

---

# DFA-RAG: Conversational Semantic Router for Large Language Model with Definite Finite Automaton

---

Yiyou Sun<sup>1,2</sup> Junjie Hu<sup>1</sup> Wei Cheng<sup>2</sup> Haifeng Chen<sup>2</sup>

## Abstract

This paper introduces the retrieval-augmented large language model with Definite Finite Automaton (DFA-RAG), a novel framework designed to enhance the capabilities of conversational agents using large language models (LLMs). Traditional LLMs face challenges in generating regulated and compliant responses in special scenarios with predetermined response guidelines, like emotional support and customer service. Our framework addresses these challenges by embedding a Definite Finite Automaton (DFA), learned from training dialogues, within the LLM. This structured approach acts as a semantic router which enables the LLM to adhere to a deterministic response pathway. The routing is achieved by the retrieval-augmentation generation (RAG) strategy, which carefully selects dialogue examples aligned with the current conversational context. The advantages of DFA-RAG include an interpretable structure through human-readable DFA, context-aware retrieval for responses in conversations, and plug-and-play compatibility with existing LLMs. Extensive benchmarks validate DFA-RAG's effectiveness, indicating its potential as a valuable contribution to the conversational agent.

## 1. Introduction

Recent advancements in machine learning, particularly in large language models (LLMs), have provided more possibilities in various fields. Their applications range from document completion to chatbots (Ouyang et al., 2022; OpenAI, 2023). These conversational agents with LLMs are

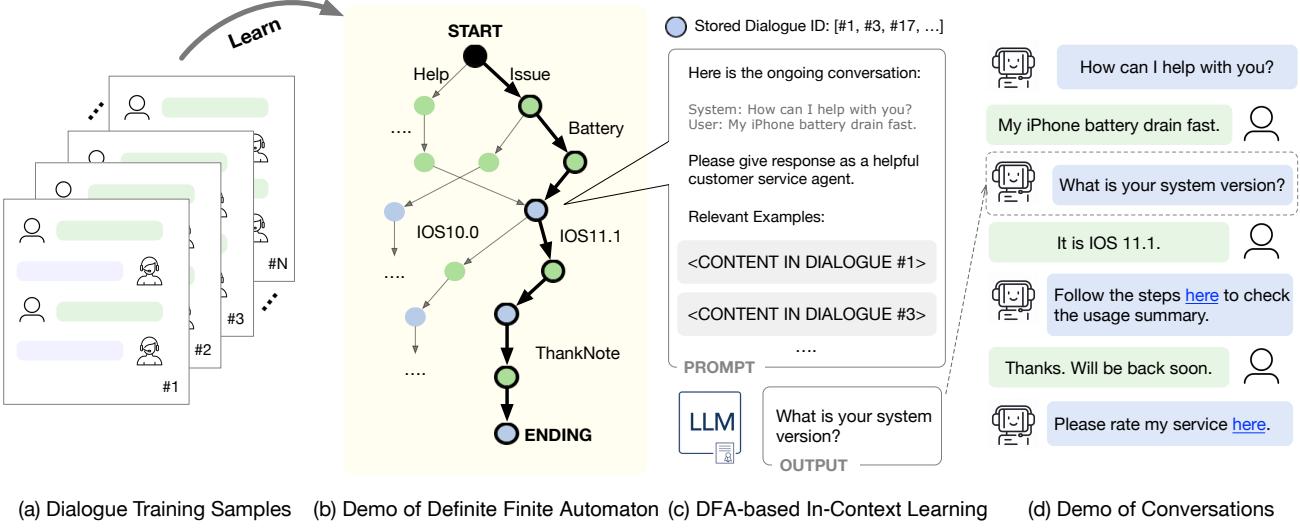
\*Equal contribution <sup>1</sup>Department of Computer Science, University of Wisconsin <sup>2</sup>NEC Laboratories America, inc., Princeton, USA. Correspondence to: Yiyou Sun <sunyiyou@nec-labs.com>, Haifeng Chen <haifeng@nec-labs.com>.

*Proceedings of the 41<sup>st</sup> International Conference on Machine Learning*, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

remarkably adaptable and have proven effective in diverse settings including customer service, marketing, education, and healthcare (Wang et al., 2023b). Compared to traditional, rule-based dialogue systems (Abdul-Kader & Woods, 2015; Hussain et al., 2019; Adamopoulou & Moussiades, 2020), LLMs offer greater efficiency, scalability, and dynamism (Medeiros et al., 2019). However, their application in practical scenarios often necessitates adherence to specific workflows or policies. For instance, an Emotional Support Chatbot (Medeiros et al., 2019) must tailor its responses to various stress conditions, while customer service bots typically follow predetermined response guidelines.

In these scenarios, popular LLMs like GPT-3.5 or GPT-4 may generate inappropriate or misleading content (Wang et al., 2023a) without specialized optimization. Fine-tuning these models, while effective, is not always viable due to open-source requirements, the need for intricate design (Ding et al., 2023), and substantial data to mitigate overfitting risks (Selvi, 2023). An alternative method, Retrieval-Augmented Generation (RAG) (Lewis et al., 2020), which references an additional knowledge base in generating a response, has been explored to address these issues. Since the generation quality of RAG is known to be sensitive to the sample selection (Gao et al., 2023), it is therefore critical to design an effective selection strategy that ensures the retrieval of the most relevant and contextually appropriate samples. This poses unique challenges to apply RAG in conversational scenarios, as one must identify partial conversational flows within historical training dialogues that closely match the current conversation context.

Targeting these challenges, we formally introduce the new framework, DFA-based retrieval-augmented generation (dubbed **DFA-RAG**). This framework assumes that a specific workflow is embedded within the training data dialogues, which we model using a Definite Finite Automaton (as illustrated in Figure 1). Acting as a Semantic Router (Horsey, 2024) like a decision-making layer, DFA-RAG routes conversations through a predefined trajectory. This ensures that the LLM adheres to the workflow encoded in the DFA. Specifically, each conversational utterance corresponds to a particular DFA state, where each state encapsulates responses from similar historical contexts. By utilizing these historical examples as retrieved samples, DFA-RAG



**Figure 1.** Illustration of the DFA-RAG Framework. (a) shows the training set with dialogues. (b) demonstrates the Definite Finite Automaton (DFA) which represents the workflow learned from the dialogues. Blue and green dots represent the states of the user and system respectively. The states are transited by keywords in conversations. (c) outlines the DFA-based In-Context Learning process, where the LLM is guided by the DFA to provide contextually relevant responses. (d) showcases sample conversations between a user and the LLM.

effectively guides the LLM to follow a deterministic DFA-based response pathway. The DFA-RAG framework offers several compelling advantages:

- 1. Interpretable Structure:** Unlike RAG (Retrieval-Augmented Generation) (Lewis et al., 2020) which fuses information on uninterpretable embedding spaces, our approach ensures responses are generated by human-readable structures (DFA), enhancing the trustworthiness of LLM-based chatbots.
- 2. Context-Aware Retrieval:** Unlike most retrieval-based strategies that treat dialogues as undivided wholes, our approach divides each dialogue into manageable segments. This granular approach allows for more precise and contextually relevant information retrieval, enhancing the chatbot’s ability to respond accurately and appropriately to various conversational nuances.
- 3. Plug-and-Play Compatibility:** The DFA, derived from training data, is designed for easy integration with any pre-trained LLM. This flexibility allows for widespread application across different LLM platforms, making DFA-RAG a versatile tool in conversational AI.
- 4. Strong Empirical Performance:** In domain-specific conversations, DFA-RAG has shown superior performance, as demonstrated by extensive evaluation benchmarks. This empirical evidence underscores the effectiveness of the proposed framework in delivering high-quality, domain-relevant responses.

## 2. Problem Setup

In our setting, we assume access to a sample set of dialogues in the application domain of interest, such as customer service or emotional support. This dataset may be pre-collected through interactions between a human agent and a customer or patient. We now formally describe the data setup.

**Data Setup.** We consider the training set with  $N$  dialogues  $\mathcal{D}_{\text{tr}} = \{x_i\}_{i=1}^N$ . Each dialogue is a sequence of utterances, denoted as  $x_i = (z_1, z_2, \dots, z_{l_i})$ , where  $l_i$  represents the length of the  $i$ -th dialogue. We assume the utterances alternate between the agent and the user.

**Goal.** At the inference stage, the LLM-based agent generates the next utterance based on the context of an incomplete dialogue  $\bar{x} = (z_1, z_2, \dots, z_j)$ :

$$\hat{z}_{j+1} \sim \mathbb{P}(\hat{z}|\bar{x}, \mathcal{D}_{\text{tr}}).$$

The objective is for the output  $\hat{z}_{j+1}$  to closely match the human agent’s response  $z_{j+1}$ .

## 3. Methodology

Human agents often provide responses that involve navigating a pre-determined landscape of options, heuristics, and potential outcomes. This process can be conceptualized as exploring a special branch of a network of interconnected paths and junctions. In contrast, current LLMs like GPT-4 lack the intrinsic ability to mimic this human-like pipeline-guided behavior in response generation. This limitation

becomes particularly pronounced in scenarios where adherence to specific workflows is paramount, such as in customer service or healthcare settings. In these cases, traditional LLMs might generate responses that are misaligned (Wang et al., 2023a) with the desired conversational trajectory or decision pathway.

To address this gap, our methodology with DFA-RAG embeds a structured pipeline within the operational paradigm of the LLM. By integrating a DFA, we propose to direct the LLM’s response generation process, aligning it more closely with the decision paths of typical human agents. In Section 3.1, we introduce how we can model the conversations as a DFA. Later in Section 3.2, we delve into details of constructing the DFA from training dialogue datasets. Subsequently, in Section 3.3, we illustrate the mechanism by which the learned DFA guides the LLM.

### 3.1. Modeling Conversations with DFA

#### 3.1.1. PRELIMINARIES OF DFA

Automata has a long history of study in theoretical computer science, linguistics, and other related fields (Minsky, 1956; Kleene et al., 1956). A deterministic finite automaton (DFA) can be specified as a tuple  $(Q, \Sigma, \delta, q_0, F)$ , where:

- $Q$  is a finite set of states
- $\Sigma$  is a finite input alphabet
- $\delta : Q \times \Sigma \rightarrow Q$  is the transition function
- $q_0 \in Q$  is the start state
- $F \subseteq Q$  is the set of accept states

The DFA processes a string of symbols from  $\Sigma$  and changes its state according to the transition function  $\delta$ . The string is accepted by the DFA if the automaton is in one of the accept states in  $F$  after processing all symbols.

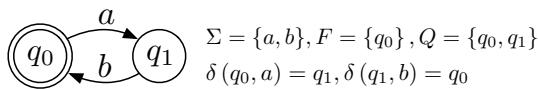


Figure 2. A demo of DFA recognizing string “ $(ab)^*$ ”.

