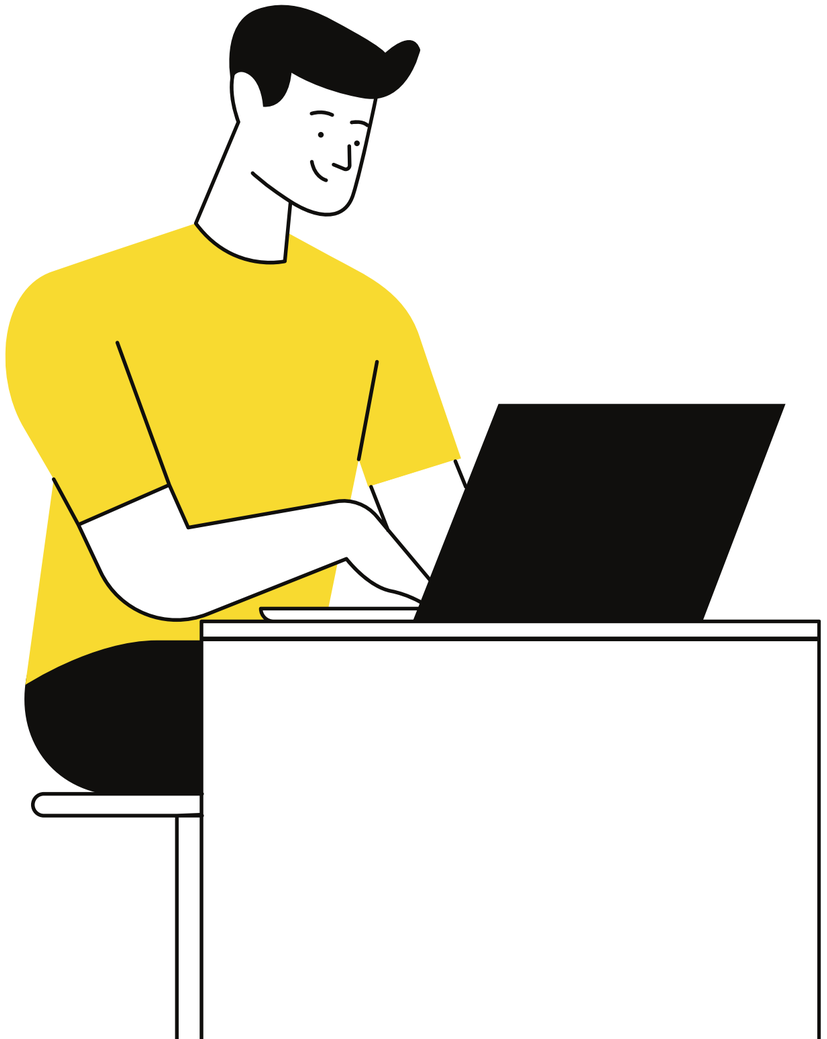
#TEAM_43

NOVEMBER 19, 2023

Proposal Idea for Retail Chatbot

Let's get visual with our processes!

Agenda

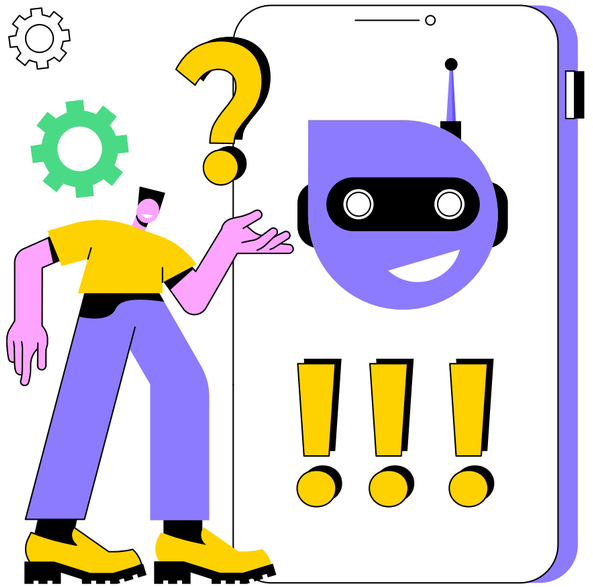


1. Overview
2. Model Architecture
3. Roadmap & Timeline
4. Conclusion

Overview

In today's *fast-paced* retail environment, customers demand quick, accurate information about products and availability

