

University of Bahrain
College of Information Technology



Leveraging Large Language Models for Trend Detection in Arabic Text

A Thesis Submitted in Partial Fulfilment of the Requirements for the
Degree of M.Sc. in Machine Learning and Computational Intelligence

Submitted By:

Student's Name: Fatema Hasan Al-Hassabi

ID Number: 20112426

Supervised By:

Dr. Hesham Al-Ammal

Dr. Fatema Al-Balooshi

(Assistant Professor)

(Assistant Professor)

University of Bahrain

Kingdom of Bahrain

April 10, 2025

DECLARATION

Acknowledgements

I would like to express my heartfelt gratitude to my supervisors, Dr. Hesham Al-Ammal and Dr. Fatema Al-Balooshi, for their invaluable guidance and support throughout this journey. Dr. Hesham, your mentorship has been truly transformative — not only inspiring me but also pushing me to think big and expand my horizons. From my senior project in 2016 to this thesis, your unwavering belief in me has been a constant source of motivation. You have consistently encouraged me to work with real data, tackle meaningful problems, and focus on creating value. Your guidance has shaped not only this thesis but also the way I approach challenges and opportunities. Dr. Fatema, thank you for your thoughtful feedback, insightful guidance, and steadfast support, which have enriched this research in countless ways.

Special thanks to Mr. Abdulla Subah for his dedicated support in managing UoB’s Benefit Advanced AI and Computing Lab. Your expertise with the cluster and setting up the LLM was crucial to this research. I also extend my gratitude to everyone involved in establishing the lab and making such AI experiments possible.

To my team, “#TeamBahrain”, I truly appreciate your trust in this idea, which was driven by the Government Services Quality Team’s dedication to better serve our community. A special thanks to Mr. Zeyad Darwish and Mr. Hussain Al-Shehab for their efforts in supporting me with the necessary resources and time to complete this project. Your understanding and help were instrumental, and for that, I am very grateful.

To my family — my mother and father who taught me the value of knowledge. Thank you for encouraging me to learn and urging me to never stop learning. Your belief in me has been my greatest strength. To my sisters, Huda and Heba—thank you for your unconditional support, encouragement, and patience.

To my dear friends, Abrar, Zahraa, and Fatima, I am grateful for your constant encouragement and for always being there through every step of this journey.

This thesis would not have been possible without the contributions, support, and encouragement of everyone mentioned above. Thank you all for being such an amazing source of support.

Abstract

The Tawasul platform serves as Bahrain’s centralized mechanism for residents to submit complaints and inquiries to government entities. This research focuses on analyzing approximately 45,000 complaints related to the Ministry of Education over five years. The dataset presents significant challenges, including linguistic diversity: primarily Arabic but also English, making it bilingual in nature. However, the performance of LLMs in this research is specifically evaluated for Arabic, as it is the dominant language in the dataset, while ensuring the models maintain the ability to handle both languages.

Additionally, as a government dataset, Tawasul’s data is highly sensitive and restricted from public access, creating challenges in utilizing public cloud-based LLMs and infrastructures. A key objective of this research is to explore privacy-preserving approaches by running LLMs on local infrastructure, ensuring compliance with data protection policies while maintaining model efficiency.

To address these challenges, we evaluated two complaint analysis methods: (1) LLaMA 2 (7B) was used for feature extraction, followed by Long Short-Term Memory (LSTM) classification, achieving an accuracy of 0.496. (2) LLaMA 3.1 (8B) with Retrieval-Augmented Generation (RAG) was employed for direct classification, significantly improving performance, with an accuracy of 0.78.

An iterative refinement of LLM prompts was implemented to extract high-quality summaries, keywords, and complaint categories. Additionally, entity recognition techniques were applied to identify school names, and advanced visualizations were developed to gain insights into complaint trends.

Our findings reveal that Method (2), LLaMA 3.1 with RAG, significantly outperformed Method (1), achieving higher accuracy, precision, and efficiency in classifying complaints and extracting relevant insights. The integration of RAG enhanced contextual alignment, reducing reliance on inconsistent human labels. Moreover, prompt engineering techniques played a crucial role in minimizing hallucinations and improving overall output quality. This result compares to human evaluation accuracy of 0.72, showing that the model performs relatively well, especially considering that human classification often includes errors.

This research makes several key contributions: (1) Demonstrating the effectiveness of LLMs in handling complex, multilingual complaint data, particularly for Arabic, using LLaMA 3.1 with RAG, achieving high accuracy (0.91) and precision. (2) Developing refined prompt engineering techniques that reduce hallucinations and enhance the reliability of generated outputs. (3) Employing a hybrid entity recognition approach, leveraging LLM-based ex-

traction and traditional matching methods to ensure accuracy. (4) Integrating advanced visualizations to provide actionable insights, enabling government stakeholders to identify trends, allocate resources, and improve public service delivery.

By showcasing the potential of LLMs in government applications, this study highlights their role in enhancing citizen satisfaction through more responsive and insightful complaint management systems, while addressing privacy concerns through localized model deployment.

Contents

DECLARATION	i
ACKNOWLEDGEMENTS	i
ABSTRACT	ii
TABLE OF CONTENTS	iv
LIST OF TABLES	vii
LIST OF FIGURES	viii
LIST OF LISTINGS	x
ACRONYMS	xi
1 Introduction	1
1.1 Background	1
1.2 Research Problem	2
1.3 Motivation	3
1.4 Objectives	4
1.5 Significance of the Research	4
1.6 Thesis Structure	5
2 Literature Review	6
2.1 Challenges in NLP for Short Text	6
2.2 Methods for Short Text Processing	7
2.3 LLMs for Trend Detection and RAG-Based Enhancements	20
3 Methodology	22
3.1 Data Collection and Preprocessing	23
3.2 LLM Installation and Setup	25
3.2.1 Initial Deployment with LLaMA 2	25
3.2.2 Transition to LLaMA 3.1	26

3.2.3	Comparison Between LLaMA and OLLaMA	27
3.2.4	LLaMA 3: Architecture and Hyperparameters	27
3.3	Retrieval-Augmented Generation (RAG)	28
3.3.1	RAG Architecture with LLaMA-3.1	28
3.4	Complaint Classification Approaches	30
3.4.1	Method 1: Classification using LLaMA 2 with LSTM	31
3.4.2	Method 2: Classification using LLaMA 3.1 with Retrieval-Augmented Generation (RAG)	31
3.5	Prompt Engineering and Task Implementation	33
3.5.1	Summarization and Keyword Extraction	33
3.6	Libraries and LLM Utilization	35
3.6.1	LLaMA 2 Implementation: LangChain and CTransformers	36
3.6.2	LLaMA 3.1 Implementation: Advanced Prompting and Retrieval-Augmented Category Prediction in ioPython	37
3.7	Evaluation Strategy	38
3.8	Visualization and Trend Detection	40
4	Results and Discussion	42
4.1	Data Exploration Findings	42
4.1.1	Dominance of “Other Complaints and Inquiries”	43
4.1.2	Presence of Unclassified or Missing Data (Indicated as “NA”)	43
4.1.3	High-Frequency Categories	43
4.1.4	Insights for Future Categorization and Trend Analysis	44
4.2	Performance Comparison: Method 1 (LLaMA2) vs. Method 2 (LLaMA3.1)	45
4.2.1	Evaluation of Summaries and Keywords	45
4.2.2	Classification Performance	49
4.2.3	Entity Recognition and Matching of Schools	60
4.2.4	Results of Monthly Summary Analysis	63
4.3	Visualization and Analysis	67
4.3.1	Monthly Analysis and Interactive Dashboard	67
4.3.2	Keyword Evolution Over Time	67
4.3.3	Geographic Heatmap of School Complaints	68
4.3.4	Monthly Summary Example	69
4.3.5	Insights and Discussion	69

5	Conclusion and Future Work	74
5.1	Conclusion	74
5.2	Future Work	75
	Bibliography	77

List of Tables

2.1	Classification Results (Normalized) for Different Stemmers and Classifiers (Alzanin et al., 2022)	9
2.2	Precision (Pr), Recall (Rec), F1, and accuracy results for Tweet Topic with Jaccard Index (JI) reported in the multi-label setting (Antypas et al., 2022)	11
2.3	Comparison of Accuracy: Double-Level CNN (DLCNN) versus Other Methods on three Datasets. Methods compared include CNN-Random (CNN-rand), CNN-Static (CNN-static), Character-level CNN (CharCNN), Word-Character CNN (WCCNN), CNN with Horizontal Embedding (CNN-HE), CNN with Vertical Embedding (CNN-VE), and Squeeze-and-Excitation CNN (SECNN) (Tang et al., 2022)	13
3.1	Summary of Research Challenges, Context, and Methodological Solutions . .	23
3.2	Comparison Between LLaMA and OLLaMA.	27
3.3	Key hyperparameters of LLaMA 3. Source: Meta AI Dubey et al. (2024). . .	28
4.1	Comparison of Summary and Keywords Relevance Scores Between LLaMA 3.1 and LLaMA 2	45
4.2	Comparison of LLaMA 3.1 and LLaMA 2 Summaries and Keywords for a Complaint Text -Example 1	46
4.3	Comparison of LLaMA 3.1 and LLaMA 2 Summaries and Keywords for a Complaint Text - Example 2	47
4.4	Comparison of LLaMA 3.1 and LLaMA 2 Summaries and Keywords for a Complaint Text - Example 3	48

List of Figures

2.1	Pipeline for text filtering to reduce noise in tweets (Antypas et al., 2022) . . .	10
2.2	Double-Level CNN (DLCNN) Overview (Tang et al., 2022)	12
2.3	Tree representation of four levels for the topics and subtopics after feeding SECA with three batches, (Abdul Aziz and Starkey, 2020)	15
2.4	Performance comparison between the four methods (KMeans, MStream, NMF, and SECA), from Abdul Aziz and Starkey (2020)	16
2.5	Tweet count timeline; Yellow boxes highlight key events, from Abdul Aziz and Starkey (2020)	17
2.6	Flowchart of LLM Application in Data Annotation Processes,from Tan et al. (2024)	18
3.1	Training process of LLaMA 2, highlighting pretraining and fine-tuning steps. Source: Meta AI (Touvron et al., 2023b).	26
3.2	Architecture of LLaMA 3, emphasizing its transformer-based design. Source: Meta AI Dubey et al. (2024).	28
3.3	RAG architecture integrating retrieval and generation for enhanced outputs. (Lewis et al., 2020).	29
3.4	Comparison of Method 1 using LLaMA2 and Method 2 using LLaMA 3.1 . . .	30
3.5	Prompt Template Using LLaMA2	34
3.6	Prompt Template Using LLaMA3.1	35
3.7	Prompt Template to Extract School Name	36
3.8	Prompt Template for Monthly Summary Analysis - Initial Version	37
3.9	Prompt Template for Monthly Summary Analysis - Refined Version	38
4.1	Distribution of Complaints by Category for Chosen Enitivity	42
4.2	Monthly Analysis Dashboard with category filters and trends.	68
4.3	Monthly Analysis Dashboard filtered by “Test and Exams Category”	69
4.4	Monthly Analysis Dashboard filtered by “Students Transfer” Category	70

4.5 Keywords per day visualization showing dynamic changes over time. 71

4.6 School complaints heatmap showing high-density complaint areas in Isa Town. 72

4.7 Monthly summary example showing trends, keywords, and summaries for
September 2023. 73

LIST OF LISTINGS

Acronyms

ARL	Arabic Light
BERT	Bidirectional Encoder Representations from Transformers
Bi-LSTM	Bidirectional Long Short-Term Memory
CNN	Convolutional Neural Network
DL	Deep Learning
DLCNN	Double-Level Convolutional Neural Network
GCN	Graph Convolutional Network
GNB	Gaussian Naive Bayes
GQA	Grouped Query Attention
HKT	Hierarchical Knowledge Tree
HPC	High Performance Computing
ISRI	Information Science Research Institute's
JI	Jaccard Index
LLM	Large Language Model
LLaMA	Large Language Model Meta AI
LSTM	Long Short-Term Memory
ML	Machine Learning
MR	Movie Reviews
MSA	Modern Standard Arabic
MTLD	Measure of Textual Lexical Diversity
NLP	Natural Language Processing

NLTK	Natural Language Toolkit
NMF	Non-negative Matrix Factorization
RAG	Retrieval-Augmented Generation
RB	RoBERTa
RF	Random Forest
SECA	Self-Evolving Contextual Analysis
SVM	Support Vector Machine
TF-IDF	Term Frequency-inverse Document Frequency
TREC	Text REtrieval Conference

Chapter 1

Introduction

1.1 Background

Effective management and analysis of citizen complaints are essential for improving public services, identifying systemic issues, and optimizing resource allocation. The Tawasul system, implemented in Bahrain, serves as a centralized complaint platform where citizens and residents report concerns to government entities (Information and eGovernment Authority, 2017). Over time, this system has accumulated a large volume of unstructured complaint data across more than 50 government authorities, making it challenging to extract meaningful insights using manual review or traditional processing methods. Additionally, as a government dataset, Tawasul's data is highly sensitive and restricted from public access, posing challenges in utilizing public cloud-based Natural Language Processing (NLP) models and infrastructures. This necessitates a privacy-preserving approach, where models are deployed locally to ensure compliance with data protection policies while maintaining computational efficiency.

A key challenge in leveraging this dataset is trend detection, which refers to understanding recurring complaints, emerging issues, and evolving patterns of public concern. Identifying trends in complaint data is complex because (1) complaints have preassigned human-labeled categories; however, variations in how employees categorize complaints introduce a degree of subjectivity, which may impact classification consistency. While these labels provide an initial structure for analysis, their reliability in accurately capturing complaint themes requires further assessment. (2) Complaint texts vary widely in structure and clarity, with some being brief or ambiguous, making it difficult to infer key concerns without advanced

NLP techniques. The lack of standardized reporting results in complaints that may omit critical details, requiring intelligent text processing models to extract meaningful insights. (3) Complaints are written in a mixture of languages and dialects, primarily Modern Standard Arabic (MSA), but also local dialects and occasionally English, necessitating sophisticated multilingual processing capabilities to ensure accurate classification and analysis.

