NOVA IMS Information Management School

MGI

Mestrado em Gestão de Informação
Master Program in Information Management

A Structured Framework for AutoML: Integrating LLMs through Comparative Experiments

Yousef Adel Hassan

Dissertation report presented as a partial requirement for obtaining a master's degree in information management.

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Statistical e Gestão de Informação
Universidade Nova de Lisboa

A STRUCTURED FRAMEWORK FOR AUTOML: INTEGRATING LLMS THROUGH COMPARATIVE EXPERIMENTS

by

Yousef Adel Hassan

Dissertation report presented as a partial requirement for obtaining a master's degree in information management specializing in Knowledge Management and Business Intelligence.

Advisor: Professor Vitor Duarte dos Santos, PhD

July 2024

ABSTRACT

This work explores the potential constructive interaction between Generative AI, specifically ChatGPT-4, and Automated Machine Learning (AutoML) frameworks. The study focuses on leveraging ChatGPT-4's capabilities within the CRISP-DM (Cross-Industry Standard Process for Data Mining) model phases to improve the efficiency and effectiveness of data-driven tasks. Through a series of experiments involving classification, regression and clustering, the research compares the performance of ChatGPT-4 in two settings: a global user perspective with general prompts and a structured approach aligned with the CRISP-DM methodology, providing a comparative benchmark.

The findings demonstrate that aligning ChatGPT-4's tasks with the CRISP-DM phases yields better performance and more comprehensive insights than the general prompt approach. The study highlights the importance of prompt engineering in optimizing ChatGPT-4's contributions to AutoML tasks, emphasizing its role in improving data preparation, model selection and the evaluation processes. Additionally, the research underscores the ethical considerations and potential challenges associated with integrating generative AI in AutoML, particularly concerning data quality, bias, and model interpretability.

KEYWORDS

Generative AI; AutoML; Prompt engineering; ChatGPT; Machine learning techniques; CRISP-DM

INDEX

1. Introduction 1
   1.1. background and problem identification 1
   1.2. Objectives 2
   1.3. Importance and relevance 2
2. Literature review 3
   2.1. Auto Machine Learning 3
      2.1.1. Concepts 3
      2.1.2. Benefits and Challenges 5
   2.2. Generative AI 5
      2.2.1. Foundational Concepts of Generative AI 5
      2.2.2. Significance of LLMs in AutoML 6
      2.2.3. Challenges and Limitations 6
   2.3. Prompt Engineering 7
      2.3.1. Prompt best practices 7
      2.3.2. Challenges and Limitations 8
   2.4. Significance of CRISP-DM in the Context of AutoML 8
3. Methodology 11
   3.1. Research Phases 11
   3.2. Summary 12
.4 Experimental Design 13
   4.1. Datasets 13
   4.2. CRISP-DM-based experiment design 13
   4.3. Global user experiment 14
   4.4. ChatGPT's AutoML features 15
   4.5. Evaluation matrix 15
5. Testing of the Use of prompt engineering in DM projects 17
   5.1. First DM experiment: Cirrhosis Patient Survival Prediction (supervised) 17
      5.1.1. Classification Experiments Assessment 17
      5.1.2. Summary 19
   5.2. Second DM experiment: Cars price prediction (supervised) 19
      .5.2.1 Regression Experiments Assessment 19
      5.2.2. Summary 21
   5.3. Third DM experiment: Mall customer classification (unsupervised) 21

5.3.1. Clustering Experiments Assessment 21
5.3.2. Summary 22
5.4. Prompt Analysis and Impact 23
5.4.1. Insights from Classification, Regression, and Clustering Experiments 23
5.4.2. Impact of Prompt Design on AutoML Performance 24
6. Results and Discussion 25
6.1. Essential differences between traditional AutoML and LLM-driven. 25
6.2. Future work 25
7. Conclusions 27
Bibliography 28

LIST OF FIGURES

Figure 1: The main feedback loops in AutoML- source (Vazquez, 2022) 3
Figure 2: CRISP-DM phases – (Martínez-Plumed et al., 2019) 9
Figure 3: Phases of the methodologies 11
Figure 4: CRISP-DM-based experiment design. 14
Figure 5: Global experiment 14
Figure 6: Feature distribution viz created by ChatGPT-4 18
Figure 7: Heatmap generated by ChatGPT-4 18
Figure 8: Cars' price distribution 20

LIST OF TABLES

Table 1: Examples of the prompt for each approach 12
Table 2: Evaluation matrix 16
Table 3: The summary of the first experiment 19
Table 4: The summary of the second experiment 21
Table 5: The summary of the third experiment 22

LIST OF ABBREVIATIONS AND ACRONYMS

API (Application Programming Interface) is a set of protocols and tools for building and interacting with software applications. APIs allow different software systems to communicate with each other.

AutoML (Automated Machine Learning) refers to a technology that automates the end-to-end process of applying machine learning to real-world problems. It simplifies tasks such as model selection, hyperparameter tuning, and data preprocessing, thereby making ML accessible to non-experts.

ChatGPT Is a generative AI model developed by OpenAI, based on the GPT (Generative Pre-trained Transformer) architecture. It generates human-like text based on prompts, facilitating tasks like conversation, content creation, and code generation.

CRISP-DM (Cross-Industry Standard Process for Data Mining) is a widely used data mining process model that outlines an iterative cycle consisting of six phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. This model ensures comprehensive and structured data mining projects across different industries.

EDA (Exploratory Data Analysis) is an approach to analyzing data sets to summarize their main characteristics, often using visual methods. EDA helps in understanding the data, detecting anomalies, and formulating hypotheses for further analysis.

GANs (Bidirectional Encoder Representations from Transformers) is a pre-trained transformer model designed to understand the context of a word in search queries. It is useful for natural language understanding tasks such as question answering and sentiment analysis.

Kaggle Is an online community and platform for data scientists and machine learning practitioners. Kaggle hosts datasets, competitions, and collaborative projects to solve data science problems.

LLMs (Large Language Models) are AI models trained on vast amounts of text data to understand and generate human language. They perform a wide range of natural language processing tasks, including translation, summarization and question answering.

1. INTRODUCTION

Automated Machine Learning (AutoML) is one of the advances in data analysis. AutoML is a comprehensive process that aims to automate the model development pipeline without requiring intensive human work and assistance. The widespread use of machine learning in several fields has highlighted the drawbacks of traditional approaches that are usually reliant on domain experts and time-consuming data operations. In light of this, AutoML has been invented to facilitate the machine learning process and eliminate the need for deep knowledge on a subject matter, thereby making complex steps of the process more feasible for non-technical users (Chauhan et al., 2020). On the other hand, large language models (LLMs) such as ChatGPT are the latest computational trend that seems capable of contesting AutoML methodologies in a more flexible manner. This study intends to investigate the constructive interaction between generative AI represented in LLMs (ChatGPT-4 in particular) and automated machine learning. to comprehend, analyze, and test ChatGPT-4 in pre-designed AutoML experiments on different algorithms with different datasets. This investigation makes it possible to identify improvements in model generation and data analysis techniques, resulting in a discussion on developing a more intuitive and effective machine learning environment along with challenges and potential ethical issues involved.

1.1. BACKGROUND AND PROBLEM IDENTIFICATION

AutoML has been identified as a game-changing technology for organizations seeking to leverage the advantages of advanced data analytics. Data mining projects of the traditional kind can be labor-intensive and involve various technical workflows, from data preparation to model building and evaluation. This challenge is addressed by AutoML, which simplifies the process by providing numerous algorithms that enable organizations to get value out of their data efficiently without being obstructed by the inherent complexity mentioned above. AutoML's value proposition lies in making machine learning accessible; by reducing the need for deep technical expertise, it emphasizes data outcomes and the interpretability of results (Chauhan et al., 2020).

The rise of generative AI, particularly Large Language Models such as ChatGPT, into AutoML procedure is an exciting development. Models of this nature have proved helpful in various fields within the machine learning process, from code generation to repetitive tasks such as model testing and hyperparameter optimization (Chauhan et al., 2020).

Although there is a clear synergy between generative AI and AutoML, the development of the studies encompassing this integration are rare and do not align with industry standards, including the popular CRISP-DM model. Critical stages of the machine learning pipeline introduced by the CRISP-DM model could leverage the powers of generative AI. However, the scope and consequences of this integration are yet to be researched.

This study aims to investigate the role and capabilities of ChatGPT in AutoML. The objective is to understand and evaluate the capabilities of ChatGPT in various machine-learning tasks. It will also explore the broader implications of integrating generative AI into AutoML and answer important questions such as: What is the best way to use ChatGPT's capabilities in AutoML? What nuances are there in prompt engineering to maximize ChatGPT's usefulness?

1.2. OBJECTIVES

This study aims to develop and present a systematic approach for utilizing ChatGPT in machine learning and data mining projects that align with the industry standard and compare it with the global-user point of view using ChatGPT-4 for the same purpose without being aligned with any industry standard. The following key objectives have been outlined to achieve the following goals:

- **Foundational Understanding:** Explore the basics of AutoML, its applications, and its importance. Investigate the potential of generative AI – research the potential of generative AI, emphasizing ChatGPT-4, within the scope of Automated Machine Learning (AutoML).
- **Prompt Engineering Exploration:** Explore the effect of prompt engineering and its role in the AutoML setting. Evaluate its consequences and the ability to enhance ChatGPT's operation under the AutoML framework.
- **Framework Development:** Develop and implement a well-defined framework for utilizing ChatGPT-4 in machine and data mining projects following CRISP-DM model phases. This would give recommendations and good practices for integrating ChatGPT into the AutoML process.
- **Tests Evaluation:** A new evaluation matrix tool will be designed to assess the performance of ChatGPT-4 in machine learning and data mining tasks. To ensure the accuracy of this evaluation, a series of experiments will be conducted to assess ChatGPT's performance. These experiments will cover a variety of data types and analytical demands. When combined, the evaluation matrix and experiments will provide a comprehensive understanding of ChatGPT's potential and highlight areas for future development.
- **Analysis and Interpretation:** Interpret the findings shedding light on potential benefits, limitations, and subtleties of the ChatGPT-4 and generative AI integration within the AutoML domain, with CRISP-DM model serving as a reference point.

