Artificial Intelligence and User Interaction

Anushya Shankar
Georgia Institute of Technology
Princeton, US
anushya.shankar@gmail.com

Allen Zhao
Rowan University
Princeton, US
zhaoa889@gmail.com

Adam Greissman
Pearl Capital Business Funding, LLC
Princeton, US
alg@udico.com

I. Introduction

The research goal was to determine if the current mix of AI-related technologies could be used to construct useful conversational agents for customer support applications using intent-fulfillment dialog. ChatGPT can engage the user in a compelling dialog about any subject. However, this project aimed to integrate both user profile data and externally sourced data into the conversation to create a meaningful and directed experience. For example, there is a fundamental difference between having a generalized discussion about mathematics and assisting a student who has known challenges to study for an upcoming exam. This principle can be applied to many domains, including customer support and what are commonly referred to as recommender applications.

The internship research team reported to the CTO of Pearl Capital, and the purpose of the project was to assess the use of AI for account management, credit applications, and follow-up for collections. However, the data and the model for those applications are proprietary so the research team was given the option to select a non-proprietary domain for the work.

Overall, the plan was to construct an artificial intelligence application that combines user profiles with external data to lead a user through directed dialogue for a specific fulfillment.

II. Background

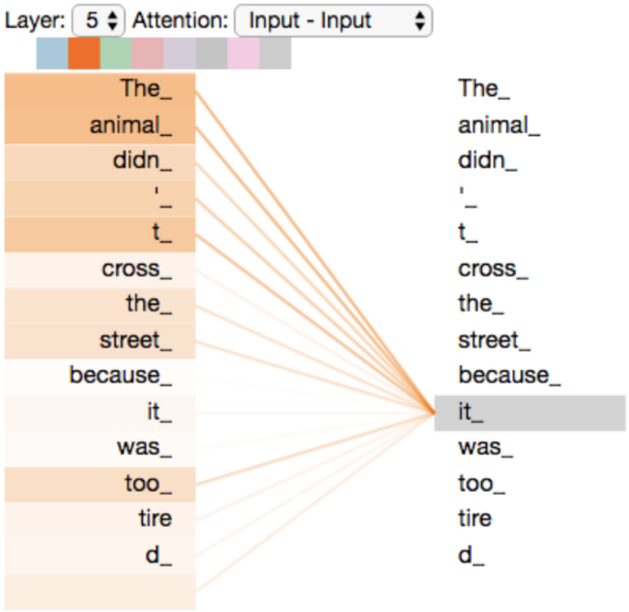
The core of large language models (LLMs) is natural language processing (NLP), the intersection of linguistics, artificial intelligence, and computer science [1]. NLP techniques facilitate communication between computer language and human language. Large language models then utilize NLP techniques within massive systems to communicate fluently with users. [2]. Almost all LLMs today use the Transformer model. This model empirically selects steps in the language understanding process and utilizes self-attention mechanisms to focus on different sections of the text input. LLMs are characterized by large amounts of textual training data, enabling them to communicate on a level comparable to that of their users [3].

The goal for the team was to test the limits of the technology by providing it with external context. This would allow for the improvement of the user experience by personalizing the text generation.

III. Journey

A. Timeline of Project

During the project, different tools and sub-goals were considered. Each step of the timeline clarified the value of the tools and what types of applications were possible, especially within the timeframe of the project.



Layer: 5 Attention: Input - Input

Fig. 1: Example of the self-attention mechanism in the Transformer model.

1) Initial Stages: Due to the scope of the project, the decision was to not make the project proprietary by using data such as credit reporting. This opened the project to a range of diverse ideas.

2) Predictive Analysis: In the early stages of the team's work with AI, the exploration was concentrated on the use of large language models (LLMs) in the predictive analysis of house prices.

3) Prompt Engineering: After predictive analysis was ruled out, a few examples were tested of fine-tuning LLMs using prompt engineering.

4) InsightEd: GPT-4 was leveraged to create an organic and directed conversation between the user and the model.

