

Simulating Task-Oriented Dialogues with State Transition Graphs and Large Language Models

Chris Samarinas*, Pracha Promthaw*, Atharva Nijasure*,
Hansi Zeng, Julian Killingback, Hamed Zamani

Center for Intelligent Information Retrieval
University of Massachusetts Amherst
United States

{csamarinas,ppromthaw,anijasure,hzeng,jkillingback,zamani}@cs.umass.edu

Abstract

This paper explores SynTOD, a new synthetic data generation approach for developing end-to-end Task-Oriented Dialogue (TOD) Systems capable of handling complex tasks such as intent classification, slot filling, conversational question-answering, and retrieval-augmented response generation, without relying on crowdsourcing or real-world data. SynTOD utilizes a state transition graph to define the desired behavior of a TOD system and generates diverse, structured conversations through random walks and response simulation using large language models (LLMs). In our experiments, using graph-guided response simulations leads to significant improvements in intent classification, slot filling and response relevance compared to naive single-prompt simulated conversations. We also investigate the end-to-end TOD effectiveness of different base and instruction-tuned LLMs, with and without the constructed synthetic conversations. Finally, we explore how various LLMs can evaluate responses in a TOD system and how well they are correlated with human judgments. Our findings pave the path towards quick development and evaluation of domain-specific TOD systems. We release our datasets, models, and code for research purposes.

1 Introduction

Task-Oriented Dialogue (TOD) systems have become increasingly popular in various domains, such as customer support, personal assistants, e-commerce, and enterprise. These systems aim to assist users in accomplishing specific goals through natural language conversations. However, building effective TOD systems requires large amounts of diverse and high-quality training data, which can be expensive and time-consuming to collect (Zhang et al., 2020). Moreover, existing datasets often lack the complexity and richness needed to develop sophisticated TOD systems that can handle a wide range of user intents, perform slot filling, address information seeking requests, and generate contextually relevant and appealing responses (Mendonça et al., 2023; Zamani et al., 2023).

Recent advancements in large language models (LLMs) have shown promising results in various NLP tasks. However, LLMs alone are not sufficient for building effective TOD systems, as they require a set of diverse task-specific training data to perform well in specialized domains (Hudecek & Dusek, 2023). We argue that trivial synthetic data generation approaches using LLMs (for example through single prompt engineering) may not provide the diverse training data required for training robust TOD systems.

To address these challenges, we introduce SynTOD, a simple yet effective framework that takes a *state transition graph* for the target TOD system and translates it into a set of synthetically generated task-oriented dialogues using LLMs. The state transition graph defines

*Equal contribution

the desired behavior of the TOD system and guides the generation of diverse, structured conversations. This approach allows for fine-grained control over the conversation structure, ensuring that the generated data covers a wide range of scenarios and edge cases. Based on our experiments, SynTOD leads up to 37% improvement in intent classification, 100% in slot filling and 30% in response relevance compared to naive single-prompt simulated conversations. By incorporating retrieval augmentation, SynTOD enables the development of TOD systems that can handle complex dialogues that involve navigation, search, result filtering, summarization, and question answering. In this work, we aim at answering the following research questions: **RQ1:** *How does data generation for TOD with a single prompt compare to graph-guided multi-prompt generation?* **RQ2:** *How do different LLMs trained on synthetic data perform on end-to-end TOD tasks?* **RQ3:** *How much synthetic training data is required to build an effective LLM-based TOD system?* and **RQ4:** *How do various LLMs evaluate response relevance in this setting, and are they correlated with human judgments?*

To address these research questions, we conduct extensive experiments in two new defined domains: cooking and e-commerce assistance. Our Synthetic datasets surpass existing TOD and conversational recommendation datasets in terms of feature richness and diversity, covering 7 different types of interactions compared to other datasets that have up to 4 (see Table 7), and allow us to train and evaluate both state tracking and various types of response generation, including attribute-related questions, comparison questions, open-domain questions, result summarization and clarifying questions. We evaluate the performance of various LLM-based TOD systems trained on our synthetic data using both automatic and human evaluations. While there exist other approaches for generating synthetic data for TOD systems, our focus on fully automated graph-guided multi-prompt generation using LLMs with minimal human input sets SynTOD apart. The only required human input is a transition graph and a set of response simulation prompts. Sampling of user intent paths using a state transition graph allows for more structured and diverse conversations. The main contributions of this work are as follows:

- We propose SynTOD, a simple framework for generating synthetic training data for end-to-end TOD systems using state transition graphs and LLMs.
- We demonstrate the effectiveness of SynTOD in two domains: cooking and e-commerce assistance, generating datasets that surpass existing TOD and conversational recommendation datasets in terms of feature richness and diversity.
- We conduct extensive experiments to evaluate the performance of LLM-based end-to-end TOD systems trained on our synthetic data, using both automatic and human evaluation, and provide insights into the four research questions posed above.
- We release two synthetic datasets generated using SynTOD, along with a curated test split that can serve as a proxy benchmark for comparing complex TOD systems. We also make our models and implementation publicly available.¹

By leveraging the power of LLMs and our synthetic data generation framework, SynTOD aims to simplify the development process of sophisticated TOD systems and enhance the user experience in real-world applications. Our approach offers a promising direction for building effective TOD systems in low-resource settings and specialized domains.

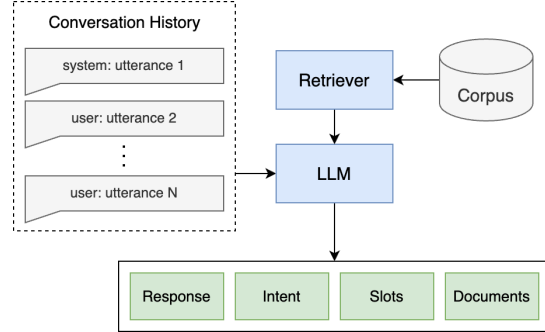


Figure 1: Overview of an end-to-end retrieval-augmented TOD system. A LLM and a retriever are the main components. A conversation history is given as input, and response, intent, slots and documents comprise the output system state.

¹Our data, models and code are available at <https://github.com/algoprog/SynTOD>.

2 Methodology

In this section, we describe SynTOD, our new framework for synthetic data generation for TOD systems. SynTOD is based on the creation of a state transition graph and the simulation of retrieval-augmented responses using LLMs. The primary goal in TOD is to facilitate a user in accomplishing a specific goal through a conversational interface. Effective TOD systems should be capable of understanding the user’s intent, filling in the necessary slots, and generating appropriate responses. Given a dialogue history $H = \{u_1, r_1, u_2, r_2, \dots, u_t\}$, where u_i and r_i represent the user’s and system’s utterances respectively at time i , a state transition graph G , and the parameters of an LLM M_θ , SynTOD should generate a system state S_t at time t : $S_t = \phi(H, G, M_\theta)$. S_t is a tuple of three components: the user’s intent I_t with related attributes (slots) V_t , and optionally some documents D_t that might be retrieved to help generate a response: $S_t = (I_t, V_t, D_t)$.

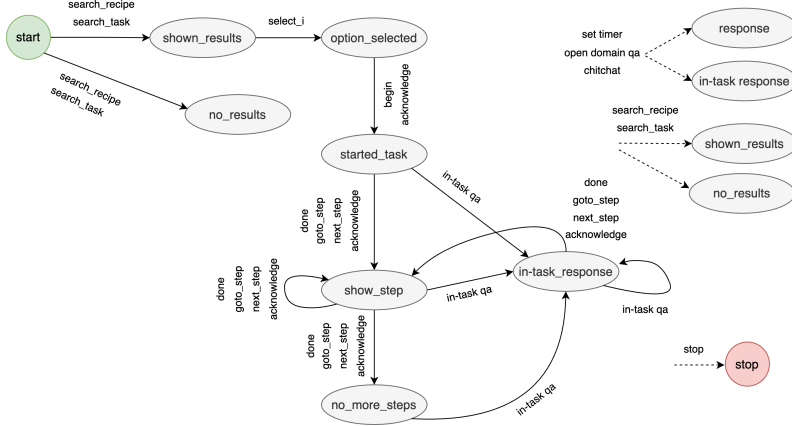


Figure 2: The state transition graph we defined for the recipe assistant domain. On the right we see transitions to nodes that are possible from any other state.

