SAPIENZA
UNIVERSITÀ DI ROMA

Feature Engineering with Large Language Models for Tabular Data Analysis

Facoltà di Ingegneria dell'Informazione, Informatica e Statistica
Master in Data Science

Edo Fejzic
ID number 2113132

Advisor	Co-Advisor
Prof. Aris Anagnostopoulos	Loris Cino

Academic Year 2024/2025

Thesis not yet defended

Feature Engineering with Large Language Modelsfor Tabular Data Analysis
Master thesis. Sapienza University of Rome

This thesis has been typeset by LaTeX and the Sapthesis class.

Author's email: fejzic.2113132@studenti.uniroma1.it

Dedicated to my family and friends for their unwavering support.

Abstract

To add

Acknowledgments

To add

Contents

Chapter 1

Introduction

The rapid evolution of machine learning and artificial intelligence has transformed how we approach data analysis and predictive modeling. Among the most significant recent developments is the emergence of Large Language Models (LLMs) as powerful tools for understanding and manipulating various forms of data, including structured tabular datasets. This thesis investigates the application of LLMs to automated feature engineering, a critical component of the machine learning pipeline that has traditionally required extensive domain expertise and manual effort.

1.1 Background and Motivation

Feature engineering, the process of creating, transforming, and selecting relevant features from raw data, represents one of the most crucial yet labor-intensive aspects of machine learning projects. Traditional approaches rely heavily on domain expertise, statistical analysis, and iterative experimentation to identify meaningful feature transformations that improve model performance. This manual process creates significant bottlenecks in data science workflows and often limits the accessibility of machine learning techniques to organizations without specialized expertise.

The advent of Large Language Models, particularly with the release of GPT-3, GPT-4, and other state-of-the-art models, has opened new possibilities for automating complex reasoning tasks that were previously exclusive to human experts. These models demonstrate remarkable capabilities in understanding context, generating coherent explanations, and applying world knowledge to novel problems. Recent research has begun exploring how these capabilities can be leveraged for structured data analysis tasks, including feature engineering and selection.

However, the application of LLMs to tabular data presents unique challenges. Unlike natural language or image data, where deep learning has achieved remarkable success through end-to-end learning approaches, tabular data analysis continues to rely heavily on carefully crafted features and domain-specific transformations. The heterogeneous nature of tabular data, with mixed data types, varying scales, and complex interdependencies, requires sophisticated reasoning that combines statistical understanding with semantic knowledge.

1.2 Research Objectives

This thesis aims to advance the understanding and practical application of Large Language Models for automated feature engineering in tabular data analysis. The research is structured around five interconnected objectives that collectively address both theoretical understanding and practical implementation challenges.

The first major objective focuses on conducting a **comprehensive literature analysis** that systematically examines existing approaches for LLM-based feature engineering. This analysis goes beyond a simple review to identify key methodological patterns that characterize successful approaches, empirical findings that establish performance baselines across different domains, and critical limitations that constrain current applications. By establishing a clear taxonomy of existing methods and understanding their relative strengths and weaknesses, this analysis provides the foundation for developing improved approaches and identifying unexplored research directions.

The second objective centers on **methodological development**, specifically the creation and evaluation of novel hybrid approaches that effectively combine the semantic reasoning capabilities of Large Language Models with traditional statistical methods for feature selection and engineering. Rather than treating LLM-based and statistical approaches as competing alternatives, this research explores synergistic combinations that leverage the complementary strengths of both paradigms. The semantic understanding and world knowledge embedded in LLMs can guide feature generation and interpretation, while statistical methods provide rigorous validation and performance optimization.

The third objective involves conducting **systematic empirical evaluation** across diverse tabular datasets to assess the effectiveness, robustness, and limitations of LLM-based feature engineering approaches under various conditions. This evaluation encompasses multiple dimensions including dataset characteristics such as size, dimensionality, and domain complexity, as well as different evaluation scenarios such as few-shot learning, domain transfer, and computational resource constraints. The goal is to establish clear performance boundaries and identify the specific conditions under which different approaches excel or fail.

