DiaSynth: Synthetic Dialogue Generation Framework for Low Resource Dialogue Applications

Anonymous ACL submission

Abstract

The scarcity of domain-specific dialogue datasets limits the development of dialogue systems across applications. Existing research is constrained by general or niche datasets that lack sufficient scale for training dialogue systems. To address this gap, we introduce **DiaSynth** - a synthetic dialogue generation framework capable of generating high-quality, contextually rich dialogues across a wide range of domains. Unlike existing frameworks, DiaSynth uses Large Language Models (LLMs) and Chain of Thought (CoT) reasoning to generate dynamic, domain-specific dialogues with simulated personas and diverse conversational features. We perform our experiments by generating synthetic data using different LLMs and few-shot examples from DialogSum and SAMSum. The pretrained language models fine-tuned on the synthetic data outperform the base models by **16.47%** on dialogue summarization, while the comparison between models fine-tuned on in-domain data and synthetic data shows that the synthetic data is able to capture **90.48%** of the performance distribution of the in-domain data on dialogue summarization. The quality of the data generated also increases as we increase the size of LLM from 3B to 8B. These results validate DiaSynth's potential as a robust alternative to traditional data collection methods. We open source the code and data generated for future research.

1 Introduction

Dialogue systems are crucial in natural language processing due to applications like customer service chatbots and virtual assistants. Their effectiveness depends on large, high-quality, domain-specific datasets. The lack of large-scale, high-quality datasets across domains like academic discussions, healthcare, and everyday conversations poses a challenge. This scarcity limits the development of dialogue systems that generalize well across domains.

Prior work (Feng et al. (2020), Zeng et al. (2020), Budzianowski et al. (2018)) collects domain-specific dialogues but often lacks depth, scale, or domain diversity. On the one hand, the conversations in a domain specific dataset are superficial and do not go deep into the domain. On the other hand, niche domain dialogue datasets, while contextually rich, often suffer from limited scale. This imbalance hinders dialogue system development in underrepresented domains, where data collection is costly and complex.

To address these problems, we introduce **DiaSynth**, a synthetic dialogue generation framework that produces contextually rich and realistic dialogues tailored to specific domains. DiaSynth, using a Large Language Model (LLM), generates high-quality conversations by simulating personas and conversation characteristics like tone and formality. With LLMs and Chain of Thoughts (CoT) Wei et al. (2024), DiaSynth generates dialogues that mimic real-world conversations for a wide range of domains. CoT plays a crucial part by building different environments for different personas (refer to Appendix D for examples) which influence varied conversations. This approach addresses data scarcity and offers a scalable, cost-effective alternative to traditional methods.

To validate the effectiveness of DiaSynth, we evaluated the framework on two criteria: the quality of the generated data and the usability of this data for downstream tasks. The results for the quality criterion showed that the data quality improved with the scale of the model. For usability, we tested DiaSynth on dialogue summarization. Models fine-tuned on DiaSynth data outperformed base versions by **16.47%** on average. Additionally, DiaSynth's synthetic data captures **90.48%** of in-domain performance, highlighting its potential as a strong alternative when domain-specific data is unavailable.

The remainder of this paper is organized as follows: Section 2 reviews related work on dialogue

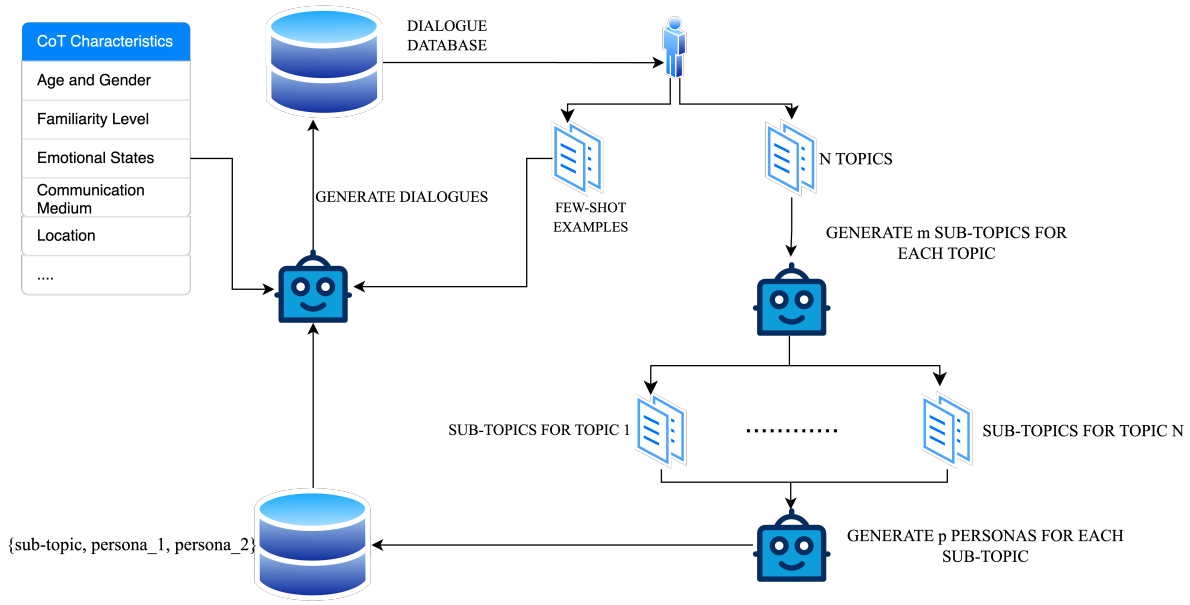

Figure 1: Overview of DiaSynth

datasets and synthetic data. Section 3 details the DiaSynth framework and methodology. Section 4 describes our experimental setup and evaluation. Section 5 compares the performance of models fine-tuned on DiaSynth data to in-domain data. Section 6 concludes with a summary of our findings and Section 7 discusses the limitations of DiaSynth and potential future directions for this research.

2 Related Work

2.1 Personality in Synthetic Data Generation

In recent years, there has been a significant increase in research focused on synthetic dialogue generation, largely driven by advancements in Large Language Models (LLMs). To generate **realistic and diverse** synthetic data, researchers have incorporated personalities, profiles, and character information when prompting LLMs to generate dialogues Han et al. (2024). By enhancing dialogue realism through the simulation of various personality profiles, utilizing the Big Five personality model, and employing structured prompts, this approach has improved task performance in models fine-tuned on these generated dialogues compared to those trained on general chit-chat datasets.