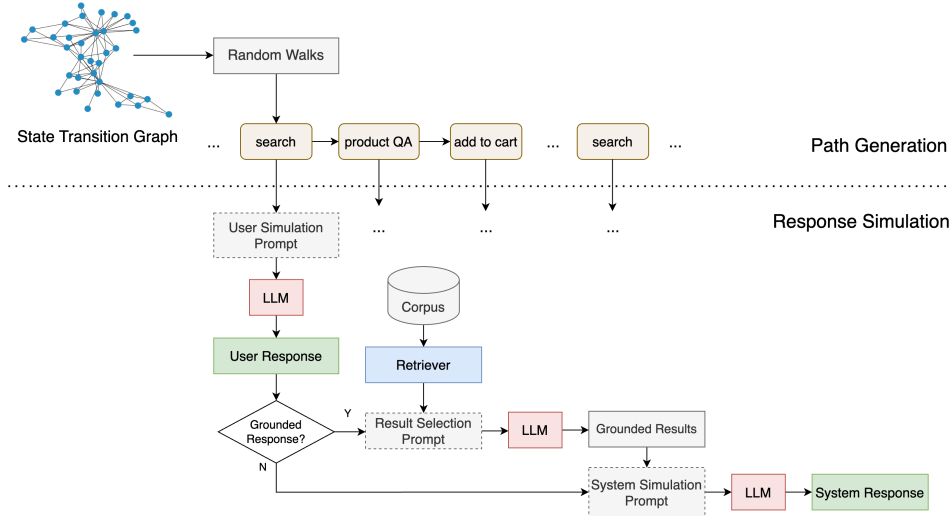


Figure 3: Overview of the SynTOD conversation simulation framework.

State Transition Graph The first step in our framework involves defining a state transition graph G . This graph encapsulates the desired behavior of the TOD system. The nodes in this graph represent the various states of the system, while the edges symbolize the user’s intents. The edges also have weights based on the transition probabilities between states. The probabilities can be defined empirically or based on some analysis of existing real

	Cooking Assistance		E-commerce Assistance	
	Train	Test	Train	Test
Total conversations	2000	300	2000	304
# utterances per conversation	35.89 (\pm 10.34)	29.08 (\pm 10.84)	23.19 (\pm 7.80)	22.30 (\pm 7.87)
# tokens per conversation	1,416 (\pm 386)	1,448 (\pm 594)	2,220 (\pm 934)	1,972 (\pm 815)
# tokens per user utterance	12.45 (\pm 7.18)	13.26 (\pm 9.06)	22.67 (\pm 14.45)	24.76 (\pm 23.26)
# tokens per system utterance	57.39 (\pm 26.65)	74.98 (\pm 36.28)	89.57 (\pm 54.78)	81.73 (\pm 53.86)

Table 1: Statistics of the graph-guided dataset for cooking and e-commerce domains.

conversations in a given domain. The graph is designed to be comprehensive, covering all possible states and transitions that the system may encounter during a conversation and serves as the backbone of the SynTOD framework, guiding the generation of synthetic conversations. An example of a state transition graph for the e-commerce assistance domain can be seen in Figure 2.

Data Generation The data generation process of SynTOD is summarized in Figure 3 and described in detail in Algorithm 1. We generate training examples T_i through a process f : $T_i = f(D_s, M, P, G)$, where D_s a collection of seed items/documents with relevant metadata, M a pre-trained LLM, P a set of response simulation prompts, one or more per user and system intent, and a state transition graph G . The training example T_i generated through the process f includes both the dialogue history H and the corresponding system state S_t for each time step t . The algorithm starts with a set of random walks generated based on the graph G . Each random walk represents a potential conversation path, with each node and edge corresponding to a system state and user intent, respectively. For each node and edge in these random walks, a large language model with a custom prompt is used to simulate a response. A pre-trained LLM is used to generate responses from either the system or the user, depending on the node or edge it is simulating. This process allows for the creation of diverse and realistic synthetic conversations that adhere to the structure defined by the state transition graph. In addition to simulating responses, the SynTOD framework also incorporates retrieval-augmentation. For example, when a user searches for items, a retriever fetches some items from a document corpus, and the LLM selects a subset of them and summarizes them for the user. To simulate the retrieval augmented responses, we use a corpus of items as seed, and then each seed item is treated as the ground truth target for selection and simulation of future responses. This grounded response generation ensures that the responses are contextually accurate and have limited hallucinations, further enhancing the realism and diversity of the synthetic conversations. Moreover, the algorithm also simulates various state variables, such as step numbers, lists, and queries. These state variables are randomly generated to add more variety to the conversations. In particular, queries are generated using special prompts that guide the LLM to create realistic search queries based on the target item metadata.

LLM Adaptation for end-to-end TOD Based on all the training examples $T_i \in T$, we learn the parameters θ of another LLM, denoted as M_θ , for generating the next system state in other real conversational contexts. For training we can employ parameter-efficient fine-tuning methods (PEFT) such as QLoRA (Detmers et al., 2023). Figure 1 shows the simplified architecture of an end-to-end TOD system powered by a single fine-tuned LLM (M_θ).

3 Experimental Setup

3.1 Datasets

For our experiments we generated datasets for two domains; cooking and e-commerce assistance. For each domain we have two splits, for training and testing. For the training split generation, we used GPT-4, while for the test split we used LLaMA 2 70B, Mistral Medium and Gemini Pro. For each domain, we generated 2000 conversations for training and 300 (100 from each model) for testing. The statistics of our synthetic dataset can be seen in Table 1. We can see that in both domains, the conversations tend to be long both in terms

Algorithm 1 SynTOD Conversation Simulation Framework

Require: State transition graph G , seed items D_s , pre-trained LLM M , response simulation prompts P , search intents I_s , retrieval model R , query generation prompt P_q

Define the state transition graph G

Generate a set of random walks W based on G

Initialize an empty set of training examples T

for each random walk $w \in W$ **do**

Sample a target document $d_t \in D_s$

Initialize dialogue history list H , list of system states S and monitored state variables V

for each node and edge $(n, e) \in w$ **do**

if previous edge $e_{prev} \in I_s$ **then**

Retrieve k relevant documents D_r for $V(query)$ using retrieval model R

$D \leftarrow D_r \cup \{d_t\}$

else

$D \leftarrow \emptyset$

end if

Generate system response r_t using M and V, D, d_t with prompt P_n

Update dialogue history $H \leftarrow H \cup \{r_t\}$

if $e \in I_s$ **then**

Generate search query q using M with prompt P_q and target item d_t

Update monitored state variables $V(query) = q$

end if

Generate user utterance u_t using M and V with prompt P_e

Update dialogue history $H \leftarrow H \cup \{u_t\}$

Update system state $S_t = (I_e, V, D)$

Update list of states $S \leftarrow S \cup \{S_t\}$

Update other monitored state variables in V (e.g. step counter, lists)

end for

Create training example $T_i = (H, S)$

$T \leftarrow T \cup \{T_i\}$

end for

return T

of conversation turns and total words, in contrast to existing conversational datasets that tend to be short in both aspects. While we used mostly commercial closed-source models to build our training and testing data with a small cost (around \$300 for recipe and \$500 for e-commerce), it might be possible to generate data of similar quality with recent open-source models such as Mixtral.

Cooking Assistance For the cooking assistance domain, we defined a system that helps the user find relevant recipes from a corpus, answers questions about them and is able to go through the instructions step by step. The system supports two types of queries, generic, when asking for general recipe recommendations based on occasion or preferences (e.g., *I want to make a dessert for Christmas*) and specific, when searching for some specific recipe (e.g., *I want to make fluffy pancakes*). We used 4000 recipes from the Tasty dataset as our corpus (Sener & Yao, 2019). The supported intents can be seen in Table 8 and an example conversation in Table 10 in the appendix.