1.3. IMPORTANCE AND RELEVANCE

The purpose of this research is to make machine learning and data mining techniques more accessible to a wider audience, thereby contributing to the democratization of advanced data analysis approaches. The study aims to demonstrate how generative AI, such as ChatGPT-4, can simplify the complex stages of data mining. Stakeholders can use the results of this research to apply more advanced analytical methods. Additionally, it may encourage those who have previously found machine learning challenging or inaccessible to participate, potentially leading to better decision-making, cost-efficiencies, and new opportunities in various sectors. The study also aims to establish realistic expectations and build users' trust and understanding of AI's role in the machine learning domain.

Furthermore, the research aims to investigate the link between generative AI and AutoML, providing information to improve the understanding of this topic. The goal is to make machine learning more intuitive and approachable, explaining the complexities of its nature to a broader audience.

2. LITERATURE REVIEW

This section will explore the related components of AutoML and generative AI. I will begin by establishing a foundational understanding of AutoML and its features, comparing these features and the added value, in comparison to using LLMs in AutoML. Subsequently, gaining a high-level understanding of generative AI is integral for this study, as it will lay the groundwork for understanding prompt engineering. A coherent understanding of these topics will aid in developing experiments used to assess the ability of generative AI in performing machine learning and data mining tasks.

2.1. AUTO MACHINE LEARNING

AutoML is a recent advancement that simplifies the design and implementation of machine learning models. It automates complex tasks such as model selection, hyperparameter tuning, and algorithm optimization, which traditionally required significant expertise and time (see figure 1). This automation enables individuals from various fields to leverage machine learning without needing in-depth technical knowledge. AutoML promotes a practical approach, making machine learning accessible to a wider audience beyond technical experts and researchers. This development represents a significant step in the evolution of machine learning, broadening the reach of advanced analytical techniques (Chen et al., 2021).

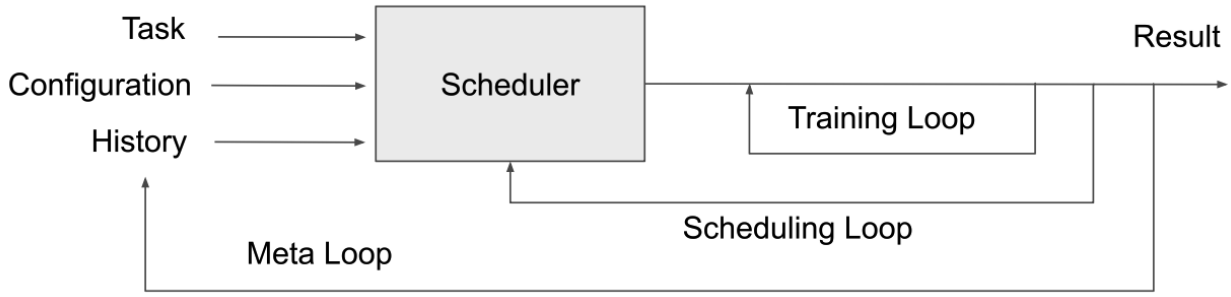

Figure 1: The main feedback loops in AutoML- source (Vazquez, 2022)

2.1.1. Concepts

AutoML is a breakthrough in machine learning, covering the entire process from data preprocessing to feature engineering, model selection, and hyperparameter tuning. This all-encompassing automation addresses diverse datasets, problems and significantly diminishes the need for domain-specific expertise. The growth and use of AutoML are driven by the need to democratize machine learning and address the gap in expertise that has been the bottleneck in the analytics field (Chen et al., 2021).

Automated Data Preprocessing (AutoDP) is a crucial component of the AutoML framework, revolutionizing the initial stages of the machine learning process. AutoDP recognizes that data quality is a critical factor in model performance. It automates several data preparation tasks that traditionally represent a large part of the analysis process often ranging between 50-80% of the entire workflow. AutoDP simplifies these complex and time-consuming tasks by introducing a data-driven solution called Auto-Prep. Auto-Prep is Python-based and automates the detection of data problems by applying relevant transformation techniques that enhance data quality and improve model performance. Auto-Prep's functionalities include auto-detecting data types, imputing missing values,

encoding categorical attributes, feature scaling, and dimensionality reduction- all customized to the dataset's specifics. This approach has been proven effective by applying it to various datasets, demonstrating that it not only simplifies the preprocessing step, however, also produces more accurate and robust machine learning models compared to manually preprocessed data-trained models (Bilal et al., 2022).

Automated Feature Engineering (AutoFE) is an essential component of AutoML that aims to simplify the identification and selection of the most relevant features for a given machine-learning task. This step typically requires substantial domain expertise, but AutoFE can automate this critical phase, significantly enhancing the machine learning model's performance.

AutoFE is comprised of three fundamental processes: feature mining, generation, and selection. Feature mining is the initial step where potential features are identified from raw data. It carries additional importance since it outlines the foundation for subsequent feature generation and selection.

Feature generation involves creating new features by transforming existing ones, often using mathematical or statistical operations. This step is vital for uncovering hidden patterns and relationships within the data that might not be apparent in their raw form.

Finally, feature selection chooses the most relevant features from the generated pool. This step is crucial to ensure the model's efficiency and prevent overfitting, which can occur when models are trained on too many or irrelevant features (Chen et al., 2021).

AutoML has a component called Automated Model and Hyperparameter Learning (AutoMHL), a significant advancement in machine learning practices. This component automates vital tasks such as algorithm selection and hyperparameter tuning. With AutoMHL, it is possible to identify the most suitable algorithm and its optimal hyperparameters for each specific data set. This efficiency involves reducing manual processes and selecting the hyperparameters that produce more accurate results. AutoMHL facilitates a data-driven, systematic approach, enabling a comprehensive exploration of the algorithmic and parameter space. By adopting this approach, the likelihood of achieving higher model accuracy and robustness increases significantly, highlighting the transformative impact of AutoMHL in the field of machine learning (Chen et al., 2021).

Automated Model Evaluation (AutoME) has become a significant part of the AutoML ecosystem. The primary focus of AutoME is on evaluating machine learning models, including model selection, hyperparameter optimization, and prediction result analysis. Most AutoML tools incorporate a three-stage pipeline, with AutoME integral to the process. The process involves training various models with different parameters and selecting the most effective model or a combination of models. A range of machine learning algorithms supports this process, often including techniques such as neural architecture search for optimizing neural network models.

Moreover, AutoME extends to model interpretation and prediction analysis, a feature found in commercialized tools. This component provides in-depth result representation through various visualization methods and analysis tools, enhancing the understanding and interpretation of the model's performance and predictions. Overall, AutoML tools, through AutoME, demonstrate

significant performance across diverse datasets, although no single tool currently outperforms all others in every aspect (Truong et al., 2019).

2.1.2. Benefits and Challenges

AutoML has had a significant impact on making machine learning accessible. It enables individuals and organizations without expertise in machine learning to use data-driven insights to make transformative decisions. AutoML also improves the efficiency of the machine learning process by automating complex tasks like data preprocessing and hyperparameter tuning. This leads to faster development of models, and, in some cases, even better performance than manually engineered models. This increased efficiency is especially important in today's digital-first business culture, where timely and accurate data analysis is essential for decision-making.

AutoML is a powerful technological innovation. However, it has its limitations that must be considered. The suitability of data for the intended outcomes is a primary concern. High-quality and relevant data are critical for the success of AutoML systems. Less-than-ideal data can significantly negatively impact model performance (Tuggener et al., 2019). The computational costs of AutoML can also be prohibitive. The search space, which includes a lot of algorithms, hyperparameters, and feature engineering techniques, is vast and presents enormous computational demands, resulting in extended training time.

Furthermore, explaining the models that AutoML generates can be challenging. While these models may have been engineered to perform, they have little inherent interpretability, which can be intimidating for non-experts. This lack of transparency can make it harder for non-technical users to welcome the approach, particularly in sectors where understanding model decision-making is critical (Tuggener et al., 2019).

2.2. GENERATIVE AI

Generative AI has emerged as an exciting development within artificial intelligence, providing a new way for humans to interact with AI through natural language. At its core, generative AI involves using a dataset to create new data views that resemble the original dataset. This field has undergone significant evolution from the earliest neural network architectures to the advanced architectures of today. The history and evolution of the foundational concepts of generative AI help to illustrate the current transformation that this field is undergoing.

2.2.1. Foundational Concepts of Generative AI

The transformation began with Generative Adversarial Networks (Goodfellow et al., 2014), which pitted a generator and a discriminator against each other to produce the most realistic samples. It continued with Variational Autoencoders which provided a way to define a probability density function over the data distribution we would like to generate. However, the transformation truly began when transformers were introduced. These are an attention-based method that allows for the capturing of long-range dependencies and seems particularly appropriate for generating language with complex, fine-grained contextual interactions that are required to produce human language that is coherent and contextually relevant (Vaswani et al., 2017).