#### 3.1.2. CONVERSATION AS TAG-SEQUENCE

DFA are widely recognized for their capability to define alphabet sequences with a specific syntax. A classic example as shown in Figure 2 is a DFA with a two-state structure that can define strings like “*ababab*” and “*abab*”. However, when it comes to modeling conversations, the challenge arises due to the complexity and variety of conversational elements. Unlike simple syntactic strings, the essence of a conversation lies predominantly at the semantic level. For instance, phrases such as “*My battery drains out fast*” and “*How come my phone can be only used for 1 hour?*” convey the same underlying issue in a customer service context, yet

they do not share a single word in common. This disparity underscores the limitations of traditional DFA models in capturing the nuances of conversational semantics.

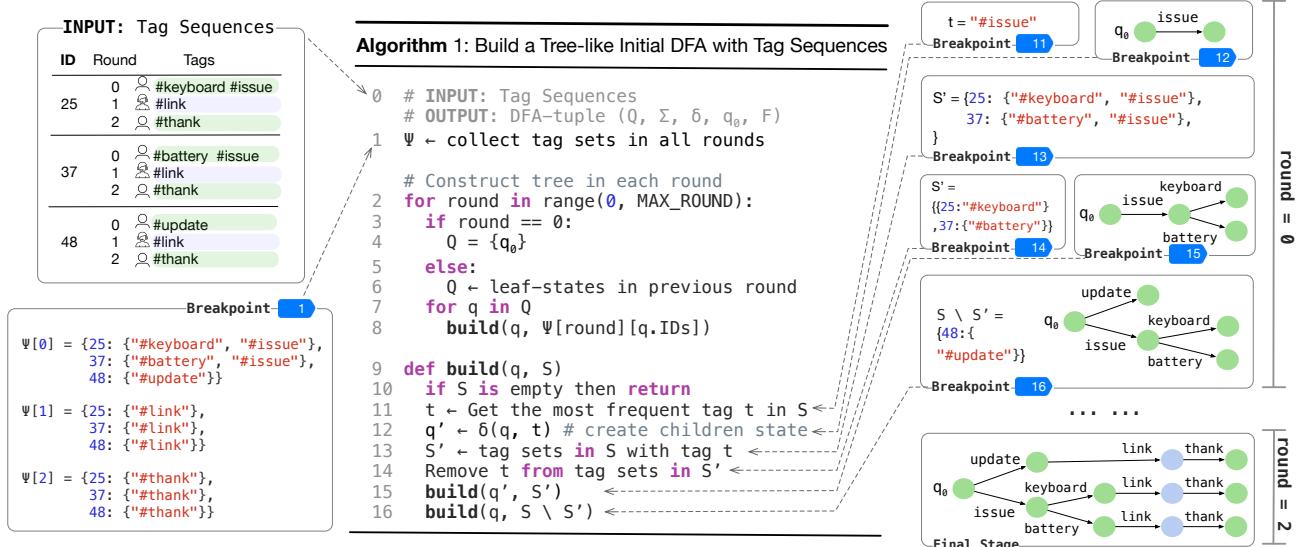
To address this, we assume each utterance in a conversation can be encapsulated by a set of “keywords” or “tags”. (We use the term “tags” in this paper). For example, the utterance “How come my phone can be only used for 1 hour?” could be succinctly represented by the tag set  $\{\#\text{issues}\text{, }\#\text{battery}\}$  in the customer service context. This method allows for the abstraction of utterances into a more manageable form, suitable for DFA modeling.

#### 3.1.3. CONVERSATION SETS AS DFA

Building upon the idea of representing individual conversations as sequences of tags, we extend this concept to model entire sets of conversations using Deterministic Finite Automata (DFAs). Recall our earlier definition of a DFA as a tuple  $(Q, \Sigma, \delta, q_0, F)$ . In this context, we adapt these components to fit our conversation modeling framework:

- **States ( $Q$ ):** Each state in the DFA represents a particular stage or context within a conversation. For instance, a state could represent the initiation of a conversation, a query about a specific issue, a response, or the conclusion of the interaction.
- **Alphabet ( $\Sigma$ ):** The alphabet in this model comprises the set of all possible tags that we identified as representative of different utterances. These tags form the basic building blocks of our conversation sequences.
- **Transition Function ( $\delta$ ):** The transition function maps a state and a tag to a subsequent state. It encapsulates the flow of conversation, determining how an utterance (through its tags) leads from one conversational context to another.
- **Start State ( $q_0$ ):** The start state represents the beginning of a conversation. It could be a generic greeting or an initial query, setting the stage for the interaction.
- **Accept States ( $F$ ):** These are the states indicating the completion of a conversation. An accept state could be reached after successfully addressing a query, reaching a satisfactory conclusion, or when the conversation naturally comes to an end.

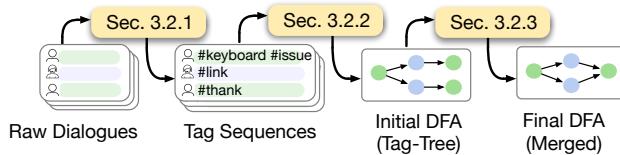
**Remark on the dialogue tracking function  $\mathcal{I}$ .** In our DFA framework, a unique functionality is incorporated where each state maintains a record of the indices of dialogues that have traversed through it. This is accomplished by tracing the trajectory of tag sequences as the conversation unfolds. Formally, this tracking is represented by the function  $\mathcal{I}(q)$ , which maps a state  $q$  to a set of dialogue IDs in  $\{1, \dots, N\}$ . Crucially, it enables us to trace back and identify the most relevant dialogue samples that can be effectively utilized in in-context learning.



**Figure 3.** Illustration of the algorithm for building a tag tree, expressed in Python-style pseudocode. **(1)** Positioned **centrally** is the core algorithm, flanked on either side by demonstrative examples that “print out” the algorithm’s internal variables at various “breakpoints”. **(2)** On the **left**, we begin by displaying the input data used in the demonstration. Following this, the first breakpoint, corresponding to *Line 1* of the algorithm, showcases the variables storing the ID-ed tag sets across different conversational rounds. **(3)** Shifting to the **right** side, we illustrate the evolution of the variables and the tree structures, starting from *round 0* with the initial state  $q_0$ . The breakpoints at *Lines 15 & 16* forego stepping into the function, instead presenting the completed tree structures directly for clarity. To simplify the illustration, we omit the iterative process for *rounds 1 & 2*, jumping straight to the final results.

### 3.2. Learn DFA from Conversations

This section delves into the methodology of learning DFA from conversation sets, with a brief outline in Figure 4. In Section 3.2.1, we detail the process of deriving tag sequences from conversational data, employing LLMs to accurately identify and extract relevant tags. Following this, Section 3.2.2 focuses on the assembly of these tag sequences into a structured tree format, laying the groundwork for DFA construction. Lastly, in Section 3.2.3, we introduce the state-merging process within the tag tree, refining the DFA to succinctly represent the dynamics of conversational flows.



*Figure 4.* Outline for Section 3.2.

### 3.2.1. EXTRACTING TAG SEQUENCES FROM CONVERSATIONS USING LLMs

In this subsection, we detail the method for converting a conversation into a sequence of tags, a crucial preprocessing step for constructing the DFA. The process hinges on the

use of the state-of-the-art LLMs, such as GPT-4, which are adept at processing natural language and extracting key information like human beings. The primary task of the LLM in this context is to preprocess each sentence of a conversation and distill it into the most representative tags.

The extraction process is guided by specific prompt parameters to ensure that the resulting tags are both concise and relevant to the conversational content. These prompts are structured to instruct LLM adhering to two main criteria: a) **Brevity**: The LLM is instructed to use as few words as possible, ideally limiting the output to no more than three words per tag. b) **Focus on Core Elements**: The summarization is directed to concentrate on actual events, issues, queries, or solutions present in the conversation. We include the full prompts in Appendix A.

By applying the prompts, the LLM processes each sentence of the conversation and outputs succinct, focused tags. These tags serve as the building blocks for constructing the DFA, encapsulating the key elements of the conversation.

### 3.2.2. TREE CONSTRUCTION WITH TAG SEQUENCES

The traditional approach to building a DFA often begins with the construction of a prefix tree, representing the sequences of alphabets that form valid strings in the language. In the case of conversation sets, however, we encounter

two significant challenges that diverge from this traditional method:

**1. Tags in each utterance can appear in various orders,** which is unlike strings in traditional DFA models that follow a fixed sequence of alphabets. This flexible ordering presents a unique challenge in constructing a prefix tag tree. For example, in a standard DFA, the string “ab” is distinctly represented by the sequence “a → b”. However, a sentence like “My iPhone battery drains fast” in a conversation can be equivalently tagged as either “#issues → #battery” or “#battery → #issues”. To efficiently manage this variability and reduce the size of the tag tree, our algorithm prioritizes tags based on their frequency of occurrence, preferring to place more frequent tags earlier in the sequence.

**2. Each tag derived from an utterance is associated with the context of a specific round in the conversation.** For instance, the tag “#address” may carry distinct semantic meanings in the user round compared to the system round. In one context, it might signify an address query, while in another, it serves to provide information. Therefore, when evaluating tag frequencies or constructing the tree, we need to focus exclusively on tags from the same conversational round. As illustrated in the left part of Figure 3, an important preprocessing step is to convert the original tag sequences to a special data structure  $\Psi$  with ID-ed tag sets across different conversational rounds.

To address these challenges, we introduce the algorithm as depicted in Figure 3. It begins with constructing a tag tree from the initial state  $q_0$  and *round 0*, detailed from *Lines 2 to 8*. For each subsequent round, the newly constructed tree’s root state becomes a child state of one of the leaf states from the previous round’s tree. The selection of leaf states  $Q$  for progression into the next round is guided by additional heuristic rules, which are elaborated in Appendix B. The procedure for building a tree from a set of tags, along with associated dialogue IDs, is expanded upon in *Lines 9 to 16*. The central strategy is recursively selecting the most frequent tag to expand the tree.

Upon execution of Algorithm 1, we yield a tree-like structure resembling a DFA. However, during our final results demonstration in the right-bottom part of Figure 3, it becomes evident that duplicated sub-sequences, such as “#link → #thank”, exist. This redundancy presents an opportunity to merge these sequences, thereby reducing the overall size of the resultant DFA.

### 3.2.3. STATE MERGING IN TAG TREE

In the process of constructing a DFA-tree from conversation sets, the variability in the tags generated by LLM for utterances with similar or identical contexts is a key challenge. For instance, tags like “#subscription” and

“#membership” might represent the same conversation topic, yet linguistically, they are distinct. Merging such states based purely on linguistic similarities is impractical due to their semantic differences. However, in a conversational context, these tags often play similar roles and are connected to related sub-structures, such as common children-tags like “#refund” or “#payment”, which are typically associated with financial aspects of subscriptions.