Moreover, integrating personas into synthetic data generation prompts Chan et al. (2024) has demonstrated that models fine-tuned on personalized synthetic data outperform some LLMs of much larger scales. The inclusion of personas in prompts provides diversity in difficulty levels and ranges within the synthetic data, enabling the models to handle situations of varying complexity.

Our approach involves persona extraction after generating subtopics related to the general topics. This enables the generated dialogues to be more in-depth and specific to those subtopics, enhancing both the scale and quality of domain-specific dialogue generation.

2.2 Prompting Task-Oriented Dialogue Generation

Prompt-based techniques have also emerged as powerful methods for generating high-quality synthetic dialogues, particularly for task-oriented dialogue systems. Steindl et al. (2023) explore the generation of synthetic dialogues from structured prompts, focusing on enhancing task-oriented dialogue systems. Their work demonstrates that prompt engineering can produce dialogues that are contextually appropriate and improve system performance by aligning synthetic data more closely with real-world requirements.

To achieve a **higher quantity, diversity, and creativity** in human-written instruction data, Wang et al. (2022) propose inputting prompts to LLMs to generate instructions based on a small set of seed human-written instructions. This approach aligns the expanded training data more closely with desired task objectives and allows for iterative improvements, producing more nuanced and effective dialogues that meet specific task demands.

Similarly, our study expands topics into

LLM	Few-shot examples	Number of Samples	Avg. number of turns	Avg. number of tokens per turn	Diversity (ROUGE-L)
Phi-3	DialogSum	1215	9.13	20.38	0.27
InternLM-2.5	DialogSum	1035	9.23	27.98	0.30
LLaMA-3	DialogSum	1154	6.86	31.99	0.29
GPT-4o	DialogSum	1375	15.16	15.96	0.29
Phi-3	SAMSum	1410	13.98	13.94	0.27
InternLM-2.5	SAMSum	1135	13.96	19.07	0.29
LLaMA-3	SAMSum	1195	10.54	20.41	0.29
GPT-4o	SAMSum	1380	15.43	13.53	0.28

Table 1: Data Statistics

subtopics, ensuring that the generated dialogues provide more in-depth and high-quality conversations. By doing so, we aim to produce synthetic data that not only covers a broader range of scenarios but also delves deeper into each topic, thereby enhancing the overall effectiveness of the dialogue systems trained on this data.

2.3 Existing Task-Oriented Dialogue Datasets

In addition to prompt-based synthetic data generation, various large-scale dialogue datasets have been instrumental in advancing task-oriented dialogue systems. Among these, the MultiWOZ dataset Budzianowski et al. (2018) is a prominent resource, providing richly annotated dialogues across multiple domains. MultiWOZ has enabled researchers to train models capable of handling complex, multi-turn interactions across diverse tasks. The nature of MultiWOZ's annotations has made it a benchmark for evaluating the performance of dialogue systems, though it is often complemented by synthetic data to introduce further diversity and variation in dialogue scenarios.

Similarly, Doc2Dial Feng et al. (2020) is another widely used dataset designed specifically for document-grounded dialogue systems. Doc2Dial includes conversations grounded in structured documents, focusing on providing users with accurate and relevant information based on their inquiries. This dataset has been instrumental in improving the ability of dialogue systems to retrieve and generate accurate responses when interacting with complex information sources. However, much like MultiWOZ, Doc2Dial's scope is limited to the predefined topics and domains covered within the dataset, which can restrict model generalizability to new or unseen situations.

To overcome the domain-specific limitations of datasets, our study adopts a synthetic data generation approach that expands existing topics into subtopics, thus providing a broader and deeper pool of conversational data. By incorporating both task-oriented prompts and personas, our generated dialogues aim to complement these datasets by offering more personalized and contextually rich conversations, thereby enhancing the robustness and versatility of dialogue systems

3 DiaSynth

DiaSynth is a synthetic dialogue generation framework designed to address the scarcity of high-quality, large-scale, domain-specific dialogue datasets. DiaSynth uses an LLM and CoT reasoning to simulate diverse, nuanced dialogues.

DiaSynth takes a list of user-provided topics to generate dialogues. The users can optionally provide few-shot examples of the format in which they want the dialogue to be generated. Directly generating dialogues from user topics would be too superficial due to their lack of specificity. To overcome this lack of specificity, we generate m sub topics for each of the n topics given by the user. Generating dialogues from the subtopics will have specificity but the dialogues will lack variety. This is because every dialogue is influenced implicitly by the personas of the people involved in the dialogue and, other characteristics such as the location, emotion and more. To enhance variety and depth, we generate p personas per subtopic and create dialogues for all persona-subtopic combinations. To further ground the dialogues in various settings and characteristics, we employ CoT reasoning during the generation process. DiaSynth employs CoT to reason about the settings and characteristics of a dialogue, which are listed in Appendix C, ensuring that the dialogues are contextually rich and realistic. This three-stage pipeline not only guarantees the quality of the generated dialogues but also allows for exponential scalability, as illustrated programmatically in Appendix B.

3.1 Subtopic Generation

Subtopic generation is a crucial step in DiaSynth's pipeline, since it enhances the specificity and depth

	coherent	error recovery	consistent	diverse	depth	likeable	understand	flexible	informative	inquisitive
					DIALOGUESUM					
Phi-3	0.9536	0.9440	0.9540	0.9534	0.9521	-0.0005	0.9353	**-3.96E-05**	0.0009	-0.0033
InternLM-2.5	0.8439	0.8313	0.8359	0.8353	0.8352	0.0048	0.8278	-0.0046	0.0042	0.0069
LLaMA-3	**0.9684**	**0.9522**	**0.9570**	**0.9596**	**0.9592**	0.0032	**0.9453**	-0.0063	0.0063	0.0105
GPT-4o	0.9525	0.9407	0.9417	0.9423	0.9425	**0.0121**	0.9368	-0.0027	**0.0085**	**0.0144**
					SAMSUM					
Phi-3	0.9161	0.9088	0.9199	0.9161	0.9130	-0.0004	0.9014	-0.0024	0.0034	-0.0040
InternLM-2.5	0.8746	0.8647	0.8734	0.8655	0.8661	0.0033	0.8582	-0.0019	0.0106	0.0028
LLaMA-3	0.9829	0.9677	0.9757	0.9712	0.9731	0.0003	0.9593	-0.0048	0.0100	0.0029
GPT-4o	**0.9939**	**0.9876**	**0.9878**	**0.9836**	**0.9824**	**0.0083**	**0.9788**	**0.0004**	**0.0141**	**0.006**