The evolution led to Large Language Models (LLMs) like ChatGPT, which are so large and fed with so much information, from their vast pre-training dataset, that they can perform a considerable number of tasks without having to be specifically trained for the task at hand. Eventually, Google's BERT used these deep bidirectional representations to set new records on various NLP tasks (Devlin et al., 2018). These foundational concepts have propelled AI to destinations in content generation that we would not have thought possible only several years ago. They have also raised interesting questions about the ethical implications surrounding content that AI generates. As we continue to harness these technologies for even more beneficial uses, the foundation of Generative AI will continue to guide us to these benefit areas rather than a roadblock that threatens to impede our progress.

2.2.2. Significance of LLMs in AutoML

AutoML has simplified the complex processes of machine learning, which includes data preprocessing to model deployment. The integration of advanced language models, such as ChatGPT, into AutoML frameworks indicates the evolution in this field. This integration can transform traditional machine-learning processes, making them more efficient and effective (Chen et al., 2023). ChatGPT's ability to generate code and comprehend context can improve AutoML systems' efficiency by generating relevant code snippets or providing insights from extensive training data. For example, ChatGPT can contribute to data exploration, feature engineering, and hyperparameter tuning. Bringing these capabilities into the realm of AutoML has an immense potential to overcome many issues introduced by traditional methods.

2.2.3. Challenges and Limitations

The advancement of Generative AI signifies the beginning of a new era of technological capability. However, progress comes with challenges and limitations that require a critical examination. The ethical and sociological impact of generative AI is profound and diverse. Its ability to produce content similar to that created by humans makes it susceptible to misinformation, deceptive media, and propaganda misuse. This necessitates the establishment of robust ethical principles and regulatory control that would ensure responsible use (Bender et al., 2021).

One of the most severe issues is bias and justice in AI systems. Since these models are trained on existing data, they risk perpetuating the same prejudices and social inequities embedded within their training data (Buolamwini & Gebru, 2018). The AI community is actively addressing this problem to reduce these biases so that AI will be fairer and more unbiased moving forward.

The use of Generative AI models and the process of training them have raised concerns about their environmental impact. These processes consume significant energy, a major contributor to carbon emissions. Therefore, the need for sustainable AI development has been emphasized, and it involves using more efficient computing technologies (Truong et al., 2019).

Generative AI has caused discussions regarding intellectual property and creative rights. The issue of technology, law, and ethics is the ownership of AI-generated content and the training data being left open. The work of Zhong et al., 2023 addresses these issues and highlights the need for attributing and watermarking AI-generated content to provide fair and open IP protections within Generative AI.

A generative AI future is not a way to the next technological leap but an AI progress that complies with ethical standards and societal values. An interdisciplinary approach to Generative AI and thoughtful

policy on how it is used to distribute benefits widely and responsibly is the only way to unlock the true power of Generative AI.

2.3. PROMPT ENGINEERING

Prompt engineering is a crucial aspect of Natural Language Processing (NLP). It involves creating and utilizing language prompts to customize and direct the outputs of artificial intelligence systems. Over the years, the field has advanced significantly from its original focus on simple input queries with regards to information retrieval systems (Muktadir, 2023). Prompt engineering began in the 1950s, and basic models such as Hidden Markov Models and Gaussian Mixture Models laid the foundation for sequence data generation in AI. The field progressed further with the development of Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU), which improved language processing capabilities (Cao et al., 2023).

In 2014, Generative Adversarial Networks (GANs) emerged as a critical technology for image generation. This, along with other techniques such as Variational Autoencoders (VAEs) and diffusion generative models, brought about a significant transformation in the industry. The introduction of transformer architecture in 2017 significant change for NLP and computer vision. This technology enabled the development of models like BERT and GPT, which marked a significant leap forward in text generation and classification tasks. As a result, multimodal models like CLIP, which combine text and image data, were created (Cao et al., 2023).

Recently, significant developments have been made in encoder-decoder language models that utilize context information and autoregressive properties. This has resulted in better performance in many tasks. Additionally, advancements in vision-language encoders and decoders have contributed to learning representations from multiple modalities, which can be reused for various tasks (Cao et al., 2023).

2.3.1. Prompt best practices

It is important to note that the diversity and complexity of human language make it difficult to create a one-size-fits-all approach to AI interaction. OpenAI emphasizes that its model is designed to comprehend various linguistic expressions. This suggests that there is no perfect way to prompt ChatGPT will work universally for everyone (OpenAI, 2024).

To effectively communicate with ChatGPT, you need to use dynamic and adaptable language. The model's flexibility allows it to understand and respond to various communications, from simple queries to complex instructions. However, this flexibility also highlights the importance of developing clear and precise prompts. A prompt is the key that enables AI to understand the task or question, resulting in more accurate and relevant responses (OpenAI, 2024).

In addition, ChatGPT chatting is iterative. The initial prompts fail to consistently produce the desired result, which calls for follow-up queries or adjustments. The iterative communication style with the AI model mirrors regular human conversational practice, often resolved through a series of exchanges (OpenAI, 2024).

Contextualizing within prompts is another critical feature of ChatGPT effective communication. Contextual information can be general or specific to the requirements of the response in terms of

format or content. Context addition enables the AI model to ground its responses in the proper structure, enhancing its relevance and accuracy (OpenAI, 2024).

Continuing prompt testing and adjusting is an essential part of this interaction. Successful prompting strategies may also need some tweaking as AI models evolve and learn. Testing the effects of various prompt structures and styles on AI performance is a critical exercise to conduct frequently. This perpetual adjustment follows the flexible nature of AI and human-language interaction (OpenAI, 2024).

Further, ethical issues help to create prompts. One should be aware of bias, disinformation, or harmful content creation. The task is to generate powerful and ethically correct prompts, adhering to the principles of justice and truth (OpenAI, 2024).

2.3.2. Challenges and Limitations

Prompt engineering, integral to harnessing the capabilities of LLMs, presents its challenges. One major challenge in prompt engineering is the issue of "hallucinations," where LLMs generate unreal or inaccurate information. This occurs when the model lacks sufficient training data or broadly generalized patterns (Chen et al., 2023).

Determining the best prompt for a specific task can be intricate. The effectiveness of a prompt can vary based on the task, data, and desired outcome. Additionally, the effectiveness of a prompt can vary between different LLMs or versions of the same model. If not crafted carefully, prompts can introduce biases or lead the model to produce outputs that might be ethically questionable. Recognizing these challenges is crucial for the effective and responsible use of prompt engineering in AutoML and the CRISP-DM model. Furthermore, staying informed about the latest developments in prompt engineering is essential for navigating these complexities (Chen et al., 2023).

2.4. Significance of CRISP-DM in the Context of AutoML

The CRISP-DM model, introduced over two decades ago, is a widely recognized process designed for data mining projects. (Figure 2) comprises six iterative phases that are industry independent.

1. **Business Understanding:** Assessing the business situation and defining the data mining goal.
2. **Data Understanding:** Collecting and exploring data and checking its quality.
3. **Data Preparation:** Selecting, cleaning, and feature engineer new features if needed data when needed to be ready for modeling step.
4. **Modeling:** Selecting the appropriate modeling technique and building the model.
5. **Evaluation:** Checking results against business objectives and interpreting outcomes. **Deployment:** Implementing the model could result in a report or software component.

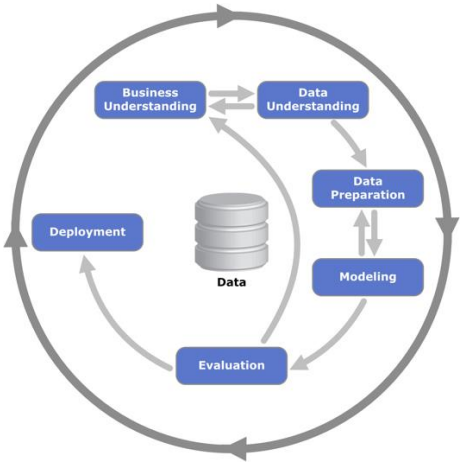

Figure 2: CRISP-DM phases – (Martínez-Plumed et al., 2019)

Integrating prompt engineering into the CRISP-DM model within the AutoML framework can result in a smoother, more effective, and more comprehensive data mining process. Utilizing the potential of LLMs through efficient prompts and following the best practices mentioned by OpenAI can enhance the data mining process. CRISP-DM aligns perfectly with these practices by breaking down the problem into logical steps, enhancing the problem-solving capabilities of the model. To understand the difference between following the CRISP-DM-aligned approach and the normal usage of the model, two main approaches can be followed:

- ChatGPT can be evaluated through more open-ended and task-oriented prompts, focusing on the end goals of typical machine learning tasks. These prompts intentionally avoid specifying the stages of the task or the methods for its execution. This method evaluates ChatGPT's practical applicability and critical thinking skills in ambiguous, real-world scenarios without specific phase-related instructions.
- Another way is to evaluate ChatGPT by aligning it with the CRISP-DM model. This can be achieved by explicitly instructing ChatGPT to follow each model phase. By doing this, ChatGPT can focus on each phase separately, extracting more context that could be relevant for data preprocessing and modeling. The goal is to determine if breaking down the process into phases could benefit ChatGPT in achieving better results than the previous open-ended approach.

In the following points, I will mention potential benefits of each stage for the data mining process and how it could provide more relevant context to ChatGPT. This will potentially improve overall performance. The emphasis is on using the nuanced capabilities of LLMs at distinct stages, from Business Understanding to Deployment, to enhance data-driven decision-making and operational efficiency.