E-commerce Assistance For the e-commerce assistance domain, we defined a system that helps users discover products from a corpus, compare, answer questions about them, add or remove them from the cart and finally buy the items in the cart. The system can handle two types of queries, generic and specific. Generic queries do not mention a specific product name (e.g., *I want to buy a smartphone*) and are followed by clarifying question from the system (e.g., *what brand would you like?*). To compile the product corpus, we combined the MAVE dataset (Yang et al., 2022) with Amazon reviews (He & McAuley, 2016), resulting in a dataset comprising 4000 products across 50 diverse categories such as *Shoes* and *Headphones & Headsets*. This dataset lacked critical information such as delivery locations. To address this, we introduced random location allocations for each product. Shopping often involves multiple simultaneous purchases or comparisons. Consequently, for the conversation simulations we included a primary product seed along with additional randomly selected products for comparison and cart additions. For cart additions we randomly sampled

products across different categories, and for comparison lists we kept the same category with the seed product. The definition of the supported intents in this domain can be seen in Table 9 and an example conversation in Table 11 in the appendix.

3.2 Training and Inference

Data Format For training an LLM for end-to-end TOD with our synthetic data, we convert the training examples to text format (similar to ChatML). In the conversations we have the following roles; system, user, suggestions, retriever and item information. The suggestions contain list of generated queries for LLM-assisted retrieval for generic queries (those queries are used to perform retrieval and append results in the retriever role), the retriever role contains result metadata such as item title, description and rating, and the item information role contains all the relevant metadata for a selected recipe or product. Each system and user utterance is separated in two segments. The first segment contains the text response while the second contains system state information in json format, which encompasses the current intent and related slots.

Fine-tuning We fine-tune the language models using the QLoRA (Dettmers et al., 2023) parameter-efficient method until convergence. The training is performed with a batch size of 2 and 8 gradient accumulation steps, resulting in a total effective batch size of 16. There are 1000 steps in total, and we evaluate the model every 200 steps. For the LoRA weight, we start with $r = 64$ and $\alpha = 16$ as the default setting. In the recipe domain, we conduct a hyperparameter search for the LoRA parameters, exploring the following variations: ($r = 128, \alpha = 16$), ($r = 64, \alpha = 32$), and ($r = 64, \alpha = 8$). The best-performing settings for the e-commerce domain are $\alpha = 32$ and $r = 64$ at the step 600 (70.59% of the training data), while for the recipe domain, $\alpha = 8$ and $r = 64$ at the steps 100 (11.76% of the training data), yield the best results.

Nearest-neighbor Intent Selection The intent prediction is performed by the LLM fine-tuned on our synthetic data. However, the generated intent text itself is not always consistent. For example, the "add_to_cart" intent could be generated as "ADD TO CART" or "select.i.add_to_cart". To address this issue and improve the stability of intent generation, we incorporate a dot product nearest neighbor approach on the generated intent,² mapping it to the closest available intent in our predefined set. This ensures that the model produces the expected intent it intended to generate.

4 Results and Analyses

RQ1: How does data generation with a single prompt compare to graph-guided multi-prompt generation? To address the first research question, we compared the diversity and coverage of the generated data using a single prompt versus graph-guided multi-prompt generation. The results demonstrate that graph-guided multi-prompt generation leads to more diverse data and better coverage of intents. Figure 4 illustrates that without the graph, some intents have near-zero frequency while others have higher frequency than desired. Additionally, using a single prompt results in less diverse utterances, as evidenced by the higher Self-BLEU (Zhu et al., 2018) scores (see Table 2). These findings highlight the importance of the state transition graph in enforcing the desired distribution of user intents and promoting utterance diversity. The model trained with data generated from the transition graph significantly outperforms the model trained without the graph in all intent classification and slot filling metrics, including micro precision, recall, and F1 score (see Table 3). It is worth noting that for slot filling evaluation, the metrics might be misleading to an extent, because in some cases the predicted slots are correct but with small variations (e.g. 'pancake recipe' instead of 'pancake'). Figures 5 and 6 in the appendix, further illustrate the improvement in accuracy across various intents when using graph-guided training data. These results demonstrate the effectiveness of the state transition graph in generating higher quality training data for intent classification and slot filling. Response relevance evaluation using human annotators, Mixtral and OpenChat 3.5 provides insights into the

²We used [all-MiniLM-L6-v2](#) as our embedding model

quality of the generated responses (see Table 4). We defined 3 relevance levels; 0 means the generated response is irrelevant or has major issues, 1 is relevant but worse than the ground truth and 2 means as good as the ground truth. For the human annotation, we asked 3 participants to give one of these labels for each response given the conversation history. We used majority voting and chose the lowest rating in case of ties. The results show that the model trained with graph-guided data achieves higher average relevance scores compared to the model trained without the graph, indicating the superiority of the graph-guided approach in generating high-quality responses.

Dataset	Graph	Self-BLEU
Cooking	✓	0.7850
	✗	0.8101
Ecommerce	✓	0.8123
	✗	0.8932

Table 2: Diversity of data generated with and without a graph.

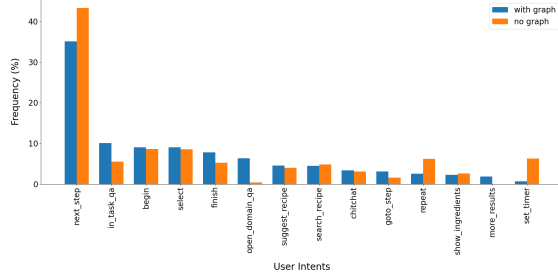


Figure 4: Frequency distribution of user intents for the recipe domain with and without state transition graph.

Metric	Cooking		E-commerce	
	no graph	with graph	no graph	with graph
intent accuracy	0.7620	0.9580*	0.6719	0.8296*
intent precision	0.7087	0.9344*	0.6534	0.8030*
intent recall	0.6591	0.9477*	0.6585	0.8435*
intent f1	0.6830	0.9410*	0.6560	0.8228*
slot precision	0.5682	0.7417*	0.3157	0.5693*
slot recall	0.6001	0.7343*	0.2789	0.5882*
slot f1	0.5561	0.7180*	0.2840	0.5690*

Table 3: Performance comparison of models for intent classification and slot filling with and without a state transition graph. Statistical significance is denoted by * ($p < 0.01$).

RQ2: How do different LLMs perform on end-to-end TOD? To investigate the second research question, we tested three different LLMs with similar architecture and number of parameters (7B) on end-to-end TOD tasks. The results, as shown in Table 5, indicate that there is no significant difference in performance across different models (Llama 2 vs. Mistral). Furthermore, instruction-tuning, also known as alignment, does not seem to have a significant effect on performance.

RQ3: How much synthetic training data is required? To address the third research question, we used the 2K generated examples (We saved 300 examples for validation and 1700 examples for training) for each domain and observed the model’s convergence using different proportions of the generated data. Our results showed that the model converges using just 11.76% of the examples for the recipe domain (200 examples) and 70.59% (1200 examples) for the e-commerce domain.

RQ4: How do various LLMs evaluate response relevance, and are they correlated with human judgments? To investigate the fourth research question, we tested two LLMs for automated response relevance evaluation (Mixtral and OpenChat 3.5) and compared their assessments with human judgments. We calculated the Cohen’s κ for each pair and we observed 17.14% for Mixtral and 10.34% for OpenChat in cooking domain and 18.29% for Mixtral and 14.9% for OpenChat in e-commerce domain. While Mixtral has much higher correlation than OpenChat with human judgments, it is still not high enough to be

considered reliable for automatic evaluation. However, we also calculated the agreement of GPT-4 annotations with human judgments and observed 64.29% for cooking domain and 43.31% for e-commerce. It is obvious that correlation of judgments increases with the number of LLM parameters, and GPT-4 or other models of that scale might be viable for automatic evaluation, however not easily reproducible.

