InsightAI: Bridging Natural Language and Data Analytics

Karthika R
Dept of Computer Science and Engineering
Toc H Institute of Science and Technology
karthika.r1229@gmail.com

Maria Toms
Dept of Computer Science and Engineering
Toc H Institute of Science and Technology
mariaktom14@gmail.com

S R Aadrash
Dept of Computer Science and Engineering
Toc H Institute of Science and Technology
aadarshsr7@gmail.com

Prabath P U
Dept of Computer Science and Engineering
Toc H Institute of Science and Technology
swethaprasadnandanam@gmail.com

***Abstract*—This project introduces an innovative application that leverages generative AI, specifically pre- trained large language models, for extracting and interpreting data from large databases, transforming it into comprehensible insights. The approach involves pre-training the model to establish a foundational understanding of language and context. Subsequently, the model is fine-tuned to specialize in database querying, learning to interpret natural language questions and translating them into precise database queries. The application further utilizes in-context learning, allowing the model to adapt and refine its understanding based on the specific context of database interactions. After retrieving the relevant data, the application employs generative AI algorithms to produce coherent, natural language answers. This process converts complex database information into easily understandable insights, bridging the gap between intricate data structures and user comprehension. To showcase this technology, the project applies these techniques to a large, synthetic dataset created using OpenAI API, simulating various customer surveys across different product segments and customer categories. For example, a user could query, "What do gold customers think about our premium broadband service?" The application would then generate and execute the appropriate database query, followed by presenting a summarized insight drawn from the data. This project not only simplifies interactions with large-scale data but also opens new avenues for advanced data analysis and informed decision-making. The combination of pre-training, fine-tuning, and in-context learning harnesses the power of pre-trained language models, enabling the application to navigate and interpret complex databases with a high degree of accuracy and efficiency.**

Keywords: ***Generative AI, Fine tuning, In-context learning, Natural language, OpenAI API, Pre- trained models, Database querying***

I. INTRODUCTION

This project introduces an innovative application that leverages generative AI, specifically pre-trained large language models, for extracting and interpreting data from large databases, transforming it into comprehensible insights. The approach involves pre-training the model to establish a foundational understanding of language and context. Subsequently, the model is fine-tuned to specialize in database querying, learning to interpret natural language questions and translating them into precise database queries. The application further utilizes in-context learning, allowing the model to adapt and refine its understanding based on the specific context of database interactions. After retrieving the relevant data, the application employs generative AI algorithms to produce coherent, natural language answers. This process converts complex database information into easily understandable in- sights, bridging the gap between intricate data structures and user comprehension. To showcase this technology, the project applies these techniques to a large, synthetic dataset created using OpenAI APIs, simulating various customer surveys across different product segments and customer categories. This project not only simplifies interactions with large-scale data but also opens new avenues for advanced data analysis and informed decision-making. The combination of pre- training, fine-tuning, and in-context learning harnesses the power of pre-trained language models, enabling the application to navigate and interpret complex databases with a high degree of accuracy and efficiency.

II. BACKGROUND

This project is based on the idea of extracting and interpreting insights from large databases. The approach is based on the use of generative AI and, in particular, the pre-training of large language models. The model is pre-trained to understand the language and the context. The model is fine-tuned to specialize in database queries, to translate natural language queries into precise queries. In context learning allows the model to adapt based on the specific interactions of the database. The application uses the power of the generative AI algorithm to transform complex database data into natural language answers. The technology is demonstrated using a synthetic dataset generated with the OpenAI APIs. This dataset simulates customer surveys across different product segments. Not only does the project simplify data interactions, but it also opens up avenues for advanced analysis and informed decisions. Pre-trained language model combines with fine- tuning and in context learning to provide an accurate and efficient navigation of the database.

III. RELEVANCE

Efficient Data Processing: Uses generative AI for large database queries. Enhances data processing efficiency.