To address this, we introduce a similarity score  $\phi_{sim}(q, q')$  between two states  $q$  and  $q'$  as follows:

$$\phi_{sim}(q, q') = \frac{\sum_{t \in \mathcal{T}(q) \cap \mathcal{T}(q')} |\mathcal{I}(\delta(q, t))| \cdot |\mathcal{I}(\delta(q', t))|}{\sum_{t \in \mathcal{T}(q)} |\mathcal{I}(\delta(q, t))| \cdot \sum_{t \in \mathcal{T}(q')} |\mathcal{I}(\delta(q', t))|},$$

where we define function  $\mathcal{T}$  as the tag sets connected to  $q$ :

$$\mathcal{T}(q) = \{t \in \Sigma \mid \exists q' \in Q, \delta(q, t) = q'\}$$

Intuitively,  $\phi_{sim}(q, q')$  calculates the ratio of common tags connected to both states  $q$  and  $q'$ . The significance of each tag is considered by counting the number of dialogues related to the connected state, denoted by  $|\mathcal{I}(\delta(q, t))|$ . This approach focuses on comparing the immediate (1-hop) children’s states. Empirical evidence suggests that comparing direct children often suffices, though the formula can be extended to compare deeper sub-tree structures if necessary.

Once the similarity score is calculated, we set a threshold  $\lambda$  to determine which node pairs should be merged. Pairs with a similarity score exceeding this threshold,  $\phi_{sim}(q, q') > \lambda$ , are considered for merging. This methodology allows us to consolidate similar conversational contexts, reducing redundancy and increasing the efficiency of the DFA-tree.

### 3.3. Conversation Generation by DFA-RAG

Once a DFA is constructed to model conversation sets, it can be integrated with a Large Language Model (LLM) to generate dynamic and coherent conversations as demonstrated in Figure 1. This process involves a cyclical interaction between the user’s inputs, the DFA’s navigational structure, and the LLM’s response generation capabilities. Here are the concrete steps involved in this procedure:

1. **Tagging User Utterances:** When a user responds, their input is first processed by the LLM, which converts it into tags. This process mirrors the approach used for tagging the training samples, ensuring consistency in tagging the conversational content.
2. **Navigating the DFA:** Based on the tags derived from the user’s input, the system navigates through the DFA with a deterministic path to identify the corresponding state. Given an input tag sequence  $\{u_0, u_1, \dots, u_T\}$ , the state transition at step  $t$  is denoted as  $q_t = \delta(q_{t-1}, u_t)$ .

*Table 1.* Results of dialogue generation quality across different base models and methods. This table reports the “Win Rate” over naive base models (GPT-4, GPT-3.5) regarding dialogue generation performance. For each method using in-context learning (RandSamp, RAG, BM25, DFA-RAG), we use 5 samples in the inference time. For FT-LLM, we perform fine-tuning using the API provided by OpenAI with standard hyperparameters. Note that the API for fine-tuning GPT-4 is not available.

Base LLM	Methods	Domains						Average
		AmazonHelp	DeltaSupport	AskPlayStation	AirbnbHelp	NikeSupport	CambridgeInfo	
GPT-4	RandSamp	69.1	84.1	57.9	78.3	45.3	67.0	66.9
	BM25	67.3	81.5	63.8	77.1	59.8	63.0	68.7
	RAG	74.4	87.0	<b>66.3</b>	72.2	57.3	66.5	70.6
	FT-LLM	-	-	-	-	-	-	-
GPT-3.5	<b>DFA-RAG (Ours)</b>	<b>78.0</b>	<b>89.9</b>	65.9	<b>80.9</b>	<b>62.6</b>	<b>68.5</b>	<b>74.3</b>
	RandSamp	70.2	83.6	61.3	69.5	58.9	57.9	66.9
	BM25	70.6	84.1	64.7	74.3	60.4	58.8	68.8
	RAG	73.8	82.9	72.4	76.6	63.3	60.6	71.6
	FT-LLM	69.7	64.6	71.7	66.1	56.8	56.1	64.2
<b>DFA-RAG (Ours)</b>	<b>78.5</b>	<b>89.8</b>	<b>72.9</b>	<b>79.1</b>	<b>70.1</b>	<b>64.9</b>	<b>75.9</b>	

In cases where  $\delta(q_{t-1}, u_t) = \emptyset$ , indicating a mismatch, the navigation process concludes, and we proceed to retrieve dialogue IDs based on the last valid state,  $q_{t-1}$ .

3. **Accessing Dialogue IDs:** Each state  $q$  in the DFA is associated with a collection of dialogue IDs  $\mathcal{I}(q)$ , which represent past conversations that have followed a similar pathway through the DFA.
4. **Compiling a Prompt for LLM:** From the dialogue IDs, the system samples 5 examples and compiles them into a prompt structured for in-context learning as shown in Figure 1(c).
5. **LLM Response Generation:** The LLM then generates a response based on the given prompt.
6. **Iterative Process:** The conversation continues with the user providing their next response based on the LLM’s output. The process reverts to **STEP 1**, with the user’s new response undergoing the same cycle of tagging, DFA navigation, and LLM response generation.

This iterative process allows for a dynamic conversation flow, where each response is contextually informed and semantically relevant.

## 4. Experimental Results

**Datasets.** We carry out experiments with dialogue datasets in extensive domains. A high-level summary of the dataset statistics is provided in Table 2.

The dataset described in the table consists of domain-specific collections, each serving distinct business communication needs. Notably, the first five datasets — AmazonHelp, DeltaSupport, AskPlayStation, AirbnbHelp, and NikeSupport, are derived from Twitter interactions (Axelbrooke, 2017), including a range of sizes from larger datasets containing over 50,000 dia-

*Table 2.* Summary statistics of the datasets.

Domains	Dataset Size	Test Size	Avg. Rounds	Source
AmazonHelp	50K	1K	6.4	Tweet
DeltaSupport	20K	1K	4.4	Tweet
AskPlayStation	10K	500	3.8	Tweet
AirbnbHelp	3K	100	4.2	Tweet
NikeSupport	1K	100	5.5	Tweet
CambridgeInfo	8K	1k	13.5	MultiWOZ

logues (AmazonHelp and DeltaSupport) to smaller ones with 1,000 entries (NikeSupport). These datasets represent unique customer service interactions pertaining to their respective single business domains.

In contrast, CambridgeInfo, an alias for the Cambridge Information Center, encompasses a broader spectrum of services. It includes dialogues related to various booking services such as train, taxi, and hotel reservations. This dataset is sourced from MultiWOZ (Budzianowski et al., 2018), and differs significantly from the Twitter-based datasets. MultiWOZ is known for its dense annotations, detailing the purpose and key information of each dialogue. This feature of MultiWOZ allows for a more nuanced analysis of dialogues in Section 4.2, particularly in assessing the success rate of interactions.

### 4.1. Generation Quality Evaluation

This section aims to evaluate the quality of generated dialogues using the DFA-RAG framework. We compare performance on major pretrained LLMs as the backbone model in generating dialogues: GPT-4 (OpenAI, 2023), GPT-3.5 (Ouyang et al., 2022). The concrete versions are *gpt-4-1106-preview* and *gpt-3.5-turbo-1106* respectively.

**Evaluation Technique.** We leverage GPT-4 for the primary evaluation, given its superior performance metrics. According to Zheng et al. (2023), GPT-4 demonstrates an 85% agreement rate with human evaluators, signifying

its reliability in assessing dialogue quality. We follow the methodology outlined in AlpacaEval (Dubois et al., 2023) with similar prompts and metrics as detailed in Appendix A. The evaluation process involves GPT-4 comparing two dialogues to determine which one is closer to the ground truth. The dialogues completed directly by their respective LLMs serve as the target to be compared. The effectiveness of each method is quantified using the “**Win Rate**” score, which reflects how often a dialogue surpasses the quality of the competitor. For a qualitative comparison, we also present example dialogues with respective comments from the GPT-4 evaluator in Appendix C.

**Baselines.** We compare diverse strategies in selecting in-context learning (ICL) examples, recognizing the sensitivity of few-shot learning to sample selection: a) Random Sampling involves randomly selecting examples from the training dataset; b) RAG utilizes *text-embedding-3-small* (from OpenAI) for generating sentence embeddings and focuses on retrieving training examples with utterances most similar to the test utterance; c) BM25 is a classical sparse retrieval method used to find training examples with utterances closely matching those in the test scenarios. Additionally, we examine another baseline, “FT–LLM” (Finetuned Large Language Model), which involves directly finetuning the LLM on the dialogues in the training set.

**Observations.** Our observations from the evaluation in Table 1 reveal some notable trends and insights: (a) the approach of fine-tuning (FT–LLM) showed underwhelming performance, which can be attributed to the limited volume of domain-specific training data, such as NikeSupport; (b) RAG demonstrates stronger performance than RandSamp and BM25, which is likely due to its more accurate retrieval of relevant samples; (c) DFA-RAG outperforms the best baseline by 4% at win-rate. Note that DFA-RAG not only facilitates the generation of high-quality dialogues but also provides a clear and interpretable structure, making it easier for humans to understand and analyze the dialogue flow.

## 4.2. Dialogue Task Evaluation

In the evaluation of dialogue systems, particularly within customer service domains, one key metric is the system’s ability to resolve the user’s inquiries and issues effectively. Our evaluation is performed on the MultiWOZ dataset (corresponds to CambridgeInfo domain in Section 4.1), a widely recognized benchmark in the realm of task-oriented dialogue generation.

In these dialogues, system responses are typically presented in a delexicalized form — specific values for certain variables like time/address within the dialogue are replaced with placeholders. In line with this practice, our model also incorporates delexicalization by replacing specific examples in in-context learning with placeholder forms.

To quantitatively measure the performance of our model, we employ standardized evaluation metrics, namely “**Inform**” and “**Success**” rates. The “Inform” rate measures the degree to which the system provides sufficient and relevant information to fulfill the user’s information needs, while the “Success” rate evaluates the system’s performance in terms of completing the user’s goal like booking a hotel. The goal is to understand the overall efficacy of the dialogue system in achieving the intended outcomes of the interaction.

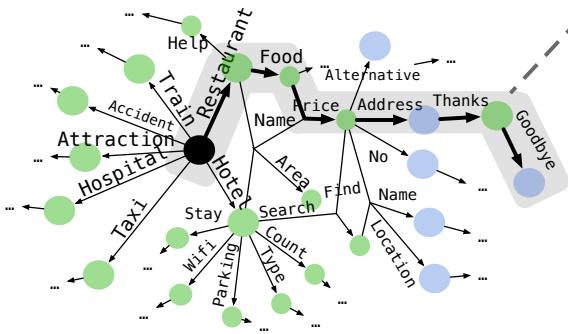
Table 3. Evaluation results on the task-oriented dialogues.

