

Synthesizing Human Conversation Data with Large Language Models

Azalea Bailey¹, Ryan Shea², Yunan Lu², Zhou Yu²
University of California, Berkeley¹; Columbia University²

References: 1. Yu X, Wu Q, Qian K, Yu Z. arXiv. 2023;2211.16773. Acknowledgements: I would like to thank Columbia University and Amazon SURE for sponsoring my research this summer.
2. Qin L, Fan W, Chen Q, et al. EMNLP. 2023;5925-5941. Furthermore, I want to say a special thank you to Ryan for giving me a truly valuable first research experience. And thanks to Dr.
3. Davidson S, Romeo S, Shu R, et al. arXiv. 2023;2309.13233. Yu for giving me a space in her lab to explore AI research.

Background

Artificial Intelligence (AI) chatbots can reduce labor costs and streamline various companies' operations. However, training these chatbots needs data, an expensive and timely item 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]:

1. Intent-Based Utterance Generation
2. Goal-Based Utterance Generation

Objective

Replicate how a human converses with an AI customer service chatbot.

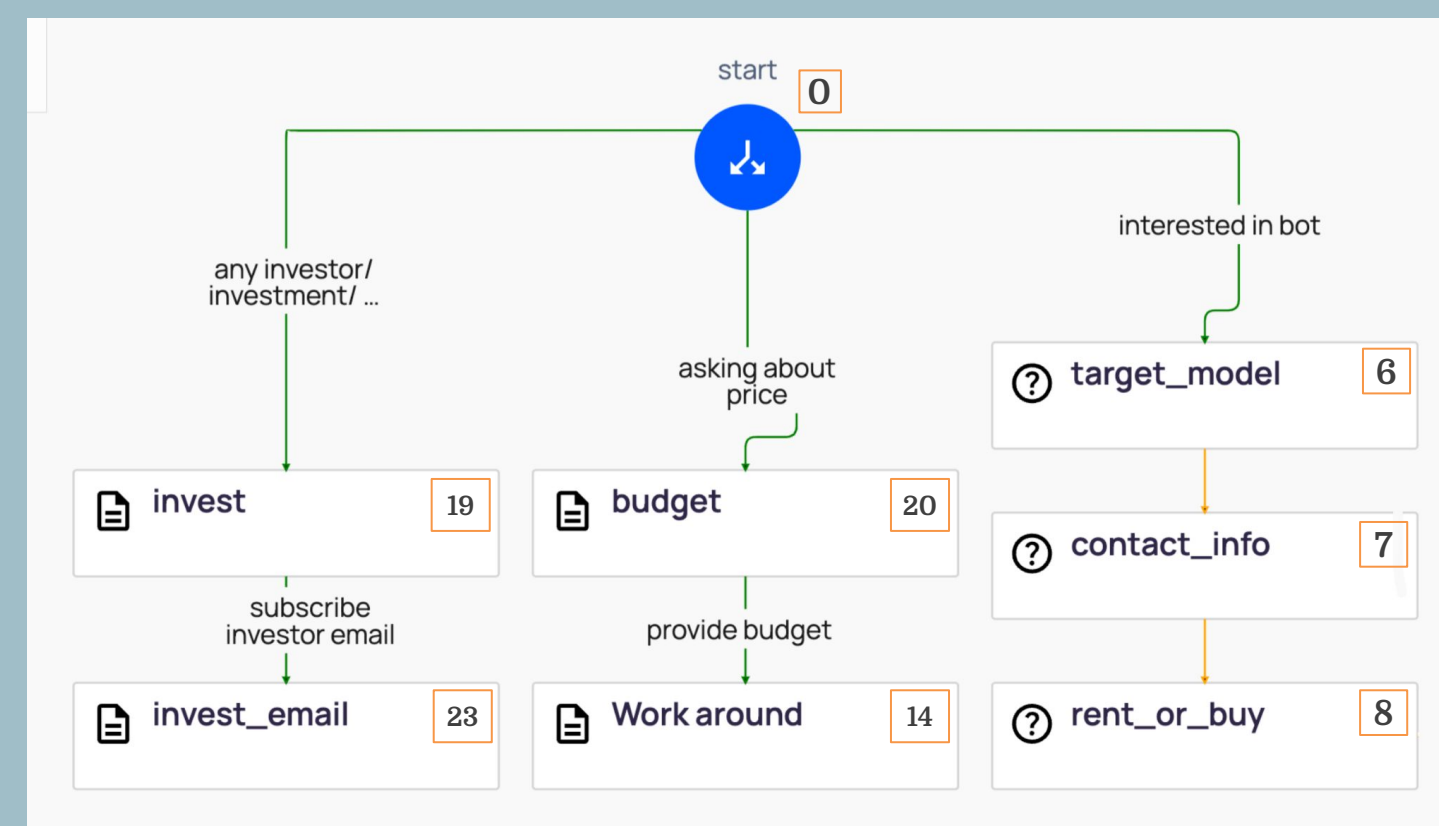
Global Edge: green edges in the first graph
New Global Edge: new green edges in the second graph
Local Edge: yellow edge in the first graph