B. Tools

1) GPT-4: Using the OpenAI API, the team engineered prompts and used GPT-4's responses to create a conversation with the user. GPT-4 is the most recent iteration of models from OpenAI, and it has surpassed much of the industry and its previous iterations in natural language processing capabilities and text generation techniques [4]. In this project, textual context is created to pass into the GPT model, which then curates a personalized experience for the user.

2) PaLM: The Pathways Language Model (PaLM) is trained in 780 billion tokens of language data [5]. It is a densely activated, autoregressive Transformer; when a model is densely activated, this means that all model parameters are used to process all input examples [6]. An autoregressive model (ARM) uses techniques that allow it to predict future data based on previous data and actions [7]. In utilizing the PaLM API from Google MakerSuite, the team was able to engineer prompts that would assist ChatGPT in adequately assisting the user.

3) Dialogflow: DialogFlow as a platform, consisting of Dialogflow ES and CX, allows developers to integrate conversational frameworks into their services. This project used ES, which provides the standard agent type best suited for simple back-and-forth conversation. CX, on the other hand, offers the advanced agent type for a more complex and nuanced model [8]. Dialogflow ES allowed for the creation of an intents-fulfillments network that helped map out the user interactions.

4) Pinecone: Pinecone allowed for the creation of a database to aggregate vector embeddings that hold information. Vector embeddings are a type of data representation that carry semantic information that the model can later draw on. Pinecone allows for organizing those vector embeddings and simplifying the query process [9]. The team leveraged Pinecone during the phase of the project where the focus was on predictive analysis; the plan was to create embeddings that represent data about each property entity.

5) EditorX: EditorX is a Wix site editor that allowed the team to construct the website and house the chatbot.

C. Predictive Analysis

The initial endeavor aimed to explore the potential utility of a large language model (LLM)'s reasoning capabilities in predicting the price of a house by considering various factors. These factors included attributes related to the properties themselves, such as material quantity and past sales records, as well as broader regional statistics, including socioeconomic trends and demographic data. An overarching objective was to extract meaningful features from textual content through sentiment analysis. However, during this phase of the project, the focus shifted toward feature extraction via time series data analysis.

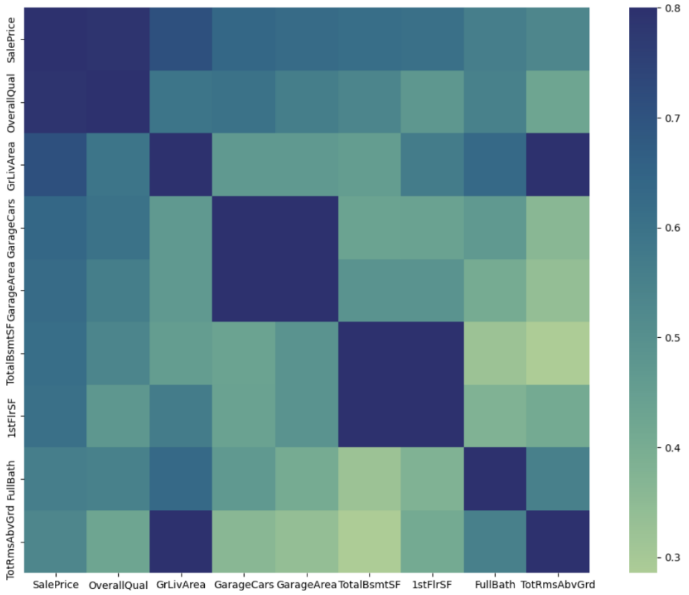


Fig. 2: Example heatmap showing factors of interest in relation to property pricing.

1) Tools: Pinecone was harnessed to create a comprehensive vector database, serving as the repository for aggregating a pool of external information about the aforementioned factors. The database would encompass a wide spectrum of relevant vector embeddings, each housing property and regional data.

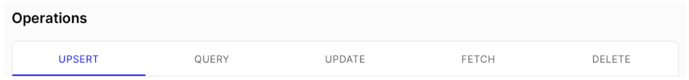


Fig. 3: Pinecone Index operations. [9]

ChatGPT's predictive capabilities and knowledge context were then utilized for prediction. By feeding the compiled vector database as external context, ChatGPT would then generate outputs that reflect a more informed and comprehensive analysis of the data.