This research focuses on complaints directed toward the Ministry of Education, comprising approximately 45,000 complaints spanning a 5-year period. The ministry was selected due to the diverse and complex challenges inherent in the education sector. These challenges often align with well-defined timelines, such as school registration periods, academic schedules, and examinations, making the data particularly suitable for effective trend analysis. Additionally, many complaints reference specific entities such as school names, creating an opportunity to explore advanced entity recognition tasks. Identifying and tracking complaints linked to specific schools enables a more localized understanding of recurring issues and facilitates targeted policy interventions. The richness and specificity of this dataset provide a valuable case study for assessing the capabilities of LLMs in handling complex, multilingual complaint data while ensuring privacy and security.

1.2 Research Problem

This research focuses on detecting trends in Arabic complaints by systematically structuring unstructured text data into actionable insights. Instead of treating categorization, summarization, keyword extraction, and entity recognition as standalone tasks, these components are leveraged together to extract trends from the complaint dataset while ensuring that the approach is adaptable to the linguistic and operational constraints of government data systems.

Trend detection in this context involves (1) Complaint Categorization, which identifies the most frequently occurring issues by assigning structured labels to complaints. (2) Summarization and Keyword Extraction, which highlight emerging topics and recurrent concerns, providing a high-level understanding of public dissatisfaction. (3) Entity Recognition, such as school names, which maps complaints to specific institutions or locations, offering localized insights into recurring problems.

Beyond the computational challenges, the sensitive nature of government complaint data requires a localized approach to model deployment. Unlike general NLP applications that rely

on cloud-based LLMs, this research explores privacy-preserving methods for running LLMs on local infrastructure, ensuring compliance with data protection policies while maintaining model efficiency. By analyzing how complaints cluster over time and which issues increase in frequency, this study aims to create a scalable, secure, and linguistically adaptive framework for trend detection, enabling government entities to extract insights without compromising privacy or data security.

1.3 Motivation

Initially, the human-labeled complaint categories were assumed to be accurate and suitable for training a supervised classification model. However, further examination revealed inconsistencies and subjectivity in how complaints were categorized, raising concerns about their reliability for trend detection. Given that trend analysis depends on structured and consistent classification, relying solely on human-labeled data could limit the ability to accurately capture emerging patterns and recurring issues.

To address this, the research explores two complementary methodologies: LLaMA 2 (7B) with LSTM classification and LLaMA 3.1 (8B) with Retrieval-Augmented Generation (RAG). The first approach, LLaMA 2 + LSTM, follows a supervised learning framework, using human-labeled categories to train a structured classification model. This approach allows for a structured assessment of how well an AI model can replicate or improve upon human classification.

Recognizing the limitations of human-labeled data, the second approach, LLaMA 3.1 + RAG, eliminates reliance on predefined labels by dynamically retrieving relevant complaint categories using a vectorized database of past complaints and category descriptions. This retrieval-based approach enables adaptive categorization, allowing the model to detect emerging complaint trends without being restricted to fixed categories. Additionally, by retrieving contextually relevant complaint information, this approach ensures that classification remains semantically meaningful rather than strictly dependent on pre-existing human interpretations.

By comparing these two approaches, this research evaluates whether LLMs can provide a more accurate and scalable alternative to human classification and how retrieval-augmented techniques can enhance trend detection in Arabic complaint analysis. While supervised learning offers structured categorization, it is inherently limited by the quality of its training labels. In contrast, retrieval-based classification provides a more flexible and adaptive method

for handling complex and evolving complaint data.

1.4 Objectives

The overarching aim of this research is to leverage LLMs to improve the analysis and understanding of complaint data in Arabic, focusing on enhancing text processing, trend detection, and categorization. Specifically, the study aims to:

1. Investigate the capabilities of LLMs, such as LLaMA 2 and LLaMA 3.1, in streamlining and enhancing the processing of Arabic language text.
2. Improve the efficiency and effectiveness of text handling by providing high-quality summaries, extracting keywords, and generating visualizations to better understand current topics and trends.
3. Evaluate and compare the effectiveness of LLM-based methods (e.g., Retrieval-Augmented Generation) versus traditional methods (e.g., clustering and classification) in handling Arabic text, particularly for topic identification and complaint categorization.
4. Explore the integration of LLMs with traditional approaches to address challenges in structured data tasks, such as entity recognition and matching.
5. Develop dynamic visualizations and dashboards that enable stakeholders to explore trends interactively, incorporating cross-filters and granular data exploration.
6. Propose methodologies for refining categorization schemas and enhancing hybrid approaches by combining domain expertise with data-driven insights.
7. Investigate the feasibility of deploying LLMs in a localized computing environment, ensuring efficient trend detection and complaint analysis while preserving data privacy and minimizing reliance on cloud-based solutions.

1.5 Significance of the Research

This research contributes to advancing NLP applications in the public service domain by:

- Demonstrating the utility of LLMs in analyzing multilingual and multidialectal datasets, particularly in Arabic.

- Addressing the limitations of human-assigned labels by introducing LLM-driven classification and trend analysis.
- Highlighting the comparative performance of different LLM-based approaches, providing insights into their strengths and limitations.
- Highlighting the effectiveness of Retrieval-Augmented Generation (RAG) in improving structured data tasks, such as school name recognition and matching.
- Providing actionable insights through interactive dashboards, enabling stakeholders to explore trends, identify recurring issues, and allocate resources more effectively.

1.6 Thesis Structure

This thesis is organized as follows:

- Chapter 2 reviews the literature on NLP, LLMs, RAG, and complaint analysis systems.
- Chapter 3 details the methodology, including data preparation, feature extraction, LSTM classification, RAG implementation, and visualization development.
- Chapter 4 presents the results of LLM-based analysis, highlighting findings from both methods, including their strengths, limitations, and comparative performance. It also discusses key insights, implications of the findings, and practical considerations.
- Chapter 5 concludes the research and outlines future directions.

Chapter 2

Literature Review

Short text processing, which refers to the task of understanding and classifying limited length texts (Karl and Scherp, 2023), has become a key area of Natural Language Processing (NLP) due to the explosion of short content platforms online, such as tweets, reviews, and social media posts. These texts present unique challenges because of their brevity, sparseness, and ambiguity (Chen et al., 2019). This literature review examines the key characteristics and challenges of short text processing, examines various techniques and applications, and discusses future directions.

2.1 Challenges in NLP for Short Text

The field of NLP encounters distinct challenges when handling short texts, creating obstacles that differentiate it from the analysis of longer texts and documents. Recent studies shed light on critical challenges actively tackled by researchers, and this section provides a summary of issues discussed in related works.

Short texts often lack the contextual depth crucial for subtle understanding and ambiguity avoidance. Tasks related to trend detection, such as topic classification, clustering, sentiment analysis, translation, and intent recognition, encounter difficulties due to the absence of vital information within the limited context (Wan et al., 2022; Amur et al., 2023). Moreover, short texts frequently incorporate informal language, slang, and abbreviations, rendering them susceptible to misinterpretation (Tang et al., 2022; Alzanin et al., 2022). The brevity and sparsity of these texts increase ambiguity, where a phrase may carry multiple potential meanings (Hu et al., 2022). Handling short text data from social media for trend and event detection (e.g. on X platform, previously known as Twitter), poses a challenge according

to Al Sulaimani and Starkey (2023). The overwhelming volume and speed of posts make processing data challenging, and misinformation and spam further contaminate the information landscape, affecting discourse quality. The dynamic nature of social media content, including new terms and evolving word contexts, introduces complexity to accurate event detection. Additionally, the multilingual nature of social media events necessitates attention to linguistic diversity for comprehensive analysis. These challenges emphasize the complexities researchers face in extracting meaningful insights from short text data. Extensive exploration reveals that short text challenges persist across diverse languages, emphasizing the universal nature of difficulties in processing such texts (Saeed et al., 2022; Sun et al., 2023; Goker and Canbuglalilar, 2018). In delving into the Arabic language, NLP encounters challenges due to the complex morphology and diverse dialects spoken by millions of people in geopolitically diverse regions. The nature of the Arabic language, coupled with pronounced variations in social and regional contexts, poses intricate hurdles for researchers (Diab et al., 2017). Another significant challenge arises from models trained on general domain data struggling in specific domains due to unique vocabulary and linguistic styles. Adapting models to new domains becomes particularly challenging for short texts, as they often struggle to generalize to new domains characterized by distinct vocabulary and linguistic styles (Li et al., 2022). This issue underscores the necessity for robust domain adaptation strategies to enhance the performance of NLP models.

2.2 Methods for Short Text Processing

This section aims to investigate methodologies employed in trend detection for short texts, encompassing diverse methods such as classification and clustering. The review delves into a selection of six papers that examine short text processing in both English and Arabic. These studies were intentionally selected for their methodological diversity and relevance to the main tasks in trend detection, including traditional machine learning approaches, deep learning architectures, and more recent work leveraging large language models (LLMs). The rationale for focusing on these six studies lies in their representation of key trends in the field, from early feature-based approaches to the integration of contextual knowledge and domain-specific models. Given the evolving nature of short text analysis, particularly in informal and multilingual settings, these works illustrate how researchers are addressing persistent challenges such as ambiguity, data sparsity, and domain mismatch. Moreover, the lack of standardized datasets and the specificity of use-cases (e.g., Arabic tweets or medical texts)

make generalizability a continuing concern in this domain.

While this section offers an in-depth review of these representative works, additional literature is referenced throughout the chapter to support broader thematic discussion. By examining these methodologies, the goal is to gain valuable insights into their effectiveness in addressing the unique challenges discussed in 2.1, with a specific focus on the context of Arabic.

In a paper titled “Short text classification for Arabic social media tweets,” Alzanin et al. (2022) proposed a system to classify Arabic tweets into different classes using classical methods. They went through the steps shown in Listing ?? like cleaning up the text, using two methods to understand the words, which are the term frequency-inverse document frequency (TF-IDF) and Word2vec, and then employing machine learning to classify tweets. In the cleaning step, they removed unnecessary symbols and words, like hashtags and mentions, and they also used two types of stemmers belonging to the NLTK library to make words simpler. The first stemming approach is based on the Information Science Research Institute’s algorithm (ISRI), and the second approach is based on the light stemming algorithm Arabic Light (ARL). The paper also used a pre-trained model for Word2vec on a large set of Arabic tweets. Researchers collected 35,627 tweets for their study and manually classified them into News, Conversation, Question, Wish, and Others. It’s worth mentioning that the datasets were collected through a search without specifying the selection criteria for local dialects of Modern Standard Arabic. However, all samples in the paper were in Modern Standard Arabic, making them suitable for selected stemmers. The study employed three classifiers, Gaussian Naive Bayes (GNB), Support Vector Machine (SVM), and Random Forest (RF). Hyperparameter tuning was conducted to optimize the performance of each classifier. For GNB, SVM, and RF, parameters such as smoothing priors, kernel types, and the number of trees in the forest were fine-tuned, respectively. Feature selection was performed after generating the TF-IDF matrix, and the k-best method was employed to select the most relevant features for the classification task. The combination of these classifiers, optimized hyperparameters, and selected features contributed to the effective classification of Arabic tweets into predefined categories. During testing, as shown in Table 2.1, RF and SVM demonstrated a strong performance, achieving a high F1 score of 0.9814. However, the study observed a negative impact of using Word2vec on the GNB classifier. The paper concludes by suggesting (i) the exploration of deep learning with larger datasets for training and (ii) the consideration of additional classes in the classification system.

Table 2.1: Classification Results (Normalized) for Different Stemmers and Classifiers (Alzanin et al., 2022)

Stemmer	Classifier	Accuracy	Precision	Recall	F1 score
ISRI	GNB	0.9843	0.9835	0.9638	0.9729
	SVM	0.9922	0.9961	0.9676	0.9812
	RF	0.9914	0.9956	0.9686	0.9814
ARL	GNB	0.9865	0.9698	0.9641	0.9669
	SVM	0.9919	0.9957	0.9675	0.9809
	RF	0.9916	0.9960	0.9679	0.9813
Word2vec	GNB	0.7452	0.6971	0.6804	0.6845
	SVM	0.9893	0.9944	0.9662	0.9796
	RF	0.9895	0.9954	0.9658	0.9799

In another study “Twitter Topic Classification” conducted by the authors Antypas et al. (2022), the focus was on identifying trends by classifying topics in social media discussions, particularly on platforms like Twitter, Snapchat, TikTok, and Instagram. The research addressed the challenge of organizing and categorizing the diverse content generated daily on these platforms. While recognizing the common practice of resorting to unsupervised topic modeling for lengthy documents, the paper introduced an approach focused on supervised tweet topic classification. This shift addressed challenges related to interpretability and eliminated dependencies on assumptions about vocabulary distribution. Additionally, the authors released two associated datasets to support the evaluation and advancement of models tailored for tweet topic classification. The authors presented the TweetTopic classification dataset construction pipeline, which involved tweet collection, data filtering, and topic annotation. The tweet collection aimed to cover a wide range of diverse topics over time by querying the Twitter API with specific keywords and time periods. The data filtering process shown in Figure 2.1 involved removing irrelevant content through text filtering techniques such as removing non-English text, incompetent sentences, and non-text data. The final dataset was annotated manually using Amazon Mechanical Turk service which allows assigning Human Intelligence Tasks to workers, streamlining the delegation of tasks that necessitate human intelligence like (Amazon Web Services, 2024). The authors categorized their dataset, which contains around 11,000 tweets, into 23 initial classes, and then reduced them to 19 after eliminating classes with less than 50 tweets. The dataset was split into two periods for a realistic evaluation of the models, and both single-label and multi-label

classification experiments were performed. Topic features were analyzed in the dataset, and various tweet statistics were extracted. Metrics, including tweet length, punctuation symbols, uppercase-to-lowercase ratio, hashtags, mentions, and emojis, were normalized. The Measure of Textual Lexical Diversity (MTLD) and the number of tweets per class were also taken into account. In the experiment section, various classifiers were used, including classical models like a traditional bag of words with SVM, a fastText classifier that utilizes pre-trained embeddings, and advanced language models such as the base BERT model along with its optimized versions, RoBERTa (RB), BERTTweet and TimeLM which specialized for social media and trained on a corpus of millions of tweets. The results in Table 2.2 indicated that domain-specific knowledge, especially models trained on Twitter data, outperformed large generic language models. The paper concluded by providing insights into potential future work, including the exploration of larger datasets for training and the consideration of additional classes in the classification system. The TweetTopic dataset, tailored to social media content, was introduced as a valuable resource for researchers and practitioners interested in social media analysis.