Domain	Evaluator	Graph	Relevance (%)			Avg. Relevance
			0	1	2	
Cooking	Human	✗	7.86	7.79	84.35	1.7649
	Human	✓	3.56	7.12	89.23	1.8558* (+5.15%)
	Mixtral 8x7B Instruct	✗	0.62	13.25	86.14	1.8553
	Mixtral 8x7B Instruct	✓	0.34	3.43	96.22	1.9587* (+5.57%)
	OpenChat 3.5	✗	1.06	7.72	91.22	1.9016
	OpenChat 3.5	✓	0.62	1.96	97.43	1.9682* (+3.50%)
E-commerce	Human	✗	18.81	14.46	67.73	1.4792
	Human	✓	1.32	5.01	93.67	1.9235* (+30.04%)
	Mixtral 8x7B Instruct	✗	0.95	3.70	95.35	1.9440
	Mixtral 8x7B Instruct	✓	0.56	1.91	97.54	1.9699* (+1.33%)
	OpenChat 3.5	✗	0.67	6.57	92.76	1.9209
	OpenChat 3.5	✓	0.33	2.28	97.38	1.9704* (+2.58%)

Table 4: Evaluation results of responses from our fine-tuned TOD model based on Llama 2 7B, using synthetic data guided with and without a transition graph for the cooking and e-commerce domains. The superscript * denotes statistical significance compared to the setting without graph in terms of two-tailed paired t-test with $p < 0.01$.

Domain	Model	Intent			Slot		
		Prec.	Rec.	F1	Prec.	Rec.	F1
Cooking	Llama 2 7B	0.9499	0.9265	0.9211	0.7508*	0.7395*	0.7249*
	Mistral 7B	0.9498	0.9344	0.9248*	0.7231	0.7154	0.7002
	OpenChat 3.5	0.9471	0.9405*	0.9213	0.7362	0.7304	0.7134
E-commerce	Llama 2 7B	0.8212*	0.8348	0.8184	0.5518	0.5692	0.5503
	Mistral 7B	0.8163	0.8405*	0.8243*	0.6046*	0.6072*	0.5984*
	OpenChat 3.5	0.8099	0.8278	0.8220	0.5764	0.5814	0.5713

Table 5: Comparison of 3 different fine-tuned LLMs for end-to-end TOD using synthetic data guided with a state transition graph for the cooking and e-commerce domains. The superscript * denotes statistical significance in terms of two-tailed paired t-test with $p < 0.01$.

5 Related Work

TOD data can be collected and created through various approaches. The most straightforward method is to have users interact with an actual system (Williams et al., 2013; Raux et al., 2005; Gasic et al., 2014). However, this approach is limited by the capabilities of the existing system. The Wizard-of-Oz (WOZ) approach (Kelley, 1984) addresses this issue by having humans play the role of the system (Wen et al., 2017; Asri et al., 2017; Budzianowski et al., 2018; Andreas et al., 2020; Byrne et al., 2019), allowing for more complex simulated behavior at the cost of increased human effort and less control over system responses. To address the limitations of human-driven data collection, some works have employed automated approaches to simulate user-system interactions. These methods generate conversation outlines based on finite state machines (Shah et al., 2018), probabilistic automata (Rastogi et al., 2020), or heuristic rules (Lin et al., 2020). The outlines are then converted

Domain	Model	Relevance (%)			Avg. Relevance
		0	1	2	
Cooking	Llama 2 7B	0.34	3.43	96.22	1.9587
	Mistral 7B	0.19	3.41	96.40	1.9621
	OpenChat 3.5	0.16	3.66	96.18	1.9602
E-commerce	Llama 2 7B	0.56	1.91	97.54	1.9699
	Mistral 7B	0.77	2.23	97.00	1.9623
	OpenChat 3.5	0.95	2.39	96.66	1.9561

Table 6: Evaluation results of text responses from 3 different fine-tuned LLMs for end-to-end TOD in the cooking and e-commerce domains using synthetic data guided with a state transition graph and Mixtral as the evaluator.

into rough natural text using templates, which are further refined by crowd-source workers to enhance diversity and naturalness. A related approach (Acharya et al., 2021) uses seed dialogues that are expanded with a simulator to create dialogues, with the option of using crowd-source workers to add response diversity. Although more cost-effective than fully human-generated dialogues, these approaches still require human intervention and may suffer from unrealistic conversational flows.

Recently, efforts have been made to minimize human effort by leveraging generative models. Mohapatra et al. (2021) used crowd-worker data to train a GPT-2 (Radford et al., 2019) model to simulate both user and system roles. LLMs have also been employed via prompting and fine-tuning to produce TOD systems (Li et al., 2022; Zhang et al., 2023; Ulmer et al., 2024). Previous datasets primarily focus on canonical TOD scenarios where users seek assistance with specific needs or tasks. To better reflect real-world conversations, recent efforts have integrated other conversation types, such as open-domain conversation, into TOD datasets (Young et al., 2022). Some datasets also incorporate more challenging tasks or contexts, such as search and recommendation (Byrne et al., 2020) and document-grounded conversations (Feng et al., 2021). Traditionally, TOD systems have relied on multiple components for intent classification, slot filling, state tracking, and handcrafted rules (Qin et al., 2023). To simplify this complex architecture, end-to-end systems have been developed where all necessary functions are trained simultaneously. Early efforts used memory networks (Bordes et al., 2017) and sequence-to-sequence (Wen et al., 2017) approaches. More recently, neural language models and LLMs have been employed (Yang et al., 2021; Hudecek & Dusek, 2023; Cao, 2023; Chung et al., 2023).

6 Conclusion & Future Work

In this paper, we introduced SynTOD, a novel framework for generating synthetic data for end-to-end retrieval-augmented task-oriented dialogue systems. Our approach leverages state transition graphs to define the desired behavior of the system and LLMs to simulate diverse and realistic conversations. We demonstrated the effectiveness of SynTOD in two domains: recipe assistance and e-commerce assistance.

The synthetic data generated using our framework exhibited superior diversity compared to data generated from a single prompt. Furthermore, the models trained on the graph-guided synthetic data significantly outperformed those trained on data generated without the graph in both intent classification and slot filling tasks as well as response relevance. We tested three different LLMs on end-to-end TOD with our synthetic data, and we didn’t observe significant differences for models of the same scale (7B parameters), and instruction tuning does not seem beneficial. In terms of convergence, we observed that 200-1000 examples might be sufficient depending on the complexity of the chosen domain. For automatic response evaluation, LLMs with larger number of parameters have higher correlation with human judgements, and based on our experiments, only GPT-4 (or potentially models of similar size) could be considered as reliable as human annotators in zero-shot setting.

Our work showcases the potential of synthetic data generation for building complex, retrieval-augmented task-oriented dialogue systems without relying on expensive real-world data collection. The SynTOD framework simplifies the development process and enhances the user experience by enabling the creation of sophisticated systems capable of handling a wide range of features and tasks. To facilitate further research and development in this area, we release the synthetic datasets used in our experiments, along with a curated test split that can serve as a proxy benchmark for evaluating retrieval-augmented task-oriented dialogue systems. We believe that our contributions will pave the way for more advanced and user-centric conversational AI systems in the future.

Acknowledgements

This work was supported in part by the Center for Intelligent Information Retrieval, in part by the Amazon Alexa Prize Competition, in part by Adobe, in part by NSF grant #2143434, and in part by the Office of Naval Research contract #N000142212688. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