Table 2: FED scores

of the dialogues that will be generated later. For each primary topic given by the user, DiaSynth generates multiple subtopics, effectively narrowing the focus of the conversation. This breakdown is necessary because the primary topics are often too general to generate contextually rich dialogues on their own. For instance, a topic like "healthcare" can be expanded into subtopics such as "doctor-patient consultations," "mental health discussions," and "medical diagnostics," each of which offers a more focused context for dialogue generation. To achieve this, DiaSynth prompts an LLM to generate the user specified number of subtopics for each primary topic. We also run a similarity check between each of the subtopics and remove subtopics that are too similar to other subtopics using a threshold.

3.2 Persona Generation

Personas of the individuals involved in a conversation are primary influencers in determining how a conversation pans out. Using random personas from persona datasets and prompting the LLM to simulate a dialogue between them about a random topic often leads to superficial dialogues that lack depth and contextual richness. To address this issue, DiaSynth generates a user-specified number of personas for each subtopic, ensuring that the personas are conditioned on the subtopic context. This conditioning prompts the LLM to create personas that are most likely to engage in a meaningful dialogue about the subtopic, such as a medical professional and a patient discussing "medical diagnostics" or a researcher and a student talking about "academic publishing." We also run a similarity check for the personas too. The conditioned persona generation is crucial because it ensures that future dialogues will not only be contextually rich but also exhibit a high level of depth. Each dialogue will be between two personas who have relevant expertise or perspectives on the given subtopic, allowing the conversation to explore nuances that would otherwise be missed in a generic dialogue setting.

3.3 Dialogue Generation

The final stage in DiaSynth's pipeline is the generation of dialogues, where all the components—subtopics, personas, and characteristics—converge to create contextually rich and realistic conversations. This step uses an LLM as the backbone and CoT as the reasoning mechanism, allowing the model to simulate dialogues that incorporate various aspects of human interaction. DiaSynth generates dialogues by pairing all persona-subtopic combinations. The process also integrates predefined characteristics (Table 11) like emotional state, formality, and familiarity to guide the flow and style. These characteristics are defined in the CoT prompt, guiding the LLM to generate realistic, contextually appropriate dialogues. The importance of CoT and the lack of it affects the quality of the dialogues, which is shown quantitatively in Appendix E.

4 Experimental Setup

In this section, we detail the experimental setup used to evaluate the effectiveness of DiaSynth. Our evaluation focuses on two criteria - quality of the dialogues generated and usability of the dialogues generated for a downstream task. Quality of the dialogues is evaluated using metrics such as FED, GPTScore, and G-Eval. We evaluate the usability of DiaSynth-generated dialogues by using summarization as the downstream task.

4.1 Quality of the dialogues

To evaluate the quality of the dialogues, we employ the metrics that have been developed for evaluating the quality of text generated by LLMs. We use the

	coherence	diversity	flexibility	understandability	inquisitiveness	consistency	informativeness	likeability	depth	error recovery
					DIALOGUESUM					
Phi-3	**0.0286**	0.0310	**0.0218**	0.0193	0.0363	**0.0369**	0.0172	**0.0213**	0.0117	0.0342
InternLM-2.5	0.0069	0.0196	0.0084	0.0061	0.0244	0.0137	0.0148	0.0050	0.0080	0.0197
LLaMA-3	0.0189	**0.0430**	0.0186	**0.0220**	**0.0415**	0.0321	**0.0318**	0.0110	**0.0201**	**0.0440**
GPT-4o	0.0039	0.0156	0.0059	0.0039	0.018	0.0080	0.0097	0.0026	0.0053	0.0135
					SAMSUM					
Phi-3	**0.0325**	0.0372	0.0260	0.0270	0.0395	0.0415	0.0201	**0.0232**	0.0126	0.0290
InternLM-2.5	0.0128	0.0408	0.0194	0.0174	0.0504	0.0306	0.0328	0.0142	0.0177	0.0407
LLaMA-3	0.0288	**0.0655**	**0.0306**	**0.0365**	**0.0622**	**0.0612**	**0.0542**	0.0168	**0.0270**	**0.0558**
GPT-4o	0.0055	0.0162	0.0094	0.0084	0.0186	0.0119	0.0129	0.0038	0.0079	0.0199

Table 3: GPTScore

following metrics:

- **FED** Mehri and Eskenazi (2020) - FED evaluates dialogue quality by comparing the probabilities of appended positive and negative utterances.
- **GPTScore** Fu et al. (2024) - GPTScore assesses dialogue quality by asking an LLM to evaluate criteria like coherence and diversity, with scores based on the probability of affirmative responses.
- **G-Eval** Liu et al. (2023) - G-Eval rates dialogue on a 1-3 scale across criteria, with final scores as a weighted average from the LLM's probability distribution.

To validate the framework across models and also domains, we generate data using three open source LLMs, one closed source LLM and also use few shot examples from two different dialogue datasets. The open sourced LLMs are - **Phi-3** Abdin et al. (2024), **InternLM-2.5** Cai et al. (2024), **LLaMA-3** Dubey et al. (2024) and the closed source LLM used is **GPT-4o**. The 8-bit quantized versions of the open source LLMs were used for faster experimentation and generation. The two different dialogue datasets that were used as few-shot examples are DialogSum Chen et al. (2021) and SAMSum Gliwa et al. (2019)

4.2 Downstream Task - Summarization

To evaluate the usability of the dialogues generated by DiaSynth, we choose summarization as the downstream task. Summarization, a key application of dialogue systems, aims to generate concise, contextually relevant summaries. We use three established evaluation metrics—QAGS, BERTScore, and ROUGE-L—to assess the performance of summarization models fine-tuned on DiaSynth-generated data.