Business Understanding: Incorporating Large Language Models (LLMs) into business understanding can give the model a wider perspective on the nature of the data; providing such context for human is usually helpful in anticipating what expected in the following steps and what the nature of the problem is being addressed.

Data Understanding: During the data understanding phase, it is vital to use LLMs to gain critical insights into data quality, identify missing values, and detect potential outliers. This is essential for ensuring the reliability of subsequent analyses. LLMs can also generate descriptive statistics or visualizations, providing a deeper understanding of data distributions and underlying patterns for the user.

Data Preparation: The data preparation stage can also greatly benefit from large language models (LLMs). ChatGPT can help with data cleaning, transformation, and feature extraction, streamlining this crucial phase of the analytics process. It can offer suggestions on handling missing data, dealing with outliers, or encoding features, which are essential for preparing the dataset for modeling.

Modeling: The modeling phase will benefit from the previous steps if ChatGPT performs better when the task is broken down into CRISP-DM. In this phase, a better understanding of the business and data context could help select a more suitable model, and clean data helps to increase trust in the results.

Evaluation: During the evaluation phase, LLMs can explain the decisions made by models in a way that is easy for humans to understand. This improves the interpretability of the models created by automated machine learning (AutoML) tools. LLMs can also assist in establishing suitable evaluation metrics or benchmarks to ensure that the developed models align with the business objectives.

Deployment: Finally, for deployment, advanced prompt engineering can enable LLMs to guide through the deployment process. This includes offering insights into model monitoring and maintenance and suggesting suitable deployment architectures.

3. METHODOLOGY

The upcoming section provides a detailed outline of the systematic approach used in the research. A mixed-methods research design, incorporating both qualitative and quantitative methods, is employed to fulfill the research objectives. The methodology offers a comprehensive understanding of the research topic, from the initial exploration to the final analysis. As illustrated in Figure 3, each phase of the research comprises specific steps to steer the research toward a thorough comprehension of ChatGPT's potential within the AutoML domain in the shade of CRISP-DM. The subsequent sections will delve into a detailed breakdown of each phase and its components.

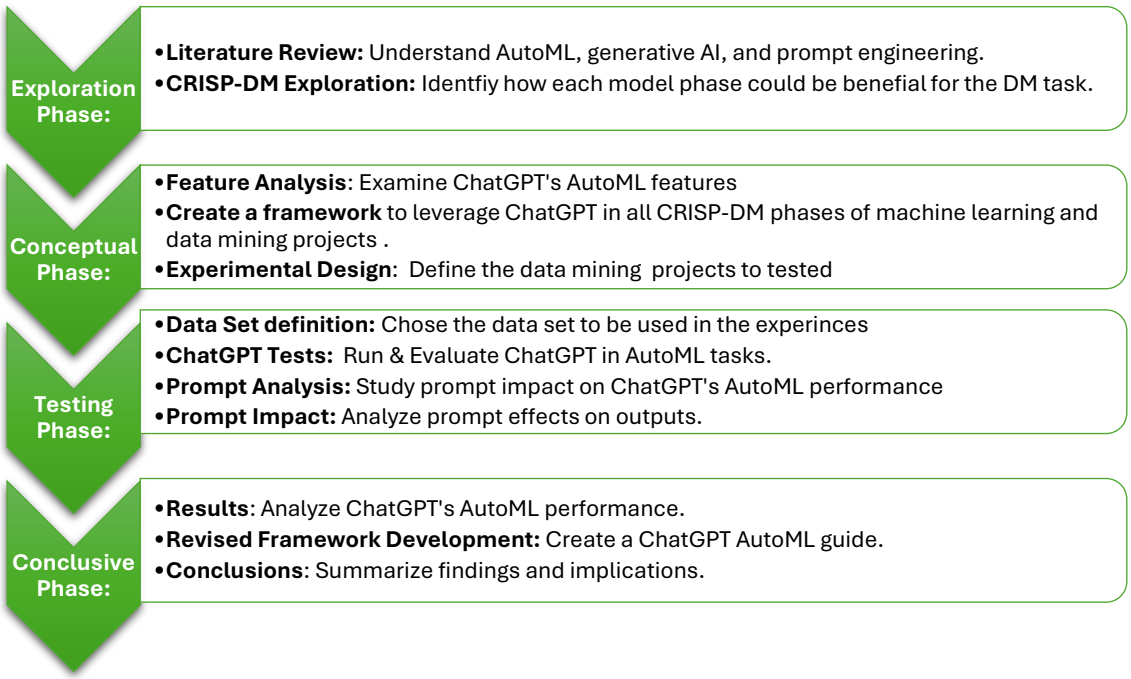

Figure 3: Phases of the methodologies

3.1. RESEARCH PHASES

Exploration Phase: The first phase begins with a thorough literature review to comprehend the use of generative AI, specifically ChatGPT, in the AutoML domain. This involves building a solid understanding of prompt engineering. Additionally, this phase aims to explore how the CRISP-DM model can improve data mining processes by examining the role of ChatGPT in each phase of CRISP-DM.

Conceptual Phase: During the conceptual phase, the study will primarily investigate ChatGPT's relevant features to AutoML. The experiment design will incorporate two perspectives: the 'global user' perspective, involving a comprehensive prompt encompassing the entire task, and an approach aligned with the CRISP-DM methodology. The objective is to compare the outcomes when tasks are segmented into CRISP-DM's logical steps versus when addressed through a holistic prompt. Additionally, this phase involves conceptualizing an evaluation matrix, which will serve as a critical tool for assessing the experiments' performance.

Testing Phase: During the testing phase, I will conduct experiments using datasets from Kaggle to evaluate the outcomes based on both the global perspective prompts and the CRISP-DM-aligned

approach. This phase will put the conceptual framework into practice by selecting appropriate datasets for the experiments. ChatGPT-4 will be used in these tests, employing the designed prompts to execute the experiments, and then analyzing the resulting outputs.

Conclusive Phase: In the final phase, the findings from the testing phase are critically analyzed, with a specific focus on rigorous statistical examination. This analysis aims to gain insights into ChatGPT's effectiveness in AutoML tasks. A comprehensive framework is formulated using these insights, providing strategic guidelines for incorporating ChatGPT into AutoML initiatives. The research concludes by reflecting on ChatGPT's potential within the AutoML landscape and proposes directions for future inquiries.

3.2. SUMMARY

After completing a literature review covering the topics of generative AI, AutoML, and prompt engineering, I will create a conceptual design for experiments using a dual approach (table 1). The approach will involve utilizing general machine learning task prompts to assess the model's problem-solving abilities from a global user perspective and directing ChatGPT to use CRISP-DM for the task to conduct a detailed analysis of ChatGPT's analytical capabilities. I will ensure that best practices in prompt engineering are followed for each approach to optimize the accuracy and relevance of the assessment.

Table 1: Examples of the prompt for each approach

Approach	Algorithm type	Prompt Example
Direct Prompts for Each CRISP-DM Phase	**Supervised algorithm (labeled data)**	***Prompt**: "We will build a machine learning model utilizing the dataset provided to construct a predictive model. The objective of this model is to accurately forecast the likelihood of a heart attack in individuals, as indicated in the column "Output," utilizing key medical parameters. The methodology followed will be CRISP-DM (Cross-industry standard process for data mining), encompassing the following steps: Business Understanding, Data Understanding, Data Preparation, Modeling, and Evaluation."*
	Unsupervised algorithm (Unlabeled data)	***Prompt**: "This dataset contains records about mall customers. My objective is to build a model that clusters these customers using the features available in the dataset. Follow the CRISP-DM model (Cross-industry standard process for data mining) to perform the task, which is broken down into the following steps: business understanding, data understanding, data preparation, modeling, and evaluation."*
General Machine Learning Task Prompts	**Supervised algorithm (labeled data)**	***Prompt:** "Utilize the provided dataset to construct a predictive model. The objective of this model is to accurately forecast the likelihood of a heart attack in individuals in column "Output," utilizing the available key medical parameters."*
	Unsupervised algorithm (Unlabeled data)	***Prompt:** "With the dataset that contains customer profiles including age, income, and purchase categories, develop a classification model to categorize customers into different loyalty tiers based on their purchasing behavior."*

4. EXPERIMENTAL DESIGN

The following section will cover the conceptual design for the experimentation I will be conducting to test the ChatGPT-4 performance following two main approaches. The first approach involves using general, open-ended prompts to evaluate ChatGPT's ability to apply instinctive knowledge in less defined situations related to the broader objectives of machine learning. The second approach directs ChatGPT to use the CRISP-DM model, following the main steps presented by the model. This methodology aims to holistically evaluate ChatGPT, from its problem-solving abilities to its accuracy in executing tasks specific to each CRISP-DM phase. A forthcoming (figures 4 and 5) will provide details on the steps involved in each approach. All of prompt used and the code generated by ChatGPT-4 will be stored in the following GitHub repository for documentation purposes.

- https://github.com/Yousef-Adel/ThesisNotebooks

4.1. DATASETS

Various datasets will be used in the experiments to test ChatGPT, enhancing the evaluation process. The datasets will be obtained from the Kaggle famous dataset to address three selected experiments. The selection will be for balanced datasets regarding the problems that need to be addressed in the preprocessing phase, as the goal is to simulate a real-life situation.