User-Friendly Interactions: Converts complex data into natural language. Facilitates user-friendly data interactions.

Advanced Data Analysis: Employs generative AI for com- plex data analysis. Allows natural language queries without technical expertise.

Informed Decision-Making: Provides understandable in- sights. Aids decision-makers without technical skills.

Adaptability through In-Context Learning: Utilizes in- context learning for adaptability. Refines understanding based on interactions.

Synthetic Dataset Demonstration: Demonstrates versatility with OpenAI APIs. Simulates scenarios like customer surveys.

IV. RELATED WORK

The Transformer architecture, introduced in a research paper by Vaswani et al., [1] revolutionized deep learning by exclusively employing attention mechanisms, eliminating the need for recurrent or convolutional layers. This innovative approach enabled the model to efficiently capture complex relationships within input sequences. The Transformer's impact transcends its immediate applications, shaping the broader landscape of neural network design and inspiring new paradigms in machine learning. Its attention mechanism has become a cornerstone in subsequent models, solidifying its status as a groundbreaking contribution to the field, with researchers and practitioners continually building upon its principles. The Transformer architecture in deep learning revolutionizes sequence modeling by introducing self-attention mechanisms, departing from recurrent and convolutional layers. Its encoder- decoder structure efficiently captures patterns and features, enhancing computational performance through parallelization during training. Particularly effective in tasks like language translation and image captioning, the Transformer's multi-head attention mechanism further improves performance by focusing on different aspects of input sequences concurrently. This innovation not only enhances pattern recognition but also streamlines parallelization, reducing training time and surpassing traditional recurrent neural networks in efficiency and effectiveness.

The research paper by Howard J et al., [2] discusses about fine tuning of Universal Language Model Fine-Tuning (ULMFiT) is a widely utilized approach in natural language processing (NLP) that involves adapting pre-trained language models, such as OpenAI's GPT series, to specific tasks through additional training on task-specific datasets. By leveraging the knowledge captured in these pre- trained models, ULMFiT significantly reduces the need for labeled data while enhancing performance in various NLP applications, including text classification and sentiment analysis. The process entails curating a labeled dataset, preprocessing the data, fine-tuning the pre-trained model's architecture, and training it on the task-specific dataset while monitoring vali- dation progress. ULMFiT's impact extends beyond technical aspects, enabling efficient knowledge transfer across domains and leading to the development of robust NLP models with reduced data requirements and broad applicability across diverse domains.

Devin J et al., introduce BERT as a new language representation model that is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. [3]. BERT uses a multi-layer bidirectional Transformer encoder, based on the original implementation by Vaswani et al. [1] The architecture is almost identical to the original Transformer, which uses self-attention mechanisms. The model is characterized by the number of layers (L), hidden size (H), and the number of self-attention heads (A). Two model sizes are primarily reported: BERTBASE (L=12, H=768, A=12, Total Parameters=110M) and BERTLARGE (L=24, H=1024, A=16, Total Parameters=340M)

The methodology introduced by Brown T et al., [4] involves the extensive training of GPT-3 across a diverse range of tasks, exposing the model to a multitude of linguistic challenges. The evaluation of the model's performance is then conducted under various conditions, each emphasizing the amount of available training data. Few-shot learning involves training the model with a small number of examples, one-shot learning involves learning from just a single example, and zero-shot learning involves learning from a task description without any accompanying examples. To introduce an element of novelty and challenge, the method includes the design of new tasks that assess GPT-3's adaptability to different linguistic contexts. These tasks may encompass activities such as generating poetry, summarizing articles, or other linguistically intricate challenges.

The framework mentioned by Raffel et al., [5] revolutionizes NLP by treating various tasks as text generation challenges, enabling a single model to tackle multiple tasks seamlessly. This paper, leads to the exploration of T5's efficiency and scalability across a spectrum of tasks, emphasizing its superiority over task-specific models.