References

- Anish Acharya, Suranjit Adhikari, Sanchit Agarwal, Vincent Auvray, Nehal Belgamwar, Arjit Biswas, Shubhra Chandra, Tagyoung Chung, Maryam Fazel-Zarandi, Raefer Gabriel, Shuyang Gao, Rahul Goel, Dilek Hakkani-Tür, Jan Jezabek, Abhay Jha, Jiun-Yu Kao, Prakash Krishnan, Peter Ku, Anuj Goyal, Chien-Wei Lin, Qing Liu, Arindam Mandal, Angeliki Metallinou, Vishal Ishwar Naik, Yi Pan, Shachi Paul, Vittorio Perera, Abhishek Sethi, Minmin Shen, Nikko Strom, and Eddie Wang. Alexa conversations: An extensible data-driven approach for building task-oriented dialogue systems. In Avi Sil and Xi Victoria Lin (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations, NAACL-HLT 2021, Online, June 6-11, 2021*, pp. 125–132. Association for Computational Linguistics, 2021. doi: 10.18653/V1/2021.NAACL-DEMOS.15. URL <https://doi.org/10.18653/v1/2021.naacl-demos.15>.
- Jacob Andreas, John Bufe, David Burkett, Charles Chen, Josh Clausman, Jean Crawford, Kate Crim, Jordan DeLoach, Leah Dörner, Jason Eisner, Hao Fang, Alan Guo, David Hall, Kristin Hayes, Kellie Hill, Diana Ho, Wendy Iwaszuk, Smriti Jha, Dan Klein, Jayant Krishnamurthy, Theo Lanman, Percy Liang, Christopher H. Lin, Ilya Lintsbakh, Andy McGovern, Aleksandr Nisnevich, Adam Pauls, Dmitriy Petters, Brent Read, Dan Roth, Subho Roy, Jesse Rusak, Beth Short, Div Slomin, Ben Snyder, Stephon Striplin, Yu Su, Zachary Tellman, Sam Thomson, Andrei Vorobev, Izabela Witoszko, Jason Andrew Wolfe, Abby Wray, Yuchen Zhang, and Alexander Zotov. Task-oriented dialogue as dataflow synthesis. *Trans. Assoc. Comput. Linguistics*, 8:556–571, 2020. doi: 10.1162/TACL\A\00333. URL <https://doi.org/10.1162/tacl.a.00333>.
- Layla El Asri, Hannes Schulz, Shikhar Sharma, Jeremie Zumer, Justin Harris, Emery Fine, Rahul Mehrotra, and Kaheer Suleman. Frames: a corpus for adding memory to goal-oriented dialogue systems. In Kristiina Jokinen, Manfred Stede, David DeVault, and Annie Louis (eds.), *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue, Saarbrücken, Germany, August 15-17, 2017*, pp. 207–219. Association for Computational Linguistics, 2017. doi: 10.18653/V1/W17-5526. URL <https://doi.org/10.18653/v1/w17-5526>.
- Antoine Bordes, Y-Lan Boureau, and Jason Weston. Learning end-to-end goal-oriented dialog. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL <https://openreview.net/forum?id=S1Bb3D5gg>.
- Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In Ellen Riloff, David Chiang, Julia

- Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 5016–5026, Brussels, Belgium, October–November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1547. URL <https://aclanthology.org/D18-1547>.
- Bill Byrne, Karthik Krishnamoorthi, Chinnadhurai Sankar, Arvind Neelakantan, Ben Goodrich, Daniel Duckworth, Semih Yavuz, Amit Dubey, Kyu-Young Kim, and Andy Cedilnik. Taskmaster-1: Toward a realistic and diverse dialog dataset. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pp. 4515–4524. Association for Computational Linguistics, 2019. doi: 10.18653/V1/D19-1459. URL <https://doi.org/10.18653/v1/D19-1459>.
- Bill Byrne, Karthik Krishnamoorthi, Saravanan Ganesh, Amit Dubey, Kyu-Young Kim, and Andy Cedilnik. Taskmaster-2. <https://github.com/google-research-datasets/Taskmaster/tree/master/TM-2-2020>, 2020.
- Lang Cao. Diaggpt: An llm-based chatbot with automatic topic management for task-oriented dialogue. *CoRR*, abs/2308.08043, 2023. doi: 10.48550/ARXIV.2308.08043. URL <https://doi.org/10.48550/arXiv.2308.08043>.
- Willy Chung, Samuel Cahyawijaya, Bryan Wilie, Holy Lovenia, and Pascale Fung. Instruct-tods: Large language models for end-to-end task-oriented dialogue systems. *CoRR*, abs/2310.08885, 2023. doi: 10.48550/ARXIV.2310.08885. URL <https://doi.org/10.48550/arXiv.2310.08885>.
- Jeffrey Dalton, Chenyan Xiong, Vaibhav Kumar, and Jamie Callan. Cast-19: A dataset for conversational information seeking. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1985–1988, 2020.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms, 2023.
- Song Feng, Siva Sankalp Patel, Hui Wan, and Sachindra Joshi. Multidoc2dial: Modeling dialogues grounded in multiple documents. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pp. 6162–6176. Association for Computational Linguistics, 2021. doi: 10.18653/V1/2021.EMNLP-MAIN.498. URL <https://doi.org/10.18653/v1/2021.emnlp-main.498>.
- Zuohui Fu, Yikun Xian, Yaxin Zhu, Yongfeng Zhang, and Gerard de Melo. Cookie: A dataset for conversational recommendation over knowledge graphs in e-commerce. *arXiv preprint arXiv:2008.09237*, 2020.
- Milica Gasic, Dongho Kim, Pirros Tsiakoulis, Catherine Breslin, Matthew Henderson, Martin Szummer, Blaise Thomson, and Steve J. Young. Incremental on-line adaptation of pomdp-based dialogue managers to extended domains. In Haizhou Li, Helen M. Meng, Bin Ma, Engsiong Chng, and Lei Xie (eds.), *INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, Singapore, September 14-18, 2014*, pp. 140–144. ISCA, 2014. doi: 10.21437/INTERSPEECH.2014-40. URL <https://doi.org/10.21437/Interspeech.2014-40>.
- Shirley Anugrah Hayati, Dongyeop Kang, Qingxiaoyang Zhu, Weiyan Shi, and Zhou Yu. Inspired: Toward sociable recommendation dialog systems. *arXiv preprint arXiv:2009.14306*, 2020.
- Ruining He and Julian McAuley. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *Proceedings of the 25th International Conference on World Wide Web, WWW '16*, pp. 507–517, Republic and Canton of Geneva, CHE, 2016. International World Wide Web Conferences Steering Committee.

- ISBN 9781450341431. doi: 10.1145/2872427.2883037. URL <https://doi.org/10.1145/2872427.2883037>.
- Vojtech Hudecek and Ondrej Dusek. Are large language models all you need for task-oriented dialogue? In David Schlangen, Svetlana Stoyanchev, Shafiq Joty, Ondrej Dusek, Casey Kennington, and Malihe Alikhani (eds.), *Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, SIGDIAL 2023, Prague, Czechia, September 11 - 15, 2023*, pp. 216–228. Association for Computational Linguistics, 2023. URL <https://aclanthology.org/2023.sigdial-1.21>.
- Meihuizi Jia, Ruixue Liu, Peiying Wang, Yang Song, Zexi Xi, Haobin Li, Xin Shen, Meng Chen, Jinhui Pang, and Xiaodong He. E-convrec: a large-scale conversational recommendation dataset for e-commerce customer service. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 5787–5796, 2022.
- Dongyeop Kang, Anusha Balakrishnan, Pararth Shah, Paul Crook, Y-Lan Boureau, and Jason Weston. Recommendation as a communication game: Self-supervised bot-play for goal-oriented dialogue. *arXiv preprint arXiv:1909.03922*, 2019.
- J. F. Kelley. An iterative design methodology for user-friendly natural language office information applications. *ACM Trans. Inf. Syst.*, 2(1):26–41, 1984. doi: 10.1145/357417.357420. URL <https://doi.org/10.1145/357417.357420>.
- Megan Leszczynski, Shu Zhang, Ravi Ganti, Krisztian Balog, Filip Radlinski, Fernando Pereira, and Arun Tejasvi Chaganty. Talk the walk: Synthetic data generation for conversational music recommendation. *arXiv preprint arXiv:2301.11489*, 2023.
- Raymond Li, Samira Ebrahimi Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. Towards deep conversational recommendations. *Advances in neural information processing systems*, 31, 2018.
- Zekun Li, Wenhui Chen, Shiyang Li, Hong Wang, Jing Qian, and Xifeng Yan. Controllable dialogue simulation with in-context learning. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pp. 4330–4347. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.FINDINGS-EMNLP.318. URL <https://doi.org/10.18653/v1/2022.findings-emnlp.318>.
- Lizi Liao, Le Hong Long, Zheng Zhang, Minlie Huang, and Tat-Seng Chua. Mmconv: an environment for multimodal conversational search across multiple domains. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pp. 675–684, 2021.
- Chien-Wei Lin, Vincent Auvray, Daniel Elkind, Arijit Biswas, Maryam Fazel-Zarandi, Nehal Belgamwar, Shubhra Chandra, Matt Zhao, Angeliki Metallinou, Tagyoung Chung, Charlie Shucheng Zhu, Suranjit Adhikari, and Dilek Hakkani-Tür. Dialog simulation with realistic variations for training goal-oriented conversational systems. *CoRR*, abs/2011.08243, 2020. URL <https://arxiv.org/abs/2011.08243>.
- John Mendonça, Patrícia Pereira, Miguel Menezes, Vera Cabarrão, Ana C. Farinha, Helena Moniz, João Paulo Carvalho, Alon Lavie, and Isabel Trancoso. Dialogue quality and emotion annotations for customer support conversations. *CoRR*, abs/2311.13910, 2023. doi: 10.48550/ARXIV.2311.13910. URL <https://doi.org/10.48550/arXiv.2311.13910>.
- Biswesh Mohapatra, Gaurav Pandey, Danish Contractor, and Sachindra Joshi. Simulated chats for building dialog systems: Learning to generate conversations from instructions. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pp. 1190–1203. Association for Computational Linguistics, 2021. doi: 10.18653/V1/2021.FINDINGS-EMNLP.103. URL <https://doi.org/10.18653/v1/2021.findings-emnlp.103>.