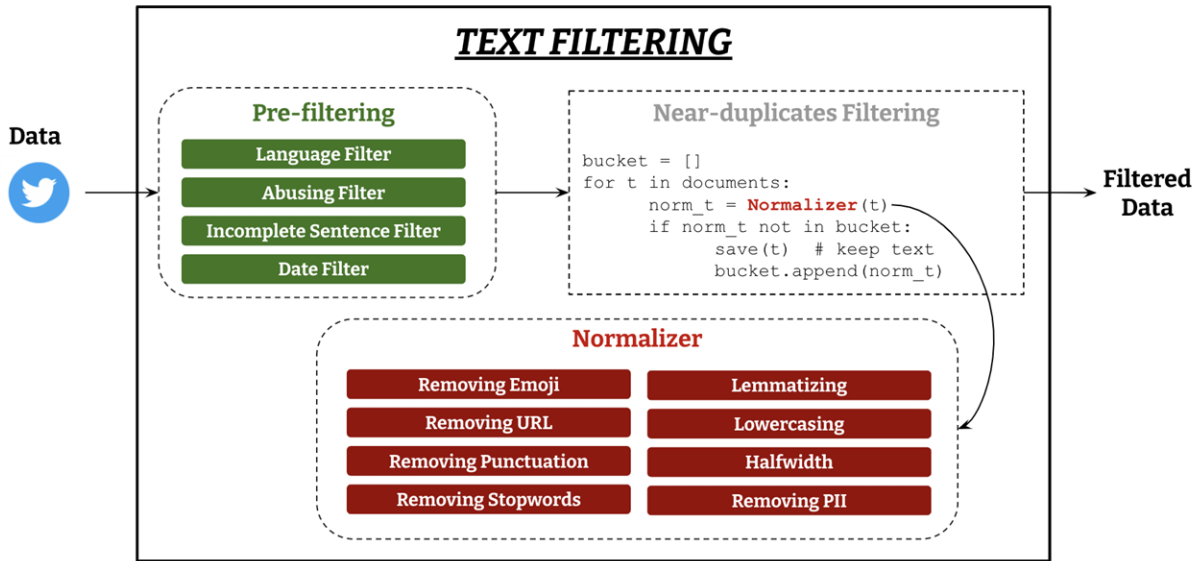


Figure 2.1: Pipeline for text filtering to reduce noise in tweets (Antypas et al., 2022)

Table 2.2: Precision (Pr), Recall (Rec), F1, and accuracy results for Tweet Topic with Jaccard Index (JI) reported in the multi-label setting (Antypas et al., 2022)

Model	Multi-label				Single-label			
	Pr	Rec	F1	Acc	Pr	Rec	F1	Acc
Baselines								
SVM	69.4	23.7	30.5	37.1	73.6	47.4	50.2	75.8
fastText	67.0	18.0	24.0	31.9	56.0	46.0	48.0	74.0
Language Models								
BERT-base	69.7	42.5	50.1	45.5	62.4	60.0	58.8	81
BERT-large	64.4	51.5	56.4	44.6	62.4	61.7	61.7	84.3
RB-base	68.5	49.2	55.8	46.5	64.8	66.7	65.6	85.9
RB-large	72.2	48.9	56.3	47.9	66.1	56.2	58.3	84.5
BERTweet	66.9	46.1	52.7	47.1	64.9	65.6	63.8	85.2
TimeLM-19	71.1	50.4	57.2	47.7	76.5	68.9	70.0	86.4

Another classification study conducted by Tang et al. (2022) presents an approach, namely the Double-Level CNN (DLCNN), to address the inherent challenges associated with short text and its sparse features and lack of context. In response to this, the DLCNN model in Figure 2.2 is structured with both superior and subordinate networks. In the superior network, which consists of 4 modules, 1) External knowledge graphs, like Probase, are employed for short text conceptualization, extracting essential concepts. 2) Word vector representations are generated through word-level feature capture, enhancing semantic understanding. 3) Bidirectional Long Short-Term Memory (Bi-LSTM) network facilitates semantic encoding, capturing contextual information bi-directionally. 4) A context-based attention mechanism combines semantic encoding and knowledge vectors for a comprehensive knowledge representation. This representation is connected with the semantic code and processed through a CNN for final classification. Simultaneously, the subordinate network, utilizing a keyword extraction algorithm, identifies short text keywords and generates corresponding word vector representations through character-level features. Results from both networks are combined for the ultimate classification effect using a linear layer. Preprocessing involves using the Probase knowledge graph for entity recognition, extracting entities and conceptualizing them based on existing knowledge. Input embedding and semantic encoding utilize Word2Vec for

the superior network and a CNN for the subordinate network. A context-dependent attention mechanism calculates the weight of concepts/keywords based on semantic information. Finally, the CNN short text classification combines results from the superior and subordinate networks. The DLCNN experiment involved testing on well-known short text datasets, including AG NEWS, Text REtrieval Conference(TREC), and Movie Reviews (MR), for text classification tasks. The comparison was made against seven other methods as in Table 2.3, namely CNN-Rand, CNN-Static, Char-CNN, WCCNN, CNN-HE, CNN-VE, and SECNN. The results demonstrated that DLCNN achieved superior accuracy across all three datasets, with particularly notable improvements on TREC (93%) and AG NEWS (87%). The model’s effectiveness was attributed to its ability to conceptualize information in short texts, emphasizing its superiority in achieving classification goals, especially in objective categories. Results on standard short text datasets reveal the superior accuracy of the DLCNN model when compared to the seven other methods. However, this emphasis on accuracy may potentially overlook other performance metrics crucial for a comprehensive understanding of the model’s capabilities. Furthermore, the study does not extensively explore the generalizability of the proposed model across various domains or types of short text, introducing uncertainties about its adaptability beyond the specific knowledge bases and datasets mentioned.

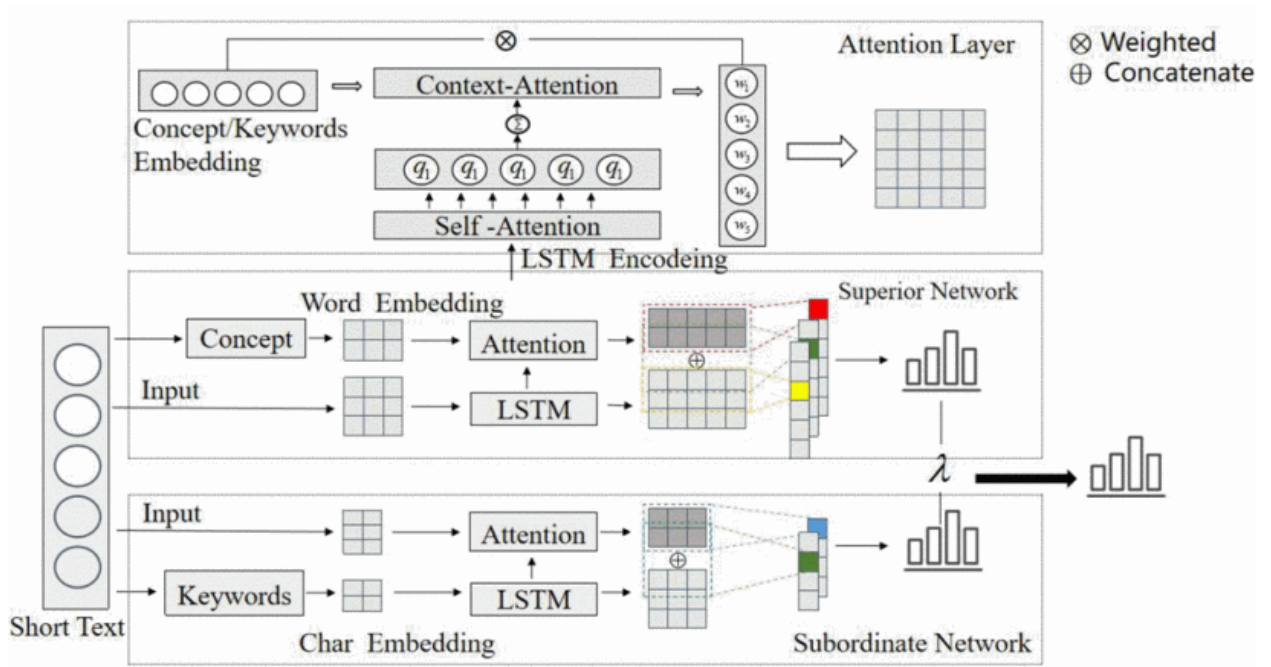


Figure 2.2: Double-Level CNN (DLCNN) Overview (Tang et al., 2022)

Table 2.3: Comparison of Accuracy: Double-Level CNN (DLCNN) versus Other Methods on three Datasets. Methods compared include CNN-Random (CNN-rand), CNN-Static (CNN-static), Character-level CNN (CharCNN), Word-Character CNN (WCCNN), CNN with Horizontal Embedding (CNN-HE), CNN with Vertical Embedding (CNN-VE), and Squeeze-and-Excitation CNN (SECNN) (Tang et al., 2022)

Methods	Movie Reviews	Text REtrieval Conference	AG NEWS
CNN-rand	76.72	86.17	84.38
CNN-static	80.77	89.26	84.38
CharCNN	76.93	76.05	78.31
WCCNN	82.95	90.68	85.76
CNN-HE	82.29	91.28	85.84
CNN-VE	82.08	91.05	85.80
SECNN	83.89	91.34	86.02
DLCNN	82.60	93.20	87.11

Further research, titled “Real-Time Event Detection Using Self-Evolving Contextual Analysis (SECA) Approach” introduces the Self-Evolving Contextual Analysis (SECA) method for event detection. In contrast to earlier studies, it advances unsupervised learning through the utilization of clustering. SECA, founded on the principles of the Contextual Analysis (CA) method introduced by Abdul Aziz and Starkey (2020), and designed to comprehend word relationships based on co-occurrence in the same context. SECA employs a tree-like structure known as a Hierarchical Knowledge Tree (HKT). The CA algorithm encompasses steps such as creating the initial Seed-HKT container, grouping words based on their presence in sources, and forming sublevel HKTs (Child-HKTs). The tree encompasses various objects, including Seed-Nodes, Seed-HKTs, Child-Nodes, Child-HKTs, Refuge-Nodes, Refuge-HKTs, Orphan-Nodes, and Paths. SECA adapts to evolving datasets by constructing a baseline tree, mapping new sources to Seed-HKT, assessing changes’ significance, and reconstructing the tree. It addresses challenges in real-time event detection by introducing SECA-Light, discarding outdated sources, and exploring clustering approaches. The direct method views nodes in Seed-Nodes and Orphan-Nodes as clusters, while the indirect method uses the tree to generate proxies for event clusters, merging those sharing top words. Figure 2.3 is a visual representation of levels of feeding SECA with new batches. In summary, CA and SECA offer adaptability in capturing evolving data patterns, enhanced by clustering methods for improved understanding of word relationships and event clusters. The experiment involved utilizing pre-processed datasets for text tokenization and implementing the SECA algorithm using C#. Baseline methods, namely KMeans, Non-negative Matrix Factorization (NMF), and MStream, were employed for comparison. SECA variants outperformed other models, as shown in Figure 2.4, showcasing adaptability and effective clustering for contextual analysis and real-time event detection. A practical application in the FIFA World Cup Final highlighted SECA’s real-world value in event detection from Twitter data, demonstrating its ability to identify key events during a football match as in Figure 2.5. A visual representation of generated trees uses ChatGPT to summarize illustrated hierarchical word relationships.

