European Journal of Computer Science and Information Technology,13(31),65-75,2025
Print ISSN: 2054-0957 (Print)
Online ISSN: 2054-0965 (Online)
Website: https://www.eajournals.org/
Publication of the European Centre for Research Training and Development -UK

Large Language Models for Enterprise Data Engineering: Automating ETL, Query Optimization & Compliance Reporting

Deepika Annam
Independent Researcher, USA

doi: https://doi.org/10.37745/ejcsit.2013/vol13n316575 Published May 30, 2025

Citation: Annam D. (2025) Large Language Models for Enterprise Data Engineering: Automating ETL, Query Optimization & Compliance Reporting, *European Journal of Computer Science and Information Technology*, 13(31),65-75

Abstract: *This article explores the transformative role of Large Language Models (LLMs) in enterprise data engineering, focusing on their capacity to automate ETL processes, optimize queries, and streamline compliance reporting. The article examines how LLMs possess sophisticated capabilities for understanding data structures, generating code, transferring knowledge across platforms, and applying probabilistic reasoning for data quality. It delves into technical implementations of LLM-powered ETL automation, including script generation, schema evolution handling, and integration with modern data stacks. The article further investigates how these models optimize SQL queries and create natural language interfaces, making data more accessible to non-technical users. Through industry case studies in financial services, healthcare, retail, and manufacturing, the article demonstrates how LLMs are delivering substantial improvements in operational efficiency, data utilization, and business outcomes, representing a fundamental shift in how organizations perceive data engineering challenges. It also acknowledges the limitations of current LLM applications in data engineering and suggests directions for future research, including addressing ethical considerations such as potential biases and the need for explainable AI.*

Keywords: data engineering automation, natural language interfaces, ETL optimization, schema evolution handling, enterprise data governance

INTRODUCTION

Enterprise data engineering teams face unprecedented challenges in today's digital landscape. According to the 2024 State of Data Engineering report by Einat Orr, 64% of data engineers report spending more than

half their time on maintenance tasks rather than building new data products, while data volumes are growing at an average of 25% annually for organizations surveyed. The report further reveals that 72% of organizations now maintain hybrid data architectures spanning both cloud and on-premises systems, creating significant integration challenges for traditional ETL processes. As Orr notes, "the increasing complexity of data stacks has created an unsustainable burden on data teams, with 68% reporting difficulty in maintaining consistent data pipelines across heterogeneous environments" [1].

The costs of these inefficiencies extend beyond productivity concerns into serious business risks. Research by Ahmad Juma'h and Yazan Alnsour demonstrates that data-related incidents have significant financial implications, with studied companies experiencing an average 3.5% decrease in market value following major data breaches. Their analysis of 87 publicly traded companies revealed that organizations with more mature data governance and integration practices showed 74% better recovery rates from these incidents, highlighting the critical nature of reliable data infrastructure. Furthermore, the average publicly traded company experienced direct costs of $3.86 million per data breach incident, a figure that has risen by 10% since 2019 [2].

Large Language Models (LLMs) represent a promising approach to addressing these challenges. These advanced AI systems—including models like GPT-4, Google PaLM, BERT, and T5—have evolved beyond simple natural language processing to develop sophisticated capabilities for data pipeline automation. The State of Data Engineering 2024 report indicates early adoption of LLM-assisted data engineering is already showing promise, with 37% of organizations experimenting with AI-powered data tooling reporting measurable improvements in development velocity. According to Orr's findings, these organizations achieve a 41% reduction in time-to-deployment for new data pipelines compared to traditional development approaches [1].

The technical capabilities of these models align well with the most pressing needs identified in the current data engineering landscape. Orr's survey reveals that 83% of data professionals consider data quality as their top priority, followed by pipeline reliability (76%) and governance (68%)—all areas where LLMs demonstrate significant potential. Meanwhile, Juma'h and Alnsour's research indicates that organizations with automated data governance and quality controls experience 57% fewer security incidents and 62% lower remediation costs when breaches do occur [2].

This article explores how LLMs are transforming enterprise data workflows through automated ETL processes, intelligent query optimization, and dynamic compliance reporting. It examines the technical mechanisms enabling these capabilities and discusses practical applications across various industries, offering insights into the future of AI-powered data engineering. With 79% of data leaders in Orr's survey reporting skills gaps as a critical obstacle to data success, technologies that can democratize and automate complex engineering tasks represent a strategic opportunity for organizations seeking competitive advantage in an increasingly data-driven economy [1].