ChatBot



Leverages datasets from Adidas, Nike, and IKEA

Offers real-time product recommendations and updates

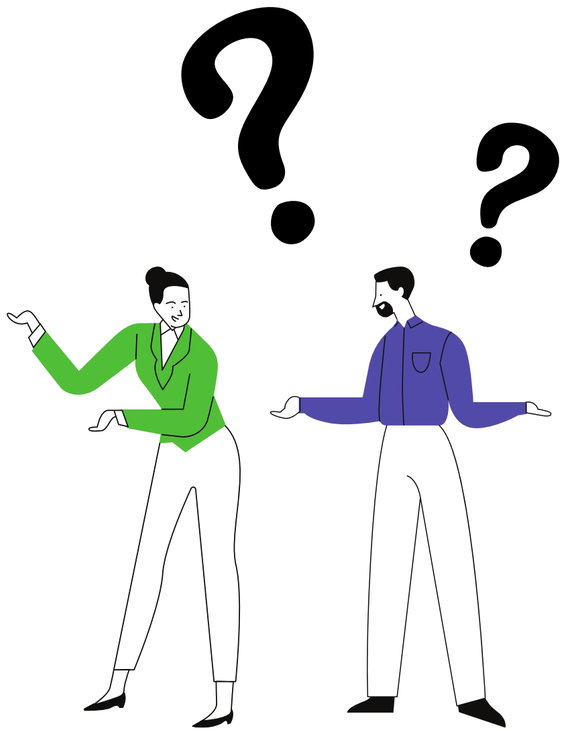
Bridges the gap between online browsing and informed purchasing decisions

Serves as a Product Inquiry

Overview

Problem *Statement*

Problems MVP seeks to solve:



- Client services are usually unavailable when it is **out of working hours**, clients' **issues can't be addressed** at all time
- If there is someone behind the screen replying to customers, there should be a **limit** for the most **conversations** they can handle **at a time**

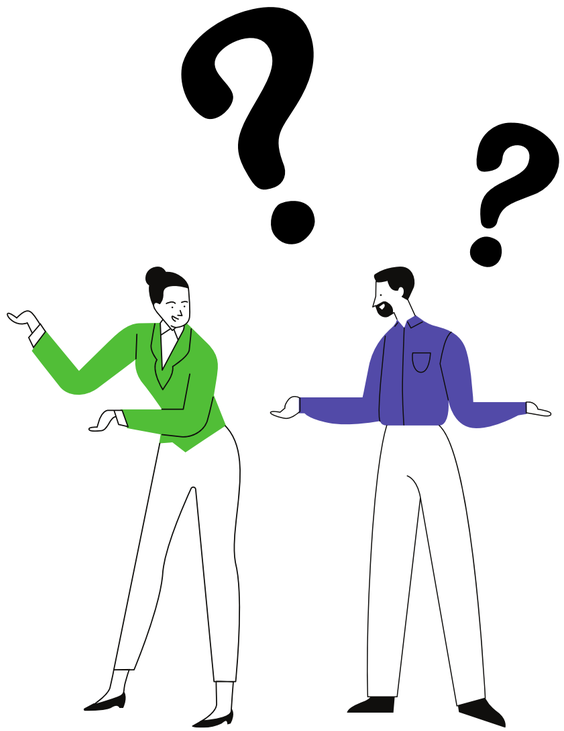
⇒ *Inefficient when the numbers of clients are large*

- Many times clients may ask about stocks, policies or even the closest store to them, which will **take some time** for agents to **search and reply**

⇒ *Time-consuming with a high probability of wrong information*

Overview

Problem *Statement*



Pain points & Inefficiency:

- **Limited Knowledge**: Chatbot might struggle to understand customers' ideas, slangs or typo
- **Limited recommendations**: Cannot fully cover various fashion styles that clients may ask for
- **Low problem solving skills**: Because it can only provide a limited range of skills that are trained

Overview

The MVP presents a cutting-edge AI-based solution, specifically designed to **revolutionize the retail shopping experience**. At its core, the system utilizes advanced **Natural Language Processing (NLP)** models, many of which are open-source and have been fine-tuned for this specific application. These NLP models enable the chatbot to understand and process user queries in natural language, **making the interaction intuitive and user-friendly**

Overview

One of the key features of our MVP is the **Real-Time Stock Update and Alerts** function. To demonstrate this capability, we've **integrated a simulated database** that represents real product inventories from **Adidas, Nike, and IKEA**

Beyond information provision, the chatbot utilizes the database to make **recommendations and decisions** based on user preferences, creating a highly interactive and responsive experience. This is achieved through the **integration of sophisticated AI algorithms** analyzing **user input, preferences, and behavior patterns**

Overview

The **solution's innovation** lies in its dual focus on **technical prowess** and meeting **specific business needs**. The application of open-source NLP models in a retail context, coupled with real-time data tracking and personalized recommendations, represents a novel approach

From a **business perspective**, the solution addresses a critical need by providing timely, accurate, and personalized shopping assistance, enhancing the customer experience and supporting effective inventory management for retailers.

Methodologies

We are going to use GPT – 2 Model and the following slides will provide information about:

- *Model description*
- *Explanations of key components*
- *Architecture of technologies that we intend to use*

LET'S BEGIN!

Methodologies

Model

Description

GPT-2 is a **transformers model pre-trained** on a very **large corpus of English** data in a **self-supervised** fashion. This means it was pretrained on the raw texts only, with no humans labeling them in any way (which is why it can use lots of publicly available data) with an automatic process to generate inputs and labels from those texts. More precisely, it was trained to guess the next word in sentences.

More precisely, **inputs** are sequences of **continuous text** of **a certain length** and the targets are the same sequence, shifted one token (word or piece of word) to the right. The model uses internally a **mask-mechanism** to make sure the predictions for the token i only uses the inputs from 1 to i but not the future tokens.

This way, the model learns an inner representation of the English language that can then be used to **extract features** useful for downstream tasks. The model is **best at** what it was pre-trained for, which is **generating texts from a prompt**.

Methodologies

Key *Components*

Components	Explanation
Transformer Architecture	Encoder-Decoder Structure: GPT-2 is based on the Transformer architecture, which consists of an encoder-decoder structure. However, GPT-2 is a decoder-only model, meaning it is designed for autoregressive language modeling
Attention Mechanism	Self-Attention Mechanism: The attention mechanism is a crucial component in Transformers. GPT-2 uses a self-attention mechanism, allowing each position in the input sequence to focus on different positions, capturing long-range dependencies
Positional Encoding	Since Transformers do not inherently understand the sequential order of data, positional embeddings are added to the input embeddings to provide information about the positions of tokens in the input sequence
Multi-Head Self-Attention	GPT-2 uses multi-head self-attention, allowing the model to attend to different positions in the input sequence with multiple sets of attention weights. This enhances the model's ability to capture different types of relationships.
Position-wise Feedforward Networks	After the attention mechanism, the model employs position-wise feedforward networks for each position, providing non-linear transformations to the input