Ground Truth States in Training?	Methods	Inform	Success
Yes	HDSA	87.9	79.4
	MarCo	94.5	87.2
	HDNO	93.3	83.4
	GALAXY	92.8	83.5
	KRLS	93.1	83.7
No	AuGPT	76.6	60.5
	MTTOD	85.9	76.5
	RSTOD	83.5	75.0
	RewardNet	87.6	81.5
	TOATOD	90.0	79.8
No	<b>DFA-RAG (Ours)</b>	93.3	90.0

In Table 3, we compared two sets of baseline methods: a) one branch of works (HDSA (Chen et al., 2019), MarCo (Wang et al., 2020b), HDNO (Wang et al., 2020a), GALAXY (He et al., 2022), KRLS (Yu et al., 2023)) assumes known dialogue states for policy optimization strategy and response generation; b) the other branch, also known as End-to-End models (AuGPT (Kulhánek et al., 2021), MTTOD (Lee, 2021), RSTOD (Cholakov & Kolev, 2022b), RewardNet (Feng et al., 2022), TOATOD (Bang et al., 2023)), rely solely on dialogue context. Our method falls into the latter category, focusing on the model’s inherent ability to generate responses without using known dialogue states. Despite this, our approach has delivered promising results across both “Inform” and “Success” metrics, exhibiting performance comparable to methods relying on ground truth dialogue states. This highlights the efficacy and adaptability of our model in handling a range of dialogue scenarios efficiently.

## 4.3. Constructed DFA Demonstrations

We display the (partial) DFA constructed from the MultiWOZ dataset in Figure 5. At the starting point of the automaton, we observe direct connections to several key domains such as “train”, “hotel”, and “taxi”. These domains represent the most frequently encountered topics in the MultiWOZ dataset. As we delve into specific domains like restaurants and hotels, a notable observation is the recurrence of similar tags such as “price”, “name”, and “area”, which are common queries from users. Another



(a) An Example Tag Pathway in Constructed DFA

ID	Raw Dialogues in Training Set
11	<p>0 USER: I want to find an <b>expensive</b> restaurant and serves <b>european</b> food. Can i also have the address, phone number and its area?</p> <p>1 SYSTEM: Eraina is a European restaurant in the centre area. Their address is <b>St. Michael's Church Trinity Street City Centre</b> and phone number 01223 355166.</p> <p>2 USER: <b>thank</b> you goodbye</p> <p>3 SYSTEM: Thank you and <b>good bye</b></p>
8238	<p>0 USER: I want a <b>moderately</b> priced restaurant and serves <b>asian</b> oriental food. What is the <b>address</b> and phone number?</p> <p>1 SYSTEM: there is an Asian Oriental place at <b>40428 King Street City Centre</b>. You can call them at 01223 518111.</p> <p>2 USER: <b>Thank</b> you goodbye.</p> <p>3 SYSTEM: Thank you for using our system! <b>Have a great day!</b></p>
7350	<p>0 USER: I want a <b>cheap</b> <b>portuguese</b> restaurant. Give me their <b>address</b>, phone number, and area.</p> <p>1 SYSTEM: Nandos is a cheap Portuguese restauran in the south area. The phone number is 01223 327908 and the address is <b>Cambridge Leisure Park Clifton Way</b></p> <p>2 USER: <b>Thank</b> you, goodbye</p> <p>3 SYSTEM: You're welcome, <b>goodbye</b>.</p>

(b) Corresponding Dialogues in the Training Set

Figure 5. DFA Results for MultiWOZ. (a) This segment of the figure illustrates a portion of the constructed DFA. The black circle indicates the starting point of the automaton. Each green circle represents a “user” state, while each blue circle denotes a “system” state. The states are interconnected by arrows, each labeled with a tag. Note that some lines are interconnected (ex. lines correspond to “name” and “area”), it means that the relevant nodes are connected in both ways. (b) A specific path within the DFA is highlighted to demonstrate its correspondence with actual dialogues traversed. In these dialogues, elements associated with the tags are emphasized in bold.

intriguing aspect of the DFA is how the same state can lead to divergent system responses based on the user’s input and the system’s capabilities. For instance, if the system successfully resolves a user’s inquiry, as illustrated in the figure, it proceeds to provide specific details like the concrete address information. Conversely, another branch emerges when the system fails to find the desired target, signified by a “no” tag. For readers’ interest, we include more DFA demos in Appendix D. Overall, the DFA illustration captures the multifaceted nature of conversational flows within the MultiWOZ dataset and serves as a potent tool for understanding and improving domain-specific dialogue systems.

## 5. Discussions and Future Work

**Handling Out-of-distribution Utterances.** In practice, it is common to encounter out-of-distribution (OOD) instances (Yang et al., 2022; Zhang et al., 2023; Sun, 2023; Sun et al., 2023; 2024) that deviate from the pre-defined trajectory in the DFA constructed from training dialogues. Unlike existing research in OOD detection (Du et al., 2024; Ghosal et al., 2024; Ming et al., 2022b;a), which typically rejects user input when an OOD instance is detected, our approach leverages the dynamic capabilities of LLMs to respond to OOD utterances based on their best judgment—a feature absent in traditional rule-based dialogue systems.

Formally, utterances are treated as OOD in dialogues where  $\delta(q_{t-1}, u_t) = \emptyset$ , indicating no viable subsequent state (a mismatch scenario), the framework concludes the current navigation process and retrieves dialogue options based on the most recent valid state,  $q_{t-1}$ . This approach addresses situations where user inputs do not match any predefined

pathway in the DFA, while still managing to navigate to the most relevant context. For example, if an unfamiliar user’s inquiry occurs—such as “booking NBA game tickets”, which might not directly correspond to a NBA tag within the DFA’s structure, our framework adeptly guides the dialogue towards a related, more general context. This is achieved by reverting to the parent state  $q_{t-1}$  which represents sports event ticket booking. This strategy ensures the DFA-RAG can maintain coherence and relevance in conversations, even when faced with partially matched or unexpected user inputs.

**Integration with External Modules.** The plug-and-play nature of our method facilitates seamless integration with existing services and algorithms. For instance, in scenarios where an input sentence’s tags fall outside the predefined tracks of the DFA, our framework can dynamically interface with web search services or other external tools via callable APIs. This enables the LLM to access a broader range of information and resources to improve the quality of generated responses.

## 6. Related Work

**Structured Dialogue System.** Traditional conversational agents operate on a fixed knowledge base, providing deterministic responses to user inputs from a pre-defined set of answers (Abdul-Kader & Woods, 2015; Bickmore et al., 2016; Jain et al., 2018; Medeiros et al., 2019; Hussain et al., 2019; Adamopoulou & Moussiades, 2020; Janssen et al., 2020; Safi et al., 2020). While reliable and consistent, they lack flexibility and struggle with unexpected queries. In contrast, ML-based conversational agents (OpenAI, 2023;

Ouyang et al., 2022), represent a more adaptive approach but can be unpredictable. Our proposed method combines the stability of traditional agents with the adaptability of ML-based systems, leveraging the strengths of both to create a more dynamic and reliable dialogue system.

**Retrieval-Augmented Generation (RAG).** RAG was originally introduced in (Lewis et al., 2020), which established an embedding database containing accessible documents, enabling the retrieval of related information in generating responses. This approach addresses key limitations of LLMs, particularly in areas of knowledge updating and reasoning transparency (Shuster et al., 2021; Yasunaga et al., 2022; Borgeaud et al., 2022; Khattab et al., 2022; Cheng et al., 2022; Wang et al., 2023c; Cheng et al., 2023). Building upon the naive RAG, more sophisticated variants have been developed. These models incorporate complex components like structured data sources. For instance, RET-LLM (Modarressi et al., 2023) constructs a personalized knowledge graph memory, extracting relation triples to enhance response generation. Similarly, SUGRE (Kang et al., 2023) utilizes Graph Neural Networks (GNN) to embed relevant sub-graphs retrieved from the knowledge graph. KnowledgeGPT (Wang et al., 2023c) generates search queries for Knowledge Bases (KB) in code format and includes predefined KB operation functions. Our algorithm can be also viewed as an extension of the RAG concept, which encodes historical dialogues into DFA and retrieves similar dialogues based on it. Our approach marks a pioneering step in conversational agent research.

**Task-oriented Dialogue System.** In this paper, we aim to enhance the reliability of LLMs for domain-specific dialogue generation. A critical subset involves task-oriented dialogue (TOD) systems, which are designed to accomplish specific tasks through interactive communication. A notable line of methods involves leveraging dialogue state annotations to train policy-based dialogue generation systems (Chen et al., 2019; Ramachandran et al., 2021; Tseng et al., 2021; Wang et al., 2020b;a; He et al., 2022; Yu et al., 2023). Alternatively, some approaches have simplified the design of TOD systems to the end-to-end (E2E) (Kulhánek et al., 2021; Lee, 2021; Su et al., 2022; Cholakov & Kolev, 2022a;b; Feng et al., 2022; Bang et al., 2023). Comparing all existing works, our DFA-RAG framework introduces a distinct approach without relying on traditional gradient-based training. This approach offers simplicity and adaptability, allowing for easy and efficient adjustments to new scenarios in a plug-and-play manner.

**Semantic Router.** The Semantic Router (Horsey, 2024; azhar, 2024; Avila, 2024; Sisodia, 2024; Hingane, 2024) represents a recent advancement in LLM technology, particularly in augmenting the capabilities of chatbots and AI assistants through the development of a predefined decision-

making layer. This router directs the decision trajectory of LLMs by evaluating semantic similarity in each branch and selecting the closest match. Typically, similarity is measured using cosine distance between the embedding vector of inquiries and the potential branches. In this paper, DFA-RAG can be seen as an evolution of the semantic router, offering several key advantages: a) DFA-RAG navigates the decision-making layer using tags instead of embedding comparison, which enhances interpretability and computational efficiency. b) Unlike decision-making layers predefined by humans<sup>1</sup>, the DFA structure within our framework is learnable from historical training data, making it more adaptable and scalable for real-world applications.

## 7. Conclusion

This paper proposes the DFA-augmented Large Language Model (DFA-RAG), integrating the adaptability of large language models with the structured approach of Definite Finite Automata. The DFA structure is learnable from the dialogues in the training set. The whole framework ensures reliable, contextually appropriate responses, addressing key limitations in current LLM applications, particularly in specialized areas like customer service. Extensive experiments are conducted to validate DFA-RAG’s effectiveness in generating pertinent dialogue content. We hope this pioneering research can pave the way for more controlled and predictable AI conversational agents without sacrificing the dynamic nature of LLMs.

## Impact Statement

The DFA-RAG framework has the potential to revolutionize how conversational agents are deployed across sectors, including healthcare, customer service, and education, by providing more accurate, context-aware, and ethically aligned interactions. This advancement could lead to broader accessibility of digital services, enhanced user experiences, and greater trust in AI technologies. Beyond the mentioned societal consequences, we do not anticipate any potentially harmful consequences to our work.

