Synthesizing Human Conversation Data with Large Language Models

Azalea Bailey[1], Ryan Shea[2], Yunan Lu[2], Zhou Yu[2]
University of California, Berkeley[1]; Columbia University[2]

Introduction: Artificial Intelligence (AI) chatbots are revolutionizing various industries [1]. Companies are motivated to invest in chatbots to reduce labor costs and streamline their operations. However, training these chatbots can be challenging due to the need for historical conversation data, which is often not readily available and costly to obtain [2]. This paper explores two methods for synthesizing conversation data between a company's AI chatbot and a Large Language Model (LLM) that mimics human dialogue [3]. The first method involves prompting an LLM to generate utterances based on the company's AI chatbot's user dialogue intent categorization tree [3]. The second method prompts an LLM to generate a user goal for the conversation to achieve based on a description of the company [3]. This study finds that LLMs can replicate human dialogue at a statistically significant level regarding conversation intent distribution and Bidirectional Encoder Representations from Transformers Score (BERTScore) [4].

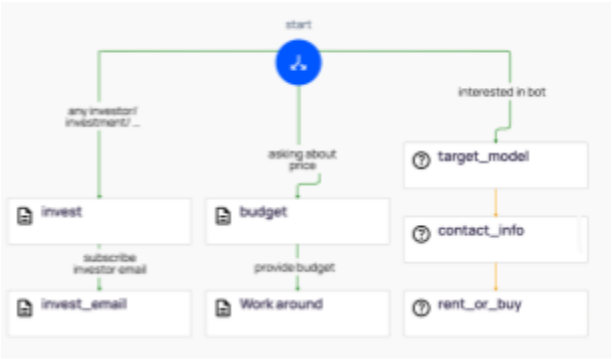

Figure 1: The graph above shows the logical flow of how the Articulate.AI chatbot responds to user utterances. The edges represent the intent that the LLM categorizes the incoming user utterance and the following node indicates the topic addressed by the Articulate.AI chatbot in its response.

Methods: To generate utterances based on a company's chatbot's user dialogue intent categorization tree, the tree was first obtained from Articulate.AI's intent tree for RichTech Robotics. The Articulate.AI chatbot pipeline takes in the user utterance and labels it with an intent so the Articulate.AI chatbot can respond accordingly as shown in Figure 1. The tree represents different paths the chatbot and user conversation can follow, the edges representing the intent of the incoming user utterance and the node representing the topic that the Articulate.AI chatbot responds with. The edge color is relevant to which path the conversation can follow. If the edge is green, any node on the graph can point to the node in question using the node in question's incoming intent. If the edge is yellow, only what is shown in Figure 1 can point to the node in question. For example, all nodes point to the "budget" node with the intent "asking about price" because that edge is green. For the "contact_info" node, only the "target_model" node points to the "contact_info" node because that edge is yellow. I will reference the green nodes on the tree as "global", the yellow nodes as "local", and the green nodes not shown on the tree as "new global." These different types of nodes make the tree highly connected, resulting in many paths that could be sampled.

To mimic human and chatbot dialogue, paths were sampled based on the distribution of paths taken by real human interactions with the RichTech customer service chatbot. Then, the intent paths were prompt engineered and given to GPT-4o using a few-shot technique based on historical conversation data. Here is the prompt:

> *Replicate the writing behavior of a human customer. You are interacting with customer service chatbot for the following company: [RichTech's about]. Humans write short questions with typos and a neutral sentiment. Here are some examples of what a human customer would type: [how much is the adam coffee robot?, Can you send info to my email, yes I need a job, want to check both proposals to rent and buy, How much does it cost a barista robot, Worker robots, Im interested in beverages robot s, hi i would like to rent out ARM, but im wondering which countries are available for rental]. Replicate the writing behavior of a human customer and begin the conversation with the following intent: [relative intent].*

Each generated utterance was fed to RichTech's customer service chatbot and the conversation history was provided to GPT-4o posing as the human each turn. I will refer to this method of data generation as Intent-Based.