- **QAGS** Wang et al. (2020) - QAGS evaluates factual consistency by generating questions from the summary and comparing answers to those from the source dialogue.
- **BERTScore** Zhang* et al. (2020) - BERTScore measures semantic similarity between generated and reference summaries using contextual embeddings.
- **ROUGE-L** Lin (2004) - ROUGE-L measures longest common subsequence (LCS) overlap between generated and reference summaries.

We fine-tune pretrained summarization models like DistilBART, BART Lewis et al. (2020), T5 Raffel et al. (2020) and LED Beltagy et al. (2020), on DiaSynth-generated dialogues and evaluate their performance using the above metrics. We evaluate the usability of DiaSynth in two key aspects: first, by assessing the performance improvement of models fine-tuned on DiaSynth-generated data compared to the pretrained models; and second, by measuring the extent to which DiaSynth-generated data reflects real-world data distribution by comparing the performance of models fine-tuned on DiaSynth data versus those fine-tuned on in-domain data.

5 Results

This section discusses the results of the data generated using DiaSynth (quality of the data and usability in downstream tasks) with different LLMs and varying few-shot examples. Specifically, we utilized Phi-3, InternLM-2.5, LLaMA-3 and GPT-4o as the LLM backbones, and the few-shot examples were sourced from DialogueSum and SAMSum datasets. These combinations allow us to evaluate the robustness and adaptability of DiaSynth across different models and few shot examples. In total, eight distinct datasets were generated using DiaSynth by pairing each LLM with the two sets

of few-shot examples, resulting in all possible combinations. For each combination, DiaSynth was provided with the same 16 broad topics and tasked with generating 6 subtopics for each topic, followed by creating 6 personas for each subtopic. The statistics of the datasets generated using DiaSynth, including the number of dialogues, average number of turns, and average number of tokens per turn, are summarized in Table 1. All the experiments were run on a single A100 GPU with the generation time ranging from 2 hours to 4 hours.

5.1 Quality of the Dialogues

The quality of the synthetic datasets produced by DiaSynth was evaluated using FED, GPTScore, and G-Eval metrics, as detailed in Tables 2, 3, and 4. The results reveal distinct variations in performance across different model and dataset configurations, reflecting the unique characteristics of each.

5.1.1 Metric Scores

FED: The FED scores in Table 2 show that LLaMA-3 and GPT-4o achieve almost a perfect score (+1) in most of the criteria, while Phi-3 and InternLM-2.5 also have decent performances. GPT-4o has a clear advantage when it comes to generating likeable dialogues while there is not much separation on other criteria.

GPTScore: Results illustrated in 3 are surprising in that GPT-4o is the worst performing model on GPTScore, which might require further research while LLaMA-3 clearly dominates the other models.

G-Eval: Table 4 highlights GPT-4o's dominance in engagingness and naturalness with perfect scores (3.0) for DialogSum, while InternLM-2.5 stands out in coherence (2.9990) and groundedness (2.9973) for DialogSum, and coherence (2.9983) and groundedness (2.9952) for SAMSum, suggesting it maintains high factual accuracy.

Dataset-Specific Performance. The contrasting performance of GPT-4o on the DialogSum and SAMSum datasets in Table 2 can be attributed to the differing structures of the dialogues in these datasets. DialogSum consists of more formal and structured dialogues, which aligns with the typical response style of GPT-4o, leading to its stronger performance. In contrast, SAMSum contains more casual, human-like conversations, which might explain GPT-4o's relatively poorer performance, as it may not adapt as well to the informal, spontaneous nature of such dialogues. Overall, while GPT-4o excels in natural and engaging dialogue, LLaMA-3 offers the most versatility, and InternLM-2.5 provides a strong alternative with high coherence and groundedness.

	engagingness	naturalness	coherence	groundedness
DIALOGUESUM				
Phi-3	2.5236	2.7238	2.6308	2.5557
InternLM-2.5	2.9995	2.9989	2.9990	2.9973
LLaMA-3	2.9987	2.9988	2.9972	2.9935
GPT-4o	**3**	**3**	**3**	**2.9975**
SAMSUM				
Phi-3	2.4623	2.6821	2.5848	2.5060
InternLM-2.5	2.9992	2.9969	**2.9983**	**2.9952**
LLaMA-3	2.9976	2.9971	2.9969	2.9916
GPT-4o	**2.9994**	**2.9977**	2.9982	2.9944

Table 4: G-EVAL

5.1.2 Strong performance of LLaMA-3

The observed superiority of LLaMA-3 over GPT-4o is surprising because an 8 billion 8-bit quantized model not only competes with but also performs better than GPT-4o in certain metrics. We hypothesize that this could be due to the way GPT-4o was trained, which might make it more constrained in its responses, whereas LLaMA-3, being an open-source model, operates with fewer restrictions. This allows LLaMA-3 to exhibit greater flexibility, diversity, and adaptability in generating dialogues, potentially explaining its better performance in certain metrics. These characteristics can be seen in criteria like 'inquisitiveness' and 'likeability' in Table 3 and, 'depth' and 'diverse' in Table 2. These results suggest that for building human-like data generation frameworks, open-source LLMs are a more suitable choice than closed-source LLMs. The minimal constraints on response formatting during the training of open-source models enable them to generate more diverse, flexible, and human-like dialogues, making them better suited for tasks requiring natural and conversational interactions.

5.2 Fine-tuning and Performance Results

To validate the usability of the synthetic data generated using DiaSynth, we fine-tuned and evaluated several pretrained language models on the task of dialogue summarization. The summaries for dialogues generated by different LLMs were created using the corresponding LLMs through prompting. The pretrained models used for evaluation include DistilBART, BART, T5, and LED.