V. HIGH LEVEL DESIGN

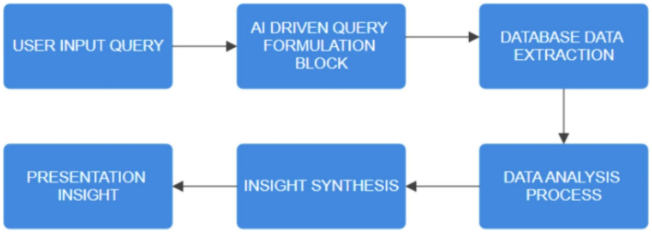

Fig. 1. Working process

Start\; The user inputs a query in natural language to initiate the process.

AI Query Formulation: The AI takes the user's natural language query and interprets it. Subsequently, it formulates a structured database query based on this interpretation.

Data Retrieval: Database Data Extraction The formulated query is executed on the database, leading to the extraction of relevant data.

Data Analysis: The extracted data undergoes a thorough analysis using various algorithms or processes. During this stage, meaningful insights and patterns are derived from the data.

Insight Generation: The results of the data analysis are synthesized into coherent, natural language

insights. This step involves translating complex data findings into understandable and informative statements.

User Output: The generated insights are presented to the user in a comprehensible format. This presentation could take the form of text, visualizations, or other suitable means to effectively convey the information.

End: The process concludes here, marking the end of the sequence.

1) MODULES PROPOSED:

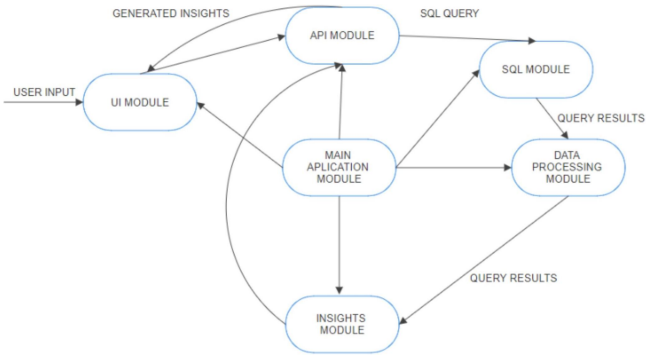

Fig. 2. Modules of the system

Module 1: User Interface:- Gradio, a Python package for creating interactive interfaces, is used to design the user interface. It enables users to interact with the system easily and enter inquiries in natural language.

Features: Takes User Input in Natural Language: Gradio allows users to enter questions in a natural language style that is conversational.

Presents Useful Data and Insights: The system uses Radio to handle user queries, extracting information from the database, analyzing it, and giving the user relevant data and insights.

Enables an Interactive and User-Friendly Experience: Gradio makes sure that the user experience is user-friendly by offering an interactive interface that allows for intuitive input and understanding of generated insights.

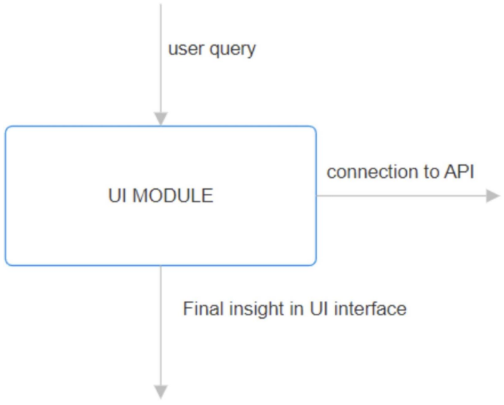

Fig. 3. UI module

Module 2: API Module:- Transmits and receives data by interacting with our application programming interface (API). Uses preset endpoints and data formats to create calls to our API based on user inputs from the UI module. Answers from the API are received and then forwarded to the SQL module for additional processing.