Methodologies

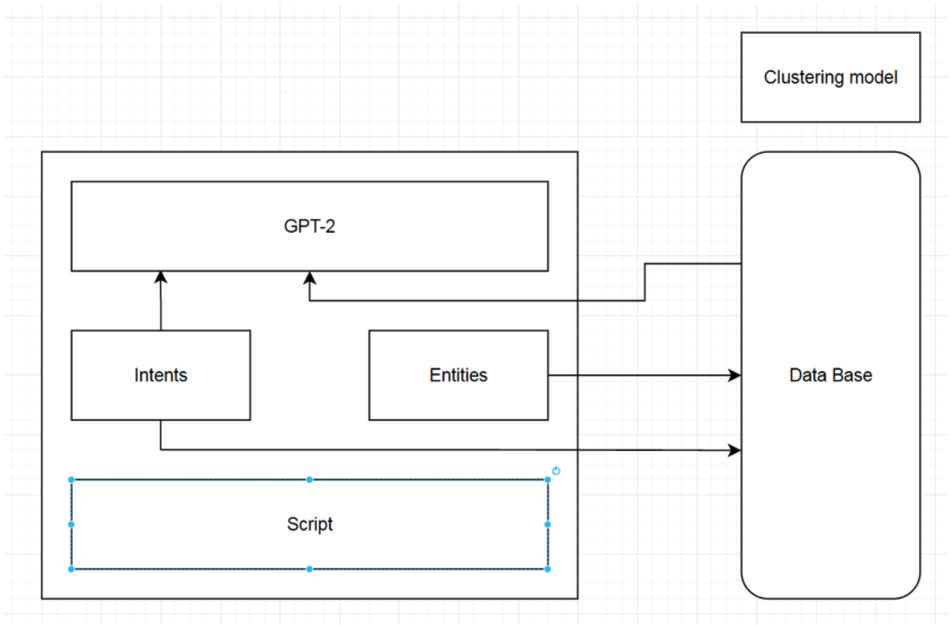
Key *Components*

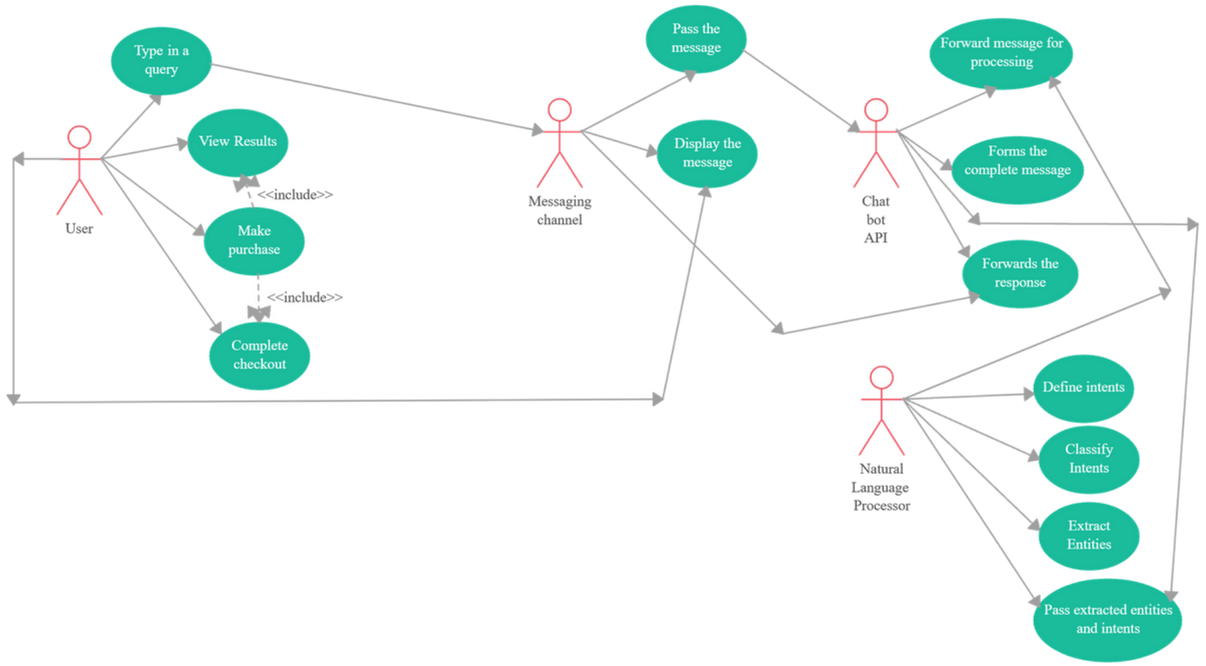
Components	Explanation
Layer Normalization	Each sub-layer in the model has a layer normalization step, contributing to the stability and faster training of the network.
Residual Connections	Residual connections (or skip connections) are used to pass the original input along with the output of the sub-layer, helping in the flow of gradients during training
Feedforward Neural Network	Position-wise Feedforward Networks: After the attention mechanism, the model employs position-wise feedforward networks for each position, providing non-linear transformations to the input
Parameterized Scaling	GPT-2 introduces a hyperparameter called the "scale" that helps control the extent of the model's influence on the output probability distribution. This parameter is important for adjusting the level of randomness in the generated text
Output Layer	Softmax Layer: The output layer of GPT-2 is typically a softmax layer, producing a probability distribution over the vocabulary. During training, the model is optimized to generate the correct next token given the context.