2) Challenges: While in the early development stages of this task, certain challenges came to the fore. It became evident that the successful execution of the task necessitated the construction of a large language model tailored explicitly to handle time series data. In this considerable undertaking, difficulty with the tools prompted the team to investigate multiple open-source large language models as potential alternatives. This exploration presented further obstacles and complexities. Ultimately, due to this roadblock and the depth of data required to train a new model, this venture was determined to be beyond the scope of the project.

D. Prompt Engineering

1) Prior Ideas: The team commenced its exploration by crafting a prompt for ChatGPT, simulating the role of an ice cream store vendor. This exercise served as a demonstrative means of delineating a path from user intent to subsequent fulfillment. The primary objective of the chatbot was to aid users in selecting an ice cream flavor tailored to their preferences, leveraging pertinent information about them. The experiment, characterized by a directed flow and limited context, acted as an ideal starting point for delving into this subject matter. Subsequently, the team undertook the development of additional models centered around themes such as baseball or horoscopes, thereby emulating the same intent-fulfillment paradigm.

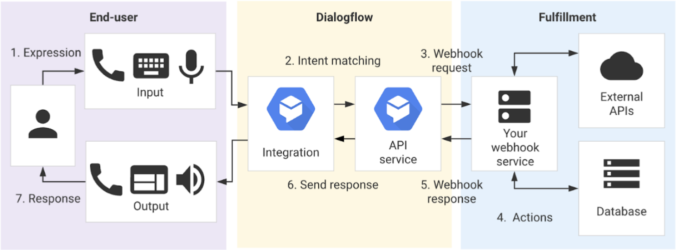


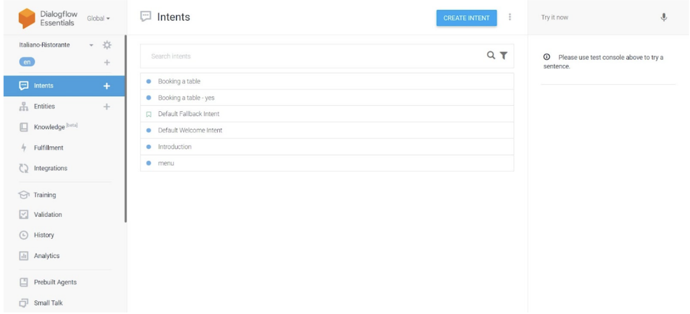
Fig. 4: Processing flow for fulfillment with typical DialogFlow implementation. [8]

Anushya's experience with teaching in both large group formats and one-on-one presented a promising opportunity for the project venture to pivot toward an educational emphasis.

Ultimately, Anushya inspired the team to create a math tutoring chatbot, as it demonstrated a meaningful fulfillment that could be extended beyond the context of this project. Not only could a tutoring-style chatbot hold the potential for broader integration within the educational sphere, but this conversation format could also be extrapolated to diverse queries and requests. Personalized education by way of artificial intelligence could function as an invaluable adjunct for teachers.

The significance of this initiative lay in the prospect of improving education and enhancing access to knowledge on a larger scale. The groundwork laid during this process offered the potential for a transformative impact on education. The task was now to implement pedagogical techniques to simulate an organic educational experience within the chatbot's functionality.

2) *Tools:* In the earlier stages of the project, Dialogflow was used to help define the conversational framework facilitating the interaction between the user and the machine. Dialogflow ES was chosen over Dialogflow CX because it mapped out a chatbot-style, question-and-answer model, which was perfect for the simplicity of the team's model. ChatGPT continued to be a core component of the project, but instead of using databases from other APIs to create the context fed into the model, external context was crafted from scratch to pass into the transformer model and receive informed output. During the project, the team also gained access to the Google PaLM family of models, introducing a new dimension to the integration process. The team's early efforts in incorporating these models were characterized by trial and error, as the goal was to be familiarized with the new API and its chatbot generation capabilities as they interfaced with OpenAI's API.