## Acknowledgment

Hu is supported by the Wisconsin Alumni Research Foundation, and by the National Institute Of Biomedical Imaging And Bioengineering of the National Institutes of Health under Award Number R01EB033782. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

<sup>1</sup>See Aurelio AI <https://www.aurelio.ai/>.

## References

- Abdul-Kader, S. A. and Woods, J. C. Survey on chatbot design techniques in speech conversation systems. *International Journal of Advanced Computer Science and Applications*, 6(7), 2015.
- Adamopoulou, E. and Moussiades, L. Chatbots: History, technology, and applications. *Machine Learning with Applications*, 2:100006, 2020.
- Avila, D. Semantic router: Enhancing control in llm conversations. *CodeGPT @ Medium*, January 2024. URL <https://blog.codegpt.co/semantic-router-enhancing-control-in-\llm-conversations-68ce905c8d33>.
- Axelbrooke, S. Customer support on twitter (kaggle competition). <https://www.kaggle.com/datasets/thoughtvector/customer-support-on-twitter>, 2017.
- azhar. Beyond basic chatbots: How semantic router is changing the game. *AI Insights @ Medium*, January 2024. URL <https://medium.com/ai-insights-cobet/beyond-basic-chatbots-how-semantic-\router-is-changing-the-game-783dd959a32d>.
- Bang, N., Lee, J., and Koo, M.-W. Task-optimized adapters for an end-to-end task-oriented dialogue system. *arXiv preprint arXiv:2305.02468*, 2023.
- Bickmore, T. W., Utami, D., Matsuyama, R., and Paasche-Orlow, M. K. Improving access to online health information with conversational agents: a randomized controlled experiment. *Journal of medical Internet research*, 18(1):e1, 2016.
- Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., Van Den Driessche, G. B., Lespiau, J.-B., Damoc, B., Clark, A., et al. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pp. 2206–2240. PMLR, 2022.
- Budzianowski, P., Wen, T.-H., Tseng, B.-H., Casanueva, I., Ultes, S., Ramadan, O., and Gašić, M. Multiwoz—a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. *arXiv preprint arXiv:1810.00278*, 2018.
- Chen, W., Chen, J., Qin, P., Yan, X., and Wang, W. Y. Semantically conditioned dialog response generation via hierarchical disentangled self-attention. *arXiv preprint arXiv:1905.12866*, 2019.
- Cheng, X., Gao, S., Liu, L., Zhao, D., and Yan, R. Neural machine translation with contrastive translation memories. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 3591–3601, 2022.
- Cheng, X., Luo, D., Chen, X., Liu, L., Zhao, D., and Yan, R. Lift yourself up: Retrieval-augmented text generation with self memory. *arXiv preprint arXiv:2305.02437*, 2023.
- Cholakov, R. and Kolev, T. Efficient task-oriented dialogue systems with response selection as an auxiliary task. *arXiv preprint arXiv:2208.07097*, 2022a.
- Cholakov, R. and Kolev, T. Efficient task-oriented dialogue systems with response selection as an auxiliary task. *arXiv preprint arXiv:2208.07097*, 2022b.
- Ding, N., Qin, Y., Yang, G., Wei, F., Yang, Z., Su, Y., Hu, S., Chen, Y., Chan, C.-M., Chen, W., et al. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 5(3):220–235, 2023.
- Dong, Q., Li, L., Dai, D., Zheng, C., Wu, Z., Chang, B., Sun, X., Xu, J., and Sui, Z. A survey for in-context learning. *arXiv preprint arXiv:2301.00234*, 2022.
- Du, X., Sun, Y., Zhu, J., and Li, Y. Dream the impossible: Outlier imagination with diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024.
- Dubois, Y., Li, X., Taori, R., Zhang, T., Gulrajani, I., Ba, J., Guestrin, C., Liang, P., and Hashimoto, T. B. Alpacafarm: A simulation framework for methods that learn from human feedback, 2023.
- Feng, Y., Yang, S., Zhang, S., Zhang, J., Xiong, C., Zhou, M., and Wang, H. Fantastic rewards and how to tame them: A case study on reward learning for task-oriented dialogue systems. In *The Eleventh International Conference on Learning Representations*, 2022.
- Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., and Wang, H. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023.
- Ghosal, S. S., Sun, Y., and Li, Y. How to overcome curse-of-dimensionality for out-of-distribution detection? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19849–19857, 2024.
- He, W., Dai, Y., Zheng, Y., Wu, Y., Cao, Z., Liu, D., Jiang, P., Yang, M., Huang, F., Si, L., et al. Galaxy: A generative pre-trained model for task-oriented dialog with semi-supervised learning and explicit policy injection. In

- Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 10749–10757, 2022.
- Hingane, A. Llm apps: Why you must know semantic router in 2024: Part 1. *Medium*, January 2024. URL <https://medium.com/@learn-simplified/llm-apps-why-you-must-know-semantic-router-in-2024-part-1-bfbda81374c5>.
- Horsey, J. Semantic router superfast decision layer for llms and ai agents, January 2024. URL <https://www.geeky-gadgets.com/semantic-router-superfast-decision-layer-for-llms-and-ai-agents/>.
- Hussain, S., Ameri Sianaki, O., and Ababneh, N. A survey on conversational agents/chatbots classification and design techniques. In *Web, Artificial Intelligence and Network Applications: Proceedings of the Workshops of the 33rd International Conference on Advanced Information Networking and Applications (WAINA-2019)* 33, pp. 946–956. Springer, 2019.
- Jain, M., Kumar, P., Kota, R., and Patel, S. N. Evaluating and informing the design of chatbots. In *Proceedings of the 2018 designing interactive systems conference*, pp. 895–906, 2018.
- Janssen, A., Passlick, J., Rodríguez Cardona, D., and Breitner, M. H. Virtual assistance in any context: A taxonomy of design elements for domain-specific chatbots. *Business & Information Systems Engineering*, 62:211–225, 2020.
- Kang, M., Kwak, J. M., Baek, J., and Hwang, S. J. Knowledge graph-augmented language models for knowledge-grounded dialogue generation. *arXiv preprint arXiv:2305.18846*, 2023.
- Khattab, O., Santhanam, K., Li, X. L., Hall, D., Liang, P., Potts, C., and Zaharia, M. Demonstrate-search-predict: Composing retrieval and language models for knowledge-intensive nlp. *arXiv preprint arXiv:2212.14024*, 2022.
- Kleene, S. C., Shannon, C. E., and McCarthy, J. Automata studies. *Princeton, NJ*, 1956.
- Kulhánek, J., Hudeček, V., Nekvinda, T., and Dušek, O. Augpt: Auxiliary tasks and data augmentation for end-to-end dialogue with pre-trained language models. *arXiv preprint arXiv:2102.05126*, 2021.
- Lee, Y. Improving end-to-end task-oriented dialog system with a simple auxiliary task. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 1296–1303, 2021.
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktaschel, T., et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.
- Li, X. and Qiu, X. Finding support examples for in-context learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 6219–6235, 2023.
- Medeiros, L., Gerritsen, C., and Bosse, T. Towards humanlike chatbots helping users cope with stressful situations. In *Computational Collective Intelligence: 11th International Conference, ICCCI 2019, Hendaye, France, September 4–6, 2019, Proceedings, Part I 11*, pp. 232–243. Springer, 2019.
- Ming, Y., Cai, Z., Gu, J., Sun, Y., Li, W., and Li, Y. Delving into out-of-distribution detection with vision-language representations. *Advances in neural information processing systems*, 35:35087–35102, 2022a.
- Ming, Y., Sun, Y., Dia, O., and Li, Y. How to exploit hyper-spherical embeddings for out-of-distribution detection? In *The Eleventh International Conference on Learning Representations*, 2022b.
- Minsky, M. Some universal elements for finite automata. *Automata studies*, 34:117–128, 1956.
- Modarressi, A., Imani, A., Fayyaz, M., and Schütze, H. RetIlm: Towards a general read-write memory for large language models. *arXiv preprint arXiv:2305.14322*, 2023.
- OpenAI. Gpt-4 technical report. 2023.
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.
- Ramachandran, G. S., Hashimoto, K., and Xiong, C. Causal-aware safe policy improvement for task-oriented dialogue. *arXiv preprint arXiv:2103.06370*, 2021.
- Safi, Z., Abd-Alrazaq, A., Khalifa, M., and Househ, M. Technical aspects of developing chatbots for medical applications: scoping review. *Journal of medical Internet research*, 22(12):e19127, 2020.
- Selvi, J. Exploring overfitting risks in large language models. <https://research.nccgroup.com/2023/05/22/exploring-overfitting-risks-in-large-language-models/>, May 2023.
- Shuster, K., Poff, S., Chen, M., Kiela, D., and Weston, J. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 3784–3803, 2021.

- Sisodia, Y. Stop chat-gpt from going rogue in production with semantic router. *Medium*, January 2024. URL <https://medium.com/@scholarly360/stop-chat-gpt-from-going-rogue-in-production-with-semantic-router-937a4768ae19>.
- Su, Y., Shu, L., Mansimov, E., Gupta, A., Cai, D., Lai, Y.-A., and Zhang, Y. Multi-task pre-training for plug-and-play task-oriented dialogue system. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4661–4676, 2022.
- Sun, Y. *Detecting and Learning Out-of-distribution Data in the Open-world: Algorithm and Theory*. The University of Wisconsin-Madison, 2023.
- Sun, Y., Shi, Z., Liang, Y., and Li, Y. When and how does known class help discover unknown ones? provable understanding through spectral analysis. *arXiv preprint arXiv:2308.05017*, 2023.
- Sun, Y., Shi, Z., and Li, Y. A graph-theoretic framework for understanding open-world semi-supervised learning. *Advances in Neural Information Processing Systems*, 36, 2024.
- Tseng, B.-H., Dai, Y., Kreyssig, F., and Byrne, B. Transferable dialogue systems and user simulators. *arXiv preprint arXiv:2107.11904*, 2021.
- Wang, B., Chen, W., Pei, H., Xie, C., Kang, M., Zhang, C., Xu, C., Xiong, Z., Dutta, R., Schaeffer, R., et al. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. *Advances in Neural Information Processing Systems*, 2023a.
- Wang, J., Zhang, Y., Kim, T.-K., and Gu, Y. Modelling hierarchical structure between dialogue policy and natural language generator with option framework for task-oriented dialogue system. *arXiv preprint arXiv:2006.06814*, 2020a.
- Wang, K., Tian, J., Wang, R., Quan, X., and Yu, J. Multi-domain dialogue acts and response co-generation. *arXiv preprint arXiv:2004.12363*, 2020b.
- Wang, S., Zhao, Z., Ouyang, X., Wang, Q., and Shen, D. Chatcad: Interactive computer-aided diagnosis on medical image using large language models. *arXiv preprint arXiv:2302.07257*, 2023b.
- Wang, X., Yang, Q., Qiu, Y., Liang, J., He, Q., Gu, Z., Xiao, Y., and Wang, W. Knowledgpt: Enhancing large language models with retrieval and storage access on knowledge bases. *arXiv preprint arXiv:2308.11761*, 2023c.
- Yang, J., Wang, P., Zou, D., Zhou, Z., Ding, K., Peng, W., Wang, H., Chen, G., Li, B., Sun, Y., et al. Openood: Benchmarking generalized out-of-distribution detection. *Advances in Neural Information Processing Systems*, 35: 32598–32611, 2022.
- Yasunaga, M., Aghajanyan, A., Shi, W., James, R., Leskovec, J., Liang, P., Lewis, M., Zettlemoyer, L., and Yih, W.-t. Retrieval-augmented multimodal language modeling. *arXiv preprint arXiv:2211.12561*, 2022.
- Yu, X., Wu, Q., Qian, K., and Yu, Z. Krsls: Improving end-to-end response generation in task oriented dialog with reinforced keywords learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12338–12358, 2023.
- Zhang, J., Yang, J., Wang, P., Wang, H., Lin, Y., Zhang, H., Sun, Y., Du, X., Zhou, K., Zhang, W., et al. Openood v1.5: Enhanced benchmark for out-of-distribution detection. *arXiv preprint arXiv:2306.09301*, 2023.
- Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*, 2023.