The fourth objective addresses **practical framework design** that bridges the gap between research prototypes and real-world deployment requirements. This involves developing systems that address genuine constraints faced by practitioners, including computational efficiency for large-scale deployment, interpretability requirements for regulated domains, and robust bias mitigation strategies that ensure fair and ethical feature generation. The framework must balance theoretical optimality with practical usability, providing clear guidelines for implementation and deployment.

Furthermore, this thesis aims also to employ LLM to do features engineering on text features. LLMs are originally designed to handle text, this characteristics can be combined to the reasoning capability to automatically design features that efficiently encode the text information.

The final objective provides **critical analysis** that offers a balanced and nuanced assessment of when and how LLM-based approaches provide genuine advantages over traditional automated feature engineering methods. This analysis identifies specific use cases and application scenarios where the unique capabilities of LLMs—such as semantic understanding and few-shot learning—provide substantial benefits, while also acknowledging scenarios where traditional statistical methods remain superior. This objective ensures that the research contributes realistic expectations and practical guidance rather than overly optimistic claims about LLM capabilities.

All the topic are really interesting, they are perfect for a thesis but they are slightly different to our work. The topics highlight 'literature review' more then a experimental work. What about LLMs for text based feature engineering? Reinforcement Learning by Machine Learning Feedback can wait for the moment.

1.3 Research Questions

This research is guided by several fundamental questions that address both theoretical understanding and practical implementation challenges, each requiring comprehensive investigation to advance the field meaningfully.

A central question concerns how different prompting strategies affect the quality and effectiveness of LLM-generated features for tabular data analysis (Prompt engineering is a mess, I suggest you to not spend too much time on it). Specifically, this research investigates the comparative performance of data-driven approaches, which provide statistical samples to guide LLM reasoning, versus text-based methods that rely primarily on semantic descriptions and contextual information. Understanding these differences is crucial because they represent fundamentally different paradigms for leveraging LLM capabilities—one emphasizing pattern recognition from examples, the other focusing on knowledge-based reasoning. The investigation extends beyond simple performance metrics to examine the types of features generated, their interpretability, and their robustness across different dataset characteristics.

Another critical research question addresses the optimal methods for integrating domain knowledge and contextual information into LLM-based feature engineering systems. This challenge is multifaceted, encompassing questions about how to effectively encode domain expertise in prompts, how to balance general statistical principles with domain-specific insights, and how to ensure that contextual information enhances rather than constrains the feature generation process. The research explores various approaches for knowledge integration, including retrieval-augmented generation techniques, structured prompt engineering, and hybrid systems that combine multiple information sources.

The research also investigates the specific conditions under which LLM-based approaches provide significant advantages over traditional automated feature engineering methods. Rather than assuming universal superiority of either approach, this investigation seeks to identify the characteristics of datasets, tasks, and resource constraints that favor one approach over another. This includes examining factors such as dataset size, domain complexity, availability of training data, computational resources, and interpretability requirements. Understanding these boundary conditions is essential for providing practical guidance to practitioners and avoiding overoptimistic claims about LLM capabilities.

A particularly important research question concerns how to effectively address bias propagation and ensure fairness in LLM-generated features. Since LLMs are trained on large-scale data that inevitably contains societal biases, there is significant risk that these biases will be encoded in generated features, potentially leading to discriminatory outcomes in downstream applications. This research investigates detection mechanisms for identifying biased features, mitigation strategies that can reduce bias without sacrificing performance, and validation approaches that can ensure fairness across different demographic groups and sensitive attributes.

Finally, the research examines the computational and scalability trade-offs associated with different LLM-based feature engineering approaches. While LLMs offer powerful capabilities, they also require substantial computational resources that may limit their practical applicability in resource-constrained environments. This investigation encompasses analysis of computational complexity, memory requirements, inference latency, and the relationship between model size and feature engineering effectiveness. Understanding these trade-offs is crucial for developing approaches that balance performance with practical deployment constraints.

