# Leveraging Large Language Model (LLM)[1] for Natural Language to SQL Query Generation in HR Analytics: A Case Study on IBM Attrition Dataset

Mayur Sinha
sinhamayur@yahoo.com

Sangram Kesari Ray
shankar.ray030@gmail.com

Khirawadhi
Khirawdhi@gmail.com

February 7, 2024

**Abstract**

This research paper explores the application of the GPT-3.5 Turbo Instruct model for the transformation of natural language queries into structured SQL queries within the domain of Human Resources (HR) analytics. The study focuses on the IBM Attrition dataset, utilizing the advanced capabilities of the GPT-3.5 Turbo Instruct model to enable efficient and intuitive querying of HR-related data.

Employing the model, we conducted experiments to assess its effectiveness in generating SQL queries from diverse natural language inputs, specifically tailored to the nuances of HR analytics questions pertaining to employee attrition within the IBM dataset. By leveraging prompt engineering, with only a few shots[2], our investigation revealed the model's capacity to accurately understand and interpret complex queries, providing SQL outputs that align with the dataset structure.

## 1 Introduction

In the realm of data-driven decision-making, the synergy between artificial intelligence (AI) and human resources (HR) analytics has become pivotal for organizations navigating workforce challenges. This study delves into the transformative potential of the GPT-3.5 Turbo Instruct model in automating SQL query generation from natural language inputs, specifically tailored to HR analytics inquiries. Focused on the IBM Attrition dataset, a cornerstone of HR-related information, our investigation harnesses the model's advanced capabilities to seamlessly translate natural language into structured SQL queries.

Through strategic prompt engineering[3], our experiments assessed the model's efficacy in generating precise SQL queries with just a few shots of instruction[4]. Notably, GPT-3.5 Turbo Instruct demonstrated a remarkable ability to interpret nuanced queries related to employee attrition, aligning SQL outputs with the dataset structure. This research illuminates not only the technical prowess of the model but also the practical implications for HR analytics. The automation of query generation promises heightened efficiency, empowering analysts to extract actionable insights effortlessly. By contributing to the evolving landscape of AI-driven analytics, this study offers organizations a promising avenue to leverage natural language processing in HR decision-making.

# 2 Methodology

## 2.1 Data engineering

We're using IBM attrition dataset[5] provided in CVS format. We're leveraging Snowflake[6] for further loading the dataset into a table. Later we'll use this table for storage and it's schema for downstream tasks, namely - synthetic questions generation and generation of SQL queries.

## 2.2 Synthetic questions generation

A dataset consisting of questions based on IBM attrition dataset has been created using gpt-3.5-turbo model. We start off with a set of 5 seed questions, we give the schema information to the model and prompt the model to generate diverse questions. We discard top 5 most similar questions in each iteration and generate more questions from the leftover, after recursively performing these steps, we stop at 100 diverse set of of natural language questions, with varying degree of complexity.

## 2.3 Dataset for evaluation

We split the questions dataset generated above into test and validation set. Later we'll use the test set to fine-tune the few shot prompt to elicit desired response from the model. We'll not be fine-tuning on the validation set to avoid the model getting biased.

## 2.4 Evaluation Metrics and Definitions

Execution Success Rate is used to assess the overall performance of the model.

$$\text{Execution Success Rate} = \frac{\text{Number of Successful Queries}}{\text{Total number of Queries}} \times 100$$

# 3 Results

## 3.1 Performance Metrics

| Dataset | Execution Success Rate (%) |
|---|---|
| Test Set | 100 |
| Validation Set | 100 |

Table 1: Execution Success Rate for Test and Validation Sets

## 3.2 Discussion of Findings

The results of our study, as depicted in Table 1, exhibit an exceptional Execution Success Rate of 100% for both test and validation sets. This indicates that the GPT-3.5 Turbo Instruct model, when primed with well-crafted prompts and a few-shot learning approach, can accurately transform natural language questions into corresponding SQL queries. This performance is particularly significant in the domain of HR analytics, where the ability to quickly and accurately access specific dataset insights can significantly enhance decision-making processes related to employee management and attrition.
Several key findings emerge from our analysis:

- **High Precision in Query Generation:** The model demonstrated a remarkable precision in understanding the nuances of the HR domain, interpreting the natural language queries, and generating the exact SQL syntax required to execute against the IBM Attrition dataset. This precision underscores the model's advanced natural language processing capabilities and its potential utility in HR analytics.