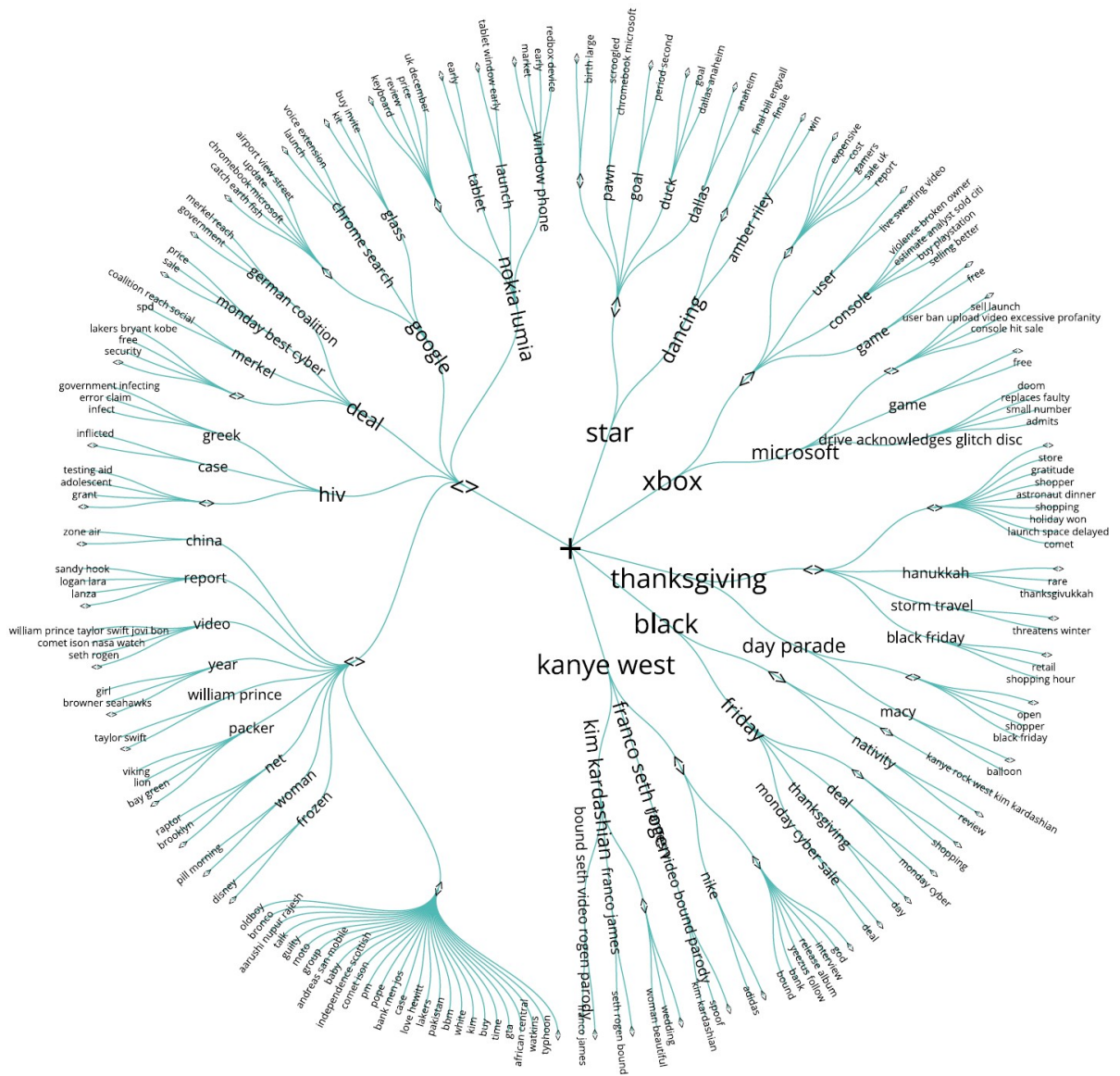


Figure 2.3: Tree representation of four levels for the topics and subtopics after feeding SECA with three batches, (Abdul Aziz and Starkey, 2020)

Avg. NMI							
A	84.77	72.74	86.32	92.73	92.70	92.74	91.83
B	85.06	72.85	86.52	94.58	94.39	94.56	93.45
C	85.07	70.17	87.24	92.80	91.87	92.43	91.52
D	87.31	71.67	89.67	93.16	93.69	93.51	93.02
E	84.90	73.51	88.64	93.08	93.29	93.79	92.21
F	80.83	73.29	84.65	90.96	90.17	90.99	89.87
G	75.36	88.95	86.87	86.21	88.30	85.94	87.93
H	73.00	88.40	80.76	77.75	87.52	84.30	88.85
	K-Means	M-Stream	NMF	SECA Direct	SECA Indirect	SECA-Light Direct	SECA-Light Indirect
Avg. F1 Score							
A	80.31	54.70	79.27	89.48	89.53	89.66	87.91
B	81.09	53.38	82.01	91.82	92.64	92.27	91.06
C	81.54	50.98	80.74	90.10	89.32	90.38	88.42
D	84.12	56.58	86.21	91.63	91.94	92.45	91.34
E	80.04	58.23	84.21	90.70	91.25	91.53	89.03
F	75.53	49.03	80.63	84.89	83.45	85.72	82.14
G	45.14	66.81	60.59	54.63	63.03	55.37	62.74
H	43.84	73.68	56.65	55.05	71.14	64.32	74.08
	K-Means	M-Stream	NMF	SECA Direct	SECA Indirect	SECA-Light Direct	SECA-Light Indirect
Avg. Purity							
A	94.12	99.78	97.68	99.87	99.89	99.97	99.91
B	94.43	99.69	96.54	99.87	99.87	99.97	99.91
C	95.13	99.96	99.61	99.87	99.87	99.96	99.91
D	96.49	99.91	99.30	99.83	99.87	99.97	99.91
E	93.51	99.74	98.38	99.91	99.85	99.97	99.91
F	82.65	99.08	79.56	99.64	99.56	99.64	99.30
G	68.16	74.27	79.56	71.67	80.36	73.67	74.03
H	77.55	85.65	85.17	75.53	89.18	81.44	89.03
	K-Means	M-Stream	NMF	SECA Direct	SECA Indirect	SECA-Light Direct	SECA-Light Indirect

Figure 2.4: Performance comparison between the four methods (KMeans, MStream, NMF, and SECA), from Abdul Aziz and Starkey (2020)

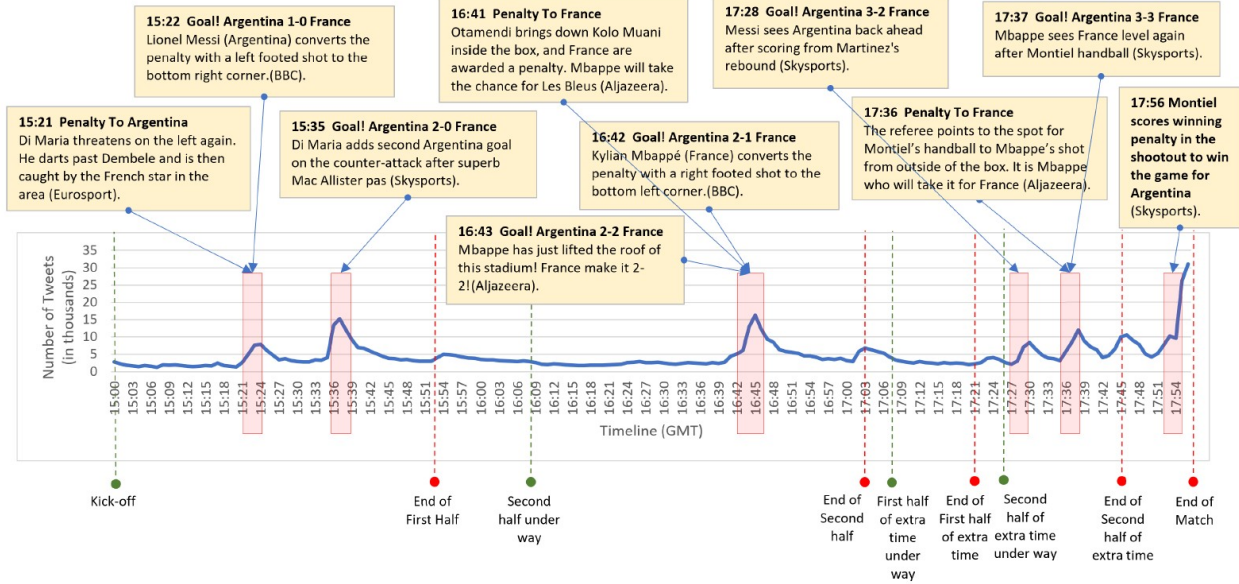


Figure 2.5: Tweet count timeline; Yellow boxes highlight key events, from Abdul Aziz and Starkey (2020)

In further studies, a recent survey has explored the use of LLMs for data annotation tasks for ML and NLP (Tan et al., 2024). The process of data annotation involves adding informative labels to raw data, which is crucial but challenging due to its complexity and labor intensity. The survey examines the application of LLMs and sets a foundational framework for LLMs for data annotation by introducing essential notations and describing the core models—annotator model A and task learner L . It details different learning scenarios such as fully supervised, unsupervised, and semi-supervised learning, illustrating how LLMs operate in each to enhance data with useful annotations. Various LLM prompting and tuning techniques like input-output prompting, in-context learning, and chain-of-thought prompting are also discussed, which help refine LLM interactions with data for optimal annotation outcomes. Following this foundation, the survey delves into three main contributions as shown in the figure 2.6. First, the LLM-Based Data Annotation section explores the capabilities of models like GPT-4 and LLaMA-2, particularly their advanced language comprehension and contextual understanding, which significantly enhance the accuracy and efficiency of data annotations. Techniques such as prompt engineering and domain-specific fine-tuning are detailed as essential for optimizing LLM outputs for specific tasks. Second, the Assessing LLM-Generated Annotations section evaluates methodologies for ensuring the quality of LLM-produced annotations, focusing on assessing accuracy and selecting high-quality op-

tions from multiple outputs. Lastly, the Learning with LLM-Generated Annotations segment investigates the impact of using LLM-annotated data to train machine learning models, emphasizing improvements in model quality and reliability.

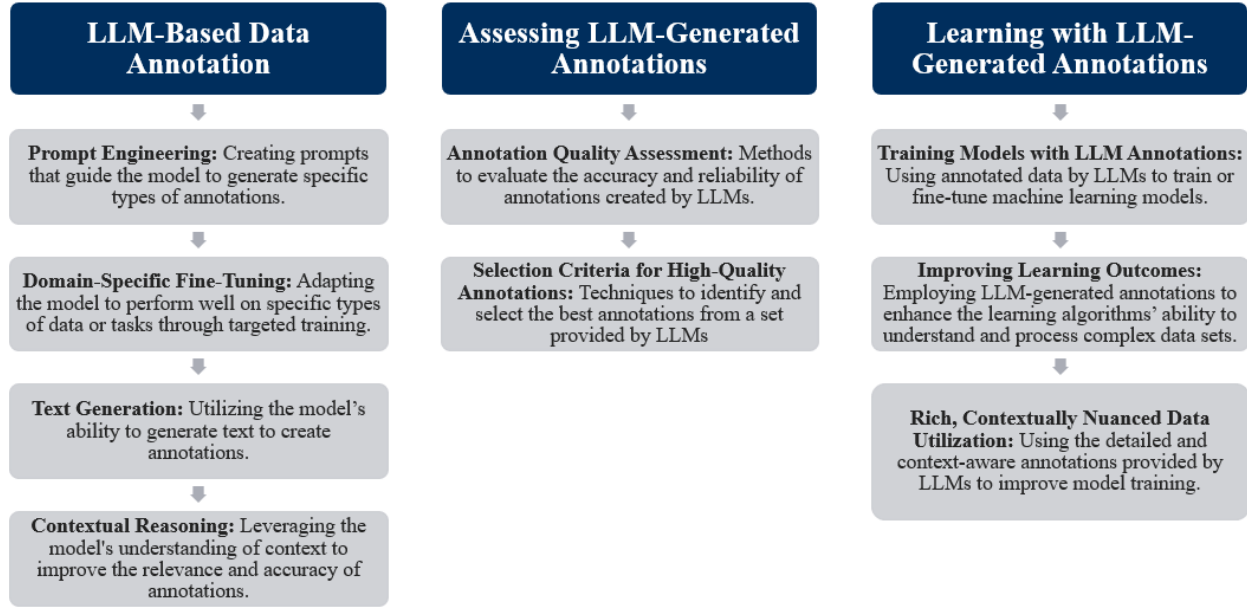


Figure 2.6: Flowchart of LLM Application in Data Annotation Processes, from Tan et al. (2024)

Building on the exploration of LLMs for data annotation tasks in various domains, another paper titled “LLMs Accelerate Annotation for Medical Information Extraction” focuses on leveraging LLMs specifically within the medical domain (Goel et al., 2023). This paper proposes a hybrid approach combining LLMs with human expertise to enhance the annotation of medical texts, significantly reducing the human workload required for generating labeled datasets. The paper begins by addressing the challenges inherent in extracting useful information from unstructured clinical notes within electronic health records. These notes often conceal vital patient-related information, making them difficult to access and interpret. Specialized NLP models are required for this task, which traditionally relies on large volumes of manually labeled data—a process that is both time-consuming and costly. To address these challenges, the paper introduces a methodology that incorporates LLMs to generate base annotations initially. These annotations are subsequently refined by medical experts, thereby creating an efficient method for generating high-quality ground truth labels for medical text annotation. This dual-step process leverages the rapid data processing capabilities of LLMs while ensuring the accuracy of annotations through expert review. The methodology intro-

duced in the paper employs Google’s LLM, PaLM 2, to initially generate base annotations for medical texts. These annotations serve as a preliminary layer of data processing, which are then refined by medical experts to ensure accuracy. This dual-step approach capitalizes on the rapid processing abilities of PaLM 2, allowing for the efficient creation of high-quality ground truth labels for medical text annotation. By integrating the capabilities of PaLM 2 with expert human review, the method not only streamlines the annotation process but also maintains the high standards required for medical data applications. This innovative use of PaLM 2 highlights the tailored approach taken to address the unique challenges of medical information extraction. The paper presents compelling results supporting the integration of LLMs with human expertise for efficient medical data annotation. PaLM 2’s initial annotations, when refined by experts, closely align with purely expert-driven annotations in terms of quality. The F1 scores for LLM annotations (0.907) and human annotations (0.910) indicate a high level of accuracy and reliability. Moreover, this method significantly reduces the median annotation time from 19.18 to 7.27 minutes, highlighting the efficiency gains achieved through the integration of LLMs. Additionally, the paper evaluates both vertical and horizontal metrics to ensure comprehensive quality assessment of the labeled data. Vertical metrics focus on individual data points, assessing the accuracy of each piece of medical information separately, while horizontal metrics evaluate the relationships between different data points, ensuring the correct linking of related fields across the dataset. This holistic approach to evaluation ensures that not only individual data points but also the relationships between them are accurately captured, providing a thorough understanding of the annotated medical information. Such integration of LLMs with human expertise is crucial for advancing patient care and healthcare operations, paving the way for further optimization and application of LLMs in diverse medical NLP tasks.