1.4 Contributions

This thesis makes several significant contributions to the field of automated feature engineering and the application of Large Language Models to structured data analysis, each addressing critical gaps in current understanding and practice.

The first major contribution is a **comprehensive survey and theoretical framework** that systematically reviews and categorizes existing LLM-based feature engineering approaches. Unlike previous reviews that focus on narrow subsets of methods, this survey provides a unified taxonomic framework that encompasses the full spectrum of current approaches Does a literature review already exist? [?] , from data-driven statistical inference methods to text-based semantic reasoning techniques. The survey identifies fundamental methodological patterns that characterize successful approaches, establishes clear performance baselines across different domains and datasets, and provides critical analysis of limitations that constrain current applications. This theoretical foundation is essential for advancing the field beyond ad-hoc experimental approaches toward principled methodological development.

The second contribution involves significant **methodological innovations** through the development of novel hybrid approaches that effectively combine the semantic reasoning capabilities of Large Language Models with the statistical rigor of traditional feature selection techniques. Rather than treating these paradigms as competing alternatives, the research develops synergistic integration strategies that leverage the complementary strengths of both approaches. These hybrid methods address key limitations of pure LLM-based approaches, such as statistical unreliability and computational inefficiency, while overcoming the semantic blindness of traditional statistical methods. The methodological innovations include novel prompt engineering techniques, integration frameworks for combining multiple information sources, and validation protocols that ensure both statistical validity and semantic meaningfulness.

The third contribution provides **extensive empirical insights** through systematic experimental evaluation that establishes clear performance boundaries and provides practical guidance for practitioners. This evaluation goes beyond simple accuracy comparisons to examine multiple dimensions of performance including robustness, interpretability, computational efficiency, and bias propagation. The empirical analysis identifies specific conditions under which different approaches excel or fail, providing nuanced guidance that helps practitioners select appropriate methods for their specific use cases. The insights are particularly valuable for understanding the trade-offs between different approaches and setting realistic expectations about LLM capabilities in feature engineering tasks.

The fourth contribution addresses the critical challenge of bias and fairness through the development of a **systematic bias analysis framework** that provides both detection mechanisms and mitigation strategies for LLM-generated features. Given the increasing recognition of bias propagation in AI systems, this framework addresses a critical gap in current LLM-based feature engineering approaches. The framework includes novel techniques for identifying potentially biased features, validation protocols for assessing fairness across different demographic groups, and practical mitigation strategies that can reduce bias without sacrificing predictive performance. This contribution is essential for enabling responsible deployment of LLM-based feature engineering in sensitive applications.

The final contribution establishes **rigorous performance benchmarking standards** through comprehensive comparison of LLM-based approaches against traditional automated feature engineering methods across diverse datasets and evaluation metrics. This benchmarking goes beyond simple accuracy comparisons to include computational efficiency, scalability, interpretability, and robustness analysis. The benchmarking framework provides standardized evaluation protocols that enable fair comparison across different methods and establish clear performance baselines for future research. This contribution is crucial for moving the field toward more rigorous experimental standards and avoiding overly optimistic claims about LLM capabilities.

In my opinion Research Objectives - Research Questions - Contribution are repetitive. What are the differences?

1.5 Thesis Structure

This thesis is organized into six chapters that systematically address the research objectives and questions outlined above:

Chapter 2 provides a comprehensive literature review examining existing approaches to LLM-based feature engineering. This chapter establishes the theoretical foundation and identifies key research gaps that motivate the current work.

Chapter 3 presents the methodological framework developed for this research, including the hybrid approaches for combining LLM reasoning with statistical methods, experimental design, and evaluation protocols.

Chapter 4 details the experimental setup, datasets used, and implementation details of the various approaches evaluated in this study.

Chapter 5 presents the empirical results of our experiments, including performance comparisons, scalability analysis, and bias assessment across different methods and datasets.

Chapter 6 summarizes the key findings, discusses their implications for both research and practice, acknowledges limitations of the current work, and outlines directions for future research.