Synthetic Conversation Example:

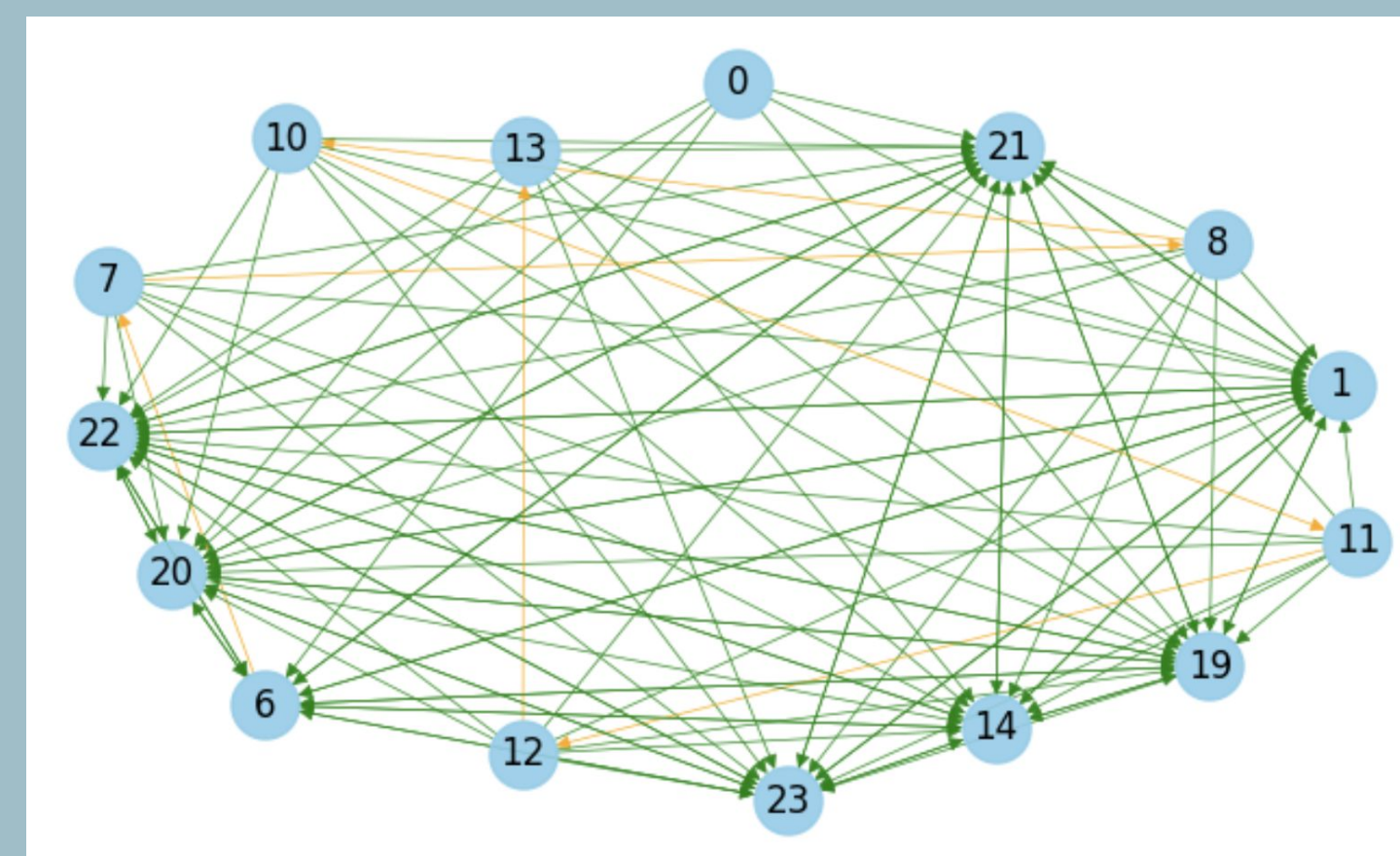
user: Hi, interested in bot. Can you give me info on ADAM n other models?
chatbot: Hi! ADAM is a versatile robot bartender...
user: Can you send info to my email. Also how much does it cost a barista robot? interesaed in ADAM n maybe Matradee. Thx
chatbot: Price? You've just hit the top-secret...
user: can you send info to my email pls. Also want to check both proposals to rent and buy. Thx. Also interested in investor updates.
chatbot: Sure! I'll send the information to your email. For details on both rental and purchase...
user: yes I need a job. Do you have any openings at the Las Vegas or Austin office?