4.2. CRISP-DM-BASED EXPERIMENT DESIGN

Based on the CRISP-DM framework, the experiment design details a systematic approach to utilizing ChatGPT in data mining tasks with various datasets. Both labeled and unlabeled data will be used for this experiment. The approach involves prompting ChatGPT to follow the CRISP-DM model phases by explicitly asking it to conduct business understanding, data understanding, data processing, data modeling, and eventually evaluating the results. The Deployment phase is not within the scope of this experiment, as it is context-dependent, and the goal is to build and run the model rather than streamline the process.

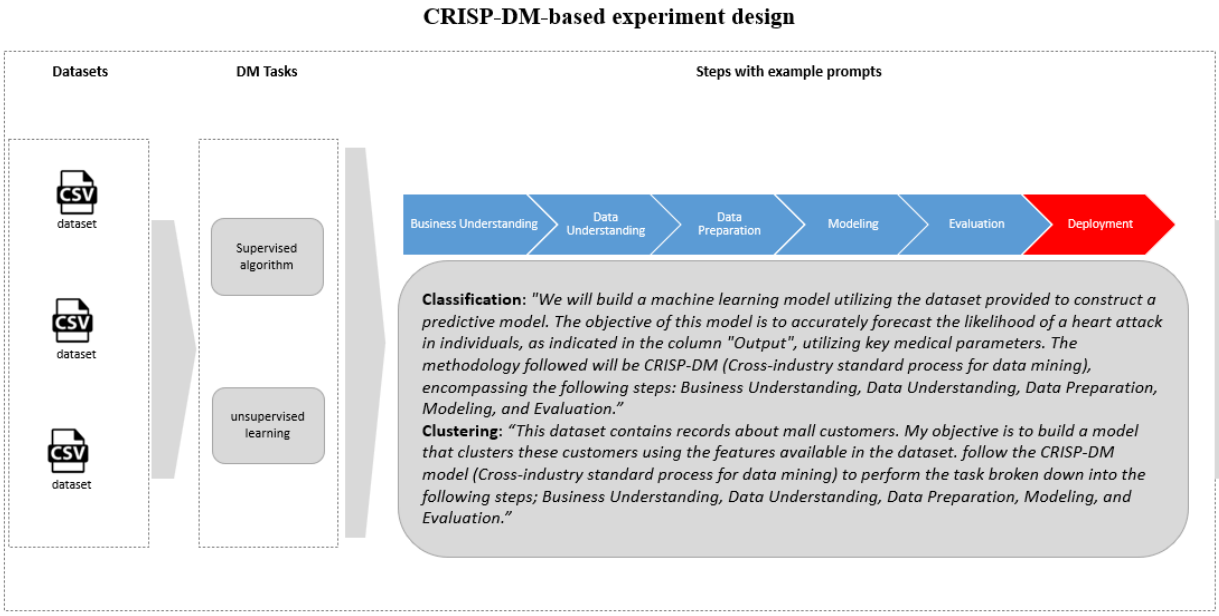

Figure 4: CRISP-DM-based experiment design.

4.3. GLOBAL USER EXPERIMENT

The global experiment is designed to evaluate the adaptability and performance of ChatGPT in data mining tasks provided by global users. It employs a general prompt strategy, allowing for a broad application without confining ChatGPT to specific procedural constraints. This approach intends to assess the model's ability to handle diverse data challenges and extract insights without the structured guidance typical of the CRISP-DM methodology.

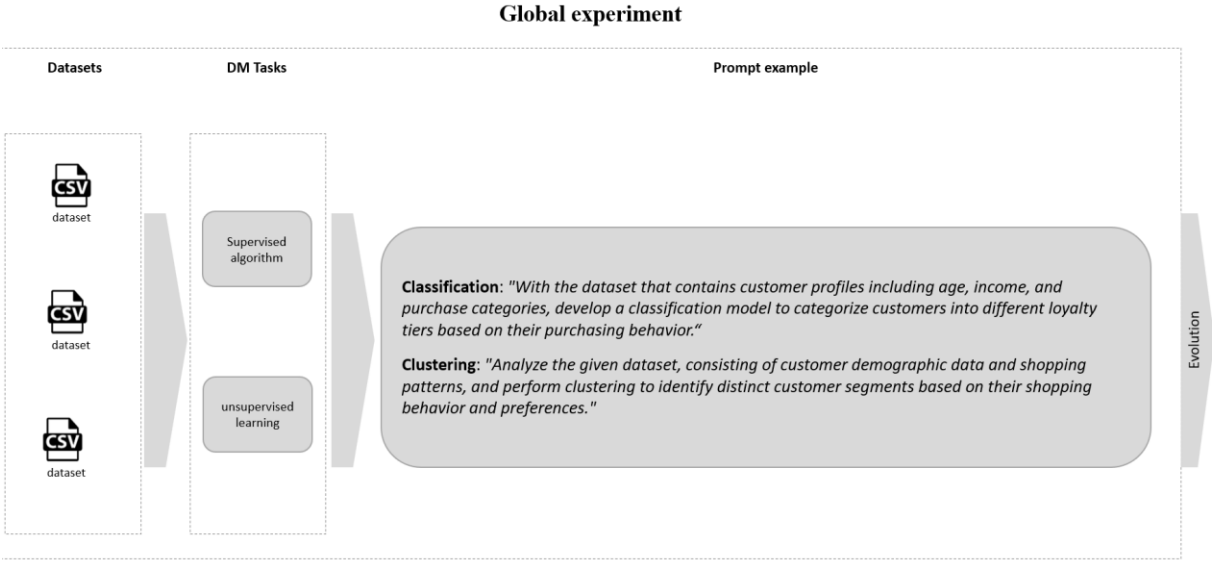

Figure 5: Global experiment

4.4. CHATGPT'S AUTOML FEATURES

Using the CRISP-DM model, ChatGPT shows a remarkable capability to retain context memory by dividing intricate problems into separate, manageable phases (Vaswani et al., 2017). This method is well-suited to the iterative aspect of data mining and machine learning projects, facilitating a more organized and targeted development process.

One noteworthy aspect of ChatGPT in AutoML is its proficiency in handling and interpreting data, particularly in commonly used formats like CSV. Its natural language processing capabilities extend beyond basic data handling to provide a deeper understanding of the data within the business or problem domain. This capacity to comprehend and integrate business context information adds valuable insights and is often challenging to achieve with traditional machine learning methods. ChatGPT seamlessly executes data manipulation and transformation, fundamental steps in any data science project, using Python's libraries and functions for data science that became popular lately like Pandas for tabular data. This streamlines the data preprocessing stage and ensures that the transformations align with the project's objectives. ChatGPT's capabilities include data visualization and leveraging Python's powerful libraries, such as Matplotlib, to produce insightful graphs and charts. This feature is essential for understanding data trends and patterns and making informed decisions in machine learning projects (OpenAI, 2024).

4.5. EVALUATION MATRIX

An evaluation matrix is a tool designed to measure the effectiveness of integrating ChatGPT within the phases of a data mining project, following the CRISP-DM model steps against using general prompts asking ChatGPT to perform the data mining tasks. In the thesis's evaluation matrix, a dual approach is used to assess the utility of ChatGPT in machine learning projects. Qualitative criteria evaluate the depth of understanding and problem-solving, with ratings from "Excellent" to "Unacceptable," which will be evaluated and justified after each experiment. Another approach for qualitative evaluation is with the dimensions of model selection and model code, evaluated on a binary scale—'Yes' for criteria met and 'No' for criteria not met—providing a straightforward assessment of ChatGPT's performance in these technical areas. On the other hand, a quantitative evaluation will be done directly by comparing the scores achieved in each experiment. The scores will vary according to the type of experiment and the algorithm used to solve the given problem. For example, in a regression problem, the metric that will be compared between two experiments will be RMSE, and for a classification problem, it will be the Silhouette Score.

Table 2: Evaluation matrix

Dimension	Criteria	Excellent (4)	Good (3)	Fair (2)	Poor (1)	Unacceptable (0)
Business Understanding	Does chatGPT provide business context, and how coherent was it?	All the business aspects are covered	Most of the business aspects are covered	Some aspects of the business are covered	Minimal Business aspects are covered	None of the business aspects are covered
Data Understanding	Does ChatGPT understand the data and identify the problems? How coherent was it?	Excellent understanding	Good understanding	Fair understanding	Poor understanding	Unacceptable understanding
Data Preparation	Does ChatGPT prepare the data correctly?	All the problems are covered.	Most of the problems are covered.	Some of the problems are covered.	Minimal of the problems are covered.	None of the problems are covered.
Modeling	Does ChatGPT select the proper model for the problem, and how does it build it?	Excellent execution	Good execution	Fair execution	Poor execution	Unacceptable execution
Evaluation	Does ChatGPT provide a clear and actionable evaluation of the model after receiving the results?	All the evaluation aspects are covered	Most evaluation aspects are covered	Some aspects of evaluation aspects are covered	Minimal evaluation aspects are covered	None of the evaluation aspects are covered
Dimension	Criteria	Yes		No		
Model selection	Evaluate choosing a relevant and effective model	A relevant model has been chosen		A relevant model has not been chosen		
Model code (Python)	The correctness of Python code produced by ChatGPT	The code is correct		The code is not correct		
Model scores						
Score	Accuracy (example)	X				
	Performance Metric (x)	X				
	Performance Metric (y)	X				
Deployment	Out of the study scope					

The test will use ChatGPT-4 from OpenAI, which provides advanced data analysis. The environment supports extensive document analysis, including CSV files. ChatGPT Enterprise's capabilities will be utilized without significant customization, adding one that highlights its out-of-the-box effectiveness in data mining tasks. The only modification will be to add a feature that ensures ChatGPT-4 takes breaks between tasks to avoid running out of time.