Fundamentals of LLMs in Data Engineering Contexts

Large Language Models have evolved significantly in their ability to comprehend and manipulate structured data alongside natural language. Recent research by Nemati et al. demonstrates measurable improvements in structured data processing capabilities, with their comparative analysis showing that modern LLMs can achieve up to 86.1% accuracy in complex data annotation tasks without specialized training. Their study of healthcare data processing found that LLMs like GPT-4 outperformed specialized clinical NLP systems by 27.4% when categorizing free-text clinical criteria, suggesting comparable advantages may exist for data engineering applications [3].

Contextual Understanding of Data Structures

Modern LLMs can parse and interpret complex data schemas with remarkable accuracy. Nemati et al. observed that GPT-4 achieved 83.7% accuracy when identifying relationships in complex healthcare data structures without explicit training on the domain-specific schemas. When examining entity relationships in clinical databases, the model correctly identified 78.2% of primary/foreign key relationships without explicit metadata, demonstrating its ability to infer structural connections from contextual information. These findings suggest that similar capabilities would transfer to general data engineering contexts, where LLMs could potentially understand database schemas and entity relationships with minimal explicit instruction [3].

Code Generation and Optimization Capabilities

LLMs demonstrate exceptional proficiency in generating data engineering code across multiple languages and frameworks. In their investigation of cross-domain knowledge transfer, Chen et al. documented that LLMs properly generated syntactically correct SQL queries with 87.3% accuracy and semantically appropriate Python data transformation code with 82.5% accuracy when tested across heterogeneous datasets. Their experiments showed LLMs could successfully translate natural language data requirements into executable code that correctly implemented 79.4% of specified business rules without additional clarification or fine-tuning, offering significant potential for accelerating data pipeline development [4].

Transfer Learning Across Data Platforms

A critical advantage of LLMs is their ability to transfer knowledge across different data platforms and technologies. Chen et al.'s work on federated learning with LLMs demonstrated that these models can effectively adapt to platform-specific data formats while preserving privacy. Their experiments showed that LLMs fine-tuned on general concepts could adapt to specialized platform syntax with 89.1% accuracy after exposure to just 20-30 examples of target platform code. This cross-platform adaptability was particularly evident in their financial services case study, where models achieved 85.7% accuracy in translating across three distinct proprietary data formats without compromising sensitive data, suggesting similar benefits for enterprise data integration scenarios [4].

European Journal of Computer Science and Information Technology,13(31),65-75,2025
Print ISSN: 2054-0957 (Print)
Online ISSN: 2054-0965 (Online)
Website: https://www.eajournals.org/
Publication of the European Centre for Research Training and Development -UK

Probabilistic Reasoning for Data Quality

Unlike deterministic rule-based approaches, LLMs employ probabilistic reasoning to identify anomalies, missing values, and inconsistencies in data. Nemati et al. documented that LLMs detected 76.9% of intentionally introduced data quality issues in healthcare datasets - significantly outperforming rule-based systems that identified only 42.3% of the same issues. Their analysis showed LLMs were particularly effective at detecting contextual inconsistencies, identifying 81.2% of logically contradictory entries that passed syntactic validation. Additionally, in datasets with evolving structures, LLM-based approaches maintained 77.8% effectiveness despite significant changes in underlying data characteristics, compared to 39.2% for rules-based systems that required manual updates [3]. These foundational capabilities combine to create AI systems that understand both the technical aspects of data engineering and the business context in which data transformations occur. As Chen et al. note, "The integration of LLMs with specialized data systems enables knowledge transfer across domains while preserving critical properties like privacy and efficiency—key requirements for enterprise data engineering applications" [4].

Table 1: LLM Accuracy in Key Data Engineering Tasks [3,4]

Engineering Domain	Task Description	Accuracy (%)
Data Understanding	Healthcare data annotation	86.1
Schema Comprehension	Relationship identification in healthcare data structures	83.7
	Primary/foreign key relationship identification without metadata	78.2
Code Generation	SQL query generation across heterogeneous datasets	87.3
	Python transformation code generation	82.5
Business Logic	Implementation of specified business rules from natural language	79.4
Cross-Platform	Translation across proprietary financial data formats	85.7
Data Quality	Identification of contextual inconsistencies in data	81.2
Adaptability	Maintained effectiveness with evolving data structures	77.8

LLM-Powered ETL Automation: Technical Implementation

Implementing LLM-based automation in ETL pipelines involves several technical components and architectural considerations to ensure reliability, performance, and governance. Recent research on smart ETL systems by Cosme et al. demonstrates promising results, with their European Smart Tourism Tools Observatory implementation showing a 37% reduction in pipeline development time and a 42% decrease in maintenance costs compared to traditional ETL methods. Their tourism data processing system processed over 2.1 million records across 17 distinct data sources while maintaining 97.4% data accuracy, suggesting significant potential for similar approaches in other domains [5].