While the reviewed studies have contributed significantly to the field of short text processing, they each fall short in addressing the combined challenges tackled in this thesis, namely, real-time complaint classification, linguistic variation in Arabic dialects, and the constraints of working with sensitive government data. Most existing work focuses on static datasets and generic classification tasks, often ignoring issues like category ambiguity, noisy labels, and the need for localized model deployment. Unlike earlier approaches that rely heavily on supervised learning or hand-crafted features, this research demonstrates how integrating LLMs with RAG and refined prompt engineering can enable more accurate and context-aware classification in resource-constrained environments. By comparing human and model performance, this study also surfaces deeper challenges in interpretability and fairness that

remain underexplored in prior literature.

2.3 LLMs for Trend Detection and RAG-Based Enhancements

Recent advances in LLMs have opened new opportunities for automating the classification and management of complaint data across various domains. Liu et al. (2024) introduced the Complaint Classification Framework (CCF), which combines real-time complaint scraping, pre-processing, and semantic compression using Yi-34B LLM, followed by classification with a BERT model. This two-stage process improved accuracy on multi-category Chinese complaint datasets and reduced reliance on manual feature engineering. While their approach emphasized long-text simplification for downstream models, our research instead focuses on short, ambiguous Arabic complaints, integrating LLMs more directly into the classification pipeline via RAG

In a different domain, Yong et al. (2024) evaluated GPT-4’s ability to generate responses to patient complaints in a hospital setting. Their comparative study found that LLM-generated replies were rated as more appropriate and empathetic than those written by human administrators. Although their focus was on response generation, not trend detection, the study highlights the capacity of LLMs to understand sensitive domain-specific language, the ability that we leverage in our work for the summarization and categorization of Arabic education complaints.

Furthermore, the work by Zhao et al. (2024) provides a comprehensive survey of RAG architectures and categorizes user queries into four levels(explicit fact, implicit fact, interpretable rationale, and hidden rationale) based on their complexity and knowledge needs. They emphasize the importance of aligning retrieval strategies with task objectives. Our method is more closely aligned with the types of ‘explicit’ and ‘interpretable rationale’ queries, where the retrieved embeddings of predefined complaint categories are integrated into prompts to improve the categorization of ambiguous or mislabeled entries in the Tawasul dataset.

These studies collectively inform the design of our enhanced methodology (Method 2), which builds upon existing LLM and RAG techniques to address unique challenges in Arabic short-text complaint analysis—such as linguistic diversity, vague or inconsistent labels, and sensitive data handling requirements. Unlike the earlier six studies reviewed in Section 2.2, which largely relied on traditional classifiers, static datasets, and supervised learning pipelines

(aligned with our initial approach, Method 1), this set of recent works demonstrates the scalability, contextual reasoning, and semantic richness of LLMs when coupled with augmentation strategies. Drawing from these insights, our Method 2 integrates LLaMA 3.1 with RAG, embedding predefined category knowledge into the retrieval process to support more accurate and context-aware classification, even in the presence of ambiguous or under-specified complaint texts.

Chapter 3

Methodology

The methodology of this research aimed to leverage LLMs to analyze trends in Arabic complaint data from the national complaint system “Tawasul”. Initially, two approaches were considered for categorizing complaints. The first method involved using LLaMA 2, where human labels were combined with LLM-based feature extraction to produce summaries, keywords, and submitter types, followed by using LSTM models for classification. However, due to performance limitations and inefficiency, the research shifted to the second method, utilizing LLaMA 3.1 with retrieval-augmented generation (RAG) to enhance results and improve efficiency. This transition allowed better utilization of LLM capabilities, replacing the dependency on human labels with embeddings and RAG for categorization, and provided faster processing of the complaint dataset.

The process began with data collection and handling complaints while addressing inconsistent category labels. LLaMA 2 and LLaMA 3.1 were used for tasks like summarizing, keyword extraction, and category prediction. RAG was implemented to enhance categorization using contextual embeddings. Tools such as LangChain and OLLaMA facilitated model deployment, with prompt engineering used to optimize results. Finally, visualization techniques were employed to illustrate trends, making the findings accessible and actionable for stakeholders.

A high-level overview of the identified research challenges and corresponding methodological solutions is summarized in Table 3.1.

Table 3.1: Summary of Research Challenges, Context, and Methodological Solutions

No.	Challenge	Context	Methodological Solution
1	Subjectivity and inconsistency in human categorization	Complaints have preassigned categories, but employee judgment introduces variability.	Used RAG and LLM embeddings for dynamic and context-aware categorization.
2	Short and ambiguous complaint texts	Complaint texts vary significantly in clarity and detail, complicating accurate interpretation.	Applied iterative prompt engineering techniques to consistently produce clear summaries, keywords, and categories.
3	Linguistic diversity (Arabic with local dialects)	Complaints submitted primarily in diverse local dialects, hindering unified processing.	Summarized complaints in English to standardize outputs, facilitating uniform trend analysis and keyword extraction.
4	Entity extraction inaccuracies (e.g., school names)	Variability in how entities, especially school names, are mentioned within complaints.	Employed targeted prompts for entity extraction, then matched extracted entities against a predefined standardized list for consistency.
5	Sensitive data restricting cloud processing	Government data privacy policies restrict external or cloud-based processing.	Locally deployed LLMs using OLLaMA and Apptainer containers to comply with data privacy requirements.

3.1 Data Collection and Preprocessing

This research involves analyzing complaints related to the Ministry of Education from the Tawasul dataset. The dataset presents challenges such as linguistic diversity, inconsistent labeling, and unstructured narratives, necessitating advanced NLP techniques for effective processing.

The complaints are written in a variety of Modern Standard Arabic (MSA), local Arabic dialects, and occasionally English. This linguistic diversity reflects the varied backgrounds of the individuals submitting the complaints, creating unique challenges for language processing

due to differences in formal and informal styles. The dataset has a maximum character limit of 1000 per complaint, which ensures that each submission is concise while still containing enough information to provide context and detail.

Each complaint is accompanied by an initial single-category label assigned by humans after the submission. These labels are meant to classify the complaint into predefined categories; however, their reliability is variable, and not all labels accurately reflect the nature of the complaints. This inconsistency presents both a challenge and an opportunity for applying LLMs to enhance categorization capabilities. Additionally, most complaints have been assigned to an 'other' category, indicating that they did not fit neatly into any of the predefined categories, leading to significant clustering of diverse concerns under a single broad label.

Further exploratory data analysis revealed notable characteristics of the dataset:

- The predefined list of categories often seemed insufficient to capture the full range of concerns raised, with many complaints sharing traits that could qualify them for multiple categories.
- An uneven distribution of categories, with a substantial proportion of complaints falling into a few dominant classes while others have very few entries.
- Conflicting labels and inconsistent categorization were evident, often requiring manual interpretation or reclassification to ensure coherent trend detection and categorization efforts.

These data inconsistencies and challenges form a key motivation for leveraging LLMs to refine and enhance the classification and trend detection processes. The LLMs' ability to understand and process mixed dialects, infer context, and generate meaningful categories is expected to address the limitations of the existing labels and provide a more structured and insightful understanding of the public's concerns.

In addition to the category labels, some complaints contain entity mentions, such as the names of specific schools, adding a further layer of complexity for entity recognition tasks. Entity extraction and normalization are critical to ensuring that these mentions are accurately identified and matched to predefined lists, allowing for analyzing specific institutions' involvement in the complaints.

Overall, the Tawasul dataset presents a unique opportunity to explore the capabilities of LLMs in handling Arabic language data, particularly in the domains of classification, summarization, and trend detection. The diversity in language, the variation of formal and

informal expressions, and the inconsistencies in labeling make it an ideal case for assessing how advanced NLP models can streamline and enhance the processing of Arabic texts to extract actionable insights.

3.2 LLM Installation and Setup

For this research, two versions of the LLaMA model were utilized—LLaMA 2 (7 Billion parameters) and LLaMA 3.1 (8 Billion parameters). These models were instrumental in processing and analyzing the Tawasul dataset. This section details their installation, architecture, and deployment setup at the University of Bahrain’s Benefit Advanced AI and Computing Lab.

3.2.1 Initial Deployment with LLaMA 2

The first approach employed LLaMA 2, a 7-billion-parameter model developed by Meta, designed for advanced NLP tasks. LLaMA 2 was selected due to its performance in open-source benchmarks and its quantization-friendly deployment capabilities. A quantized version in GGUF format was installed on the university’s cluster using the LLaMA.cpp framework.

GGUF is an updated version of GGML, originally developed by Georgi Gerganov. The name “GG” stands for “Georgi Gerganov” while “ML” represents machine learning. GGUF improves on GGML by enhancing flexibility and supporting additional features while maintaining compatibility with older versions. It allows reduced computational requirements by representing model weights in lower precision (e.g., 8-bit integers), optimizing the use of hardware resources (IBM, 2024). The installation of LLaMA 2 required careful management of dependencies and integration with the high-performance computing (HPC) environment. Batch processing was employed to handle the dataset efficiently, but processing times per batch were considerable, ranging between 1.5 to 2 hours. The installation of LLaMA 2 was managed using LLaMA.cpp, a core library developed to handle GGUF models. LLaMA.cpp provides utilities for running models, conversion, and integration across different environments, allowing LLaMA models to be deployed on a wide range of hardware, from high-performance GPUs to consumer-grade CPUs.

LLaMA 2, like other transformer-based models, utilizes a transformer architecture comprising multiple encoder-decoder layers. The architecture includes self-attention mechanisms, which allow the model to focus on different parts of the input sequence when making predictions. Specifically, LLaMA 2 incorporates grouped-query attention (GQA) to enhance

efficiency and accuracy by reducing the number of attention heads while preserving performance (Ainslie et al., 2023). The model also features an increased context length of up to 4k tokens, which is double the context length of its predecessor, LLaMA 1, which supported up to 2k tokens (Touvron et al., 2023a). This extended context length allows LLaMA 2 to better capture long-range dependencies in the input text, improving its ability to understand complex linguistic structures.

Figure 3.1 illustrates the training pipeline and architecture of LLaMA 2. This model leverages advancements such as GQA and increased context length (up to 4k tokens), which improved its inference and contextual understanding (Touvron et al., 2023b).

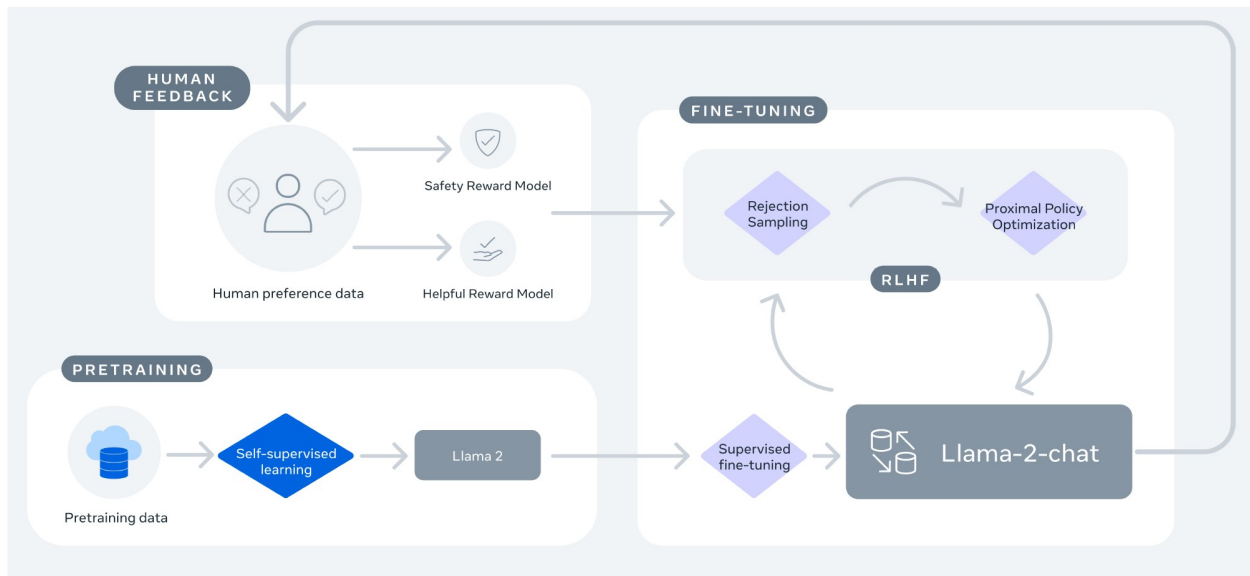


Figure 3.1: Training process of LLaMA 2, highlighting pretraining and fine-tuning steps. Source: Meta AI (Touvron et al., 2023b).

3.2.2 Transition to LLaMA 3.1

To address the limitations observed with LLaMA 2, such as extended processing times, the study transitioned to LLaMA 3.1 (8B). LLaMA 3.1, part of Meta AI's latest iteration of large language models, introduces several architectural enhancements to improve inference efficiency and scalability. These include **grouped-query attention (GQA)**, designed to optimize attention mechanisms, and an **enhanced tokenizer** capable of supporting a 128k vocabulary size (Dubey et al., 2024).

The deployment of LLaMA 3.1 utilized the **OLLaMA framework**, a platform specifically designed to optimize and manage the execution of LLaMA models. OLLaMA simplifies

deployment by enabling containerized execution through Docker, further facilitated by Apptainer within the university’s high-performance computing (HPC) infrastructure. This setup significantly reduced batch processing times, bringing them down to approximately 5–6 minutes. Such efficiency was achieved through quantization techniques and streamlined resource allocation, which are hallmarks of the OLLaMA framework (Ollama AI, 2024).