- Libo Qin, Wenbo Pan, Qiguang Chen, Lizi Liao, Zhou Yu, Yue Zhang, Wanxiang Che, and Min Li. End-to-end task-oriented dialogue: A survey of tasks, methods, and future directions. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pp. 5925–5941. Association for Computational Linguistics, 2023. URL <https://aclanthology.org/2023.emnlp-main.363>.
- Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019. URL <https://api.semanticscholar.org/CorpusID:160025533>.
- Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pp. 8689–8696. AAAI Press, 2020. doi: 10.1609/AAAI.V34I05.6394. URL <https://doi.org/10.1609/aaai.v34i05.6394>.
- Antoine Raux, Brian Langner, Dan Bohus, Alan W. Black, and Maxine Eskénazi. Let’s go public! taking a spoken dialog system to the real world. In *INTERSPEECH 2005 - Eurospeech, 9th European Conference on Speech Communication and Technology, Lisbon, Portugal, September 4-8, 2005*, pp. 885–888. ISCA, 2005. doi: 10.21437/INTERSPEECH.2005-399. URL <https://doi.org/10.21437/Interspeech.2005-399>.
- Fadime Sener and Angela Yao. Zero-shot anticipation for instructional activities. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019.
- Pararth Shah, Dilek Hakkani-Tür, Bing Liu, and Gökhan Tür. Bootstrapping a neural conversational agent with dialogue self-play, crowdsourcing and on-line reinforcement learning. In Srinivas Bangalore, Jennifer Chu-Carroll, and Yunyao Li (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 3 (Industry Papers)*, pp. 41–51. Association for Computational Linguistics, 2018. doi: 10.18653/V1/N18-3006. URL <https://doi.org/10.18653/v1/n18-3006>.
- Dennis Ulmer, Elman Mansimov, Kaixiang Lin, Justin Sun, Xibin Gao, and Yi Zhang. Bootstrapping llm-based task-oriented dialogue agents via self-talk. *CoRR*, abs/2401.05033, 2024. doi: 10.48550/ARXIV.2401.05033. URL <https://doi.org/10.48550/arXiv.2401.05033>.
- Tsung-Hsien Wen, David Vandyke, Nikola Mrksic, Milica Gasic, Lina Maria Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve J. Young. A network-based end-to-end trainable task-oriented dialogue system. In Mirella Lapata, Phil Blunsom, and Alexander Koller (eds.), *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 1: Long Papers*, pp. 438–449. Association for Computational Linguistics, 2017. doi: 10.18653/V1/E17-1042. URL <https://doi.org/10.18653/v1/e17-1042>.
- Jason D. Williams, Antoine Raux, Deepak Ramachandran, and Alan W. Black. The dialog state tracking challenge. In *Proceedings of the SIGDIAL 2013 Conference, The 14th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 22-24 August 2013, SUPELEC, Metz, France*, pp. 404–413. The Association for Computer Linguistics, 2013. URL <https://aclanthology.org/W13-4065/>.
- Hu Xu, Seungwhan Moon, Honglei Liu, Bing Liu, Pararth Shah, and Philip S Yu. User memory reasoning for conversational recommendation. *arXiv preprint arXiv:2006.00184*, 2020.

- Li Yang, Qifan Wang, Zac Yu, Anand Kulkarni, Sumit Sanghai, Bin Shu, Jon Elsas, and Bhargav Kanagal. Mave: A product dataset for multi-source attribute value extraction. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining, WSDM '22*, pp. 1256–1265, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391320. doi: 10.1145/3488560.3498377. URL <https://doi.org/10.1145/3488560.3498377>.
- Yunyi Yang, Yunhao Li, and Xiaojun Quan. UBAR: towards fully end-to-end task-oriented dialog system with GPT-2. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pp. 14230–14238. AAAI Press, 2021. doi: 10.1609/AAAI.V35I16.17674. URL <https://doi.org/10.1609/aaai.v35i16.17674>.
- Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. Webshop: Towards scalable real-world web interaction with grounded language agents. *Advances in Neural Information Processing Systems*, 35:20744–20757, 2022.
- Tom Young, Frank Xing, Vlad Pandeale, Jinjie Ni, and Erik Cambria. Fusing task-oriented and open-domain dialogues in conversational agents. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelfth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pp. 11622–11629. AAAI Press, 2022. doi: 10.1609/AAAI.V36I10.21416. URL <https://doi.org/10.1609/aaai.v36i10.21416>.
- Hamed Zamani, Johanne R. Trippas, Jeff Dalton, and Filip Radlinski. Conversational information seeking. *Foundations and Trends® in Information Retrieval*, 17(3-4):244–456, 2023. ISSN 1554-0669. doi: 10.1561/15000000081. URL <http://dx.doi.org/10.1561/15000000081>.
- Xiaoying Zhang, Baolin Peng, Kun Li, Jingyan Zhou, and Helen Meng. SGP-TOD: building task bots effortlessly via schema-guided LLM prompting. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pp. 13348–13369. Association for Computational Linguistics, 2023. URL <https://aclanthology.org/2023.findings-emnlp.891>.
- Zheng Zhang, Ryuichi Takanobu, Qi Zhu, MinLie Huang, and XiaoYan Zhu. Recent advances and challenges in task-oriented dialog systems. *Science China Technological Sciences*, 63(10):2011–2027, Oct 2020. ISSN 1869-1900. doi: 10.1007/s11431-020-1692-3. URL <https://doi.org/10.1007/s11431-020-1692-3>.
- Kun Zhou, Yuanhang Zhou, Wayne Xin Zhao, Xiaoke Wang, and Ji-Rong Wen. Towards topic-guided conversational recommender system. *arXiv preprint arXiv:2010.04125*, 2020.
- Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. Txygen: A benchmarking platform for text generation models. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR '18*, pp. 1097–1100, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356572. doi: 10.1145/3209978.3210080. URL <https://doi.org/10.1145/3209978.3210080>.

A Appendix

In this section we provide some additional details about the features of our synthetic dataset compared to existing TOD and conversational recommendation datasets, detailed definitions of intents and slots we used for our experiments, confusion matrices for the intent classification task and two sample conversations from our best fine-tuned models.

Dataset	ARQ	ODQ	CQ	RC	RS	CLQ	ST
ReDial (Li et al., 2018)	✓					✓	✓
GoRecDial (Kang et al., 2019)	✓					✓	
TG-ReDial (Zhou et al., 2020)	✓	✓					
MGConvRex (Xu et al., 2020)	✓	✓					✓
INSPIRED (Hayati et al., 2020)	✓	✓				✓	✓
MMConv (Liao et al., 2021)	✓					✓	✓
COOKIE (Fu et al., 2020)	✓					✓	✓
E-ConvRec (Jia et al., 2022)	✓					✓	✓
MultiWOZ (Budzianowski et al., 2018)	✓	✓				✓	✓
CAsT 19 (Dalton et al., 2020)	✓		✓			✓	✓
WebShop (Yao et al., 2022)	✓	✓		✓		✓	✓
SGD (Rastogi et al., 2020)	✓			✓		✓	✓
Synthetic Music Data (Leszczynski et al., 2023)	✓					✓	✓
SynTOD (this work)	✓	✓	✓	✓	✓	✓	✓

Table 7: Comparison of our dataset derived using the SynTOD framework with existing TOD and conversational recommendation datasets. Dataset features: ARQ (Attribute Related Question), ODQ (Open Domain Question), CQ (Comparison Questions), RC (Retrieval Component), RS (Result Summarization), CLQ (Clarifying Questions), ST (State Tracking).