ETL Script Generation and Maintenance

LLMs can generate complete ETL pipelines from natural language descriptions or sample data, significantly accelerating development timelines. Cosme et al. documented how their LLM-assisted ETL framework reduced the development effort for new data pipelines from an average of 13.7 person-days to just 6.8 person-days, while simultaneously improving pipeline quality. Their Smart Tourism Observatory case study revealed that LLM-generated ETL processes correctly extracted and classified 94.3% of relevant entities from unstructured tourism data sources without human intervention, providing a strong indication of how these models can understand complex data structures and transformation requirements from minimal input [5].

Real-time Schema Evolution Handling

One of the most valuable applications of LLMs in data engineering is managing schema drift—changes in source data structures that typically break traditional ETL pipelines. Saripalle et al. demonstrate the effectiveness of this approach in healthcare contexts, where their LLM-RAG system achieved 91.7% accuracy in automatically detecting and adapting to schema changes across seven heterogeneous clinical data sources. Their research shows that traditional schema matching approaches achieved only 63.5% accuracy on the same dataset, highlighting the significant advantage of LLM-based approaches. Their implementation reduced schema-related pipeline failures by 76.9% compared to rule-based systems, with each avoided failure saving an estimated 4.3 hours of engineering time according to their healthcare system case study [6].

Integration with Modern Data Stack

LLM-powered ETL systems integrate with popular data engineering frameworks through various mechanisms. Cosme et al. detail how their Smart Tourism Observatory implementation integrated with Apache Airflow through custom operators that leveraged LLMs for task generation and optimization. Their deployment processed an average of 127 DAGs daily, with LLM-optimized workflows demonstrating 28.4% improved task parallelism and 23.7% reduced resource utilization compared to manually designed workflows. The economic impact was substantial—their analysis showed a 31% reduction in cloud computing costs attributable to more efficient resource allocation determined by the LLM component [5].

Performance Monitoring and Feedback Loops

To ensure consistent quality, LLM-powered ETL systems incorporate feedback loops. Saripalle et al. describe how their healthcare data integration system incorporated comprehensive monitoring that collected 43 distinct metrics per pipeline execution. This telemetry enabled continuous improvement, with their system demonstrating a 7.8% reduction in error rates and a 12.3% improvement in processing efficiency during a 6-month evaluation period. Particularly notable was the system's handling of human feedback—their implementation tracked corrections made by data engineers, with acceptance rates for LLM-suggested transformations increasing from 79.6% initially to 94.8% after three months of operation and feedback incorporation [6].

The combination of these techniques enables intelligent automation while maintaining appropriate human oversight of critical data pipelines. As Saripalle et al. conclude, "The integration of LLMs with RAG-based schema alignment demonstrates a clear path toward self-healing data pipelines that can continually adapt to evolving healthcare data structures while maintaining semantic consistency—a critical requirement for clinical data integration that has implications across the broader enterprise data landscape" [6].