Each chapter builds upon the previous ones to develop a comprehensive understanding of how Large Language Models can be effectively leveraged for automated feature engineering while addressing the practical challenges and ethical considerations that arise in real-world applications.

Chapter 2

Literature Review

The intersection of Large Language Models (LLMs) and tabular data analysis represents an emerging and rapidly evolving research domain. This chapter provides a comprehensive review of the current state-of-the-art in LLM-based feature engineering and selection, examining methodological approaches, empirical findings, and critical limitations that inform our research direction.

2.1 Feature Engineering

The first section is to introduce the feature engineering for tabular data.

2.1.1 Preprocessing (Not sure)

2.1.2 Feature Creation

2.1.3 Feature Selection

Mention also dimensionality reduction.

2.1.4 Others if needed...

2.1.5 AutoML

Feature engineering is tedious are time consuming. To face this challenge several framework have been developed. These techniques are called AutoML.

2.2 Foundations of LLM-Based Feature Engineering

The application of Large Language Models to tabular data analysis has emerged from the recognition that traditional feature engineering approaches, while statistically sound, often fail to leverage the rich semantic knowledge that humans naturally apply when understanding data [?]. Unlike computer vision or natural language processing domains where deep learning has achieved remarkable success through end-to-end learning, tabular data analysis has remained heavily dependent on manual feature engineering and domain expertise [?].

Recent advances in large language models, particularly with the development of models like GPT-4 [?] and their demonstrated few-shot learning capabilities, have opened new possibilities for automating feature engineering tasks. The fundamental premise underlying this research direction is that LLMs, trained on vast corpora of text that include descriptions of datasets, domain knowledge, and analytical procedures, can be prompted to generate meaningful feature transformations that would traditionally require human expertise.

2.3 Taxonomies and Methodological Frameworks

In this section can help to have a Figure like this

2.3.1 Data-Centric Categorization

Li et al. [?] propose a fundamental taxonomy that categorizes LLM-based feature selection methods into two distinct paradigms based on the type of information provided to the model:

Data-driven methods operate by providing specific data samples to LLMs, enabling them to perform statistical inference and correlation analysis. In this approach, the LLM receives feature values paired with target variable values, formatted as few-shot examples within the prompt. The model is then expected to infer relationships and assign importance scores based on observed patterns in the provided samples. Formally, given a dataset d with m samples, data-driven methods construct sample pairs $SP_i = \{(n^j_{f_i}, n^j_y)\}$ for feature f_i and target variable y, where $j \in \{1, ..., m\}$.

The prompt construction follows the pattern:

$$P^{Data}_{f_i} = \text{prompt}(C, SP_i) \tag{2.1}$$

where C represents the instruction context and SP_i contains the sample pairs for feature f_i.

Text-based methods, in contrast, leverage the extensive semantic knowledge embedded within LLMs by incorporating detailed dataset and feature descriptions into

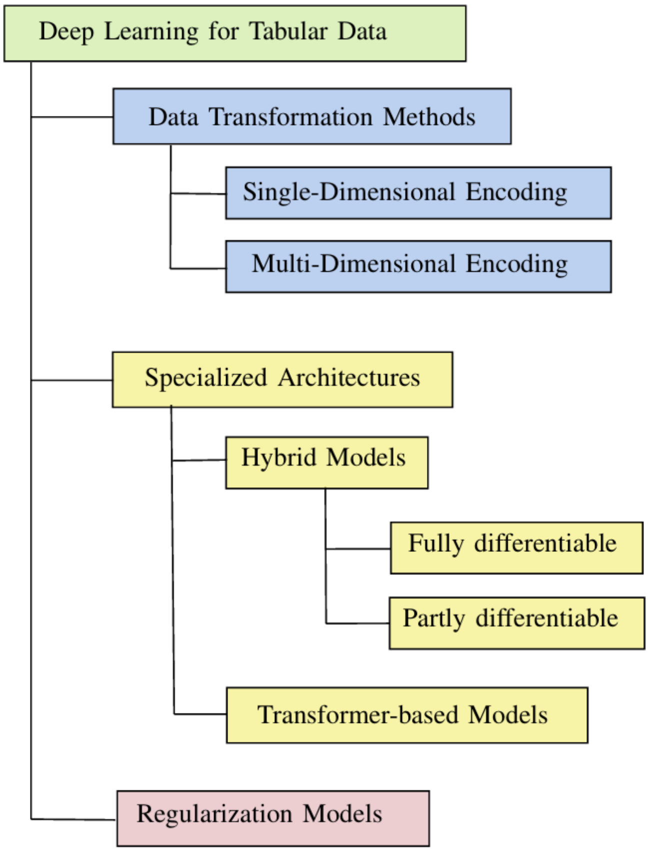
Deep Learning for Tabular Data
Data Transformation Methods
Single-Dimensional Encoding
Multi-Dimensional Encoding
Specialized Architectures
Hybrid Models
Fully differentiable
Partly differentiable
Transformer-based Models
Regularization Models