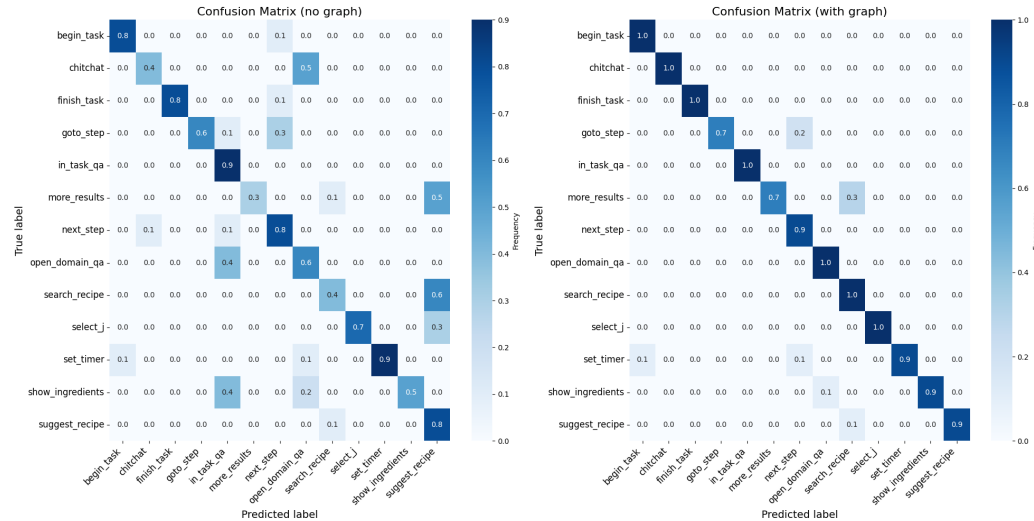


Figure 5: Intent classification confusion matrix for the cooking assistance domain without and using a state transition graph with the SynTOD framework. Using the graph-guided data generation approach, we observe significant improvements in intent classification across all intents.

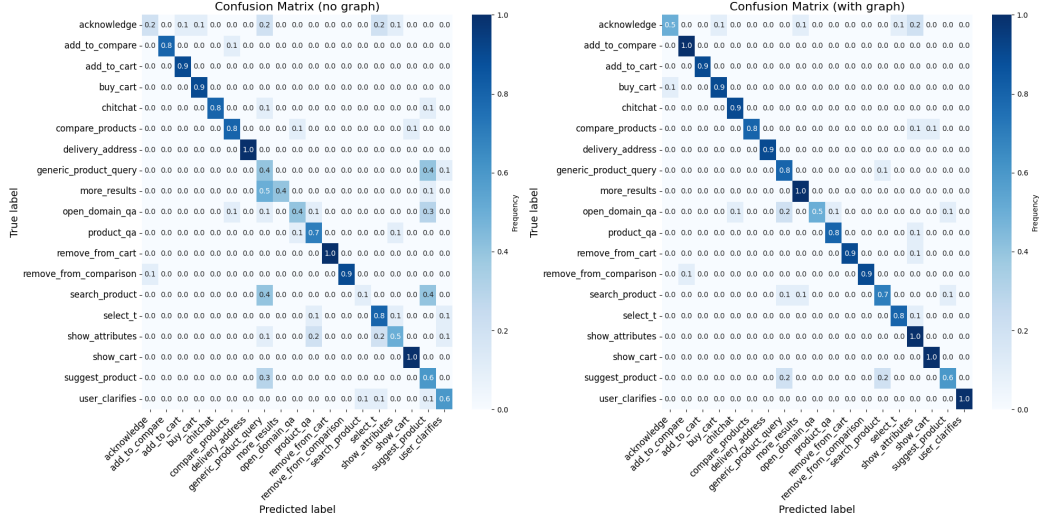


Figure 6: Intent classification confusion matrix for the e-commerce domain. Like in the cooking assistance domain, using the graph-guided data generation approach, we observe significant improvements in intent classification across all intents.

User Intent (slots)	Definition
search_recipe(query)	User searches for a specific recipe
suggest_recipe(query)	User asks for recipe recommendation based on occasion, cuisine etc.
more_results	Shows more recipe results
select(i)	Selects a recipe from a list
show_ingredients	Asks to display the ingredients for a recipe
begin	Starts the recipe guidance
next_step	Requests the next step in the recipe
goto_step(i)	Jumps to a specific step in the recipe
repeat	Asks to repeat the current step or a previous step
finish_task	Finishes the recipe guidance
open_domain_qa	Asks open-domain questions
in_task_qa	Asks questions related to recipe instructions
chitchat	Engages in general conversation not related to the task
set_timer(duration)	Requests to set a timer for a specific duration

Table 8: Cooking Assistance User Intents and Slots we defined for our experiments.

User Intent (slots)	Definition
search_product(query, attributes_list)	Searching a specific product
suggest_product(query, attributes_list)	Asking about some product
generic_product_query	Very generic product query that needs clarification
user_clarifies(query, attributes_list)	Clarifying bot's question regarding a product.
more_results	Asking bot to present more results
select(i)	Selecting a product from the display
show_attributes(product_id)	Asking to describe general features of a product
acknowledge	Acknowledging the reply from taskbot
open_domain_qa(question)	User asking an open domain question
product_qa	Asking questions related to a product
chitchat	Engages in general conversation
show_cart	Asking to show the cart
buy_cart	Asking to buy products present in the cart
add_to_cart(product_id)	Asking to add a product to the cart
remove_from_cart(product_id)	Asking to remove a product from the cart
compare_products	Asking to compare products in compare list
delivery_address(address)	Asking availability of a product in certain country
remove_from_comparison(product_id)	Asking to remove products from compare list
add_for_comparison(product_id)	Asking to add product in compare list

Table 9: Ecommerce Assistance User Intents we defined for our experiments.