3.2.3 Comparison Between LLaMA and OLLaMA

While LLaMA is a foundational model developed by Meta AI, OLLaMA serves as a complementary framework focused on efficient deployment.

Table 3.2 highlights the differences between LLaMA and OLLaMA, focusing on their distinct roles. LLaMA is a foundational large language model designed for training and inference tasks, requiring GPU or TPU hardware for optimal performance. In contrast, OLLaMA is a deployment and optimization framework tailored for efficient execution of LLaMA models, supporting both CPU and GPU hardware. Key features of OLLaMA include built-in support for quantization to reduce resource usage and containerization through Docker and Apptainer for streamlined deployment. While LLaMA emphasizes architectural advancements, OLLaMA focuses on enhancing deployment efficiency.

Table 3.2: Comparison Between LLaMA and OLLaMA.

Feature	LLaMA	OLLaMA
Type	Foundation LLM	Deployment and optimization framework
Focus	Training and inference	Efficient deployment of LLaMA models
Hardware Support	GPU/TPU	CPU and GPU
Quantization	Optional	Supported for efficiency
Containerization	Not included	Built-in (Docker and Apptainer)
Optimization Techniques	Architectural	Deployment-focused

3.2.4 LLaMA 3: Architecture and Hyperparameters

LLaMA 3 represents a leap forward in the evolution of transformer-based architectures. Its design emphasizes inference efficiency, scalability, and support for large-scale datasets. Key architectural advancements include:

- **Grouped-Query Attention (GQA):** Reduces computational overhead during inference.

- **Optimized Tokenizer:** Expands support for up to 128k unique tokens, enabling better handling of diverse text datasets.

Table 3.3 summarizes the key hyperparameters of the LLaMA 3 family of models, ranging from the smallest 8B version to the largest 405B version.

Table 3.3: Key hyperparameters of LLaMA 3. Source: Meta AI Dubey et al. (2024).

Model	8B	70B	405B
Layers	32	80	126
Model Dimension	4,096	8,192	16,384
Attention Heads	32	64	128
Context Length	4k	4k	4k
Vocabulary Size	128k	128k	128k

Figure 3.2 illustrates the architecture of LLaMA 3, highlighting its token embedding, self-attention, and autoregressive decoding mechanisms.

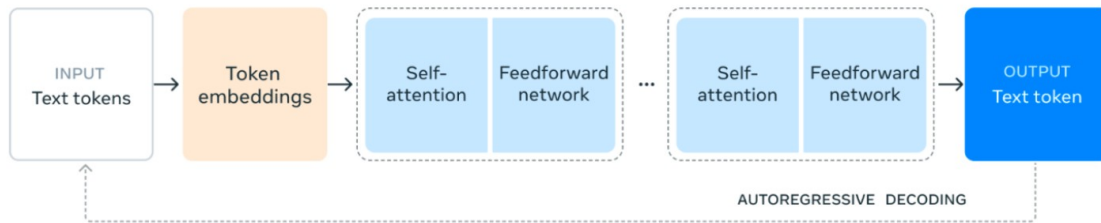


Figure 3.2: Architecture of LLaMA 3, emphasizing its transformer-based design. Source: Meta AI Dubey et al. (2024).

3.3 Retrieval-Augmented Generation (RAG)

RAG was employed in this study to enhance the classification and categorization capabilities of LLaMA models. The RAG architecture integrates a retrieval mechanism with the generative model, grounding the outputs in external knowledge sources (Lewis et al., 2020).

3.3.1 RAG Architecture with LLaMA-3.1

RAG combines the capabilities of embedding models, vector databases, and language models to generate responses grounded in external knowledge. This methodology allows for efficient

handling of large-scale datasets by retrieving relevant chunks of information and using them to generate responses. RAG provides a solution for overcoming the limitations of LLMs, such as limited token input size and the inability to access dynamic or domain-specific information during inference. Figure 3.3 illustrates the RAG pipeline, which combines document embeddings, retrieval mechanisms, and generative responses.

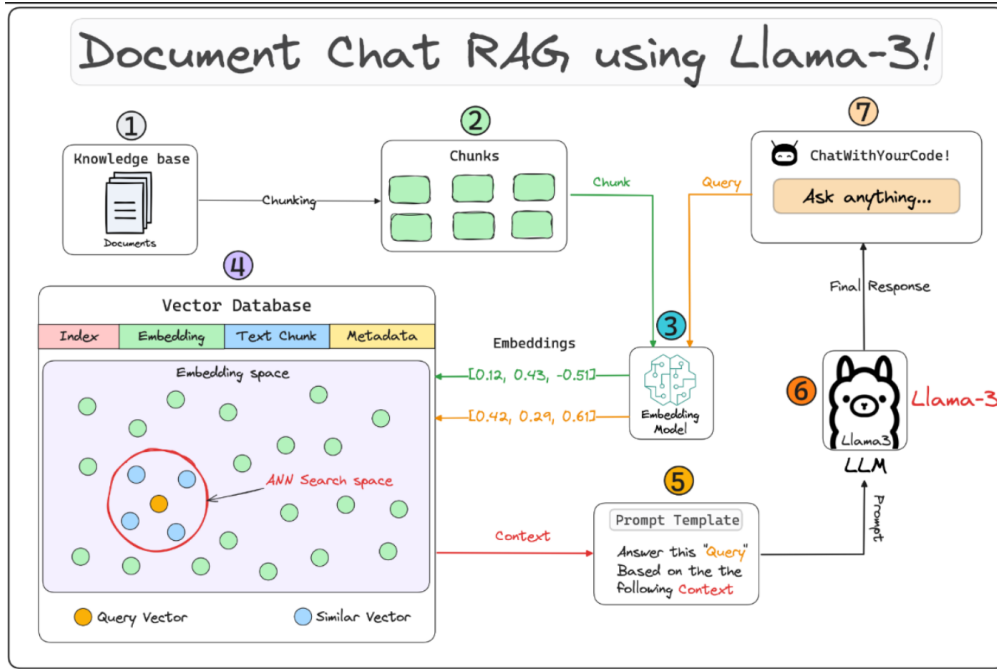


Figure 3.3: RAG architecture integrating retrieval and generation for enhanced outputs. (Lewis et al., 2020).

1. **Knowledge Base Preparation:** Documents are chunked into smaller text segments to ensure compatibility with embedding models and the LLM's context window.
2. **Embedding Generation:** Each chunk is processed through an embedding model, generating numerical vectors that capture semantic information.
3. **Vector Database Indexing:** The embeddings are stored in a vector database, enabling efficient similarity searches during query processing.
4. **Query Execution:** A user query is converted into an embedding vector and matched against the stored embeddings to retrieve the most relevant chunks.
5. **Contextual Prompting:** Retrieved chunks are integrated into the prompt sent to the LLaMA-3.1 model, ensuring the generated response is contextually grounded.

6. **Response Generation:** LLaMA-3.1 processes the prompt and retrieves the relevant chunks to produce a final response.

This architecture leverages the LLaMA-3.1 model for generating accurate, relevant, and context-aware outputs while addressing the constraints of standalone LLMs. The combination of LLaMA-3.1 and RAG ensures the scalability of the system to handle domain-specific tasks and dynamic data. By combining LLaMA models with RAG, this research leveraged the strengths of both approaches, delivering superior classification, summarization, and trend analysis results compared to traditional methods.

3.4 Complaint Classification Approaches

In this research, two distinct approaches were used for complaint classification, as shown in Figure 3.4. The classification process underwent significant changes in response to the challenges and limitations observed during the initial phase, resulting in the development of a more sophisticated method in the later stages.

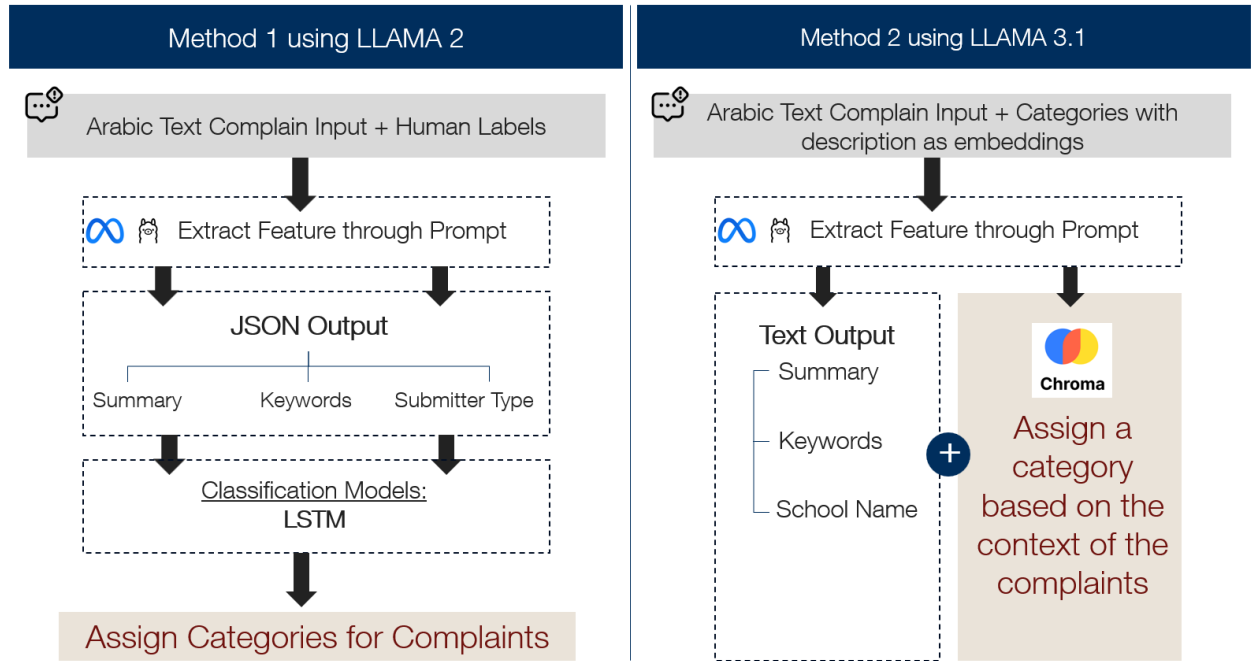


Figure 3.4: Comparison of Method 1 using LLaMA2 and Method 2 using LLaMA 3.1

3.4.1 Method 1: Classification using LLaMA 2 with LSTM

The initial classification approach involved leveraging LLaMA 2 to extract structured features from the Arabic complaint dataset, including concise summaries (s_i), keyword sets (k_i), and submitter types (u_i), using carefully engineered prompts. These features were returned in structured JSON format, simplifying the pipeline for downstream classification tasks. Each complaint t_i was thus transformed into a structured feature representation $f_i = \{s_i, k_i, u_i\}$, where s_i is a short English summary, k_i is a set of 5–10 relevant keywords, and u_i is a categorical label representing the submitter type (e.g., teacher, parent, student).

For the classification task, these textual and categorical features were encoded as input sequences and passed to a supervised learning model based on a LSTM network. The LSTM processed the input sequence $\{x_1, x_2, \dots, x_T\}$ and produced a final hidden state h_T , which was then mapped to a predefined category label using a softmax classifier:

$$\hat{y} = \text{softmax}(Wh_T + b)$$

where W and b are trainable parameters. The model was trained using a cross-entropy loss function:

$$\mathcal{L} = - \sum_{c=1}^C y_c \log(\hat{y}_c)$$

where C is the total number of complaint categories, and y_c is the ground truth label.

Despite this structured approach, the model faced significant challenges. Many human-labeled complaint categories were inconsistent or overly broad (e.g., “Other”), reducing the effectiveness of supervised learning. The extracted features, while semantically rich, lacked the granularity required for clear class separation. Additionally, the training process was computationally intensive — processing a batch of 200 records required between 1.5 and 2 hours — and the classification accuracy plateaued below acceptable thresholds. These limitations led to the development of a more scalable and context-aware solution using LLaMA 3.1 and Retrieval-Augmented Generation (RAG), discussed in the next section.

3.4.2 Method 2: Classification using LLaMA 3.1 with Retrieval-Augmented Generation (RAG)

To address the limitations observed in Method 1, the research transitioned to a more advanced approach that integrates LLaMA 3.1 with RAG) for the complaint classification task. In this

method, LLaMA 3.1 was used to directly generate structured outputs—namely summaries (s_i), keywords (k_i), and predicted categories ($\hat{y}_i$)—by grounding the model’s generation process in a retrieval-based context, rather than relying solely on human-assigned labels.

At the core of this approach is the integration of semantic embeddings and vector retrieval mechanisms. A predefined set of C category descriptions $\mathcal{D} = \{d_1, d_2, \dots, d_C\}$ was curated and stored in a JSON file. Each category description d_j was encoded into a dense vector representation $v_j = \text{Embed}(d_j)$ using the same embedding model employed by the RAG pipeline. These embeddings were indexed using Chroma, a high-performance vector store supporting similarity search via cosine distance. During inference, each input complaint t_i was transformed into an embedding $q_i = \text{Embed}(t_i)$, and the top- k most semantically relevant category vectors were retrieved from the Chroma index:

$$\mathcal{R}(q_i) = \{d_{j_1}, d_{j_2}, \dots, d_{j_k}\} \quad \text{where } \cos(q_i, v_j) \text{ is maximized}$$

These retrieved category descriptions $\mathcal{R}(q_i)$ were appended to the prompt provided to LLaMA 3.1, enabling the model to generate a grounded category prediction $\hat{y}_i$ that aligns with the semantic context of the complaint. The prompt was structured to instruct the model to choose the most relevant category based on the retrieved context.