## A. Prompts

**Prompts used for extracting tags.** This prompt template is used in Section 3.2.1 to make sure the resulting tags are both concise and relevant to the conversational content.

```

1 # Task Description
2
3 You are helping me compress the following dialog with customer service into the following
4 form:
5
6 <id> <User/System>: <compressed phrase>
7
8 You will have to follow several principles:
9 1. Please use words as few as possible, ideally no more than 3 words.
10 2. The summarization needs to focus on the actual events/issues/queries/solutions.
11
12 # Example
13
14 Input:
15 "0 User: What is going on with my keyboard... fix it"
16 Output:
17 "0 User: #keyboard #issue"
```

**Prompts used for generating response.** This prompt template is used in Section 3.3 to generate the response given the example dialogues.

```

1 # Task Description
2 You are a helpful service agent. Please help me fill in the system response in a dialogue.
3 Please note that key information is encoded in the dialogue.
4
5 The dialogue is with the format:
6
7 [ID] [USER/SYSTEM]: [UTTERANCE]
8
9 Here is a list of related example dialogues you can use for reference.
10
11 {examples}
12
13 # Remarks:
14
15 1. Please directly generate the completed dialogue according to the format in the example.
16 2. (**IMPORTANT**) Please make sure the generated utterance ID is consistent with the
   original input!
```

**Prompts used for evaluation.** This prompt template is used in Section 4.1 to evaluate the response quality and win rate calculation. This prompt is modified from AlpacaEval (Dubois et al., 2023).

```

1
2 I'll provide you with task prompts given to these models and their corresponding outputs.
3 Your task is to assess these responses, and select the model that produces the output
4 that is most smooth and consistent with the ground truth dialogue.
5 Please note that it is very important for model to provide response in a **similar
   style and content**.
6
7 ## Instruction
8
9 {
10     "instruction": """Please act as a helpful customer service agent and complete the
   following dialogue: """,
11     "input":
12     """
13         {task_input}
```

```

13     """
14     "ground truth answer": """
15     {raw_diag}
16     """
17 """
18 }
19
20 ## Model Outputs
21
22 Here are the unordered outputs from the models. Each output is associated with a
23 specific model, identified by a unique model identifier.
24
25 {
26   {
27     "model_identifier": "m",
28     "output": """
29 {pred_cmp_diag}
30     """
31   {
32     "model_identifier": "M",
33     "output": """
34 {pred_diag}
35     """
36   }
37 }
38
39 ## Task
40
41 Evaluate the models based on the quality and relevance of their outputs, and select
42 the model that generated the best output.
43 Answer by first providing a concise explanation and then end your answer by providing
44 the model identifier of the best output.
45 We will use the last character of your output `output[-1]` as the name of the best
46 model, so make sure you finish with the token of the model identifiers and nothing
47 else: 'm' or 'M' (no quotes, no dots, no backticks, no new lines, ...).
48 For example:
49
50   ### Concise explanation
51   ...some text...
52
53   ### Which is best, m or M?
54   M
55
56 Now is your turn.
57
58 ## Your answer: "Concise explanation" followed by "Which is best, m or M?"
```

## B. Experiment Details

In this section, we delve into the specifics of implementing the DFA learning process from training data, as outlined in Section 3.2, and the intricacies of conversation generation, discussed in Section 3.3.

**A threshold for determining the “splittable” states.** Figure 3 illustrates the algorithm for constructing the DFA-tree. However, the practical application of this algorithm involves several nuanced considerations. One key aspect is the handling of nodes/states that correspond to only a limited number of dialogues in the training dataset. Constructing a sub-tree under such nodes/states may not yield significant insights due to their limited number of tracked dialogues. Therefore, the tree-building function, as depicted in Figure 3, is only pursued if the number of tracked dialogue IDs at a node/state exceeds a certain threshold:  $|\mathcal{I}(q)| > \tau$ . For our experiments, we have set this threshold,  $\tau$ , to 5.

**State-merging threshold.** In Section 3.2.3, we illustrate that once the similarity score is calculated, we set a threshold  $\lambda$  to determine which node pairs should be merged. In practice, we set  $\lambda = 0.1$  and by observation, 5% of nodes get merged

with the threshold set.

**Examples limits in In-Context Learning (ICL).** In alignment with the practice in in-context learning (Li & Qiu, 2023; Dong et al., 2022), we limit the maximum number of examples used in in-context learning to 5. This means that if the number of tracked dialogue IDs for a state, denoted as  $|\mathcal{I}(q)|$ , exceeds 5, we randomly select 5 dialogues from  $|\mathcal{I}(q)|$  for further processing.

## C. Dialogue Generation Results Demonstrations

In this section, we conduct a qualitative analysis of dialogue generation outcomes using the baseline algorithms outlined in Table 1. The dialogues are derived from the AirbnbHelp dataset, with all baseline models utilizing GPT-3.5 as the underlying large language model (LLM) for generation. Our objective is to compare each model's output with the ground truth dialogue from the dataset, aiming for generated dialogues that closely mirror the original interactions. Additionally, we include insights from a GPT-4 evaluator to shed light on the relative performance of each model.

### C.1. Ground Truth Dialogues

1. User: I just received a mail asking me to give feedback on a trip that I didn't make. Error at your end or has my account been hacked?
2. System: Hey Nick, thank you for bringing this to our attention. We've escalated a case to the appropriate team and a Case Manager will follow up with you for further assistance. We appreciate your patience in the meantime.
3. User: I presume you got my message via the Help Center contact form?
4. System: Yes, we've also noticed that your case manager has just sent you an email with all the instructions to recover your account. Please let us know if you need further assistance. Thanks.
5. User: Well, I changed my password, and my account seems to be working. I canceled the trip and it said I would get a refund, but whoever made the trip didn't use my credit card, so I didn't pay!
6. System: Hey Nick, we've looked into it and everything looks good. You shouldn't have to upload your ID again, as we see your account is still fully verified. Please let us know if there's anything else we can do for you. Thank you.

### C.2. GPT-3.5 Generation Results

Below, we display the dialogue completions generated by GPT-3.5, without using any additional prompts:

1. User: I just received a mail asking me to give feedback on a trip that I didn't make. Error at your end or has my account been hacked?
2. System: It seems like there might have been some misunderstanding. I'll need to investigate this further. **Could you please provide me with the email address that received the feedback request?**  
**(Comment from GPT-4: It asks for the email address from the user, although presumably, the system should already have this information since the user received an unwanted email from the service.)**
3. User: I presume you got my message via the Help Center contact form?
4. System: Yes, I received your message through the Help Center contact form. Thank you for reaching out to us that way.
5. User: Well, I changed my password, and my account seems to be working. I canceled the trip and it said I would get a refund, but whoever made the trip didn't use my credit card, so I didn't pay!
6. System: I'm glad to hear that you changed your password and that your account is working. Our team will investigate the unauthorized trip and ensure that you are not charged for it. Rest assured that we will take care of the refund process for you.

These results, while informative, reveal certain limitations in capturing the nuances of the original conversation, as will be discussed in the evaluator's comments.

### C.3. DFA-RAG Generation Results (Ours)

We next present results from the DFA-RAG approach. The GPT-4 evaluator offers more favorable comments for these outputs, highlighting an improved alignment with the ground truth dialogue:

1. User: I just received a mail asking me to give feedback on a trip that I didn't make. Error at your end or has my account been hacked?
2. System: Hi there, it sounds like there may have been some confusion with your account. We'll investigate this further and get back to you with more information. Thank you for bringing this to our attention.
3. User: I presume you got my message via the Help Center contact form?
4. System: Our team is currently reviewing your message and will provide you with an update as soon as possible. Thank you for your understanding.
5. User: Well, I changed my password, and my account seems to be working. I canceled the trip and it said I would get a refund, but whoever made the trip didn't use my credit card, so I didn't pay!
6. System: I'll ensure that your account is thoroughly checked to confirm its security and functionality. We appreciate your proactive steps, and we'll keep you updated on any developments.

(Comment from GPT-4: This model does not ask for additional information that it should already have and instead focuses on confirming that the issue will be investigated and the user will be updated. This is closer to the "ground truth" response mentioned in the instruction, where the system acknowledges the receipt of the message and provides assurance without requesting information it should already possess. In addition, this model uses language that is more in line with the professional customer service manner shown in the ground truth dialogue, such as thanking the user multiple times and indicating appreciation of the user's proactive measures.)

### C.4. RAG-Based Generation Results

Results from the RAG (Retrieval-Augmented Generation) model are also presented. This model retrieves dialogues that closely match the context in the embedding space, using these selections as prompts for generation. However, the generated responses often inaccurately focus on requesting the user's email, failing to directly address the user's initial request:

1. User: I just received a mail asking me to give feedback on a trip that I didn't make. Error at your end or has my account been hacked?
2. System: Hi there, we're sorry to hear about this. Can you please DM us with your account email so we can look into this for you and provide further assistance? Thank you.
3. User: I presume you got my message via the Help Center contact form?
4. System: Thanks for confirming that. We'll review your account and ensure everything is in order. If you have any further concerns, please don't hesitate to reach out to us.
5. User: Well, I changed my password, and my account seems to be working. I canceled the trip and it said I would get a refund, but whoever made the trip didn't use my credit card, so I didn't pay!
6. System: We understand your concern and will definitely look into it for you. Please DM us your email address so we can verify your account and address any potential security issues. Thank you.

(Comment from GPT-4: It did not address the final user query. Furthermore, the model started well by asking for a Direct Message (DM) for privacy, it did not fully reassure the user regarding the data breach or provide clear next steps concerning the unauthorized trip.)