- **Effectiveness of Prompt Engineering:** Our experiments highlighted the critical role of prompt engineering in leveraging the GPT-3.5 Turbo Instruct model's capabilities. By carefully designing the prompts and providing a few examples of the desired output, we were able to guide the model towards producing highly accurate SQL queries, even for complex questions involving multiple dataset fields and conditions.

- **Potential for Streamlining HR Analytics Workflows:** The study's findings suggest that integrating LLMs like GPT-3.5 Turbo Instruct into HR analytics workflows could significantly streamline the process of data querying. By enabling HR professionals to generate SQL queries through natural language, organizations can reduce the time and technical expertise required to derive insights from HR data, thus making data-driven decision-making more accessible.

However, it is important to note some considerations and limitations:

- **Dependence on Quality of Input Data:** The model's performance is contingent upon the quality and structure of the input dataset. In cases where the dataset is poorly structured or contains ambiguous fields, the accuracy of the generated SQL queries may be compromised.

- **Need for Domain-Specific Tuning:** While the model showed excellent performance on the IBM Attrition dataset, its effectiveness may vary across different HR datasets or analytics questions. This suggests a potential need for domain-specific tuning or prompt adjustment to maintain high levels of accuracy across diverse HR analytics applications.

- **Ethical and Privacy Considerations:** When applying LLMs in HR analytics, organizations must navigate ethical and privacy considerations, ensuring that the use of such technologies complies with data protection regulations and respects employee privacy.

In conclusion, the findings from our study affirm the potential of GPT-3.5 Turbo Instruct to revolutionize HR analytics by providing an efficient means for converting natural language queries into SQL. This capability not only enhances analytical efficiency but also democratizes access to data-driven insights within organizations. Nonetheless, achieving optimal results requires careful consideration of dataset quality, prompt engineering, and adherence to ethical standards.

## 3.3 Significance and Implications

The findings from our investigation into the application of the GPT-3.5 Turbo Instruct model for generating SQL queries from natural language inputs within the HR analytics domain, specifically focusing on employee attrition, carry profound implications for both the field of data analytics and human resource management. The implications of this study are multi-faceted, touching on technical innovation, operational efficiency, and strategic decision-making in organizations.

- **Technical Innovation and Advancement:** The successful application of GPT-3.5 Turbo Instruct for SQL query generation represents a significant leap in the use of large language models (LLMs) for data querying. This underscores the potential of LLMs to bridge the gap between complex data systems and end-users, facilitating a more intuitive interaction with data.

- **Operational Efficiency in HR Analytics:** By enabling HR professionals to generate SQL queries through natural language, the technology can significantly reduce the barrier to accessing and analyzing data. This efficiency gain not only speeds up the data analysis process but also allows HR personnel to focus more on strategic decision-making rather than spending excessive time on technical query formulation.

- **Enhanced Decision-Making Capabilities:** With more accessible data analytics, organizations can make quicker and more informed decisions regarding workforce management, employee retention strategies, and other critical HR functions. The ability to rapidly analyze attrition-related data means that HR departments can proactively address issues, improve employee satisfaction, and reduce turnover rates.

- **Democratization of Data:** This study illustrates the potential for democratizing access to complex data analytics, making it possible for individuals without technical SQL knowledge to effectively engage with and extract insights from databases. This broader accessibility can foster a culture of data-driven decision-making across all levels of an organization.

- **Future Research and Development Directions:** The promising results of this study open avenues for further research into the optimization of LLMs for specific domains, ethical AI use, and the development of more advanced models capable of handling even more complex queries and datasets.

In conclusion, the application of GPT-3.5 Turbo Instruct in HR analytics for SQL query generation from natural language queries holds significant promise for enhancing operational efficiencies, improving decision-making processes, and democratizing data analytics. This study not only showcases the technical capabilities of the model but also highlights the practical benefits and strategic advantages that can be gained by integrating such technologies into organizational workflows.

# 4 Data Availability

The code can be downloaded from here: https://github.com/b1nch3f/auto-sql

# References

[1] https://arxiv.org/pdf/2302.13971.pdf.

[2] https://www.promptingguide.ai/techniques/fewshot.

[3] https://python.langchain.com/.

[4] https://www.pinecone.io/learn/series/langchain/langchain-prompt-templates/.

[5] https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset.

[6] https://docs.snowflake.com/en.