industry 4.0 Circular and Agile Manufacturing and Individualized Consumer Preferences (H2020); Real-time Condition-Based Maintenance for Adaptive Aircraft Maintenance Planning (H2020); Real-Time Monitoring of Ambient Air Quality with Low-Cost Nano-Sensors (Sudoe); and Natural Language Processing (P2020). She skilled at managing different sized projects and scientific entrepreneurships, involving people with different backgrounds, namely faculty, students, alumni, and companies. She has author and co-author of four books, circa 30 journal articles and 120 conference papers. She is a scientific committee and paper reviewer of several conferences and journals. She is a Board Member of the Portuguese Association of *Pattern Recognition*, a Senior Member of the IEEE Computational Intelligence Society, and the Past-Chair of the IEEE Portugal Section.

HUGO GONÇALO OLIVEIRA is currently an Associate Professor with the Department of Informatics Engineering, University of Coimbra (UC), and a Researcher with CISUC. He has participated in European and nationally-funded projects in computational creativity (CC), natural language processing (NLP), and responsible artificial intelligence, and led the development of computational resources, methods and systems, mostly for Portuguese language. He is the author of circa 150 papers in peer-reviewed journals and conferences, regularly part of the program committee of the main NLP and CC venues, and the Executive Editor of the journal *Linguamática*. He won the best doctoral thesis in the area of Computational Processing of Portuguese (2011–2014), and a special recognition in education, by the World Cultural Council, in 2022.