### C.5. RandSample Generation Results

Finally, we discuss outcomes from the RandSample method, which randomly selects dialogues from the training dataset for prompt generation. This approach sometimes leads to generated dialogues that contain hallucinated content, such as mismatched usernames, illustrating the pitfalls of random selection:

1. User: I just received a mail asking me to give feedback on a trip that I didn't make. Error at your end or has my account been hacked?
2. System: We appreciate you reaching out, **Simona**. It seems there may have been a mix-up. Please DM us your email address so we can look into this further for you.
3. User: I presume you got my message via the Help Center contact form?
4. System: We understand your concern, **Simona**. We assure you that we will investigate this and get back to you with an update.
5. User: Well, I changed my password, and my account seems to be working. I canceled the trip and it said I would get a refund, but whoever made the trip didn't use my credit card, so I didn't pay!
6. System: We'll definitely look into this for you, **Simona**. Please DM us your email address so we can check the status of your account and the refund for the canceled trip. Thank you.

**(Comment from GPT-4: This model consistently addresses the user by the name “Simona,” which does not align with the ground truth dialogue where the user is referred to as Nick. )**

Each of these methodologies offers unique insights into the challenges and opportunities present in dialogue generation. Through the lens of the GPT-4 evaluator’s comments, we gain a nuanced understanding of how each model performs relative to the complexities of real-world conversation dynamics.

## D. More DFA Results Demonstrations

We include the constructed DFA for each dataset in Figure 6, Figure 7, Figure 8, Figure 9, Figure 10, Figure 11, we have constrained the maximum depth to four levels due to spatial constraints, and have chosen to display only those nodes that correspond to a minimum of ten dialogues in the training set, as indicated by  $|\mathcal{I}(q)| > 10$ . It’s important to note that, owing to the limitations of our tree-like visualization tool, the ‘looping’ structures typical of DFAs, characterized by nodes with multiple incoming edges, are not depicted here. Additionally, for clarity in this visualization, tags are shown attached to the nodes; however, in the actual DFA model, these tags are associated with the edges leading to each state/node. This visualization is intended to provide an overarching view of the DFA’s structure. Readers should be mindful of these distinctions to accurately understand and interpret the DFA results.

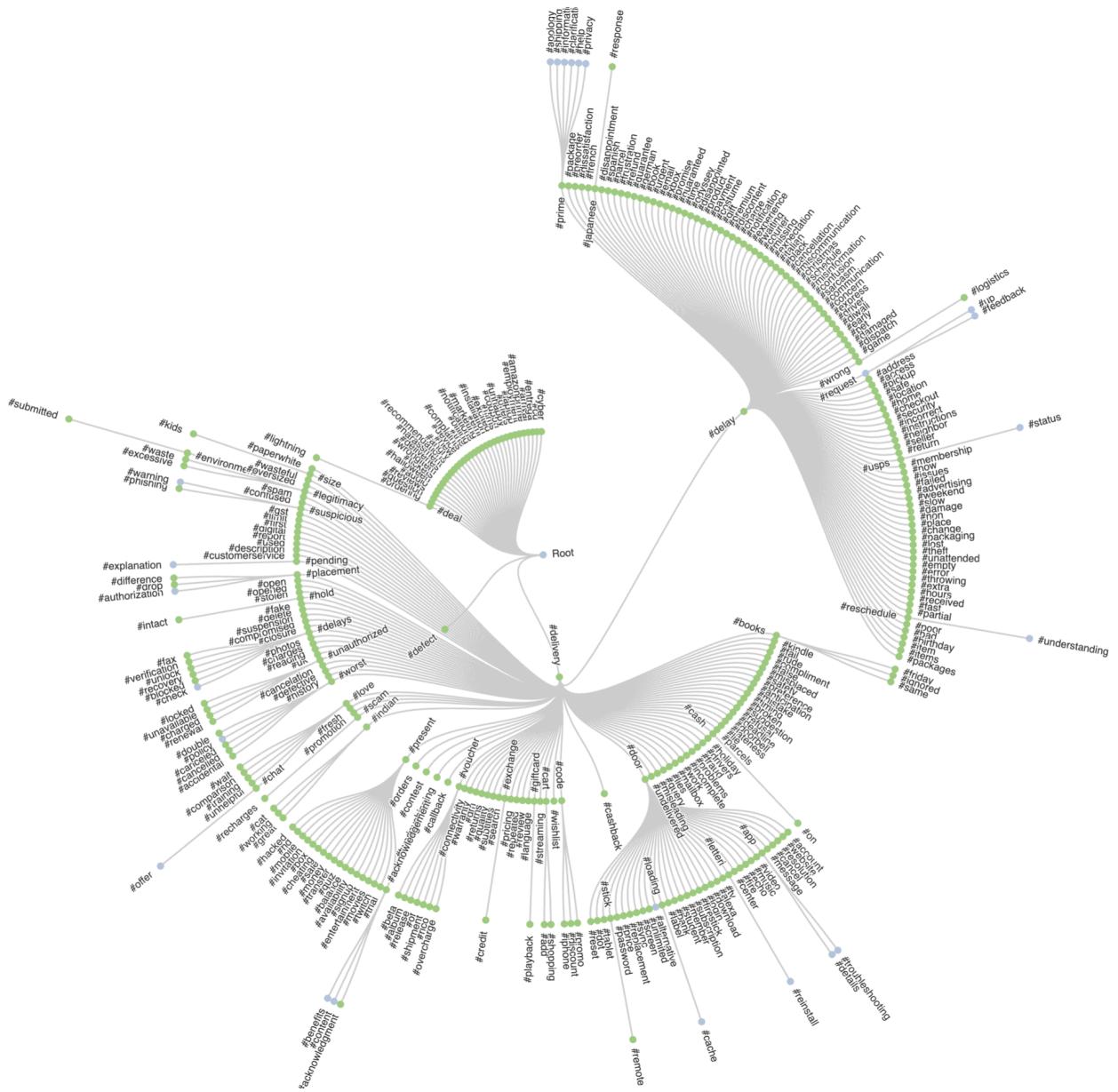
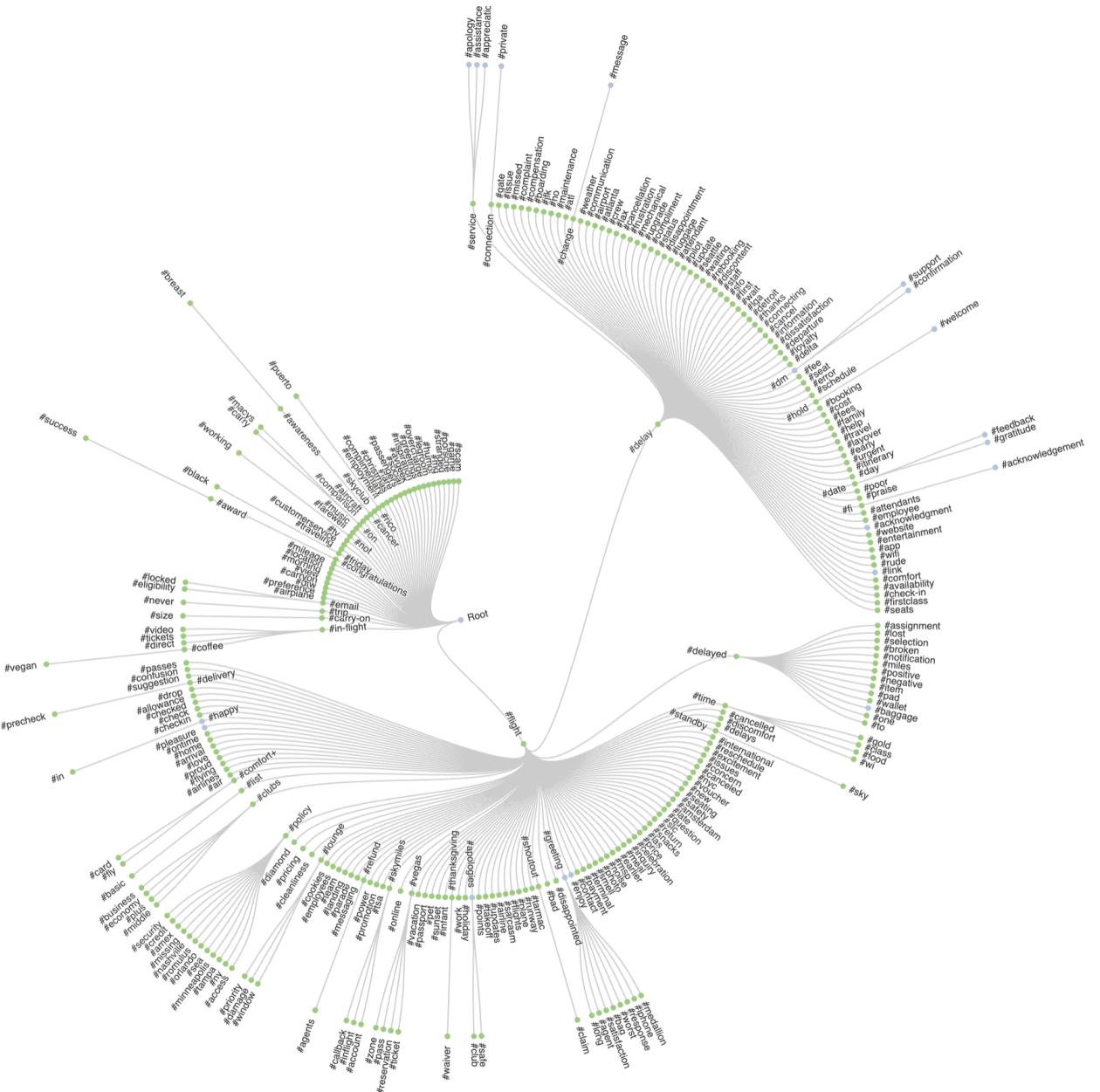
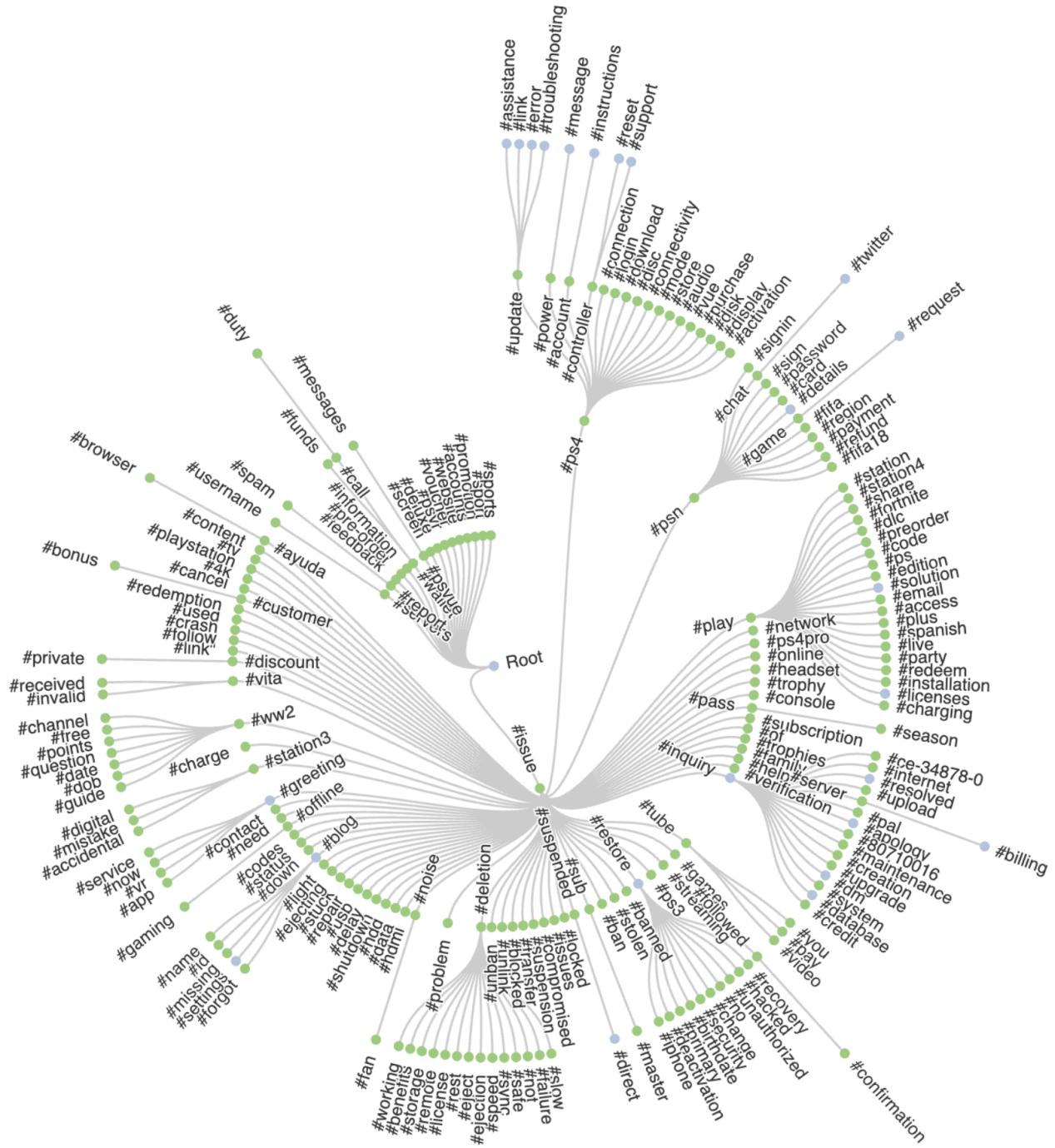