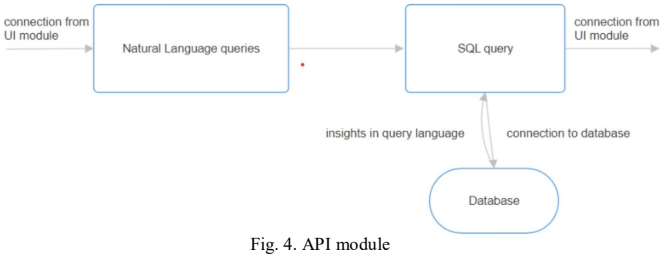

Fig. 4. API module

Module 3: SQL Module:- In charge of running SQL queries on the database. Receives input from the API module, usually consisting of conditions and query arguments. Makes a connection to the database with the right authorization and credentials. Carries out the SQL query and accesses the database to get the generated dataset.

Fig. 5. SQL module

Module 4: Data Processing Module:- Receives the dataset retrieved by the SQL module.

Module 5: Insights Module:- Utilizes the processed data to derive meaningful insights using the API.

Module 6: Main Application Module:- Serves as the system's orchestrator, arranging how the various modules communicate with one another. Controls logic and data flow in response to system events and user interaction.

VI. METHODOLOGY

The primary challenge at hand is the intricate and con- voluted nature of extracting meaningful insights from large databases. Currently, this process is marked by a high level of complexity, especially for users lacking technical expertise. The hindrance lies in the complexity of database querying, which acts as a barrier, preventing non-technical users from efficiently interacting with and deriving insights from extensive datasets. The overarching goal of the project is to simplify and streamline this intricate process by harnessing the power of generative AI. The focus is on making data interactions more accessible and user-friendly, effectively bridging the existing gap between the complexity of database structures and the comprehension level of non-technical users. One of

the key issues to address is the adaptability of the solution to changing contexts. In this context, the project emphasizes the importance of in-context learning. The ability to adapt to evolving circumstances and dynamic data requirements is crucial for the long-term effectiveness of the proposed solution. To illustrate the practicality and benefits of the project, a synthetic dataset is utilized as a demonstrative tool. This showcases the potential applications across various industries, emphasizing the impact on informed decision-making and advanced data analysis. The synthetic dataset serves as a tangible representation of how the proposed solution can be effectively employed in real-world scenarios. In essence, the project's problem analysis revolves around the complexities of database interactions, focusing on the challenges faced by non-technical users. The proposed solution aims to simplify this process through generative AI, with a strong emphasis on adaptability and practical applications in diverse industries.

VII. SYSTEM STUDY

The revolutionary application's system study centers on a thorough examination of its features, architecture, and operational procedures. The pre-training phase, in which a big language model, like GPT-3, is used to impart a basic understanding of language and context, is thoroughly examined at the outset of the study. The exploration of various textual material is part of this step, which helps the model to pick up on complex language patterns and semantic subtleties. The next step, which involves fine-tuning the model to specialize in the particular task of database querying, is the main focus. In order to create user-friendly database interactions, this entails training the model to comprehend natural language questions and convert them into accurate and efficient database queries. In the system study, the in-context learning component which promotes dynamic adaptability receives special focus. This stage entails learning how the model adjusts its understanding according to the particular database interaction scenario. Through constant learning from user inputs and dynamic database structures, the model guarantees precision and applicability when producing insights. The research explores the mechanics behind this process of adaptive learning, providing insight into the model's ability to adjust to changing
user needs and scenarios.

The system study also examines how generative AI algorithms are applied to generate answers that are coherent and written in plain language. This analysis highlights the need of bridging the gap between technical data structures and user comprehension by encompassing the complicated procedures involved in translating complex database information into clearly understood insights. The research investigates the model's capacity to provide responses that are contextually appropriate, guaranteeing that the information provided to users is correct and presented in a way that is understandable to a wide range of people.

The system research includes a detailed analysis of the application's performance on a sizable synthetic dataset created with OpenAI APIs. This entails evaluating the application's adaptability to various settings and simulating user surveys for various product categories and customer segments. The study evaluates the application's adaptability in handling synthetic data and offers information about its possible effectiveness when used with datasets from the actual world.