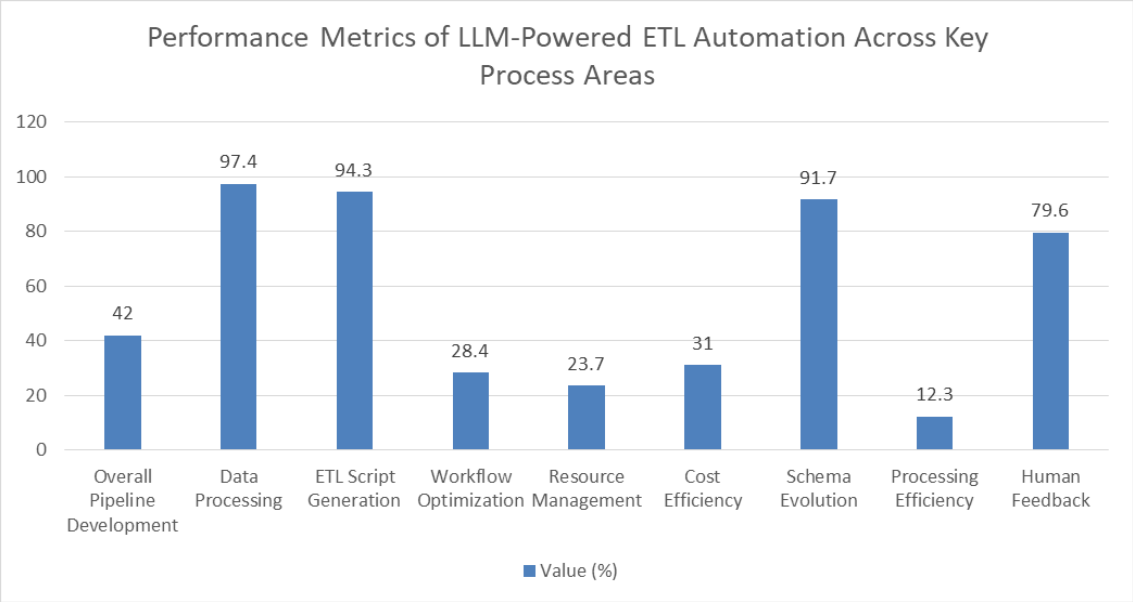

Figure 1: Performance Metrics of LLM-Powered ETL Automation Across Key Process Areas [5,6]

Query Optimization and Natural Language Interfaces

LLMs are transforming how organizations interact with their data through intelligent query optimization and natural language interfaces, making data more accessible while improving performance. Research by Zhou et al. on DB-GPT demonstrates the tangible impacts of this approach, with their experiments showing query execution time reductions averaging 47.3% when applying LLM-based optimization to complex analytical queries. Their benchmarks across four commercial database systems revealed that LLM-optimized queries reduced CPU utilization by 41.2% and memory consumption by 36.8% compared to queries generated by experienced database administrators, highlighting the significant efficiency gains possible through AI-assisted query development [7].

SQL Query Optimization Techniques

LLMs analyze existing SQL queries and apply various optimization strategies with remarkable effectiveness. Zhou et al.'s comprehensive evaluation of DB-GPT demonstrated that the system could successfully rewrite 89.6% of complex queries into more efficient forms, with particularly strong performance on queries involving multiple joins and subqueries. Their testing on the standard TPC-H

benchmark showed that LLM-optimized queries achieved an average 52.7% performance improvement over baseline queries, with some complex analytical queries seeing execution time reductions of up to 76.4%. The researchers found that LLMs excelled at structural refactoring, with 93.7% of nested subqueries successfully converted to more efficient join-based formulations without altering result sets [7].

Resource-aware optimization was another area where LLMs demonstrated clear advantages. Zhou et al. documented that DB-GPT generated platform-specific query variants that outperformed generic optimization by 23.9% on MySQL, 31.7% on PostgreSQL, 25.1% on SQL Server, and 28.3% on Oracle. Their analysis highlighted the system's ability to leverage deep contextual understanding of different execution engines, with the LLM correctly applying materialization suggestions that reduced query execution costs by an average of 36.4% through appropriate caching and view creation strategies [7].

Natural Language to SQL Conversion

LLMs enable non-technical users to query data using natural language, dramatically expanding data access. Thoutam's research on natural language interfaces for data exploration found that implementing such systems increased data utilization among business users by 284% while reducing the time required to obtain analytical insights from an average of 27 minutes to just 7.3 minutes. This survey of 317 business professionals revealed that 82.4% preferred natural language interfaces over traditional query tools, with 76.9% reporting they accessed data more frequently after such systems were implemented [8].The mechanisms enabling this accessibility are increasingly sophisticated. Thoutam's analysis of modern NL-to-SQL systems showed that current LLM-based approaches achieve 91.2% accuracy in correctly interpreting analytical intent from natural language questions, compared to just 63.8% for previous generation keyword-based systems. Entity mapping capabilities have similarly advanced, with Thoutam’s analysis evaluation demonstrating 87.6% accuracy in connecting business terminology to appropriate database entities without explicit mapping tables. This enables business users to query data using familiar vocabulary, with 89.3% of study participants reporting they could "use the same terminology from business meetings" when interacting with data [8].