Figure 6. DFA Results Demonstration for AmazonHelp.



*Figure 7. DFA Results Demonstration for DeltaSupport.*



*Figure 8. DFA Results Demonstration for AskPlayStation.*

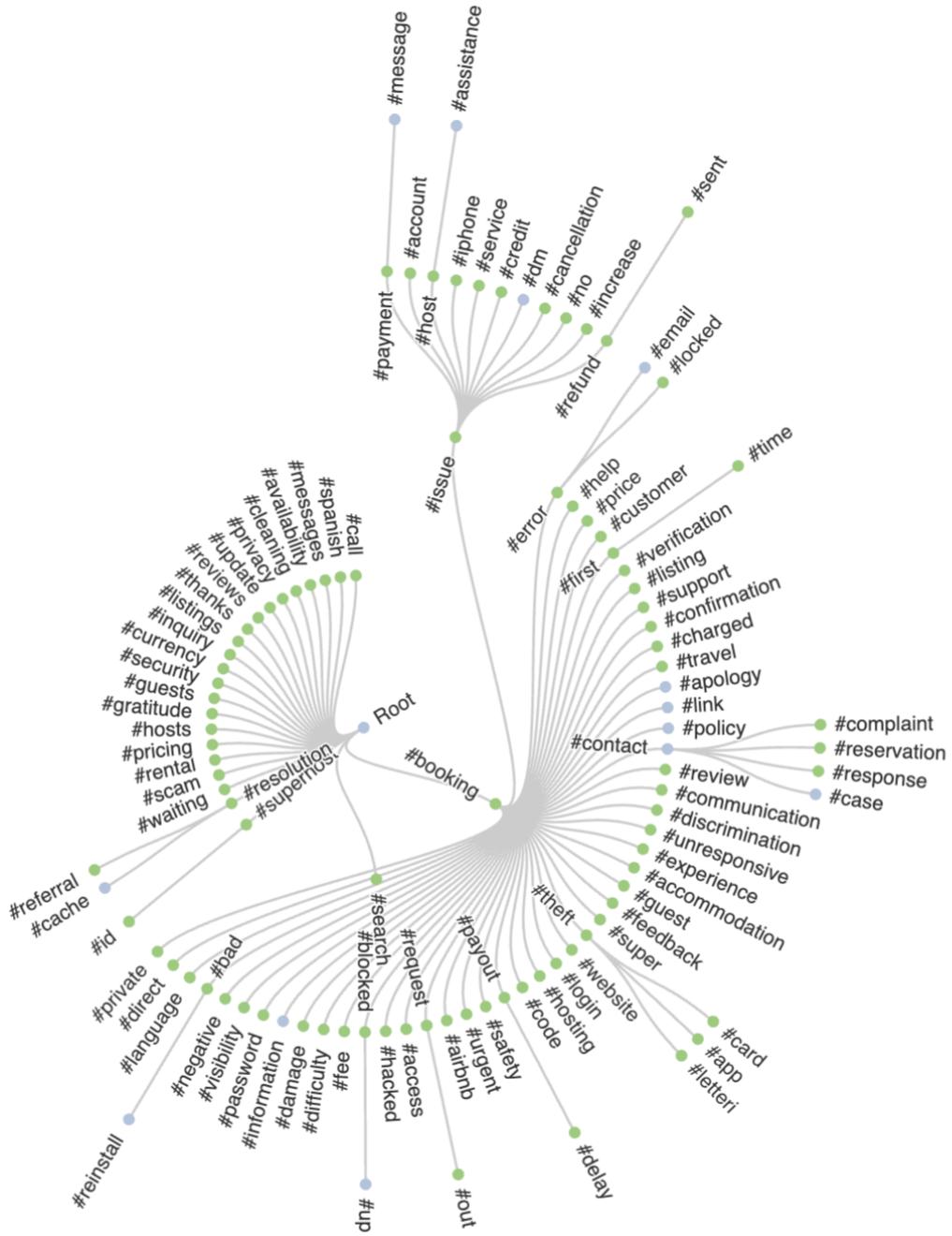


Figure 9. DFA Results Demonstration for AirbnbHelp.

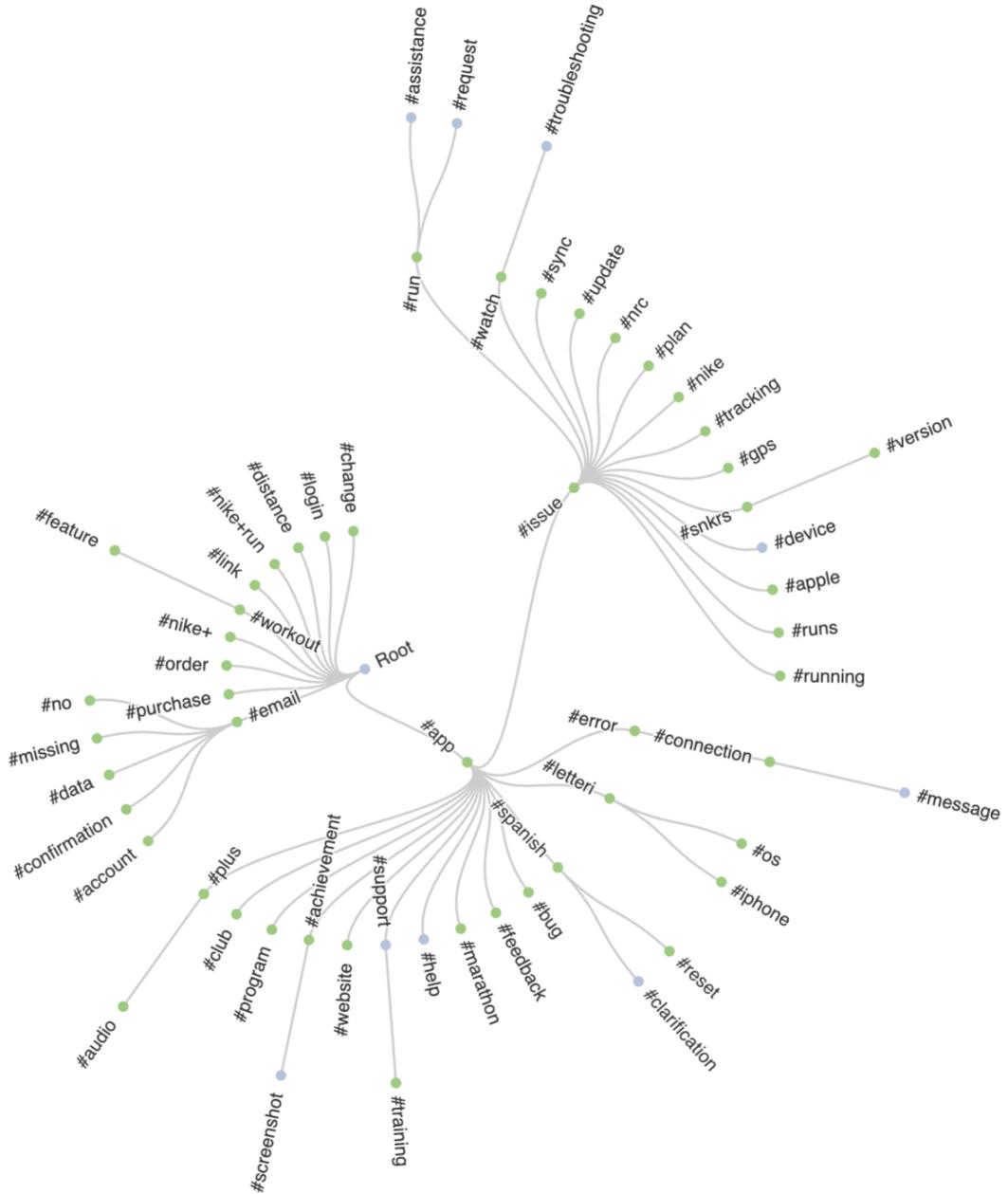


Figure 10. DFA Results Demonstration for NikeSupport.



**Figure 11.** DFA Results Demonstration for CambridgeInfo/MultiWOZ. We extend the depth limitation to eight in the demo due to the complicated branching in this demo.