The experimental setup is designed as follows:

Model	Before Finetuning			Finetuning on In-Domain Data		
	QAGS	BERTScore	ROUGE-L	QAGS	BERTScore	ROUGE-L
DIALOGSUM						
distillbart-cnn	0.6134	0.5093	0.1950	0.5586	0.7005	0.3367
bart-base	0.7007	0.5274	0.1375	0.4789	0.6868	0.2969
t5-base	0.5901	0.5491	0.1812	0.4766	0.6953	0.2986
led-base-16384	0.8261	0.5471	0.1634	0.4872	0.7084	0.3165
SAMSUM						
distillbart-cnn	0.6627	0.5500	0.2394	0.6041	0.6849	0.3578
bart-base	0.7563	0.4389	0.1765	0.5302	0.6520	0.3049
t5-base	0.5574	0.4190	0.1237	0.5460	0.6448	0.3000
led-base-16384	0.7429	0.4310	0.1812	0.5440	0.6522	0.3175

Table 5: Performance of models before and after finetuning on in-domain data

Model	Phi-3			InternLM-2.5			LLaMA-3			GPT-4o		
	QAGS	BERTScore	ROUGE-L	QAGS	BERTScore	ROUGE-L	QAGS	BERTScore	ROUGE-L	QAGS	BERTScore	ROUGE-L
DIALOGUESUM												
distillbart-cnn	0.6588	0.5778	**0.2187**	0.6420	0.6008	0.2167	0.6586	0.6161	0.2040	**0.6713**	**0.6242**	0.2014
bart-base	0.5355	0.5958	**0.2029**	0.5418	**0.6212**	0.1897	**0.5825**	0.6033	0.1789	0.5590	0.6039	0.1769
t5-base	0.5937	0.5949	**0.2047**	0.5825	0.5941	0.1878	0.6034	0.6172	0.1959	**0.6305**	**0.6319**	0.2044
led-base-16384	0.5358	0.6129	**0.2109**	0.5189	0.6027	0.1606	0.5697	0.6302	0.1999	**0.5791**	**0.6308**	0.1989
SAMSUM												
distillbart-cnn	0.6585	0.5931	0.2262	0.6388	0.6066	0.2422	**0.6849**	**0.6029**	**0.2374**	0.6757	0.6029	0.2291
bart-base	0.5648	0.5665	0.2146	0.5435	0.5663	0.2021	**0.6132**	**0.5899**	**0.2345**	0.5707	0.5808	0.2154
t5-base	0.5905	0.5397	**0.2085**	0.5457	0.5193	0.1854	**0.6412**	0.5054	0.1976	0.6023	**0.5419**	0.1979
led-base-16384	0.5883	0.5477	0.2289	0.5457	0.5615	0.2167	**0.5917**	**0.5785**	**0.2390**	0.5738	0.569	0.2298

Table 6: Performance after finetuning on synthetic data

- Metrics are reported on the validation and test sets of DialogSum and SAMSum.
- To evaluate DiaSynth-generated data, we compared models fine-tuned on DiaSynth data with their base versions (no fine-tuning).
- In-domain training sets were randomly sampled to match the size of the DiaSynth-generated data, enabling fair comparison.
- The experiment aimed to quantify performance improvement of DiaSynth-fine-tuned models and assess alignment with in-domain data distributions.
- Models were fine-tuned for 2 epochs with a learning rate of `5e-5` and a warmup of `50` steps.

The results presented in Tables 5 and 6 present the performance of the base models, models fine-tuned on in-domain data and models finetuned on DiaSynth generated data. Models finetuned on DiaSynth data generally improves the performances from the BERTScore and ROUGE-L metrics. Surprisingly, for some models (LED and BART) the QAGS scores were higher than the models fine-tuned on DiaSynth. On further exploration, we found out that these models extracted multiple sentences from the given dialogue instead of generating a summary which led to high QAGS scores. Comparing models finetuned on in-domain data to those finetuned on DiaSynth data reveals that DiaSynth finetuning generally enhances factual accuracy, with BERTScore and ROUGE-L scores remaining comparable. The disparity in BERTScore and ROUGE-L results may be due to format variations. Models fine-tuned on in-domain data were evaluated on summaries that matched the training format closely, while DiaSynth-fine-tuned models were trained on LLM-generated summaries and evaluated on human-generated summaries, leading to minor format mismatches. Comparison between the different LLMs from Table 6, shows that GPT-4o is better at generating dialogues and summaries that are formal in nature while LLaMA-3 and open source LLMs would be better for generating dia-

logues that are informal and casual in nature.

$$\% \text{ Improvement} = \frac{\text{After - Before}}{\text{Before}} \tag{1a}$$

$$\% \text{ Coverage} = \frac{\text{Score DiaSynth}}{\text{Score In-domain}} \tag{1b}$$

Model	% Improvement	% Covered
distilbart-cnn	10.96	88.81
bart-base	9.21	90.6
t5-base	7.59	93.67
led-base-16384	2.14	89.68

Table 7: Summarization results on DialogSum

Model	% Improvement	% Covered
distilbart-cnn	6.07	87.25
bart-base	16.12	94.35
t5-base	30.04	87.36
led-base-16384	15.25	90.91

Table 8: Summarization results on SAMSum

To assess the percentage improvement and percentage coverage of the distributional characteristics of the in-domain data by the synthetically generated data, we use Equations 1a and 1b respectively. We use the scores of models finetuned on LLaMA-3 generated data because of its dominance in both quality and usability. Across the 24 reported results, the overall coverage percentage of the LLaMA-3 generated data is **90.48%**. Notably, the QAGS scores of models fine-tuned on synthetic data surpass those of models trained on in-domain data, suggesting that synthetic data can match or even exceed in-domain data performance in some aspects. Excluding QAGS, the coverage percentage is calculated to be **77.07%**. In addition to the average percentages, we also present the model wise percentage improvement and coverage in Table 7 and 8. The results presented are with respect to the dialogues generated using LLaMA-3 and they illustrate clear improvements for every model, highlighting that even with moderate LLMs of small scale (3B - 8B), high-quality synthetic dialogue datasets can be effectively created across different domains and different dialogue formats.

6 Conclusion

In this paper, we introduced DiaSynth, a synthetic dialogue generation framework capable of producing high-quality, contextually rich dialogues across a wide range of domains. Our experiments demonstrated that models fine-tuned on DiaSynth-generated data exhibit significant improvements in downstream tasks, as evidenced by substantial increases in BERTScore and ROUGE-L compared to their base models. These results highlight the potential of DiaSynth as an effective tool for generating dialogue data, particularly in domains where training data is scarce.

Furthermore, our analysis showed that different LLMs excel in different dialogue structures, with LLaMA-3 performing better for informal dialogues and GPT-4o for more structured settings. This insight suggests that leveraging open-source LLMs may be more advantageous for generating human-like conversational data. Despite certain limitations, such as varying LLM performance across dialogue types and knowledge gaps in zero-shot generation, DiaSynth presents a promising approach to dialogue data generation and offers a valuable resource for future advancements in building more sophisticated and adaptable dialogue systems.