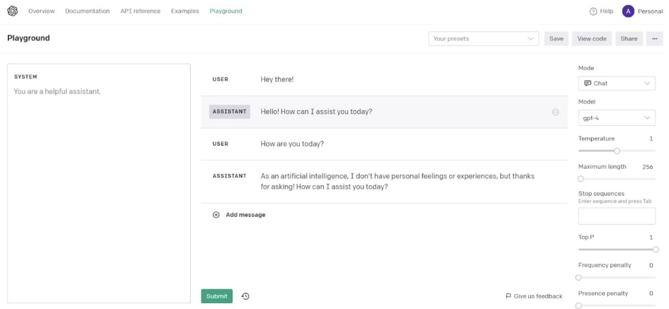


Booking a table
Booking a table - yes
Default Fallback Intent
Default Welcome Intent
Introduction
menu

Please use test console above to try a sentence.

Fig. 5: DialogFlow playground for developers. [8]

3) *Challenges:* Some of the challenges faced in this phase of the project necessitated minor adjustments and adaptations. Initially, Dialogflow enabled the mapping of each of the intents and fulfillments within the conversation. However, it also presented limitations when attempting to personalize the interaction and introduced delays when a user interacted with it. Consequently, the decision was made to remove its implementation to improve the overall quality of the product. Secondly, prompt engineering with GPT-3 posed considerable difficulties once the team opted for a tutoring-oriented approach. GPT-3 encountered challenges in accurately answering questions or tracking the subtleties inherent in natural conversation which would be second nature to any tutor. However, many of those smaller impediments were resolved when the team gained access to GPT-4 through the waitlist. This iteration of ChatGPT greatly improved the quality of the final product.



SYSTEM
You are a helpful assistant.

USER Hey there!

ASSISTANT Hello! How can I assist you today?

USER How are you today?

ASSISTANT As an artificial intelligence, I don't have personal feelings or experiences, but thanks for asking! How can I assist you today?

Mode: Chat
Model: gpt-4
Temperature 1
Maximum length 256
Stop sequences
Top P 1
Frequency penalty 0
Presence penalty 0

Fig. 6: GPT-4 playground for developers.

E. Development Process

1) *Profiles:* The original concept of InsightEd revolved around leveraging existing student data to provide personalized insight into what teaching style would be best suited for them. Since this wasn't feasible to replicate on a smaller scale, the decision was to, instead, create four distinct profiles for users to choose from with pre-loaded data.

Each of the profiles encompasses detailed information about grades, the class that pertains to the chatbot, extracurriculars,

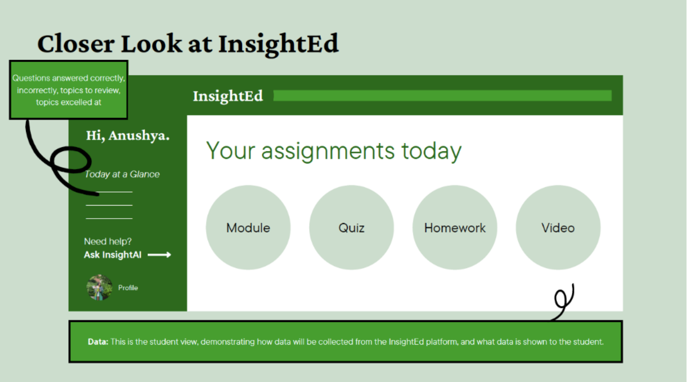

Closer Look at InsightEd
Questions answered correctly, incorrectly, topics to review, topics excelled at
InsightEd
Hi, Anushya.
Today at a Glance
Your assignments today
Module
Quiz
Homework
Video
Need help?
Ask InsightAI
Profile
Data: This is the student view, demonstrating how data will be collected from the InsightEd platform, and what data is shown to the student.

Fig. 7: Initial approach to InsightEd, as applied in a preexisting school system.

etc. They also each align with a high school stereotype, implicitly informing the user of more niche details. The four users are John Smith, a nerd, Jane Doe, an athlete, Mary Jones, an artist, and Peter Brown, a gamer. The four students each have their interests and characteristics that feed into their niches, allowing the user to assume their personality easily.

Fig. 8: The four user profiles.