5. TESTING OF THE USE OF PROMPT ENGINEERING IN DM PROJECTS

The test will be conducted using the main two categories of data mining algorithms: supervised and unsupervised. After the experiments in the ChatGPT-4 interface are finished, the prompts and the generated text and code by ChatGPT-4 will be incorporated with Jupyter notebooks as a reference for the experiment details and steps. For each dataset and problem, two experiments will be conducted, with one aligned with the CRISP-DM model by explicitly asking ChatGPT-4 to follow the phases of the model, and the other will be executed with a general prompt asking ChatGPT-4 to build a machine learning model to solve the given problem. As mentioned before, the goal is to determine if using the CRISP-DM model with ChatGPT-4 could lead to any improvement in the overall performance and the results.

5.1. FIRST DM EXPERIMENT: CIRRHOSIS PATIENT SURVIVAL PREDICTION (SUPERVISED)

The dataset "Heart Attack Analysis & Prediction" is available on Kaggle. It is designed to facilitate the development of machine learning models for predicting the risk of heart attacks. The dataset includes various clinical and health-related features, making it a valuable resource for creating predictive models related to cardiovascular health. Using a medically focused dataset, we can explore the application of classification algorithms in a critical healthcare context.

5.1.1. Classification Experiments Assessment

In the business understanding phase, ChatGPT-4 showed clear insights into the business in both experiments. However, it provided a more detailed analysis of the CRISP-DM-aligned experiment. ChatGPT-4 effectively summarized the data on a column-by-column basis during the data understanding phase. The two experiments had a significant difference in the depth of data understanding. In the global experiment, ChatGPT-4 performed essential steps such as checking for missing values and ensuring well-distributed data without providing evidence. Conversely, the CRISP-DM-aligned experiment involved more thorough data preparation, including detailed analysis of missing values, correlation assessment between features via a heatmap, and in-depth investigation of each feature's distribution. In the CRISP-DM-aligned experiment, ChatGPT-4 visualized the data and feature correlations (refer to Figures 7 & 8).

Regarding model selection, both experiments chose models appropriate for the problem: 'Logistic Regression' in the global experiment and 'Random Forest' in the CRISP-DM aligned experiment. Notably, the CRISP-DM-aligned experiment achieved a higher accuracy of 86.9%, compared to 85.25% in the global experiment. This improvement is probably attributed to the more comprehensive data understanding and preparation in the CRISP-DM-aligned experiment.

The CRISP-DM-aligned experiment provided a much deeper evaluation of the results, assessing the model performance using the mentioned matrices and reassessing the business objectives considering the model performance. In the global user experiment, ChatGPT provided a short evaluation and explanation for the results.

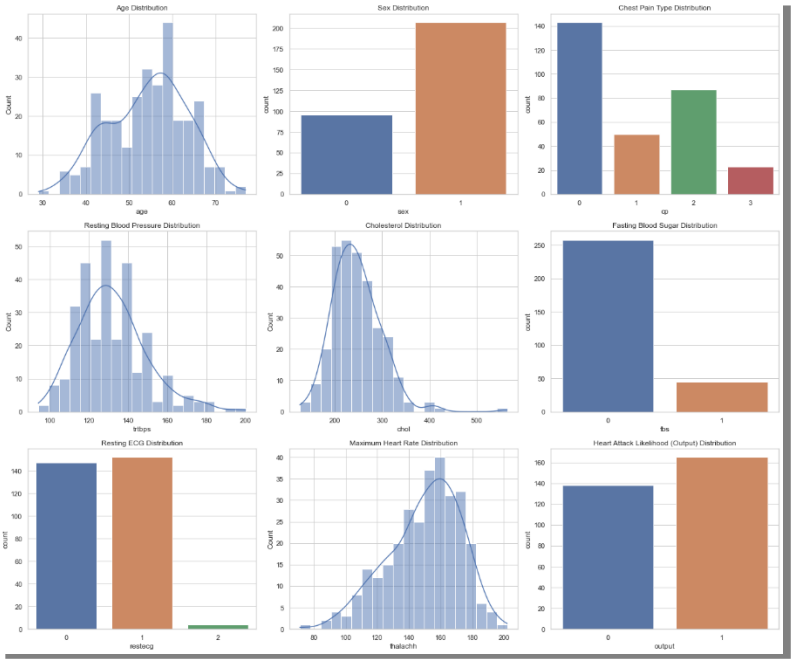

Figure 6: Feature distribution viz created by ChatGPT-4

Another visualization created by ChatGPT-4 highlights the model's tendency to produce more visuals when it's aligned with the CRISP-DM model. This alignment is more useful in helping the user identify patterns and validate the logic that the model is following, and perhaps make specific modifications for more expert users. Here chatGPT-4 provided a heatmap that shows the correlation between the existing variables in the data.

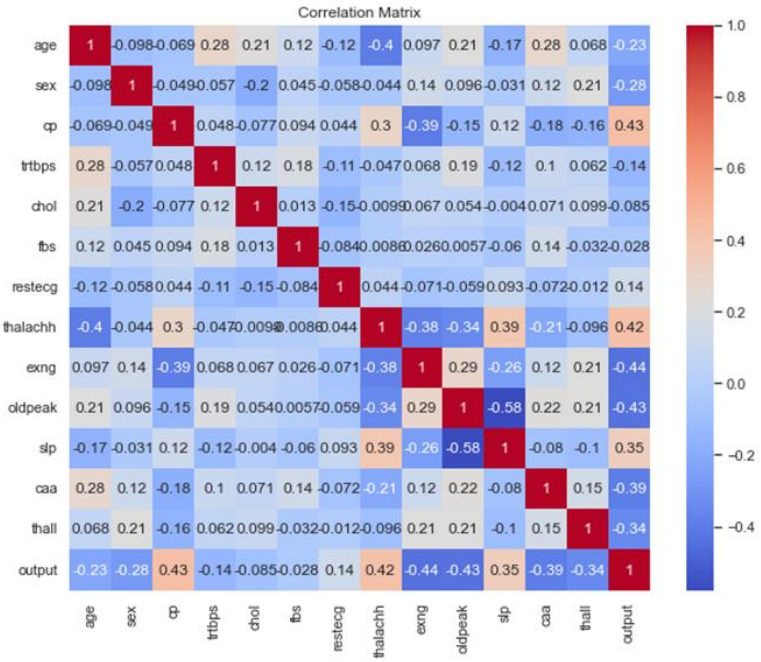

Figure 7: Heatmap generated by ChatGPT-4

5.1.2. Summary

The following table will summarize the findings of the conducted experiment providing a score on the evaluation matrix based on the analysis conducted in the previous part.

Table 3: The summary of the first experiment

Dimension	CRISP-DM experiment	Global experiment
	Score: Excellent (4), Good (3), Fair (2), Poor (1), Unacceptable (0)	
Business Understanding	4	3
Data Understanding	3	2
Data Preparation	4	3
Modeling	3	2
Evaluation	4	3
Dimension	Boolean: Yes (positive), No (Negative)	
Model selection	Yes	Yes
Model code (Python)	Yes	Yes
Metric	Metric value	
Accuracy	86.9%,	85.25%
Precision	88%	87.10%
Recall	88%	84.38%
F1 score	88%	Non

5.2. SECOND DM EXPERIMENT: CARS PRICE PREDICTION (SUPERVISED)

The dataset that will be used for this experiment is available on Kaggle, has been specially curated to help develop machine-learning models dedicated to predicting car prices. This dataset includes a comprehensive range of features encompassing various car specifications. These elements make it an essential resource for constructing accurate and efficient predictive models within the automotive industry. Focusing on this automotive dataset provides an opportunity to delve deeply into the application of regression algorithms.

5.2.1. Regression Experiments Assessment

During the business understanding phase, the global experiment skipped insights into the business and investigated the dataset. In contrast, the CRISP-DM experiment with ChatGPT-4 thoroughly understood the business and proposed more relevant business questions related to car pricing. In the data understanding phase, the global experiment provided a brief overview of the data. At the same time, the CRISP-DM approach delivered detailed information about the data, including the number of records, columns, and a classification of the dataset's attributes into numerical and categorical features. ChatGPT-4 also identified the potential for feature engineering, such as extracting the brand name from the car name.

In the data preparation phase, the global experiment examined existing features for potential engineering and conducted exploratory data analysis (EDA). ChatGPT-4 created visuals to check the distribution of prices using a histogram and the correlation between numerical features by

constructing a heatmap (Figure 9). The final proposed data preparation steps included scaling the numerical data and encoding the categorical features. In the CRISP-DM-aligned experiment for data preparation, ChatGPT-4 began by checking for missing values, then scaled the numerical data and encoded the categorical variables but decided against engineering the car name as an additional feature. Unlike the previous experiment, ChatGPT-4 generated graphs in the global experiment rather than the CRISP-DM, which indicates that using CRISP-DM does not contribute to that.

Before proceeding to modeling, both experiments split the data into 80% for training and 20% for testing, but it might be worth reconsidering different ratios due to the limited number of records in this dataset. In the modeling phase, ChatGPT-4 chose the random forest regressor for both experiments, which is suitable for this problem.