Figure 2.1. Esempio di tassonomia.

prompts. These approaches construct prompts using dataset descriptions (des_d) and feature-specific descriptions (des_{f_i}):

$$P_{f_i}^{Text} = \text{prompt}(C, des_d, des_{f_i}) \tag{2.2}$$

This methodological distinction proves crucial for understanding the relative strengths and limitations of different approaches, as empirically demonstrated across multiple evaluation studies.

2.3.2 Iterative Feature Generation Frameworks

Han et al. [**?**] introduce a more sophisticated framework that employs iterative feature generation specifically optimized for few-shot learning scenarios. Their approach addresses the fundamental challenge of limited training data by systematically leveraging LLMs' world knowledge to create meaningful feature representations through multiple generation cycles.

The iterative process operates through four interconnected phases that systematically build upon each other to achieve optimal feature generation outcomes.

The initial phase involves comprehensive context provision, where dataset descriptions and carefully selected representative samples are provided to the LLM to establish robust semantic understanding of the domain and task requirements. This phase is crucial because it sets the foundation for all subsequent feature generation by ensuring that the model has sufficient contextual information to generate meaningful and relevant features. The context provision goes beyond simple data descriptions to include domain-specific knowledge, task objectives, and examples that illustrate the types of patterns and relationships that are important for the specific application.

Subsequently, the framework focuses on intelligent feature generation, where the LLM leverages both its embedded world knowledge and the provided context to generate novel features through sophisticated domain knowledge application and pattern recognition capabilities. During this phase, the model draws upon its extensive training to identify potentially useful transformations, combinations, and derived features that might not be obvious from purely statistical analysis. The generation process is guided by the contextual understanding established in the previous phase, ensuring that generated features are both theoretically sound and practically relevant to the specific domain and task.

The framework then implements rigorous performance evaluation, where generated features undergo systematic validation using downstream classifiers through comprehensive cross-validation protocols. This evaluation phase is essential for determining which generated features actually provide predictive value and should be retained for the final feature set. The evaluation goes beyond simple accuracy metrics to consider factors such as feature stability, interpretability, and potential for overfitting, ensuring that only genuinely useful features are incorporated into the final model.

Finally, the process establishes adaptive iterative refinement, where the entire procedure repeats with updated context that incorporates lessons learned from previous iterations, including performance feedback and insights about which types of features proved most effective. This refinement process allows the system to continuously improve its feature generation capabilities by learning from both successes and failures in previous iterations. The iterative nature ensures that the feature engineering process can adapt to the specific characteristics of each dataset and task, rather than relying on generic feature generation strategies.

This framework demonstrates particular effectiveness in scenarios with limited training data (4-64 samples), where traditional feature engineering approaches typically struggle due to insufficient statistical power for reliable feature selection.

2.4 Empirical Findings and Performance Analysis

2.4.1 Comparative Effectiveness of Methodological Approaches

Extensive empirical evaluation across multiple studies reveals several consistent patterns regarding the relative effectiveness of different LLM-based approaches for feature engineering tasks.