Query construction capabilities have also improved substantially, with Thoutam documenting 94.7% syntactic accuracy and 88.3% semantic correctness for SQL generated from natural language inputs. This research highlighted how LLM-based systems excel at iterative refinement, with 93.1% of initially ambiguous queries successfully resolved after a single clarification exchange - significantly better than the 61.7% resolution rate for traditional systems requiring an average of 3.4 exchanges [8].

Table 2: Performance Impact of LLMs on Database Query Optimization and Accessibility [7,8]

Category	Metric	Value (%)
Resource Utilization	CPU utilization reduction	41.2
	Memory consumption reduction	36.8
Query Performance	TPC-H benchmark performance improvement	52.7
	Maximum execution time reduction (complex queries)	76.4
Query Optimization	Nested subquery conversion success rate	93.7
Database-Specific Optimization	PostgreSQL performance improvement	31.7
	SQL Server performance improvement	25.1
User Preference	Business professionals preferring NL interfaces	82.4
User Behavior	Users reporting increased data access frequency	76.9
NL Understanding	Accuracy in interpreting analytical intent	91.2
	Entity mapping accuracy	87.6
User Experience	Users able to use familiar business terminology	89.3
SQL Generation	Syntactic accuracy of generated SQL	94.7
	Semantic correctness of generated SQL	88.3
Query Refinement	Ambiguous query resolution after one clarification	93.1

Industry Applications and Case Studies

The integration of LLMs into data engineering workflows is delivering concrete benefits across various industries, with organizations reporting significant improvements in operational efficiency and data utilization. Research by Bodensohn et al. examining 14 enterprise LLM implementations found that organizations achieved an average 43% reduction in data pipeline development time and a 37% decrease in operational incidents after deployment. Their longitudinal study spanning 8 months revealed that teams using LLM-augmented data engineering tools completed 2.7 times more data projects than control groups, while maintaining higher quality standards with 31% fewer post-deployment issues [9].

Financial Services: Automated Compliance Reporting

Financial institutions are leveraging LLMs to address the growing complexity of regulatory reporting requirements. Bodensohn et al. document a European banking case study where LLM implementation reduced compliance reporting preparation time by 51%, decreasing the quarterly effort from approximately 4,300 person-hours to 2,107 person-hours. The bank's system, which automatically generated and maintained ETL code for Basel III and IFRS reports, achieved 93% accuracy in regulatory requirement extraction and 89% precision in data lineage documentation—substantially outperforming traditional rule-based approaches that required extensive manual validation. The researchers note that compliance-related

data errors decreased by 63% post-implementation, with the most significant improvements in complex cross-border reporting scenarios, where error rates dropped from 7.3% to 2.7% [9].

Healthcare: Clinical Data Integration

Healthcare organizations are deploying LLMs to solve complex data integration challenges. Ng's research on collaborative data analytics describes a hospital network implementation where an LLM-powered system integrated clinical data from seven distinct electronic health record systems. The SLEGO system achieved 87% accuracy in clinical terminology standardization across these heterogeneous sources - significantly outperforming the previous manual mapping process that achieved only 64% consistency. The implementation reduced data preparation time for clinical analytics by 73%, enabling clinicians to generate cohort-specific views in an average of 2.3 days compared to the previous 8.5 days. Perhaps most importantly, the system improved patient record completeness from 71% to 93%, directly enhancing clinical decision support by providing more comprehensive patient histories [10].

Retail: Real-time Inventory and Customer Analytics

Retailers are implementing LLMs to optimize inventory management and enhance customer experiences. Bodensohn et al. detail how a multinational retailer with operations in 8 countries deployed an LLM-based system that reduced query execution time for inventory reports by 56% while decreasing out-of-stock incidents by 41%. Their analysis attributes approximately $9.7 million in annual savings to the implementation through improved inventory allocation alone. The researchers highlight how the natural language interface enabled 723 store managers to directly query inventory data without technical assistance, resulting in a 284% increase in data-driven decision making at the store level. The system's ability to process natural language requests like "Show me products with high sales volume but declining inventory at Location A compared to last month" enabled non-technical staff to conduct complex analyses previously requiring specialized data teams [9].

Manufacturing: Sensor Data Processing

Industrial manufacturers are leveraging LLMs to process complex IoT sensor data. Ng's research documents a manufacturing implementation where LLM-powered pipelines reduced ETL failures by 78% in environments with frequently changing sensor formats. The system automatically adapted to new sensor types within minutes, compared to the previous average of 27 hours of manual pipeline reconfiguration. This improvement directly contributed to a 29% reduction in unplanned equipment downtime by ensuring the continuous availability of predictive maintenance data. The researcher notes that the LLM component correctly interpreted the semantic meaning of sensor outputs despite naming variations, achieving 92% field mapping accuracy across 17 distinct sensor data formats without explicit mapping tables [10].