In the evaluation phase, key performance metrics such as Root Mean Squared Error (RMSE) and R-squared (R^2) were compared between the CRISP-DM-aligned and global experiments. The CRISP-DM-aligned experiment's RMSE, derived from its Mean Squared Error (MSE), is approximately 1843.70, slightly lower than the global experiment's testing RMSE of 1896.18, suggesting superior predictive accuracy. Furthermore, the R^2 score for the CRISP-DM-aligned experiment on testing data is 0.957, marginally higher than the global experiment's 0.946, indicating a better explanation of the variance in the dependent variable. Overall, the CRISP-DM-aligned experiment demonstrates slightly improved performance in both RMSE and R^2 during the testing phase, indicating greater accuracy and explanatory power.

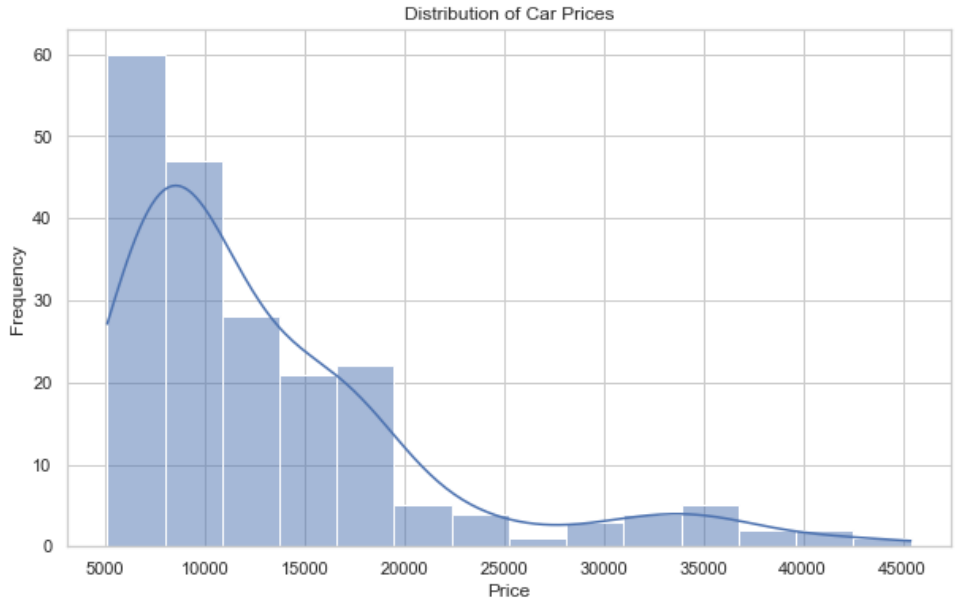

Figure 8: Cars' price distribution

5.2.2. Summary

The following table will summarize the findings of the conducted experiment providing a score on the evaluation matrix based on the analysis conducted in the previous part.

Table 4: The summary of the second experiment

Dimension	CRISP-DM experiment	Global experiment
	Score: Excellent (4), Good (3), Fair (2), Poor (1), Unacceptable (0)	
Business Understanding	3	1
Data Understanding	4	3
Data Preparation	3	2
Modeling	4	4
Evaluation	4	3
Dimension	Boolean: Yes (positive), No (Negative)	
Model selection	Yes	Yes
Model code (Python)	Yes	Yes
Metric	Metric value	
RMSE	1843.7	1896.18
R^2 Score	0.957	0.946

5.3. THIRD DM EXPERIMENT: MALL CUSTOMER CLASSIFICATION (UNSUPERVISED)

The dataset used in this experiment is available on Kaggle and consists of unlabeled data, indicating using unsupervised learning algorithms for analysis. It includes various attributes related to mall customers, covering different aspects of each individual. The lack of labels directs our attention toward algorithms that identify patterns without predefined categories. Using these characteristics, the study aims to group customers into distinct clusters, using the rich dataset to reveal the underlying structure of customer behavior and preferences.

5.3.1. Clustering Experiments Assessment

ChatGPT-4 did not provide any prior business insight during the business understanding phase. Instead, it proceeded directly to understanding the data. It is clear from this and previous experiments that ChatGPT-4 does not offer business understanding without explicit prompts. This contrasts with the CRISP-DM-aligned experiment, where it provided a coherent context of business understanding. In the data understanding phase, both experiments thoroughly explained the dataset's columns, clarifying each column's contents.

In the modeling phase, both experiments used elbow methods to identify the number of clusters. The CRISP-DM-aligned experiment was notably superior in scenarios requiring in-depth analysis and actionable insights. Unlike the global user experiment, which focused primarily on fitting the model and assigning labels, the CRISP-DM-aligned experiment not only assigned cluster labels but also delved into a comprehensive examination of cluster centroids. This was done within the context of original features and involved evaluating the distribution of data points across clusters. Such a detailed

approach significantly enhances the interpretability and practical application of clustering results, proving highly valuable for strategic decisions like targeted marketing efforts.

Regarding clustering quality, the CRISP-DM-aligned experiment outperformed the global user experiment. This is evidenced by superior metrics in the Silhouette Score (0.35 vs. 0.317) and the Calinski-Harabasz Index (97.26 vs. 71.21). The higher Silhouette Score indicates that the CRISP-DM-aligned experiment achieves better cohesion within clusters and separation from neighboring clusters, ensuring a more appropriate grouping of data points. Similarly, the greater Calinski-Harabasz Index suggests that its clusters are not only more densely packed but also more distinct from each other compared to those in the global user experiment. These metrics demonstrate that the CRISP-DM-aligned experiment provides a more effective and coherent clustering solution.

The evaluation for the CRISP-DM-aligned experiment provides superior information, offering a concise and insightful interpretation of performance metrics. This evaluation clearly explains the implications of the Silhouette Score and Calinski-Harabasz Index, effectively communicating the model's success in creating well-separated and dense clusters and suggesting clear paths for future improvements. In contrast, the evaluation for the global user experiment is less explicit in its implications for model performance and potential enhancements. The evaluation of the CRISP-DM-aligned experiment stands out for its directness and usefulness, guiding stakeholders through the results and indicating practical steps for refinement. This approach embodies a more actionable and understandable way of explaining clustering effectiveness.

5.3.2. Summary

The following table will summarize the findings of the conducted experiment providing a score on the evaluation matrix based on the analysis conducted in the previous part.

Table 5: The summary of the third experiment

Dimension	**CRISP-DM experiment**	**Global experiment**
	Score: Excellent (4), Good (3), Fair (2), Poor (1), Unacceptable (0)	
Business Understanding	5	0
Data Understanding	4	4
Data Preparation	5	5
Modeling	5	2
Evaluation	4	2
Dimension	**Boolean: Yes (positive), No (Negative)**	
Model selection	Yes	Yes
Model code (Python)	Yes	Yes
Metric	**Metric value**	
Silhouette Score	0.35	-
Calinski-Harabasz Index	97.26	71.21

5.4. PROMPT ANALYSIS AND IMPACT

This section discusses how well-structured prompts affect the results of automated machine learning (AutoML) processes in ChatGPT. The experiments focused on classification, regression, and clustering tasks. The outcomes reveal that when ChatGPT-4 follows the CRISP-DM model phases, it performs better in all three experiments. The following sub-section provides a summary comparison that clarifies this improvement in the CRISP-DM-aligned experiments.

5.4.1. Insights from Classification, Regression, and Clustering Experiments

Classification Experiment (5.1.1): In the classification task, the CRISP-DM methodology resulted in a final accuracy of 86.9%, while the global experiment only achieved an accuracy of 85.25%. Apart from the results, there were significant differences in the details provided by ChatGPT-4 in both experiments. For example, in data preparation for modeling, the CRISP-DM aligned model provided a more detailed explanation and better preprocessed the data. ChatGPT-4 looked at the data distribution for all numerical variables, analyzed the correlation between the variables, and provided graphs to explain what each graph represents (see Figure 7: Feature distribution). On the other hand, in the global experiment, ChatGPT-4 only judged the data distribution based on the data summary and did not check for the correlation between existing variables.

Another example is that ChatGPT-4 provided a clear explanation and justification for the preprocessing decisions in the CRISP-DM aligned model, including feature selection and the method used for handling categorical variables. It is unclear if ChatGPT-4 took care of these elements in the global experiment. However, mentioning these justifications and explanations is essential for the user interacting with ChatGPT-4.

Regression Experiment (5.2.1): After comparing two experiments for predicting car prices, it was found that the CRISP-DM methodology had a slightly higher R-squared value of 0.957 compared to the global experiment approach, which had a value of 0.946. This suggests that both models performed similarly, but the CRISP-DM method had a slightly better fit.

The CRISP-DM approach followed a structured process that included explicit phases such as data preparation and model evaluation, which was beneficial for systematic analysis and clarity. On the other hand, the global approach was less structured in its execution, particularly in the preliminary data handling and feature engineering stages.

It was observed that the CRISP-DM experiment was more aligned with thorough business understanding and detailed data preparation, which could be crucial for practical applications and scalability in real-world scenarios. Therefore, the CRISP-DM methodology might provide a slight edge for future experiments aiming at business deployment due to its structured approach and clear evaluation metrics.

Clustering Experiment (5.3.1): The CRISP-DM and global user approaches for clustering can be compared by analyzing their Calinski-Harabasz Index scores that indicate the clustering quality. The CRISP-DM approach recorded a higher Calinski-Harabasz Index score of 97.26, while the global user approach scored 71.21. This indicates that the clusters generated by CRISP-DM are more distinct and better separated than those produced by the global user approach.

In addition to the Calinski-Harabasz Index, the CRISP-DM approach also reported a Silhouette score of 0.35, which indicates that the clusters generated by this approach are well-separated and cohesive. This combined metric approach highlights the effectiveness and robustness of CRISP-DM in achieving successful clustering outcomes.