To sum up, the system research offers a thorough and organized analysis of the functionality and architecture of the creative application. It clarifies the nuances of the pre-training, fine-tuning, and in-context learning stages, highlighting their contributions to the development of a flexible, intuitive, and perceptive database interface and interpretation system. The basis for comprehending the application's features, restrictions, and prospective directions for future development is this thorough examination.

VIII. CONCLUSION

This innovative project represents a significant stride in the realm of data analytics by harnessing the capabilities of generative AI, particularly pre-trained large language models. Seamlessly integrating advanced language understanding with database querying, it offers a transformative approach to extracting and interpreting insights from vast datasets. Employing a threefold strategy of pre-training, fine-tuning, and in- context learning, the model establishes a robust foundational understanding of language while adapting and refining its knowledge to specific database interactions. This dynamic learning process enables the application to navigate complex data structures with high accuracy and efficiency, bridging the gap between intricate database information and user com- prehension. By converting complex data into coherent, easily understandable insights, it enhances accessibility to large- scale data and empowers intuitive interactions with databases, eliminating the need for specialized query languages. The practical application of these techniques to a synthetic dataset demonstrates the project's versatility and adaptability, show- casing its ability to translate user queries into precise database queries and present summarized insights. Beyond simplifying interactions with large-scale data, this project opens new horizons for advanced data analysis and informed decision- making, offering a potent tool for organizations seeking actionable insights from their data. As it paves the way for a more user-friendly and efficient approach to data exploration, it stands as a testament to the transformative potential of state- of-the-art generative AI technologies in the field of data analytics.

ACKNOWLEDGMENT

We wish to record our indebtedness and thankfulness to all who helped us prepare this paper titled InsightAI : Bridging Natural Language and Data Analytics and present it in a satisfactory way. We would like to express our gratitude to the Management of Toc H Institute of Science Technology, our HoD, mentors and guides for their continuous support throughout this project. Their commitment to academic excellence has provided a conducive environment for learning and exploration. We would like to thank our parents for all the support and efforts they have put for us in our educational pathway. We would also like to express our thanks to any others who contributed in any capacity, big or small, to the successful completion of this project. Once again, thank you to everyone involved for making this project a fulfilling and enriching experience. Your support has been a cornerstone of our academic journey, and we are truly grateful.

REFERENCES

[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I., 2017. Attention is all you need.31st Conference on Neural Information Processing Systems (NeurIPS 2017) held in Long Beach, California, USA 3.

[2] Howard, J. and Ruder, S., 2018. Universal language model fine-tuning for text classification. Advances in neural information processing systems, 33, pp.1877-1901.

[3] Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2018. Bert: Pre- training of deep bidirectional transformers for language understanding.
NAACL-HLT 2019: Minneapolis, MN, USA - Volume 1

[4] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. and Agarwal, S., 2020. Language models are few-shot learners. Advances in Neural Information Processing Systems 33 (NeurIPS 2020)

[5] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W. and Liu, P.J., 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research, 21(1), pp.5485-5551.

[6] Petroni, F., Rockta¨schel, T., Lewis, P., Bakhtin, A., Wu, Y., Miller, A.H. and Riedel, S., 2019. Language models as knowledge bases?. arXiv preprint arXiv:1909.01066

[7] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. and Sutskever, I., 2019. Language models are unsupervised multi task learners. OpenAI blog, 1(8), p.9.

[8] Hospedales, T., Antoniou, A., Micaelli, P. and Storkey, A., 2021. Meta- learning in neural networks: A survey. IEEE transactions on pattern analysis and machine intelligence, 44(9), pp.5149-5169.

[9] Xian, Y., Lampert, C.H., Schiele, B. and Akata, Z., 2018. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. IEEE transactions on pattern analysis and machine intelligence, 41(9), pp.2251-2265.