The classification output was further refined using a scoring mechanism where the semantic similarity between the complaint embedding and the retrieved candidates was used to filter or rank category options:

$$\hat{y}_i = \arg \max_{d_j \in \mathcal{R}(q_i)} \cos(q_i, v_j)$$

This RAG-based architecture effectively mitigated several limitations observed in Method 1. It addressed label inconsistencies by anchoring predictions in standardized, curated descriptions, and eliminated the need for retraining on inconsistent human-labeled data. Moreover, the LLM was able to interpret and classify diverse complaint texts—including those in Modern Standard Arabic and local dialects—by leveraging retrieval-based grounding to improve both contextual understanding and classification fidelity.

From a performance perspective, Method 2 demonstrated significant improvements. Batch processing time was reduced to 5–6 minutes per 200 records, a dramatic improvement over the 1.5–2 hours required by the earlier method. More importantly, classification accuracy and consistency improved substantially, particularly for ambiguous or previously misclassified complaints. By incorporating domain-specific knowledge dynamically through retrieval, this

method offered a scalable and intelligent solution for handling large volumes of complaint data in real time.

3.5 Prompt Engineering and Task Implementation

Prompt engineering is a critical aspect of effectively utilizing LLMs, as the design and structure of prompts directly impact the quality and relevance of the outputs generated. Effective prompt engineering involves carefully crafting the input to guide the language model toward producing accurate, contextually relevant, and actionable outputs. By clearly defining the task, context, format, and constraints within the prompts, researchers can significantly enhance the performance of LLMs for specific use-cases, reducing ambiguities and minimizing undesirable behaviors such as hallucinations. In this research, prompt engineering was systematically employed to enhance the performance of LLaMA 2 and LLaMA 3.1 models for tasks including summarization, keyword extraction, category prediction, and entity recognition—particularly the extraction of school names. The iterative process involved continuous evaluation and adjustments to prompts based on the quality of the outputs generated, thereby optimizing model performance and reliability.

3.5.1 Summarization and Keyword Extraction

For LLaMA 2, initial prompts were designed to elicit precise and structured information. These prompts explicitly requested the model to: (1) provide a concise, high-level summary of each complaint, (2) identify the submitter type from a predefined set (parent, student, teacher, employee, or unspecified), (3) extract the school name if mentioned, and (4) list relevant keywords separated by commas. The structured JSON format required by these prompts facilitated automated parsing and efficient downstream processing, enhancing the utility of model outputs for subsequent analysis. For the specific structure used in the initial prompt design with LLaMA 2, refer to Figure 3.5.

Building upon the initial prompt engineering approach, the prompts for LLaMA 3.1 were further refined to enhance the clarity and effectiveness of the outputs. These enhanced prompts, shown in Figure 3.6, provided explicit guidelines for maintaining brevity, precision, and contextual accuracy. Unlike the strict JSON format used with LLaMA 2, the LLaMA 3.1 prompts employed a structured yet human-readable textual response format, facilitating easier integration and interpretation. Additionally, the use of OLLaMA and Apptainer tools allowed for rapid evaluation and iteration of prompts, significantly improving efficiency.

```

# Template for building the PROMPT
instruction = "Provide a very high-level summary of the complaint
submitted to ministry of education: {complaint_context}\nIdentify the
submitter type must be 1 of the following: parent,student,teacher,
employee,unspecified\n school name if it's mentioed\nList the keywords
(if any) separated by commas: keyword1,keyword2,..etc"

system_prompt = "Don't respond as a human or assistant; just follow the
instructions below and generate a response in english accordingly, and
Output the answer in JSON with literal keys (summary, submitter, school,
keywords). Only output JSON without extra word"

template = get_prompt(instruction, system_prompt)

prompt = PromptTemplate(input_variables=["complaint_context"],
template=template)

```

Figure 3.5: Prompt Template Using LLaMA2

Entity Recognition for School Name Extraction

Prompt engineering was also specifically applied to the task of extracting and accurately matching school names from complaint texts. Recognizing initial challenges related to variations in naming and potential ambiguities, prompts were designed to explicitly instruct the model to match extracted school names to a predefined reference list, thereby improving the consistency and accuracy of entity recognition. This refined prompt approach is illustrated in Figure 3.7.

Monthly Summary Analysis

The iterative prompt engineering approach was similarly applied to monthly summaries of complaints. Initially, prompts for monthly analysis (Figure 3.8) requested summaries, keywords, and trends, but outputs were sometimes overly verbose or lacked specificity. To address this, prompts were refined to emphasize conciseness, explicit instruction on keyword extraction (targeting 10–15 terms), and clarity on trend identification. These adjustments significantly improved output consistency, precision, and actionable value for stakeholders.

The refined prompt design (Figure 3.9) included explicit constraints to prevent hallucina-

```
def query_llm_for_complaint_details(complaint_text, model):
    prompt_template = """
    You are an AI assistant tasked with understanding complaints about
    education in Arabic. Carefully review the complaint below and extract
    the following details in **exactly the following format**:

    Complaint: {complaint_text}

    Instructions:
    - School Name: [Extract the school name here or write 'No school' if no
    school is mentioned]
    - Summary: [Write a one-sentence summary of the complaint in English
    here]
    - Keywords: [Extract exactly 3 to 5 relevant keywords, separated by
    commas]
    - Category: [Predict the most appropriate category for the complaint]

    **Ensure that your output strictly follows this format:**
    - School: [school name here]
    - Summary: [Summary of the complaint]
    - Keywords: [Keyword1, Keyword2, Keyword3]
    - Category: [Category here]
    """
```

Figure 3.6: Prompt Template Using LLaMA3.1

tions, instructing the model to strictly base its outputs on explicitly mentioned information in the provided complaint data. This strategic prompt design ensured high accuracy and reliability of the outputs, aligning closely with best practices in prompt engineering (Reynolds and McDonell, 2021; Ji et al., 2023).

the provided complaint data, thereby minimizing any fabricated content.

3.6 Libraries and LLM Utilization

This section discusses the Python libraries, tools, and software used to implement and run LLaMA 2 and LLaMA 3.1. Leveraging various open-source Python libraries established an effective pipeline for text analysis and trend detection in Arabic complaints. The implementation of LLMs was tailored to meet the specific requirements of this research, with each model setup evolving to improve the overall analysis quality and efficiency.

```

prompt = f"""
You are an AI assistant tasked with matching school names extracted from
complaint records to their unified official names.

Extracted School Name: "{extracted_name}"

School List:
{school_names_text}

**Instructions:**
- Find the unified school name from the provided list that most closely
matches the extracted name.
- Ignore descriptors like 'Primary', 'Boys', or 'Girls' often found in
government school names, as they are not part of the official name.
- Provide only the unified school name without additional descriptors or
assumptions.

**Output**:
Unified School Name: [Provide the exact match here]
"""

```

Figure 3.7: Prompt Template to Extract School Name

3.6.1 LLaMA 2 Implementation: LangChain and CTransformers

The primary tools employed for the LLaMA 2 model were Python libraries LangChain and CTransformers. These libraries were integral in defining and executing prompt engineering tasks, managing LLM interactions, and handling the resulting outputs.

The LangChain library was used to construct and manage prompt templates, which formed the core of the model’s input pipeline. Specifically, the PromptTemplate module from LangChain was used to create well-defined prompts that effectively guided the model in producing desired outputs. This modularity allowed for easier prompt customization, testing, and optimization throughout the research (Wang et al., 2023).

The CTransformers library provided an interface for efficiently executing quantized versions of LLMs, supporting multiple devices such as CUDA-enabled GPUs. By using the CTransformers LLM interface in Python, we were able to load and execute LLaMA 2 in an optimized manner. The library allowed efficient usage of hardware, reducing computational costs and enhancing overall inference speed (Gerganov, 2023)

Using PromptTemplate in combination with CTransformers, the LLaMA 2 model was

```

prompt = f"""
You are analyzing complaints submitted in {year}-{month}. Below is the list of complaints for this period:

Complaints:
{complaints_text}

Please analyze the complaints and provide the following information **in exactly this format**. Strictly follow the structure given below without adding any extra text.

**Required Format:**

**Summary**
Summary: [Provide a concise summary of the main issues, starting with "Summary:"]

**Keywords**
Keywords: [List 6 to 10 keywords separated by commas, starting with "Keywords:"]

**Trends**
Trends: [Describe any observed trends or patterns, starting with "Trends:"]

**Important:** Make sure to follow this format exactly. Include **only** the required information in this structure, and do not add extra sections or explanations. For example:
- Summary: [Your summary here]
- Keywords: [Keyword1, Keyword2, Keyword3, ..., Keyword10]
- Trends: [Your trends here]
"""

```

Figure 3.8: Prompt Template for Monthly Summary Analysis - Initial Version

directed to provide summaries, keywords, and submitter type information for each complaint. The integration of these Python libraries created a flexible yet powerful interface for handling prompt-based LLM queries and collecting the output for further analysis.

3.6.2 LLaMA 3.1 Implementation: Advanced Prompting and Retrieval-Augmented Category Prediction in ioPython

The LLaMA 3.1 implementation used a more sophisticated software stack in Python to not only run the model but also enable retrieval-augmented category prediction. To utilize LLaMA 3.1, the LangChain-OLLaMA integration in Python was employed. This setup used various components to efficiently extract information and handle more complex queries:

ChatOLLaMA: Used for managing LLaMA 3.1 prompts and responses. ChatOLLaMA facilitated interaction with the model by providing an easy way to run prompts while ensuring that responses adhered to expected formats and were generated efficiently.

Chroma Vector Store: The Chroma vector store was utilized as a repository for storing embeddings. Embeddings are low-dimensional representations of text data that enable efficient similarity searches. The Chroma library was used to store and retrieve embeddings of predefined categories and descriptions, which helped to accurately determine the category of a complaint.

OLLaMAEmbeddings: OLLaMAEmbeddings were used to represent the text as embeddings. These embeddings were then queried using similarity measures to find the most

```

prompt = f"""
You are an AI assistant helping to summarize a large volume of complaints received in {year}-{month} by the Ministry of Education in Bahrain.
This summary will be used for high-level visualization, so focus on capturing the main themes and recurring issues without adding any extra
information.

Complaints:
{complaints_text}

**Please follow these instructions carefully:**

1. Do not hallucinate. Only include information that is explicitly mentioned in the complaints provided. Avoid adding assumptions,
unrelated details, or any content not present in the text.
2. Provide a comprehensive summary by identifying the most frequently mentioned issues across the complaints. Capture all major themes
and recurring topics, including both prominent and less common but significant concerns, for a balanced overview.
3. List Keywords: Identify 10 to 15 unique keywords that accurately reflect the most common and relevant topics or issues.
4. Identify Trends: Instead of repeating the summary, analyze patterns by recognizing distinct topics or areas of concern that recur.
Describe high-level trends, focusing on patterns that would be useful for visualization purposes.

Provide the output in this exact format:

Summary
Summary: [Provide a comprehensive overview of the most frequently mentioned issues, starting with "Summary:"]

Keywords
Keywords: [List 10 to 15 keywords separated by commas, starting with "Keywords:"]

Trends
Trends: [Describe high-level patterns or recurring areas of concern observed in the complaints, starting with "Trends:"]

Important: Follow this format strictly. Include only the required information, and ensure all information is based directly on the provided
complaints.

Example output format:
- Summary: [Comprehensive summary here]
- Keywords: [Keyword1, Keyword2, ..., Keyword15]
- Trends: [Distinct trends and patterns observed here]
"""

```

Figure 3.9: Prompt Template for Monthly Summary Analysis - Refined Version

relevant categories for each complaint.

Retrieval-Augmented Generation (RAG) for Category Prediction:

In this research, RAG was used to predict categories. Predefined categories and short descriptions were stored in a JSON file. These categories were then embedded using Chroma and queried during prompt execution to classify the complaints. This ensured that the LLM's category predictions were grounded in a predefined knowledge base, significantly improving the reliability of the classification. The use of Chroma and RetrievalQA enabled accurate matching between the complaint context and predefined categories, thus mitigating the risk of misclassification (Lewis et al., 2020).

3.7 Evaluation Strategy

The evaluation of the model outputs was conducted using both manual assessment and automated metrics, depending on the type of output. This combined approach allowed for a thorough analysis of the quality, relevance, and accuracy of the information generated by the LLaMA 2, LLaMA 3.1 models, and LSTM classification model.

Manual Evaluation Approach:

- **Summaries and Keywords:** In this study, we assess the relevance of English summaries and keywords to the original Arabic descriptions by using multilingual embeddings generated from the LaBSE (Language-agnostic BERT Sentence Embedding) model. LaBSE is well-regarded for its performance in cross-lingual tasks and enables the comparison of Arabic descriptions with English text in a unified semantic space Feng et al. (2020). The process involved generating embeddings for the Arabic “Complaint Text” the English “summary,” and the English “keywords.” Cosine similarity was then used to quantify the semantic relevance between these components. The mean, median, standard deviation, minimum, and maximum of the cosine similarity scores were calculated to provide a comprehensive understanding of the alignment between the Arabic complaint text and the English summaries and keywords. This approach ensured that we had a numerical representation of the quality of the translations, reflecting their semantic accuracy and consistency.
- **Categories (LSTM Model):** The LSTM-based classification was evaluated using standard classification metrics such as precision, recall, F1-score, and accuracy.
- **Categories (LLaMA 3.1 with RAG):** A manual evaluation was conducted to verify the correctness of the categories predicted by LLaMA 3.1 using Retrieval-Augmented Generation. Evaluators were provided with the same predefined list of categories and their textual descriptions that were embedded and used in the model prompts. These human evaluators were individuals already familiar with the Tawasul complaint classification scheme, having previously categorized complaints as part of their official responsibilities. They reviewed the model’s predictions and marked whether each classification was accurate or not based on contextual appropriateness and semantic alignment with the complaint content.
- **Trends:** Identified trends were reviewed manually to ensure they accurately reflected recurring concerns without hallucinations or misrepresentation of complaint intent. Reviewers cross-checked the grouped summaries, keywords, and frequencies to validate the themes against the original data.