Despite the numerical advantage of the CRISP-DM approach, both methodologies demonstrated rigorous data handling and preprocessing. However, CRISP-DM excelled in detailed visualization and thorough preprocessing, including effective categorical encoding and feature scaling. Also, the Elbow Method was used to determine the optimal number of clusters. These detailed visual aids and explanations enhance the interpretability of the clustering process, fostering a deeper understanding of the data.

On the other hand, the global user approach followed similar preprocessing steps but did not include additional metrics like the Silhouette score, which could have provided further insights into the cohesion and separation of clusters.

5.4.2. Impact of Prompt Design on AutoML Performance

After conducting three different experiments, it became evident that the prompts given to ChatGPT have a significant impact on its ability to perform AutoML tasks. By breaking down the process into CRISP-DM phases, ChatGPT-4 can delve into each phase, resulting in more insights that enhance overall performance. For each phase:

1. **Business Understanding:** The first phase of developing ChatGPT-4 involved providing more context to the assistant. This was aimed at enabling better decision-making in the following stages.
2. **Data Understanding and Preparation:** The second phase involved better data preprocessing done by ChatGPT-4. Understanding the data better led to improved data preparation.
3. **Modeling and Evaluation:** In the third phase, structured prompts were used to guide the selection of suitable models and evaluation metrics. This ensured that the AutoML process was tailored to specific project goals.

The experiments revealed a clear connection between the design of prompts and the effectiveness of AutoML tasks carried out by ChatGPT. This analysis emphasizes the importance of prompt engineering in enhancing AutoML performance and establishes a groundwork for creating compelling, prompt guidelines that can result in more precise, efficient, and business-aligned outcomes in data mining projects.

6. RESULTS AND DISCUSSION

6.1. ESSENTIAL DIFFERENCES BETWEEN TRADITIONAL AUTOML AND LLM-DRIVEN.

The literature review section for AutoML (2.1 Auto Machine Learning) mentions that the traditional method for processing data involves a pipeline with underlying relatively complex logic for each phase in each data mining task. For instance, Automated Data Preprocessing (AutoDP) is a Python code that takes in the data and performs the cleaning process based on the predefined logic in the code. However, this method lacks flexibility and understanding, leading to poor preprocessing in cases not pre-encoded in the underlying logic. On the other hand, the LLMs-driven method generates the preprocessing Python code based on the data and business understanding, making it more flexible. Although imperfect, it covers wider cases and can solve more complex problems than traditional methods.

Another critical aspect that LLMs excel in compared to traditional methodologies is the ability to provide results and evaluations using human language, which can be less intimidating for non-expert users. In the traditional AutoML methods, there is a lack of transparency presented in explanation with human language that ChatGPT-4 can provide for the results that the traditional methods cannot give; this concern is critical, especially in sectors where understanding model decision-making is of utmost importance (Tuggener et al., 2019).

The aspects mentioned above are points where LLMs can outperform the traditional methods, yet they are far from perfect. The exceptional performance of ChatGPT-4 would probably encourage more non-expert people to conduct data mining projects, but given the downside and limitations mentioned in (2.2.3 Challenges and Limitations), these non-expert users could lead to undetectable failures. A critical concern in developing AI systems is the potential for bias. This is because these models rely on existing data that may contain domain disinformation in the data they are trained on (Buolamwini & Gebru, 2018). Furthermore, one of the primary obstacles in prompt engineering is the problem of "hallucinations." This refers to situations where LLMs produce either inaccurate or non-existent information. This issue arises when the model lacks adequate training data or broadly generalizes patterns. That could be a problem with non-experts as it would be harder to decide when they should trust the output and when they should not.

Also, there is a difference in the determinism of the output; in the traditional methods, the same input will always result in the same output as the underlying logic will re-perform the same decisions in each iteration, which is not the case with LLMs-based AutoML as the output of these models are not predicted even if the same input exactly is given due to the complexity of the technology used for these models. It could introduce a problem if the user is trying to get many iterations for a given task, as the comparison could be more complicated due to using different techniques for each iteration.

6.2. FUTURE WORK

The current work only involved limited experimentation with ChatGPT-4 using selected algorithms. There's an opportunity for future work to experiment with other algorithms to further analyze performance. Another potential area for future work is to use additional benchmarks to compare ChatGPT-4's performance with other classical frameworks like SEMMA and TDSD.

It would be beneficial to conduct a comprehensive comparative analysis of ChatGPT-4 with other Language Models in the context of AutoML to evaluate their strengths, weaknesses, and unique contributions to the field, for example, comparing it with Gemini from Google. Additionally, creating new methods to evaluate performance from different angles, focusing on other aspects of data mining projects, would be advantageous.

Lastly, incorporating real data from various industries could provide insight into the real synergy between AutoML using Language Models and real-world operations, helping to adjust exceptions in using these technologies in real business scenarios.

7. CONCLUSIONS

My research has achieved its objectives by first establishing a foundational understanding of the various components involved, including AutoML, generative AI, and prompt engineering. The potential of LLMs was investigated, precisely their capabilities in completing a data mining task and how to use prompt engineering properly to achieve the intended goal. An evaluation matrix was developed to assess the performance of the conducted experiments both quantitatively and qualitatively.

Three experiments were conducted using diverse datasets in the ChatGPT-4 environment, employing both CRISP-DM-aligned and global user settings. Two experiments involved supervised learning with labeled data, while the third focused on unsupervised learning with unlabeled data. Meticulous analysis after each experiment filled the evaluation matrix and compared the two approaches.

The results showed that the CRISP-DM-aligned experiments provided better outcomes, with clear evidence that breaking down the task into CRISP-DM phases and allowing ChatGPT-4 to deeply analyze each phase separately led to better insights and results. Aligning the process with CRISP-DM also provided more detailed feedback and analysis from ChatGPT-4, which could benefit users.

Based on this research, it is recommended to use a structured and phase-driven approach when using LLMs like ChatGPT-4 for data mining tasks. The CRISP-DM model provides a proven framework for organizing and executing such projects, and its integration with ChatGPT-4 can significantly improve the quality and depth of analysis.

Future research could explore how effective ChatGPT-4 is in handling more complex data mining situations, such as dealing with imbalanced datasets or integrating domain-specific knowledge. Additionally, studying the impact of different prompt engineering techniques on ChatGPT-4's performance could further enhance its application in AutoML (more on future work mentioned in Results and Discussion)

In summary, this research highlights the transformative potential of integrating AI, particularly LLMs like ChatGPT-4, with structured methodologies like CRISP-DM to revolutionize the field of data analytics. By adopting a systematic and context-aware approach, we can unlock new possibilities for data-driven decision-making and knowledge discovery.

BIBLIOGRAPHY

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610-623). https://doi.org/10.1145/3442188.3445922

Bilal, M., Ali, G., Iqbal, M. W., Anwar, M., Malik, M. S. A., & Kadir, R. A. (2022). Auto-PreP: Efficient and automated data preprocessing pipeline. IEEE Access, 10, 107764–107784.

https://doi.org/10.1109/access.2022.3198662Buolamwini, J., & Gebru, T. (2018, January). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on Fairness, Accountability and Transparency (pp. 77-91). PMLR.

Cao, Y., Li, S., Liu, Y., Yan, Z., Dai, Y., Yu, P. S., & Sun, L. (2023). A comprehensive survey of AI-generated content (AIGC): A history of generative AI from GAN to ChatGPT. arXiv preprint arXiv:2303.04226.

Chauhan, K., Jani, S., Thakkar, D., Dave, R., Bhatia, J., Tanwar, S., & Obaidat, M. S. (2020, March). Automated machine learning: The new wave of machine learning. In 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA) (pp. 205-212). IEEE.

Chen, B., Zhang, Z., Langrené, N., & Zhu, S. (2023). Unleashing the potential of prompt engineering in large language models: A comprehensive review. arXiv preprint arXiv:2310.14735.

Chen, Y. W., Song, Q., & Hu, X. (2021). Techniques for automated machine learning. ACM SIGKDD Explorations Newsletter, 22(2), 35-50.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In Advances in Neural Information Processing Systems (Vol. 27).

Martínez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernández-Orallo, J., Kull, M., Lachiche, N., & Flach, P. (2019). CRISP-DM twenty years later: From data mining processes to data science trajectories. IEEE Transactions on Knowledge and Data Engineering, 33(8), 3048-3061.

Muktadir, G. M. (2023). A brief history of prompt: Leveraging language models. arXiv preprint arXiv:2310.04438.

OpenAI. (2024). Six strategies for getting better results. OpenAI. Retrieved May 26, 2024, from https://platform.openai.com/docs/guides/prompt-engineering/six-strategies-for-getting-better-results

Truong, A., Walters, A., Goodsitt, J., Hines, K., Bruss, C. B., & Farivar, R. (2019, November). Towards automated machine learning: Evaluation and comparison of AutoML approaches and tools. In

2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI) (pp. 1471-1479). IEEE.

Tuggener, L., Amirian, M., Rombach, K., Lörwald, S., Varlet, A., Westermann, C., & Stadelmann, T. (2019, June). Automated machine learning in practice: State of the art and recent results. In 2019 6th Swiss Conference on Data Science (SDS) (pp. 31-36). IEEE.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (Vol. 30).

Vazquez, H. C. (2022). A general recipe for automated machine learning in practice. In Lecture notes in computer science (pp. 243–254). https://doi.org/10.1007/978-3-031-22419-5_21