Methodologies

Technologies

intend to use





Core functionality

Technologies
intend to use

Personalized Product Finder Objective: Utilize detailed product descriptions to help users find products that match their preferences.

⇒ *Implementation:*

- Integrate a conversational AI that ⇒ processes natural language inputs to interpret user descriptions.
- Implement a recommendation system using dataset attributes to match products with user descriptions.

Core functionality

Technologies
intend to use

Product Inquiry Assistant Objective: Answer user queries regarding product details, stock availability, and more.

⇒ *Implementation:*

- Design the chatbot to access real-time data from the dataset for accurate information.
- Enable the chatbot to provide comparative product information upon request.

Core functionality

Technologies
intend to use

Real-Time Stock Update and Alerts Objective: Inform users about the availability of products and stock changes.

⇒ *Implementation:*

- Integrate a real-time tracking ⇒ system for product availability updates.
- Implement a notification system for user alerts based on their interest and previous queries.

Metrics

- User Engagement:
  - Number of chatbot uses within a specific time period.
  - Average time each user interacts with the chatbot.
  - User feedback ratio after interacting with the chatbot (conversion rate).
- Product Search Accuracy:
  - Number of times users find desired products through the chatbot.
  - Ratio of successful product search queries to the total number of queries.
- Real-Time Inventory Updates:
  - Speed of inventory updates in the system compared to reality.
  - Number of times users receive accurate notifications about inventory status.
- Customer Satisfaction:
  - Customer ratings after using the chatbot (via surveys or direct feedback).
  - Ratio of successful issue resolution from the first interaction.
- System Reliability:
  - Chatbot operating time (uptime) versus downtime.
  - Number of system incidents or errors within a specific time period.

Chat Coherence:

- Evaluate the extent to which chatbot responses align with user queries.
- Ratio of conversations where users do not report inconsistencies or irrelevance in responses.

Chat Fluency:

- Measure the naturalness and readability of the language used by the chatbot in conversations.
- Ratio of conversations where users feel the storytelling is smooth, without interruptions or difficulties in understanding.

To assess these two metrics, the following methods can be used:

- Automated Analysis:
  - Utilize natural language processing tools to evaluate the coherence and fluency of chatbot responses. These tools may include coherence scoring algorithms and fluency scoring algorithms.
- User Interaction Review:
  - Collect and analyze user feedback on their experience when interacting with the chatbot. This can be done through surveys, feedback after conversations, or specific quality review sessions.
- Quality Review Sessions:
  - Organize review sessions where language and AI experts assess the quality of conversations, focusing on coherence and fluency.

These methods provide a comprehensive evaluation of the chatbot's language capabilities and overall user experience.

Roadmap and Timeline

Date	Result
7/11	Data Acquisition and a Data Lake
10-12/11	Data pipeline
20/11	Intent Classification
25-30/11	Entities and Database
7/12	Script
8-10/12	Test, evaluate, make a story
10/12	MVP Deployment

Conclusion

In summary, our MVP stands out as an innovative solution that adeptly combines AI technology with practical retail applications, setting a new standard in digital customer service and product management in the retail industry.