Li et al. [**?**] demonstrate through comprehensive experiments across diverse datasets that text-based feature selection consistently outperforms data-driven methods in scenarios with limited training data. This superiority manifests across multiple critical dimensions that collectively establish the robustness of text-based approaches. Text-based methods consistently achieve higher performance metrics, including Area Under the Receiver Operating Characteristic curve (AUROC) for classification tasks and Mean Absolute Error (MAE) for regression tasks, across different dataset types ranging from medical and financial domains to social and behavioral datasets. The stability advantage is evident in the significantly lower variance in performance across different data availability settings, indicating that text-based approaches maintain consistent quality regardless of the specific sample size or data distribution characteristics. Finally, the robustness of text-based methods is demonstrated through their consistent performance regardless of specific dataset characteristics such as dimensionality, class balance, or feature correlation structures, making them more reliable for practical deployment across diverse application domains.

The authors report that text-based approaches using GPT-4 achieve performance comparable to traditional methods like Minimum Redundancy Maximum Relevance (MRMR) selection and Recursive Feature Elimination (RFE), while requiring no training data.

Both techniques must be introduced. Maybe an introduction to standard feature engineering methodologies can be useful.

A significant finding across multiple studies is the strong correlation between LLM size and feature engineering effectiveness, particularly for text-based approaches. Li et al. observe clear scaling laws where larger models (GPT-4 vs. ChatGPT vs. LLaMA-2) demonstrate progressively better feature selection capabilities. This scaling behavior is less pronounced for data-driven methods, suggesting that semantic understanding rather than pure computational capacity drives effectiveness in feature engineering tasks.

A consistent limitation identified across studies is the degradation of data-driven methods as sample size increases beyond 64-128 samples. Li et al. attribute this phenomenon to LLMs' well-documented struggles with processing long sequences [**?**], which constrains the practical applicability of data-driven feature selection in real-world scenarios with abundant data.

2.4.2 Domain-Specific Applications and Specialization

The literature reveals particular promise for LLM-based feature engineering in specialized domains where traditional statistical methods may miss important semantic relationships.

Medical and biomedical applications represent a particularly promising area for LLM-based feature engineering. Li et al. introduce Retrieval-Augmented Feature Selection (RAFS) specifically designed for medical applications involving high-dimensional genomic data. RAFS addresses the challenge of domain-specific terminology by retrieving metadata from authoritative sources such as the National Center for Biotechnology Information (NCBI). In experiments with The Cancer Genome Atlas (TCGA) Lung Adenocarcinoma dataset, RAFS demonstrates significant improvements over random feature selection while maintaining privacy by avoiding direct data sharing.

The RAFS approach demonstrates measurable improvements across multiple evaluation metrics, achieving an Antolini's Concordance score of 0.6566 compared to 0.6516 for random feature selection, indicating better discrimination capability in survival prediction tasks. The method also shows improvement in the Integrated Brier Score, achieving 0.1830 versus 0.1833 for random selection, suggesting better calibration of predicted survival probabilities over time. Additionally, RAFS achieves superior D-Calibration performance with a score of 1.7666 compared to 2.0255 for random selection, indicating more reliable probability estimates across different risk groups.

These metrics are not common, they should be introduced too.

Han et al. demonstrate particular strength in complex domains requiring domain-specific knowledge, including medical prediction tasks and game-theoretic scenarios. Their framework achieves state-of-the-art performance across 13 diverse datasets, with particularly notable improvements in scenarios where traditional feature engineering struggles due to limited domain expertise availability.

2.5 Critical Perspectives and Limitations

2.5.1 Systematic Biases and Overly Simplistic Features

Küken et al. [?] provide essential critical analysis that tempers overly optimistic assessments of LLM capabilities in feature engineering. Their comprehensive evaluation across 27 datasets using multiple state-of-the-art models (GPT-4o-mini, Gemini-1.5-flash, Llama3.1-8B, Mistral7B-v0.3) reveals several systematic limitations.