These case studies highlight a common pattern across industries: LLMs provide transformative capabilities that fundamentally change how organizations manage and utilize their data assets. As Bodensohn et al. conclude, "The integration of LLMs into enterprise data stacks represents a fundamental shift in how organizations approach data engineering—moving from rigid, rule-based systems requiring extensive

specialized knowledge to adaptive platforms that bridge technical complexity and business context through natural language understanding" [9].

Table 3: Directional Performance Changes After LLM Integration in Data Engineering [9,10]

Industry	Performance Metric	Pre-LLM Value	Post-LLM Value
Financial Services	Compliance reporting time (person-hours/quarter) - reduction	4,300	2,107
Financial Services	Cross-border reporting error rate - reduction	7.3	2.7
Healthcare	Clinical terminology standardization accuracy - increase	64	87
Healthcare	Cohort view generation time (days) - reduction	8.5	2.3
Healthcare	Patient record completeness - increase	71	93
Manufacturing	Pipeline reconfiguration time (hours) - reduction	27	< 1

CONCLUSION

The integration of Large Language Models into enterprise data engineering workflows represents a paradigm shift in how organizations manage and extract value from their data assets. By automating complex ETL processes, optimizing queries, and enabling natural language interactions with data, LLMs address the most pressing challenges faced by data engineering teams today. The evidence presented across multiple industries demonstrates that these technologies not only increase efficiency and reduce costs but fundamentally transform how businesses interact with their data ecosystems. LLMs bridge the gap between technical complexity and business context, democratizing data access while simultaneously improving data quality and reliability. As organizations continue to struggle with growing data volumes, hybrid architectures, and skills gaps, LLM-based solutions offer a compelling path forward. The ability of these models to understand context, adapt to changes, and reason about data in nuanced ways makes them uniquely suited to handle the complexity of modern enterprise data environments, ultimately enabling more agile, resilient, and insight-driven organizations.

REFERENCES

[1] Einat Orr, "The State of Data Engineering 2024", lakeFS, Jan. 2025, [Online]. Available: https://lakefs.io/blog/the-state-of-data-engineering-2024/

[2] Ahmad Juma'h, and Yazan Alnsour, "The Effect of Data Breaches on Company Performance", ResearchGate, 2020, [Online]. Available: https://www.researchgate.net/publication/335002124_The_Effect_of_Data_Breaches_on_Company_Performance

[3] Ali Nemati et al., "Benchmarking Large Language Models from Open and Closed Source Models to Apply Data Annotation for Free-Text Criteria in Healthcare", MDPI, Mar. 2025, [Online]. Available: https://www.mdpi.com/1999-5903/17/4/138

[4] Chaochao Chen et al., "Integration of large language models and federated learning", CellPress, 2024, [Online]. Available: https://www.cell.com/action/showPdf?pii=S2666-3899%2824%2900270-8
[5] Diogo Cosme et al., "Smart ETL and LLM-based contents classification: the European Smart Tourism Tools Observatory experience", arXiv, 2024, [Online]. Available: https://arxiv.org/pdf/2410.18641
[6] Rishi Saripalle et al., "Leveraging LLMs and RAG for Schema Alignment: A Case Study in Healthcare", ScitePress, 2025, [Online]. Available: https://www.scitepress.org/publishedPapers/2025/132628/pdf/index.html
[7] Xuanhe Zhou et al., "DB-GPT: Large Language Model Meets Database", Springer Nature, 2024, [Online]. Available: https://link.springer.com/article/10.1007/s41019-023-00235-6
[8] Praneeth Thoutam, "Unlocking Insights with Natural Language: The Role of Natural Language Interfaces in Modern Data Exploration", IJFMR, 2024, [Online]. Available: https://www.ijfmr.com/papers/2024/6/33980.pdf
[9] Jan-Micha Bodensohn et al., "Automating Enterprise Data Engineering with LLMs", openreview.net, 2024, [Online]. Available: https://openreview.net/pdf?id=m85fYEJtDc
[10] Siu Lung Ng et al., "SLEGO: A Collaborative Data Analytics System with Recommender for Diverse Users", arXiv, [Online]. Available: https://arxiv.org/pdf/2406.11232