2) External Data and Time: One sub-goal was to incorporate real-time information into the conversation to demonstrate a genuine integration of tools, considering that ChatGPT's training was conducted in 2021. However, after obtaining access to PaLM through the developer waitlist, it was determined that PaLM lacks the capability to access live Google Search like Bard does. Therefore, the team was unable to move forward with this step. One of the main issues with the system responses was lag time. When users entered messages, generating a response took a considerable amount of time, which was impeding the overall user experience. To resolve this, unnecessary implementations were removed to reduce the strain on the system. This step included removing DialogFlow, as mentioned earlier.

3) Context: Many of the techniques employed in engineering the prompts and achieving certain conversation outcomes relied on persistent trial and error. Subtle modifications in language exerted a notable influence on the machine's responses. For instance, GPT-4 required multiple examples with various edge cases to equip the model for a natural conversation. Iterating through minor changes in language and instruction helped to further elucidate the mechanics of the large language model. The intricacies of the training involved were reflected in the meticulousness of the prompt engineering process.

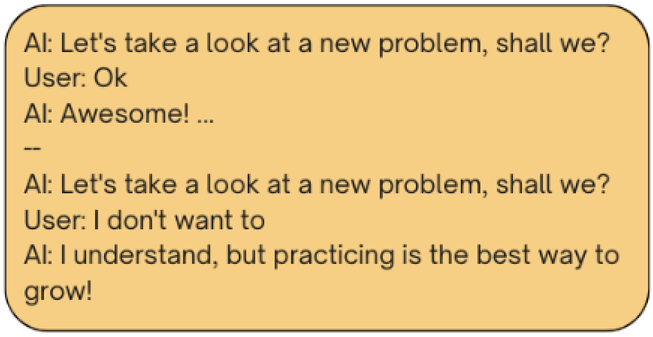

AI: Let's take a look at a new problem, shall we?
User: Ok
AI: Awesome! ...
--
AI: Let's take a look at a new problem, shall we?
User: I don't want to
AI: I understand, but practicing is the best way to grow!

Fig. 9: Example of Conversation Training.

IV. Results

A. *InsightEd*

The final product is InsightEd, an AI-driven tutoring tool designed to provide learners with a personalized teaching experience. Powered by GPT-4, InsightEd engages in natural and engaging conversations with users. The user demo for InsightEd simulates a typical student profile by incorporating pre-set information about the learner complemented by user input. Leveraging prior data and contextual information associated with each profile, the tool provides an interactive and challenging environment to learn math. It also prioritizes the review of errors and fostering growth in the learning process.

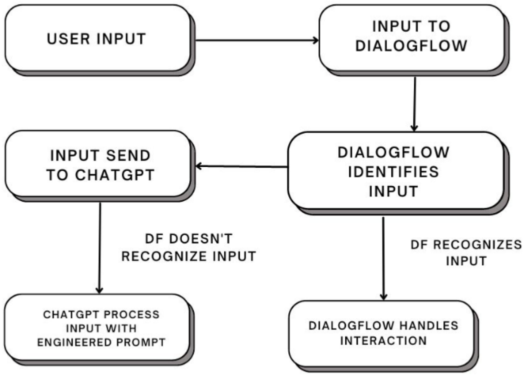
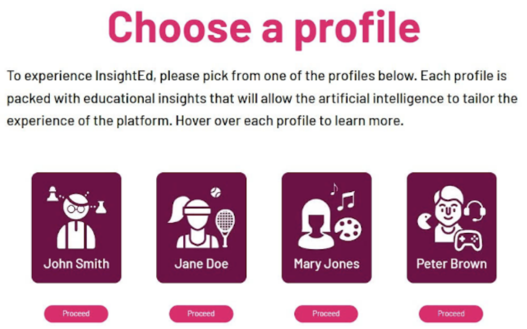

USER INPUT
INPUT TO DIALOGFLOW
DIALOGFLOW IDENTIFIES INPUT
INPUT SEND TO CHATGPT
DF DOESN'T RECOGNIZE INPUT
DF RECOGNIZES INPUT
CHATGPT PROCESS INPUT WITH ENGINEERED PROMPT
DIALOGFLOW HANDLES INTERACTION

Fig. 10: Flowchart of the full conversation process.