LLMs demonstrate a tendency to repeatedly select small subsets of features with disproportionately high frequency, potentially overlooking important but less obvious relationships. This bias manifests as over-reliance on features that align with common patterns in training data rather than dataset-specific optimal features. Furthermore, analysis reveals that LLM-generated features often represent overly simplistic transformations that fail to capture complex, non-linear relationships present in tabular data. This limitation suggests that while LLMs excel at understanding semantic relationships, they may lack the sophisticated statistical reasoning required for optimal feature engineering.

Contrary to more optimistic reports, Küken et al. find that LLM-based methods show marginal or no significant improvement over established automated feature engineering approaches like OpenFE across many datasets. This finding highlights the importance of rigorous evaluation and realistic performance expectations when assessing the practical utility of LLM-based feature engineering approaches.

2.5.2 Computational and Practical Constraints

Several studies identify significant practical limitations that affect the real-world applicability of LLM-based feature engineering.

The iterative nature of most LLM-based approaches requires substantial computational resources, with Hollmann et al. [?] reporting average processing times of approximately 5 minutes per dataset for their CAAFE framework. While this may be acceptable for research contexts, it raises questions about scalability to production environments. Different LLMs exhibit varying levels of effectiveness for feature engineering tasks, with no single model consistently outperforming others across all dataset types. This inconsistency complicates the development of robust, generalizable approaches.

Current LLMs face constraints in processing datasets with large numbers of features due to context length limitations. Hollmann et al. address this by restricting evaluation to datasets with fewer than 20 features to remain within GPT-4's token limits. These practical constraints highlight the need for more efficient approaches that can handle larger datasets while maintaining computational feasibility.

2.6 Advanced Frameworks and Methodological Innovations

Some Figure that represent the workflow and some example prompt can help to understand the key differences of the techniques.

2.6.1 CAAFE: Context-Aware Automated Feature Engineering

Hollmann et al. [?] present the most comprehensive framework to date for LLM-based feature engineering, addressing many limitations identified in earlier work through sophisticated design choices and rigorous validation procedures.

CAAFE incorporates explicit human oversight mechanisms that prevent execution of potentially harmful or biased transformations. This design choice addresses growing concerns about AI bias propagation while maintaining the benefits of automated feature generation. The framework systematically incorporates multiple types of contextual information to enhance the quality and relevance of generated features. Dataset descriptions provide essential domain context that helps the LLM understand the underlying data generation process and the specific characteristics of the target domain. Feature metadata offers semantic meaning by explaining the conceptual significance of each variable, enabling the model to make more informed decisions about appropriate transformations. Sample data illustrates distributional characteristics, providing the LLM with concrete examples of data patterns and ranges that inform realistic feature generation strategies. Additionally, performance feedback from previous iterations creates a learning loop where the system can adapt its feature generation approach based on empirical validation results, progressively improving the quality of generated features through iterative refinement.

Unlike approaches that directly apply LLM-generated features, CAAFE employs rigorous cross-validation to ensure that only features providing genuine predictive improvement are retained. This validation mechanism provides protection against overfitting and spurious feature generation. Furthermore, CAAFE generates human-readable Python code for feature transformations, enabling manual inspection and modification. This transparency addresses interpretability concerns while allowing domain experts to understand and potentially refine generated features.

Empirical evaluation demonstrates that CAAFE achieves meaningful performance improvements, with average ROC AUC increasing from 0.798 to 0.822 across evaluated datasets when using TabPFN as the downstream classifier.

2.6.2 Bias Mitigation and Ethical Considerations

The literature increasingly recognizes the critical importance of addressing bias propagation and ethical concerns in LLM-based feature engineering systems.

Hollmann et al. identify three primary levels where bias can manifest in LLM-based feature engineering systems. At the model level, biases originate from the training data and model parameters, reflecting historical prejudices and discriminatory patterns present in the large-scale text corpora used to train these language models. At the feature generation level, biases are introduced through feature selection and transformation processes, where the model may systematically favor certain types of transformations or patterns that encode societal stereotypes or discriminatory associations. Finally, at the downstream classifier level, biases affect final model predictions through the interaction between potentially biased features and the classification algorithm, amplifying discriminatory signals that may have been subtle or implicit in the original feature space.