7 Limitations

Despite the promising results, our approach has some limitations. Firstly, different LLMs exhibit varied performance based on the dialogue structure, with certain models like LLaMA-3 performing better for more informal dialogues (e.g., SAMSum) and others like GPT-4o excelling in structured, formal dialogues (e.g., DialogSum). This indicates that there is no single model that can universally handle all types of dialogue structures, but a single SOTA model can give stable and decent results like LLaMA-3.

Secondly, the generation process may suffer from a lack of knowledge on certain topics, especially in cases where the LLMs were not sufficiently trained on those domains. Additionally, our framework relies on zero-shot generation for personas and sub-topics, which, while flexible, can sometimes result in less coherent or less accurate persona simulation, as it is not fine-tuned for specific contexts.

Since LLMs power DiaSynth, hallucinations and compute-need are two inherent limitations. We present a detailed hallucination study in Appendix A, which indicates that though the hallucinations rates are acceptable, there is still scope for improvements. These limitations suggest directions for future work, such as combining LLMs to leverage

their strengths or incorporating more topic-specific training to enhance knowledge coverage.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiao wen Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, Aijia Guo, Qipeng Guo, Conghui He, Yingfan Hu, Ting Huang, Tao Jiang, Penglong Jiao, Zhen Jin, Zhikai Lei, Jiaxing Li, Jingwen Li, Linyang Li, Shuaibin Li, Wei Li, Yining Li, Hongwei Liu, Jiangning Liu, Jiawei Hong, Kaiwen Liu, Kui-Jie Liu, Xiaoran Liu, Chen Lv, Haijun Lv, Kai Lv, Li Ma, Runyuan Ma, Zerun Ma, Wenchang Ning, Linke Ouyang, Jiantao Qiu, Yuan Qu, Fukai Shang, Yunfan Shao, Demin Song, Zifan Song, Zhihao Sui, Peng Sun, Yu Sun, Huanze Tang, Bin Wang, Guoteng Wang, Jiaqi Wang, Jiayu Wang, Rui Wang, Yudong Wang, Ziyi Wang, Xing Wei, Qizhen Weng, Fan Wu, Yingtong Xiong, Chao Xu, Rui Ze Xu, Hang Yan, Yirong Yan, Xiaogui Yang, Haochen Ye, Huaiyuan Ying, Jia Yu, Jing Yu, Yuhang Zang, Chuyu Zhang, Li Zhang, Pan Zhang, Peng Zhang, Ruijie Zhang, Shuo Zhang, Songyang Zhang, Wenjian Zhang, Wenwei Zhang, Xingcheng Zhang, Xinyue Zhang, Hui Zhao, Qian Zhao, Xiaomeng Zhao, Fen-Fang Zhou, Zaida Zhou, Jingming Zhuo, Yi-Ling Zou, Xipeng Qiu, Yu Qiao, and Dahua Lin. 2024. Internlm2 technical report. *ArXiv*, abs/2403.17297.

Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2024. Scaling synthetic data creation with 1,000,000,000 personas. *arXiv preprint arXiv:2406.20094*.

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021. DialogSum: A real-life scenario dialogue summarization dataset. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5062–5074, Online. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Song Feng, Hui Wan, Chulaka Gunasekara, Siva Patel, Sachindra Joshi, and Luis Lastras. 2020. doc2dial: A goal-oriented document-grounded dialogue dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8118–8128, Online. Association for Computational Linguistics.

Robert Friel and Atindriyo Sanyal. 2023. Chainpoll: A high efficacy method for llm hallucination detection. *ArXiv*, abs/2310.18344.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2024. GPTScore: Evaluate as you desire. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6556–6576, Mexico City, Mexico. Association for Computational Linguistics.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.

Ji-Eun Han, Jun-Seok Koh, Hyeon-Tae Seo, Du-Seong Chang, and Kyung-Ah Sohn. 2024. PSYDIAL: Personality-based synthetic dialogue generation using large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13321–13331, Torino, Italia. ELRA and ICCL.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore. Association for Computational Linguistics.

Shikib Mehri and Maxine Eskenazi. 2020. Unsupervised evaluation of interactive dialog with DialoGPT. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 225–235, 1st virtual meeting. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Sebastian Steindl, Ulrich Schäfer, and Bernd Ludwig. 2023. Generating synthetic dialogues from prompts to improve task-oriented dialogue systems. In *KI 2023: Advances in Artificial Intelligence: 46th German Conference on AI, Berlin, Germany, September 26–29, 2023, Proceedings*, page 207–214, Berlin, Heidelberg. Springer-Verlag.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online. Association for Computational Linguistics.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2024. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, Hongchao Fang, Penghui Zhu, Shu Chen, and Pengtao Xie. 2020. MedDialog: Large-scale medical dialogue datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9241–9250, Online. Association for Computational Linguistics.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A Hallucination Study

In addition to evaluating the quality and usability of dialogues produced by DiaSynth, we conducted a study on the phenomenon of hallucinations within the generated dialogues. Hallucinations in language models refer to instances where the output contains misleading or incorrect information or situations where the model repeats the same content. To evaluate the occurrence of hallucinations, we compared the generated dialogues with their respective summaries and assessed them using two well-known hallucination benchmarks: **SelfCheckGPT** (Manakul et al., 2023) and **ChainPoll** (Friel and Sanyal, 2023). This analysis provides insights into the prevalence of hallucinations and informs strategies for improving dialogue quality in future iterations of DiaSynth. The results are presented in Tables 9 and 10.

A.1 SelfCheckGPT

SelfCheckGPT quantifies the self-consistency of LLM outputs by examining agreement across multiple outputs from the same prompt. This assessment reveals potential inaccuracies through metrics like SelfCheck-BertScore.

The SelfCheck-BERTScore results for various models show that hallucination levels are at worst around 25%, which is acceptable but still indicates areas for improvement. Across both datasets, **Phi-3** demonstrates the most robustness, likely due to its pretraining on structured, textbook-like data, which may contribute to greater consistency and factual accuracy.

A.2 ChainPoll

ChainPoll utilizes a chain-of-thought prompting approach to identify hallucinations by iteratively polling the model with structured reasoning prompts. This method systematically detects both open-domain and closed-domain hallucinations, where lower scores indicate fewer hallucinations.