B. *Demo Walkthrough*

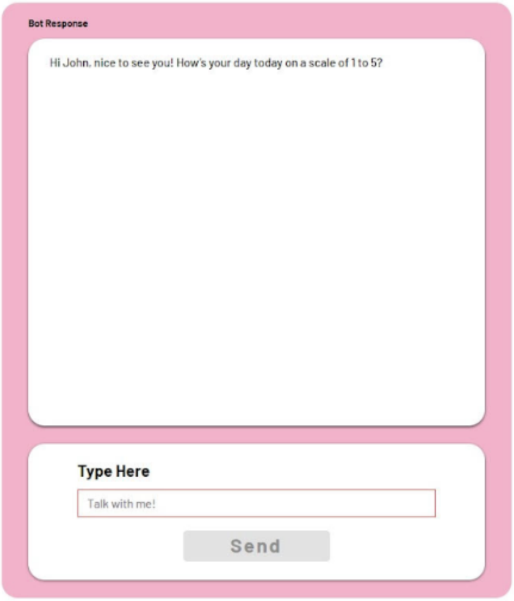
When the user first enters the website (insighted-demo.com), they are presented with the following screen where they can access each of the profiles and make an informed decision on which one to assume when conversing with the artificial intelligence model.


Choose a profile

To experience InsightEd, please pick from one of the profiles below. Each profile is packed with educational insights that will allow the artificial intelligence to tailor the experience of the platform. Hover over each profile to learn more.

Fig. 11: Profile selection screen.

Once the user decides on a profile, they will press the “Proceed” button and be redirected to their first look at the chatbot.



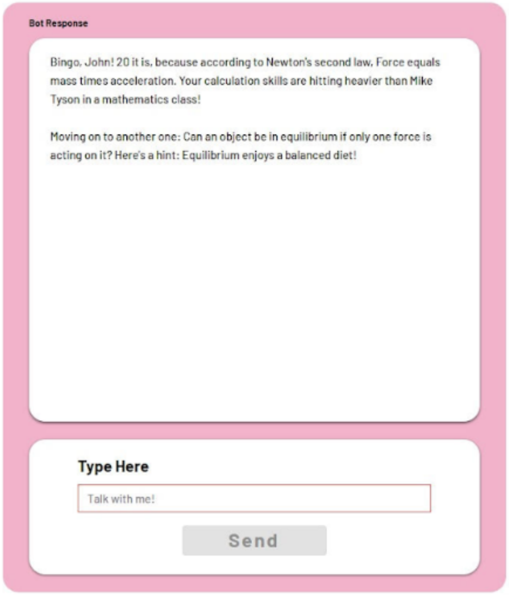
Hi John, nice to see you! How's your day today on a scale of 1 to 5?

Fig. 12: Introduction from chatbot.

This is where the chatbot initiates the conversation with the user to commence the intent-fulfillment paradigm. The chatbot asks a series of questions that will range in difficulty based on topics established in the early parts of the conversation. If the user answers correctly, the chatbot congratulates them and then moves to a question of increased difficulty.

If the user answers incorrectly, however, the chatbot allows them to try again until they’ve reached the answer. It will provide feedback and attempt to help them through the steps to try and ascertain the challenges.

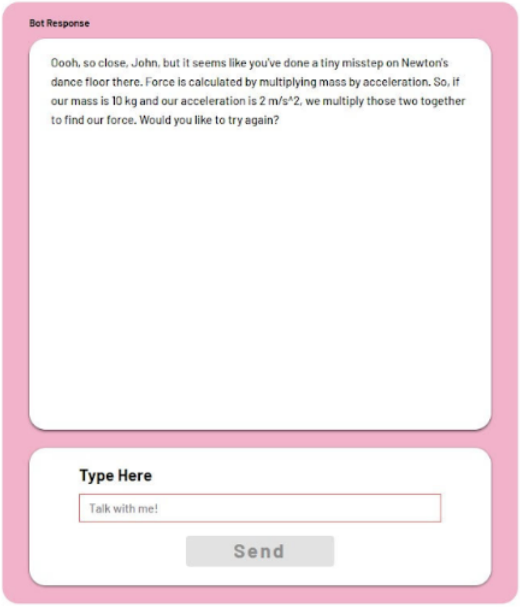
Once the user either answers correctly or gives up, the chatbot reiterates the correct answer and then adjusts the following problem to an easier difficulty with similar topics. This demonstrates a capacity for flexibility and adaptation that mirrors a real tutor.