2.6.3 Hybrid Approaches: Combining LLM Reasoning with Traditional Methods

A promising research direction emerges from hybrid approaches that seek to combine the semantic reasoning capabilities of Large Language Models with the statistical rigor of traditional data-driven feature selection methods. Li and Xiu [**?**] introduce the LLM4FS framework, which represents a significant methodological innovation in this space.

LLM4FS operates by providing approximately 200 data samples (typically representing 20% or less of the total dataset) directly to LLMs, instructing them to apply traditional data-driven techniques such as random forest, forward sequential selection, backward sequential selection, recursive feature elimination (RFE), minimum redundancy maximum relevance (MRMR), and mutual information (MI). This approach leverages the contextual understanding capabilities of LLMs while maintaining the statistical reliability of established feature selection methods.

The hybrid framework addresses fundamental limitations of pure LLM-based approaches by incorporating the robustness of traditional methods while overcoming their semantic limitations. Unlike purely data-driven LLM approaches that struggle with long sequences, or purely text-based approaches that may lack statistical grounding, the hybrid method provides a principled way to combine complementary strengths from both paradigms.

Comprehensive evaluation across multiple datasets (Bank, Credit-G, Pima Indians Diabetes, Give Me Some Credit) demonstrates that LLM4FS achieves superior performance compared to both standalone LLM-based methods and traditional approaches when evaluated individually. The framework shows particular effectiveness when using state-of-the-art models like DeepSeek-R1, which demonstrates performance comparable to GPT-4.5 while offering significantly better cost-efficiency.

The authors report that DeepSeek-R1 exhibits consistently strong performance across all evaluated datasets, with output costs approximately 50% of GPT-o3-mini and only 1.5% of GPT-4.5. This cost-effectiveness, combined with robust performance, makes the hybrid approach particularly attractive for practical deployment scenarios where computational budget constraints are significant.

A key advantage of the hybrid approach is its demonstrated stability in feature selection across different data availability scenarios. Unlike pure LLM-based methods that may exhibit inconsistent performance across different sample sizes or dataset characteristics, LLM4FS maintains reliable performance through the incorporation of established statistical methods. The framework shows particular stability in the 10%-30% feature selection range, which is often the most practically relevant scenario for real-world applications.

The hybrid approach achieves an advantageous trade-off between computational efficiency and performance quality. While pure LLM-based methods may require extensive iterative prompting and traditional methods may require full dataset analysis, LLM4FS achieves competitive performance with reduced computational overhead by leveraging the reasoning capabilities of LLMs to guide the application of efficient traditional methods on smaller data samples.

Several complementary approaches are proposed to address these multiple sources of bias, each targeting different aspects of the feature generation and validation pipeline. Explicit bias detection operates through systematic analysis of generated feature explanations, requiring the LLM to provide detailed justifications for why specific features are considered important, thereby making potentially biased reasoning visible to human reviewers. Human oversight mechanisms establish mandatory checkpoints requiring manual approval of generated transformations before they can be applied to the dataset, ensuring that domain experts can identify and reject features that may encode inappropriate biases or discriminatory patterns. Cross-validation filtering implements automated screening processes that systematically discard features leading to biased outcomes across different demographic groups, using fairness metrics to identify features that disproportionately impact protected classes. Finally, transparency requirements mandate that all feature generation processes produce interpretable, human-readable code and explanations, enabling post-hoc analysis and audit trails that support accountability and bias detection in deployed systems.

The authors provide a concrete example using a synthetic dataset predicting doctor vs. nurse classification based on names. Their analysis reveals how GPT-4 generates features that rely on gender-associated name endings, demonstrating how societal biases can be inadvertently encoded in automated feature engineering systems.

Chapter 3

Methodology

Chapter 4

Experimental Setup

Chapter 5

Results and Analysis

Chapter 6

Conclusions and Future Work