Intent-Based

1. Sample intent paths



*the actual number of paths that could be sampled



2. Prompt GPT-4o to generate utterances based on sampled intent paths

What Worked:

- Clear instructions
 - Humans write short questions with typos and a neutral sentiment.
- Providing examples
 - Here are examples of what a human would type: [how much is the adam coffee robot?, ...]

What Didn't:

- Trusting GPT-4o to how the average person types

Methodology

Goal-Based

1. Prompt GPT-4o to generate a goal based on company description

Generate a goal representing what a customer or a potential customer would have based on the description of the following company: [company description]

2. Prompt GPT-4o to generate utterances based on goal → end conversation if goal

What Worked:

- Clear instructions
 - Replicate the writing behavior of a human customer and begin the conversation with a question to answer the following goal: [goal]
- Providing examples

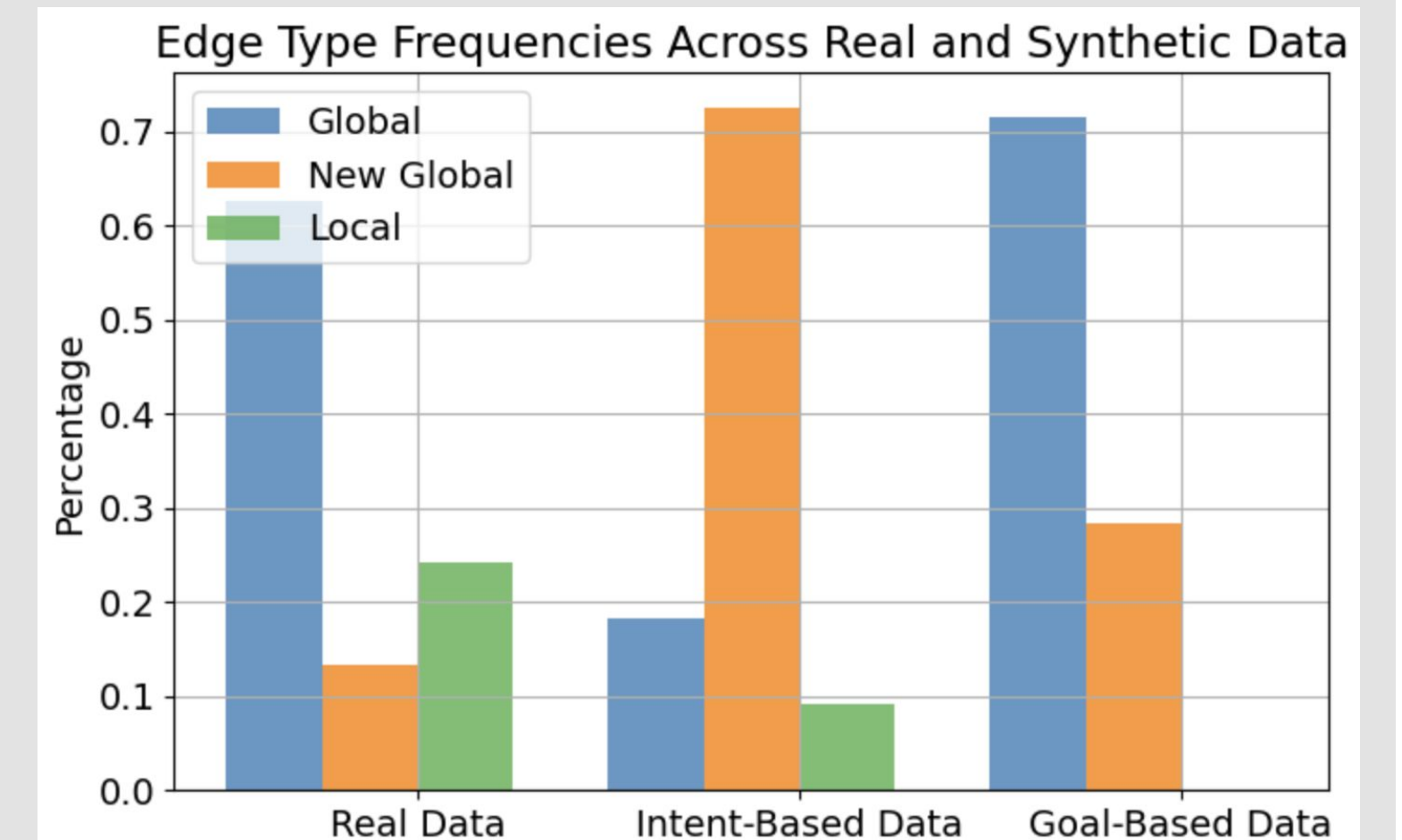
What Didn't:

- Providing too much information about the company

Synthetic Conversation Example:

user: Hi, looking to streamline my business with ur robots, what options do you have?
chatbot: Hi! We offer worker robots like ADAM (for beverages), delivery robots like Matradee, and cleaning robots like DUST-E SX. How can I as