Bingo, John! 20 it is, because according to Newton's second law, Force equals mass times acceleration. Your calculation skills are hitting heavier than Mike Tyson in a mathematics class!

Moving on to another one: Can an object be in equilibrium if only one force is acting on it? Here's a hint: Equilibrium enjoys a balanced diet!

Fig. 13: Correct answer dialog.



Oooh, so close, John, but it seems like you've done a tiny misstep on Newton's dance floor there. Force is calculated by multiplying mass by acceleration. So, if our mass is 10 kg and our acceleration is 2 m/s^2, we multiply those two together to find our force. Would you like to try again?

Fig. 14: Correct answer dialog.

V. Conclusion

This endeavor is a demonstration that GPT-4 can be customized and applied to a specific context, allowing for specialization and increased quality of output.

Having said that, creating customer support agents for account inquiries or outbound calling follow-up would not be an easy task. The intent-fulfillment flow needs to be carefully orchestrated in the prompt, and tools that are external to the prompt, such as DialogFlow, introduced complexity but did not add value with respect to managing the narrative arc of the conversation.

Due to the restricted timeframe allocated for the successful completion of this project, numerous development ideas required adjustments to ensure that the work could be accomplished within the prescribed timeframe.

In specific relation to InsightEd, one significant limitation involved the lack of pre-aggregated student data that the team could utilize as a backbone for the model's insights. This led to the creation of four base profiles, as mentioned earlier, that the model could then converse with based on previously injected information. These profiles were required to be general enough that any given user could assume them and, in conjunction, detailed enough that the conversation could adequately elucidate the model's personalization capability.

References

[1] J. P. Bharadiya, "A comprehensive survey of deep learning techniques natural language processing," *European Journal of Technology*, vol. 7, no. 1, pp. 58–66, 2023.

[2] T. Teubner, C. M. Flath, C. Weinhardt, W. van der Aalst, and O. Hinz, "Welcome to the era of chatgpt," *Business and Information Systems Engineering*, vol. 65, pp. 95–101, 2023.

[3] J. U. et al., "Transformer: A novel neural network architecture for language understanding," 2017. Google Research Blog.

[4] OpenAI, "Gpt-4 technical report," 2023.

[5] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, P. Schuh, K. Shi, S. Tsvyashchenko, J. Maynez, A. Rao, P. Barnes, Y. Tay, N. Shazeer, V. Prabhakaran, E. Reif, N. Du, B. Hutchinson, R. Pope, J. Bradbury, J. Austin, M. Isard, G. Gur-Ari, P. Yin, T. Duke, A. Levskaya, S. Ghemawat, S. Dev, H. Michalewski, X. Garcia, V. Misra, K. Robinson, L. Fedus, D. Zhou, D. Ippolito, D. Luan, H. Lim, B. Zoph, A. Spiridonov, R. Sepassi, D. Dohan, S. Agrawal, M. Omernick, A. M. Dai, T. S. Pillai, M. Pellat, A. Lewkowycz, E. Moreira, R. Child, O. Polozov, K. Lee, Z. Zhou, X. Wang, B. Saeta, M. Diaz, O. Firat, M. Catasta, J. Wei, K. Meier-Hellstern, D. Eck, J. Dean, S. Petrov, and N. Fiedel, "Palm: Scaling language modeling with pathways," 2022.

[6] R. Liu, Y. J. Kim, A. Muzio, and H. H. Awadalla, "Gating dropout: Communication-efficient regularization for sparsely activated transformers," 2022.

[7] E. Hoogeboom, A. A. Gritsenko, J. Bastings, B. Poole, R. van den Berg, and T. Salimans, "Autoregressive diffusion models," 2022.

[8] Google, "Dialogflow documentation," 2023.

[9] Pinecone, "What is a vector database?," 2023.