The ChainPoll scores indicate that hallucination levels on these models are generally low, with the best performance seen by **GPT-4o** on SAMSum, which achieves the lowest score of 0.120, suggesting minimal hallucinations. On the other hand, **LLaMA-3** scores higher at 0.237 on SAMSum, indicating more frequent hallucinations. These findings highlight different models' strengths in generating accurate and reliable dialogues.

A.3 Implications for DiaSynth

The results from both SelfCheckGPT and ChainPoll evaluations suggest that DiaSynth, when leveraging models like Llama 3, is capable of generating dialogues with relatively low hallucination rates. However, specific models show variability in performance across datasets, indicating that further enhancements, such as fine-tuning or incorporating additional guardrails, could improve DiaSynth's robustness in generating reliable dialogues across diverse domains.

LLM	ChainPoll	SCGPT-BERTScore
Phi-3	0.198	0.791
InternLM-2.5	0.199	0.726
LLaMA-3	0.205	0.793
GPT-4o	0.178	0.765

Table 9: Hallucination calculation for DialogSum few-shot data

LLM	ChainPoll	SCGPT-BERTScore
Phi-3	0.154	0.785
InternLM-2.5	0.159	0.716
LLaMA-3	0.237	0.733
GPT-4o	0.120	0.742

Table 10: Hallucination calculation for SAMSum few-shot data

B Scalability of DiaSynth

This section illustrates the scalability of DiaSynth with a python program and examples.

```python
def calculate_total_dialogs(n, m, p):
    """
    Calculate the total number of dialogs generated.

    Parameters:
    n (int): Number of topics
    m (int): Number of subtopics per topic
    p (int): Number of personas per subtopic

    Returns:
    int: Total number of dialogs generated
    """
    # Calculate the total number of subtopics
    total_subtopics = n * m

    # Calculate the number of dialogs for each subtopic using combinations of personas
    dialogs_per_subtopic = (p * (p - 1)) // 2

    # Total dialogs generated across all subtopics
    total_dialogs = total_subtopics * dialogs_per_subtopic

    return total_dialogs
```

Listing 1: Calculation of Total Dialogs Generated

B.1 Example 1

- Number of topics (n): 10
- Number of subtopics per topic (m): 5
- Number of personas per subtopic (p): 3

Calculation:

$$\text{Total subtopics} = n \times m = 10 \times 5 = 50$$
$$\text{Dialogues per subtopic} = \frac{p \times (p-1)}{2} = \frac{3 \times 2}{2} = 3$$
$$\text{Total dialogues} = 50 \times 3 = 150$$

This setup generates **150 dialogues**.

B.2 Example 2

- Number of topics (n): 20
- Number of subtopics per topic (m): 4
- Number of personas per subtopic (p): 5

Calculation:

$$\text{Total subtopics} = 20 \times 4 = 80$$
$$\text{Dialogues per subtopic} = \frac{5 \times 4}{2} = 10$$
$$\text{Total dialogues} = 80 \times 10 = 800$$

This setup generates **800 dialogues**.

B.3 Example 3

- Number of topics (n): 15
- Number of subtopics per topic (m): 6
- Number of personas per subtopic (p): 10

Calculation:

$$\text{Total subtopics} = 15 \times 6 = 90$$
$$\text{Dialogues per subtopic} = \frac{10 \times 9}{2} = 45$$
$$\text{Total dialogues} = 90 \times 45 = 4050$$

This setup generates **4050 dialogues**.

B.4 Scaling Observations

- **Linear Scaling with Topics and Subtopics**: Increasing the number of topics or subtopics results in a linear increase in the total number of dialogues, making it straightforward to expand the scope of dialogue generation.
- **Exponential Scaling with Personas**: The number of dialogues scales exponentially as the number of personas increases because each additional persona allows for more combinations, making the framework highly scalable for complex scenarios.
- **Practical Use Case**: For specific domains like academic, healthcare, or business, these parameters can be adjusted to generate thousands of dialogues to fit the needs of various applications such as training chatbots, virtual assistants, or dialogue-based assessments.

These examples illustrate DiaSynth's potential for rapid and scalable generation of dialogues, which can be tailored to different domains by simply adjusting the input parameters.

C Characteristics for the conversation

Table 11 shows different characteristics that we let the LLMs reason and decide using CoT. Before generating the dialogues, the LLMs are prompted to first reason about the various characteristics list for the dialogue given the topic and the personas and how the LLMs reason are illustrated in Appendix D.

Characteristic	**Description**
Age and Gender	Defines demographic details, influencing style and tone.
Familiarity Level	Affects formality and depth based on relationship between speakers.
Emotional States	Impacts tone and flow based on emotions (e.g., happy, sad).
Formality Level	Determines level of politeness or casualness.
Duration of the Conversation	Suggests the intended length and complexity of dialogue.
Communication Medium	Defines the medium (e.g., face-to-face, phone), influencing style.
Topic of the Conversation	Guides the content and direction of the dialogue.
Location of the Conversation	Adds context influencing formality and content.
Agreement or Disagreement	Drives dialogue dynamics based on agreement level.
Natural Dialogue Features	Adds authenticity with fillers, pauses, and slang.

Table 11: Characteristics of the Dialogue for CoT Prompt

D Example CoT environments

Topic	AI-powered content generation for marketing and advertising
Persona 1	A marketing manager at a mid-sized e-commerce company looking to leverage AI-generated content to increase their social media presence
Persona 2	A startup founder creating an AI-powered content generation platform for small businesses and entrepreneurs to create high-quality marketing materials
Environment created by CoT	1. **Age and gender of both personas**: Both personas are likely to be in their 30s or 40s, with the marketing manager being a woman and the startup founder being a man. This is because the marketing manager is likely to have more experience in the industry, while the startup founder is likely to be younger and more tech-savvy. 2. **How familiar the speakers are with each other**: The marketing manager and the startup founder are likely to be acquaintances, having met at industry events or conferences. They may have a professional relationship but not a close personal one. 3. **Emotional states of the speakers:** The marketing manager is likely to be curious and open-minded, while the startup founder is likely to be enthusiastic and optimistic. This is because the marketing manager is looking for new ways to leverage AI-powered content generation, while the startup founder is passionate about his product. 4. **Formality level:** The conversation is likely to be semi-formal, with a mix of professional and casual language. 5. **Duration of the conversation**: The conversation is likely to last around 15-20 minutes, as they discuss the potential applications of AI-powered content generation for marketing and advertising. 6. **Communication medium:** The conversation is likely to take place in-person at a conference or industry event, or via video conference. 7. **Topic of the conversation:** The topic of the conversation is AI-powered content generation for marketing and advertising, with a focus on its potential applications and benefits. 8. **Location of the conversation**: The conversation is likely to take place at a conference or industry event, or at the startup founder's office. 9. **Agreement or disagreement on the viewpoints of the conversation topic**: The marketing manager and the startup founder are likely to have differing opinions on the potential applications and benefits of AI-powered content generation, but they will also be open to learning from each other and finding common ground. 10. **Incorporation of natural human dialogue characteristics:** The conversation will include natural human dialogue characteristics such as fillers (e.g., "umm", "uh"), pauses, and slang where appropriate.