system: Welcome to MarunaChef, your friendly recipe taskbot developed by CIIR at UMass Amherst! Whether you're a beginner cook or an experienced chef, I'm here to assist you in finding and preparing delicious recipes. From quick weekday meals to impressive dishes for special occasions, I've got you covered. Just let me know what ingredients you have or the dish you're craving, and I'll provide you with step-by-step instructions to create culinary delights. Let's get cooking! intent: start
user: I'm having a party this weekend and I want to serve some amazing dips. Can you help me find some unique and easy dip recipes? intent: search_recipe
system: Here are some delicious dip recipes you might enjoy! 1. **Taco Dip**: A spicy and savory party appetizer that's a crowd-pleaser, with a high rating of 4.7. It's loaded with melty cheese, spicy poblanos, and savory beef. Ready in just 25 minutes! 2. **Buffalo Chicken Dip**: This creamy dip is a fan favorite, boasting a stellar rating of 4.9. It's packed with shredded chicken, hot sauce, and plenty of melty cheese. Also ready in 25 minutes! 3. **Chocolate Brownie Dip**: For the sweet tooth out there, this decadent dip is perfect for sharing or indulging solo. It's a quick 5-minute recipe with a rating of 3.5. Which one tickles your taste buds? Let me know if you'd like to see the full recipe for any of these, or if you'd like more options. intent: show_results
user: I'd like to go for the recipe with the highest rating please. intent: select.i
system: Great choice! Buffalo Chicken Dip is a classic crowd-pleaser. Would you like me to show you the list of required ingredients? And when you're ready, just let me know, and we'll dive right into the first step. intent: option_selected
user: What is the weather like in London right now? intent: acknowledge_task
system: I'm sorry for any confusion, but as an AI developed by CIIR at UMass Amherst, I don't have real-time data capabilities and cannot provide current weather updates. I recommend checking a trusted weather service for the most accurate information. intent: started_task
user: This recipe seems interesting, let's not waste any more time and get started. intent: acknowledge_step
system: Alright, let's dive right in! The first step is to preheat your oven to 350°F (180°C). While it's warming up, we can proceed with the next steps. intent: show_step
user: The oven is preheated. Now what? intent: done_step
system: In step 2 of the recipe, you'll be mixing together the shredded chicken, cream cheese, and buffalo sauce. This combination creates a creamy and spicy filling for the dip. It's important to evenly distribute the ingredients for the perfect flavor. When you're ready, let me know and we can move on to the next step! intent: show_step
user: I'm done mixing the chicken, cream cheese, and buffalo sauce. intent: done_step
system: Step 3 of the Buffalo Chicken Dip recipe is all about getting that mixture into a cast iron skillet and making it smooth with a spoon. This step helps to distribute the flavors evenly and ensures a creamy texture throughout. So, when you're ready to continue, let me know! And here's a fun tip: Using a cast iron skillet allows for even heating, giving your dip a lovely golden crust. Enjoy cooking! intent: show_step
user: What exactly does the fifth step of this recipe involve? intent: goto_step_5
system: Step 5 of this mouthwatering recipe is where the magic happens! After it's baked to perfection, you get to top your Buffalo Chicken Dip with fresh scallions for an extra burst of flavor. Don't forget to grab your favorite chips for dipping! Are you ready to continue, or would you like a tip or fun fact? intent: no_more_steps
user: I've followed the last step and the dip is complete. intent: finish_task
system: Congratulations on completing the mouthwatering Buffalo Chicken Dip! You did an incredible job following the recipe! Is there anything else I can assist you with? Feel free to ask any further questions or explore more delicious recipes. Happy cooking! intent: start
user: That's all I needed, I'll let you know how it turns out! intent: task_complete
system: Great job! I hope you enjoyed making this recipe. Enjoy your meal and don't hesitate to reach out if you'd like to try another recipe. Bon Appétit! intent: end

Table 10: Cooking Assistance Sample Conversation

system: Hello! Welcome to MarunaShopper, your personal taskbot designed by CIIR at UMass Amherst to assist you in finding the perfect items in our product catalog. How can I help you shop today? intent: start
user: I'm looking for something to protect my iPod with a bit of style. Can you help me find a suitable case? intent: generic_product_query
system: Absolutely, I'd be happy to help you find a stylish case for your iPod! To narrow down the options, could you tell me which iPod model you have? Also, do you have any preferences when it comes to material, color, or any specific design elements like patterns or prints? Would you like any additional features such as a stand, pockets for cards, or extra shock absorption? intent: clarifying_questions
user: Oh, I have the 4th Generation iPod touch and I love a bit of uniqueness in my accessories. I'm really into animal prints and my favorite color combo is black and white. Something with a bit of bling wouldn't hurt either, as long as it's not too over the top. And I need good scratch protection, of course. Could you find me something like that? intent: user_clarifies query: iPod touch 4th Generation case black white animal print bling attributes.List: [{"Color": "Black and White"}, {"Compatibility": "iPod touch 4th Gen"}]
system: Take a look at our stylish options for your iPod touch 4th Generation case needs! 1. Step up your game with the *eForCity Snap-On Rubber Coated Case for iPod touch 4G (Zebra/Bling Bow and Heart)*. This eye-catching case not only provides durable... It's a perfect fit for your device and allows easy ... 2. For those who love patterns, the *Green Zebra TPU Rubber Skin Case Cover for iPod Touch 4th Generation 4G 4* might be... Feel free to add any of these cases to your cart... iPod touch a new look while ensuring it stays protected. Which one catches your eye? intent: show_results
user: The Snap-On case with the rubber coating and that unique design is my choice. intent: select.i
system: Absolutely, the Zebra Bling Case for the iPod touch 4G sounds like a stylish choice! You can take a closer look at its features, add it to your cart for purchase, or if you're still deciding, you can compare it with other cases. What would you like to do next? intent: option_selected
user: Include the eForCity Snap-On Case with Zebra and Bling Bow design in the comparison list to see how it stacks up against the other options. intent: add_for_compare title: eForCity Snap-On Rubber Coated Case for iPod touch 4G product.id: 3
system: Certainly! I've added the eForCity Snap-On Rubber Coated Case for the iPod touch 4G with the chic Zebra and Bling Bow design to your comparison list. You can now easily compare its style, durability, and user rating against your other choices. If you need detailed comparisons on specific attributes like compatibility, protection level, or how it might complement your style, feel free to ask! intent: system_response_add_for_compare
user: Can you suggest a protective case with a unique design for my Apple iPod Touch 4th Generation that also includes a screen protector and has high user ratings? intent: suggest_product query: iPod Touch 4th Gen case screen protector high rating attributes.List: [{Compatibility : Apple iPod}, {Feature : Screen Protector}, {Rating : 5.0}]
system: Discover the perfect companions... Feel free to add any of these to your cart, compare them side by side, or proceed to purchase to give your iPod Touch 4th Gen the protection it deserves. intent: show_results
user: I'd like to choose the protector that comes with a cleaning cloth and has precise cutouts for the buttons and ports. intent: select.i
system: Great choice! The Purple Butterfly Tree Case for the iPod Touch 4 offers both style and protection. You can view its detailed attributes, add it to your cart for later, proceed to purchase it now, or add it to your comparison list. What would you like to do next? intent: option_selected
user: Can you provide details on the material composition of the hard plastic used in the Snap on Hard Cover Protector Case, and whether the rubberized coating extends to all surfaces of the case? intent: product.qa product.id: 2
system: Certainly! The *Purple Butterfly Tree Polka Rubberized Snap on Hard Cover Protector Faceplate Case* is crafted with high-quality..., protecting your Apple Ipod Touch 4th Generation from scratches, dents, and chips. The hard plastic is reinforced... Would you be interested in adding this protective and stylish case to your cart, or is there anything else I can assist you with today? intent: product.qa_system_response
user: Include the Purple Butterfly Tree Case in the comparison list to evaluate it alongside the other options. intent: add_for_compare title: Purple Butterfly Tree Polka Rubberized Snap on Hard Cover Protector Faceplate Case product.id: 2
system: Absolutely, the Purple Butterfly Tree Polka Rubberized Snap-on Hard Cover Protector Faceplate Case for the Apple iPod Touch 4th Generation will be added to your comparison list... If you need further assistance or have more items to compare, please let me know! intent: system_response_add_for_compare
user: Can you provide a detailed comparison between these iPod touch cases, focusing on their compatibility, style, protection level, and user ratings? intent: compare_products
system: Certainly! Let's compare the two iPod Touch 4th Generation cases you're interested in: *Purple Butterfly Tree Polka Rubberized Snap on Hard Cover Protector Faceplate Case for Apple Ipod Touch 4 4th Gen.* - This case offers... - It is designed to protect your device from scratches, dents, and chips while... intent: show_comparison product.ids: [1,2]
user: Please remove the item with the Purple Butterfly Tree design from our comparison list. intent: select.i_remove_from_compare title: Purple Butterfly Tree Polka Rubberized Snap on Hard Cover Protector Faceplate Case product.id: 1
system: Absolutely, the Purple Butterfly Tree Polka Rubberized Snap ... has been removed from your comparison list. If there's anything else... intent: system_response_remove_from_compare
user: Could you find me the South Korea flag patch with a gold border that can be ironed on? intent: search_product
...
user: Please add the South Korea Flag patch to my shopping cart. intent: add_to_cart
system: ... intent: system_response_added_to_cart
user: Could you please show me the items currently in my cart? intent: show_cart
system: Here are the items in your shopping cart so far: 1. *SOUTH KOREA FLAG w/GOLD BORDER/Iron On Patch Applique/South Korean National Flag* Please let me know if you would like to proceed to checkout or continue shopping! intent: shown_cart
user: I'm ready to make the purchase. intent: buy_cart

Table 11: E-commerce Assistance Sample Conversation.