user: Can you send me info on ADAM and Matradee to my email?
chatbot: Sure! We're creating a support ticket now. Please provide your email address and phone...
user: Sure, my email is example@email.com and phone number is 123-456-7890. Also how much does Matradee cost?
chatbot: Price? ...Do you have a budget? The...
user: Thanks! Can you also send details on buying vs renting ADAM?
chatbot: "Sure! To rent ADAM, it's \$10,000/day in Las Vegas and \$15,000/day elsewhere (min 2 days). To buy, please contact us for a quote..."

Results



Average BERTScore against Real Convo:

- Real Convo Data: 84.85%
- Intent-Based: 71.14%
- Goal-Based: 84.33%

BERTScore Distribution p-values

- Intent-Based: p-value < .01, significant
- Goal-Based: p-value > .05, not significant
 - Intent-Based is more similar to human dialogue than Goal-Based

Edge Type Frequency p-values

- Intent-Based: p-value > .05, not significant
- Goal-Based: p-value < .01, significant
 - Goal-Based is more similar to human dialogue than Intent-Based

Conclusion

- Our metrics shows that GPT-4o does a good job at replicating human dialogue when interacting with a customer service chatbot
- Qualitatively, Goal-Based conversations are more human-like than Intent-Based, but neither replicate human dialogue perfectly