Figure 2: CoT Example 1

Topic	Recommendations for book clubs and reading groups
Persona 1	A retired English literature professor who has read over 500 books and is always looking for new titles to recommend to her book club.
Persona 2	A stay-at-home mom who loves to read and is looking for book recommendations that she can share with her friends and family.
Environment created by CoT	1. **Age and gender of both personas**: Both personas are adults, with the retired English literature professor being an older adult (around 60-65 years old) and the stay-at-home mom being a younger adult (around 35-40 years old). The professor is a woman, and the stay-at-home mom is a woman as well. 2. **How familiar the speakers are with each other**: The speakers are acquaintances, having met through a book club or a reading group. They have a friendly and casual relationship, but not extremely close. 3. **Emotional states of the speakers:** The retired professor is in a peaceful and calm stat having just finished reading a book and looking forward to discussing it with others. The stay-at-home mom is in a relaxed and curious state, looking for new book recommendations to share with her friends and family. 4. **Formality level:** The conversation will be informal, with a friendly and casual tone. 5. **Duration of the conversation:** The conversation will be relatively short, lasting around 0-15 minutes. 6. **Communication medium:** The conversation will take place through a phone call or video call. 7. **Topic of the conversation:** The topic of the conversation will be book recommendation for a reading group or book club. 8. **Location of the conversation**: The conversation will take place remotely, with the speakers being in different locations. 9. **Agreement or disagreement on the viewpoints of the conversation topic**: The speakers will likely agree on some book recommendations and disagree on others, leading to a lively and engaging discussion. 10. **Incorporation of natural human dialogue characteristics**: The conversation will include fillers (e.g., "umm", "uh"), pauses, and slang where appropriate to make it sound natural and realistic. The retired professor, being an expert in English literature, will likely dominate the conversation with her vast knowledge and recommendations. The stay-at-home mom will ask questions and share her own thoughts and opinions, seeking to learn more about the books and authors. The conversation will be a mix of book recommendations, discussions, and personal anecdotes.

Figure 3: CoT Example 2

These examples illustrate how CoT sets the various dialogue characteristics defined in Appendix C. As can be seen in the examples, the characteristics for the conversation are completely changed based on the personas and the topics, resulting in more grounded conversation generation. Future works can further explore how CoT can be used to further break down to generate even more realistic dialogues.

E Ablation study on the use of CoT

In this section of the appendix, we present the impact of CoT in the dialogues generated by DiaSynth. We generate two datasets without CoT, using Phi-3 using DialogSum and SAMSum as few-shot examples with 8 topics. Tables 12, 13, 14 and 15 compare the scores of dialogues generated with and without CoT for the FED score and GPTScore and it can be clearly that CoT tends to increase the quality of the dialogues generated by DiaSynth.

We hypothesize that this improvement in quality is due to allowing the LLM to set diverse characteristics for the dialogue before generating the dialogue. This illustrates that either manually setting the relevant context or letting the LLM on its own to set the relevant context, we get better outputs, as adding relevant context lowers the probabilities of sequences that are not useful. The lower FED scores of CoT generated dialogues in Table 14, might be because of the CoT generated dialogues being longer in length but it needs further research.

Criteria	Without CoT	With CoT
Coherent	0.9507	**0.9521**
Error Recovery	0.938	**0.9424**
Consistent	0.9424	**0.9523**
Diverse	0.9431	**0.952**
Depth	0.9451	**0.9506**
Likeable	**0.0088**	-0.0003
Understand	0.9317	**0.9338**
Flexible	-0.0013	**0.0001**
Informative	**0.0032**	0.0009
Inquisitive	**0.0095**	-0.003

Table 12: FED scores of dialogues generated with and without CoT for DialogSum few-shot

Criteria	Without CoT	With CoT
Coherence	0.0052	**0.0284**
Diversity	0.0120	**0.0303**
Flexibility	0.0074	**0.0215**
Understandability	0.0056	**0.019**
Inquistiveness	0.0162	**0.0362**
Consistency	0.0091	**0.0366**
Informativeness	0.0099	**0.0168**
Likeability	0.0029	**0.0209**
Depth	0.0062	**0.0115**
Error Recovery	0.0142	**0.0242**

Table 13: GPTScore of dialogues generated with and without CoT for DialogSum few-shot

Criteria	Without CoT	With CoT
Coherent	**0.9667**	0.9125
Error Recovery	**0.9577**	0.9051
Consistent	**0.9621**	0.9163
Diverse	**0.9586**	0.9124
Depth	**0.9589**	0.9094
Likeable	**0.0059**	-0.0003
Understand	**0.9503**	0.8978
Flexible	**-0.0022**	-0.0023
Informative	**0.0110**	0.0035
Inquisitive	**0.0030**	-0.0037

Table 14: FED scores of dialogues generated with and without CoT for SAMSum few-shot

Criteria	Without CoT	With CoT
Coherence	0.0076	**0.0324**
Diversity	0.0172	**0.0372**
Flexibility	0.0116	**0.0259**
Understandability	0.0104	**0.0272**
Inquistiveness	0.0201	**0.0389**
Consistency	0.0149	**0.0415**
Informativeness	0.0148	**0.0200**
Likeability	0.0041	**0.023**
Depth	0.0078	**0.0124**
Error Recovery	0.0163	**0.0289**

Table 15: GPTScore of dialogues generated with and without CoT for SAMSum few-shot