To synthesize utterances based on a user goal, GPT-4o was prompted to generate a user goal based on a description of RichTech Robotics. This goal was prompt engineered and given to GPT-4o using a few-shot technique to synthesize human dialogue. Here is the prompt:

> *Replicate the writing behavior of a human customer. You are interacting with a customer service chatbot for the following company: [RichTech's about]. Humans write short questions with typos and a neutral sentiment. Here are some examples of what a human customer would type: [how much is the adam coffee robot?, Can you send info to my email, yes I need a job, want to check both proposals to rent and buy, How much does it cost a barista robot, Worker robots, Im interested in beverages robot s, hi i would like to rent out ARM, but im wondering which countries are available for rental]. Finally, respond to this utterance keeping the following goal in mind: [generated user goal].*

Each generated utterance was fed to RichTech's customer service chatbot and the conversation history was provided to GPT-4o posing as the human each turn. I will refer to this method of data generation as Goal-Based.

	Type	Global	New Global	Local
0	Real Data	0.626466	0.132328	0.241206
1	Intent-Based Data	0.182857	0.725714	0.091429
2	Goal-Based Data	0.715789	0.284211	0.000000

Figure 2: Table showing edge frequencies across different data

To measure the similarity between each method's generated conversations and real human chatbot interactions, a BERTScore was calculated between 20 real conversations and 20 randomly selected synthetic conversations for each data generation method. A BERTScore between 20 different real conversations was also calculated. Additionally, the percentages of global, new global, and local nodes for each dataset were calculated. The total edges used to calculate the edge type percentages for the real, Intent-Based, and Goal-Based data was 597, 175, and 190, respectively.

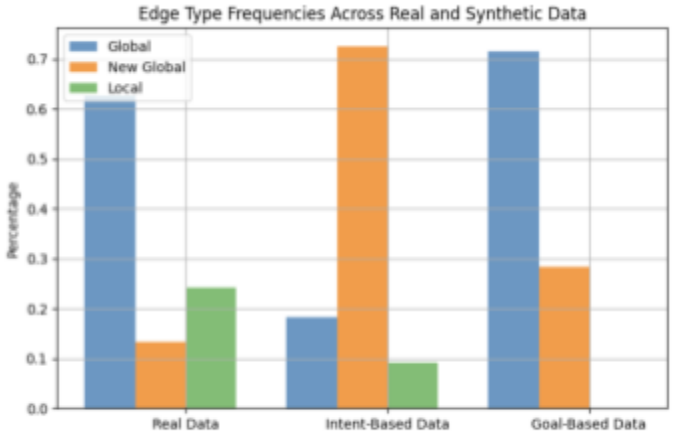

Figure 3: Graph of edge frequencies across different data

Results: The average BERTScore between the real and Intent-Based data was 71.14%. The average BERTScore between the real data and the Goal-Based Data was 84.33%. The average BERTScore between the two different sets of real conversation data was 84.85%. The p-value between the BERTScores for the real and the Intent-Based data is < .01 and the p-value for the Goal-Based data is > .05. The results of the edge type percentages are shown in Figure 2 and 3 [5, 6]. The Goal-Based Data had edge frequencies similar enough to the real data to have p-value < .01 while the Intent-Based Data had p-value > .05.

Conclusion: The BERTScore between the Intent-Based and the real data as well as the similar edge type frequencies between the real and the Goal-Based data indicate that an LLM can successfully mimic human dialogue with a customer service chatbot. However, looking at the conversations qualitatively, the LLM better replicates human dialogue patterns with an overarching goal rather than a series of intents. The Intent-Based conversations are choppy and jump topics quickly while the Goal-Based data appears more natural and humanlike. In the future, it would be interesting to explore combining the two data generation methods to synthesize less conversationally choppy dialogue data.

References:

1. Yu X, Wu Q, Qian K, Yu Z. arXiv. 2023;2211.16773.
2. Qin L, Pan W, Chen Q, et al. EMNLP. 2023;5925-5941.
3. Davidson S, Romeo S, Shu R, et al. arXiv. 2023;2309.13233.
4. Zhang T, Kishore V, Wu F, et al. ICLR. 2020.
5. Reback J, McKinney W, et al. Zenodo. 2020.
6. Harris CR, Millman KJ, van der Walt SJ, et al. Nature. 2020;585:357-362.

Acknowledgements: I would like to thank Columbia University and Amazon SURE for sponsoring my research this summer. Furthermore, I want to say a special thank you to Ryan for giving me a truly valuable first research experience. And thanks to Dr. Yu for giving me a space in her lab to explore AI research.