Manual evaluation was essential for LLM-generated outputs, especially due to the complexity of working with bilingual and multidialectal texts (Modern Standard Arabic and local dialects). The use of trained evaluators, who were already involved in complaint classification, ensured reliable benchmarking of LLM performance relative to human standards.

3.8 Visualization and Trend Detection

The final step in the research workflow was visualizing the processed data and detecting trends in the complaint dataset. This stage aimed to present the extracted insights in a way that was both clear and easy to interpret, enabling stakeholders to understand recurring concerns, emerging patterns, and overall trends within the complaint data effectively. To achieve this, a combination of JavaScript libraries, specifically D3.js and Chart.js, was utilized to create interactive and engaging visual representations of the data. These visualization tools allowed for multiple perspectives on the trends identified during the LLM analysis, ensuring a holistic understanding of the results.

The following visualization methods were applied:

1. **Monthly Analysis Chart:** This chart displayed the volume of complaints over time, allowing for the identification of key periods where complaint numbers peaked, such as in August and September. Understanding these spikes was crucial for linking the frequency of complaints to possible seasonal events or policy changes.
2. **Category Filtering:** A Category Filter was integrated into the dashboard, allowing users to select specific complaint categories, such as “Student Transfers” or “Teacher Issues.” This enabled stakeholders to conduct a more targeted investigation into specific concerns. The ability to filter by category added an important layer of granularity, helping to understand the distribution of different complaint types over time.
3. **Bar Charts and Line Graphs:** These visualizations were used to illustrate trends across different categories of complaints. For example, monthly counts of complaints by type were plotted to identify which challenges were recurring or changing across the year. This approach was instrumental in highlighting seasonal variations and provided stakeholders with a clearer understanding of the dynamics of various complaint types.
4. **Monthly Summary Section:** The Monthly Summary provided stakeholders with concise insights for each selected month, including a summary of complaints, relevant keywords, and trends. This allowed for easy identification of recurring concerns and evolving trends, making it easier to communicate findings to decision-makers.
5. **Word Clouds:** Word clouds were generated from the keywords extracted by the LLMs. This visualization offered an intuitive and visually appealing way to identify the most

frequently occurring topics within the complaints dataset. By presenting these keywords in a larger font size depending on their frequency, word clouds enabled stakeholders to quickly grasp the dominant concerns without delving into detailed data.

6. Heatmaps: Heatmaps were employed to visualize the geographic concentration of complaints, particularly for complaints related to school locations. The intensity of color indicated areas with a higher number of complaints, effectively highlighting potential hotspots that might require additional attention or intervention. This spatial analysis provided valuable insights that couldn't be captured through text analysis alone.

The combination of these interactive visualization tools and filtering options ensured that the stakeholders had the ability to explore trends dynamically and focus on specific challenges as needed. This visual approach was crucial for effectively implementing trend analysis and ensuring that the findings were accessible, understandable, and actionable for a wide range of users.

Chapter 4

Results and Discussion

4.1 Data Exploration Findings

The initial exploration of the Tawasul complaint dataset provides insights into the distribution and characteristics of various complaint categories, as illustrated in the Figure 4.1. This analysis reveals patterns in complaint frequencies, highlights areas for data quality improvement, and suggests potential refinements for future categorization efforts.

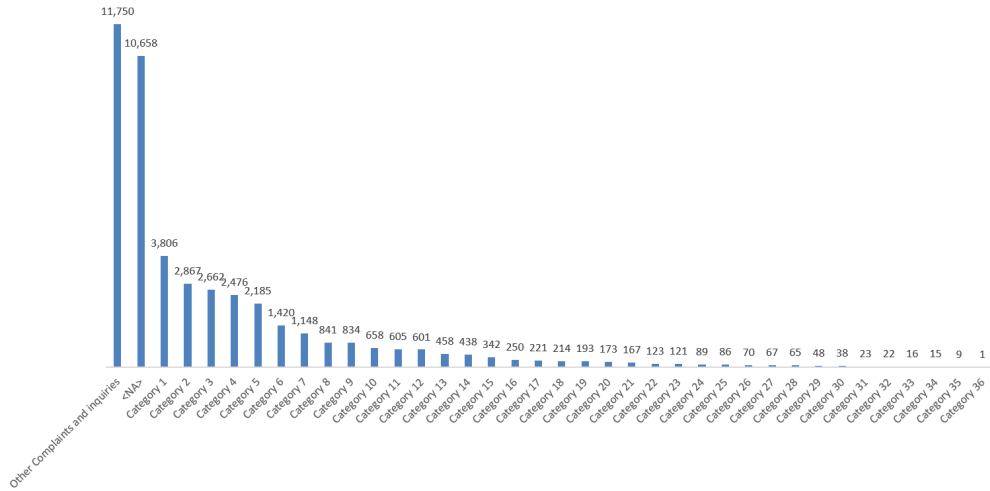


Figure 4.1: Distribution of Complaints by Category for Chosen Entity

4.1.1 Dominance of “Other Complaints and Inquiries”

The dataset is significantly skewed toward the “Other Complaints and Inquiries” category, which contains 11,750 complaints, accounting for a substantial portion of the overall dataset. This category’s prominence indicates that a large number of complaints did not fit neatly into predefined categories, pointing to limitations in the existing categorization schema. This catch-all category potentially includes a diverse range of issues, making it difficult to draw specific insights. The large volume of complaints in “Other” suggests an opportunity to re-evaluate and possibly expand the predefined categories or utilize advanced NLP techniques to reassign complaints to more descriptive categories. By refining this classification, stakeholders can gain a clearer picture of recurring issues and concerns within the dataset.

4.1.2 Presence of Unclassified or Missing Data (Indicated as “NA”)

The second most common category in the dataset is labeled as “NA,” containing 10,658 complaints. This category represents complaints that were submitted without a specified category, which could indicate gaps in the data entry or classification process by humans. The high number of complaints in “NA” suggests a need for improved data quality controls to ensure all entries are appropriately categorized. Addressing these missing classifications would enhance the dataset’s consistency and allow for more accurate trend analysis. Re-classifying these complaints using LLMs could provide more meaningful insights and reduce potential biases in the categorization process.

4.1.3 High-Frequency Categories

After “Other Complaints and Inquiries” and “NA”, the next most frequent categories—Category 1 with 3,806 complaints, Category 2 with 2,867 complaints, and Category 3 with 2,662 complaints—also represent a significant portion of the dataset. These high-frequency categories highlight recurring themes or issues that are likely of primary concern to the public and may correlate with common challenges faced by the Ministry of Education. These categories should be further analyzed to identify specific trends, seasonal spikes, or commonalities that might inform policy changes or resource allocation. Understanding these top categories in more detail will help prioritize interventions that could address the most frequently reported issues.

Long Tail of Niche Categories

Beyond the high-frequency complaint categories, the dataset displays a long tail of categories with relatively low counts, representing more specialized or less common concerns. This distribution pattern reveals that while some issues affect a broad audience, others are unique to specific groups or circumstances. These niche categories may include operational, administrative, or specialized educational topics that, while not frequently reported, provide insight into particular areas that require attention.

The presence of a long tail in the data emphasizes the need for advanced NLP techniques to ensure that even low-frequency complaints are accurately captured and analyzed. Although each of these niche categories has fewer complaints, the issues they represent could be of critical importance to specific stakeholders, such as certain student demographics or regions. Addressing these smaller but significant categories can offer a comprehensive view of the public’s concerns and help identify potential areas for targeted policy adjustments or resource allocation.

The diversity of issues within these niche categories also demonstrates the complexity and range of challenges within the education sector. By leveraging LLMs for in-depth analysis, this research aims to ensure that less common concerns are not overlooked in favor of high-frequency issues, thus supporting a more nuanced understanding of the dataset and allowing for thorough trend analysis.

4.1.4 Insights for Future Categorization and Trend Analysis

The exploration of complaint distributions in the Tawasul dataset underscores the importance of accurate categorization for effective trend analysis. The reliance on a broad “Other” category, combined with the presence of unclassified data in the “iNAi” category, suggests a need to refine the categorization framework to better capture the diversity of complaints. Advanced NLP methods, such as using large language models (LLMs) to perform reclassification or sub-categorization, could enhance the dataset’s usability. Additionally, focusing on the high-frequency categories can provide actionable insights into widespread issues, while niche categories may highlight emerging or specialized concerns. A more structured and consistent categorization approach will support the Ministry of Education and other stakeholders in identifying trends, allocating resources, and addressing citizen concerns more effectively.

4.2 Performance Comparison: Method 1 (LLaMA2) vs. Method 2 (LLaMA3.1)

The results from Method 1 and Method 2 offer critical insights into the effectiveness of different approaches for categorizing and analyzing complaint data within the context of Arabic language processing. The evaluation of Method 1 and Method 2 focused on two key performance aspects: classification accuracy and the quality of generated summaries and keywords. Each method was assessed to determine its effectiveness in handling the diverse and complex complaint data in Arabic, with distinct outcomes that highlight the comparative strengths and weaknesses of each approach.

4.2.1 Evaluation of Summaries and Keywords

The relevance scores presented in the table 4.1 were computed using the LaBSE model, which provides cosine similarity metrics for multilingual sentence embeddings. These scores objectively indicate the similarity between summaries/keywords and the original text.

Table 4.1: Comparison of Summary and Keywords Relevance Scores Between LLaMA 3.1 and LLaMA 2

Metric	Summary Relevance		Keywords Relevance	
	LLaMA 3.1	LLaMA 2	LLaMA 3.1	LLaMA 2
Mean	0.677450	0.711930	0.610944	0.629152
Median	0.676911	0.724378	0.609627	0.631556
Standard Deviation	0.073878	0.100643	0.077578	0.079901
Minimum	0.224693	0.290582	0.247235	0.307975
Maximum	0.923501	0.972844	0.978261	0.920019

However, after conducting a manual human evaluation, it was observed that LLaMA 3.1 generally provides more accurate and contextually relevant summaries than LLaMA 2. This subjective preference may be attributed to LLaMA 3.1’s nuanced handling of language complexities and specific terminology.

Due to the intricate nature of language, especially in multilingual contexts, capturing all subtle differences in relevance can be challenging through automated cosine similarity alone.

The human assessment also noted that LaBSE may not fully capture certain contextual nuances that were better represented in LLaMA 3.1’s outputs.

To illustrate these differences, the following examples showcase summaries generated by both LLaMA 3.1 and LLaMA 2:

In the first example, Table 4.2, although LLaMA 2 achieved higher relevance scores, LLaMA 3.1 generated a summary that more accurately captures the context of the complaint. LLaMA 3.1 better reflects the student’s concern about failing a specific subject and their inquiry about exam eligibility. However, it still falls short in precisely identifying that the complaint is related to a math course and that the question is specifically about repeating the exam, not advancing to the next academic year. Despite these limitations, LLaMA 3.1 provides a summary closer to the intended meaning of the complaint.

Table 4.2: Comparison of LLaMA 3.1 and LLaMA 2 Summaries and Keywords for a Complaint Text -Example 1

Complaint Text		
السلام عليكم، اختي رايحة صف ثاني ثانوي (ادي) ورسبت في مقرر رياض ١٥١ فهل بتكون مشمولة في الدور الثالث تاريخ ٧		
LLaMA	LLaMA 3.1	LLaMA 2
Summary	A student complains about failing a subject and inquires if they will be allowed to proceed to the next year.	A parent has submitted a complaint to the Ministry of Education regarding their child’s second-grade class (English) and the teacher’s inability to provide adequate support during the third term (July 9th).
Summary Relevance	0.501995206	0.516551554
Keywords	education, student, complaint, school, year	registration, certificates, graduation, authorities, thanking
Keywords Relevance	0.498060286	0.544336319

In the second example, Table 4.3, LLaMA 3.1 successfully interprets the complaint regarding the student’s inability to access their results, providing a specific and contextually accurate summary. In contrast, LLaMA 2 generates a general summary that does not accurately reflect the complaint’s context. The keywords generated by LLaMA 3.1—such as

“accessibility,” “results,” “online system,” and “student issue”—all directly reflect the core issues in the complaint. However, it’s clear that the model did not fully recognize that the “online system” referred to here is specifically “edu.net”. Notably, LLaMA 3.1 demonstrates a robust understanding of the content even when faced with extra characters, typos, and local dialect variations, handling these imperfections without requiring pre-processing.

Table 4.3: Comparison of LLaMA 3.1 and LLaMA 2 Summaries and Keywords for a Complaint Text - Example 2

Complaint Text		
\@ \@ ؟\@ /ممكن تفعلونها		
LLaMA	LLaMA 3.1	LLaMA 2
Summary	The student is unable to access their results.	a complaint was submitted to the ministry of education regarding a specific incident or issue.
Summary Relevance	0.562816501	0.42122072
Keywords	accessibility, results, online system, student issue	incident, issue
Keywords Relevance	0.614499807	0.534162164

Table 4.4 compares the summaries and keywords generated by LLaMA 3.1 and LLaMA 2 for a complaint about malfunctioning air conditioning in a school. While LLaMA 3.1 provides a more specific and contextually accurate summary—emphasizing the extreme heat in classrooms and its impact on the students’ learning environment—LLaMA 2 delivers a more generalized summary that highlights “poor air conditioning in all girls’ schools” but lacks specificity to the direct concerns raised in the complaint. Despite the contextual superiority of LLaMA 3.1, its relevance scores are lower than those for LLaMA 2, indicating a limitation in the relevance scoring metric. This discrepancy likely stems from the inherent complexity

of Arabic, including its diverse phrasing, dialectical variations, and unique linguistic features, which automated metrics may struggle to fully capture.

When it comes to keywords, both models produce largely similar outputs, with LLaMA 3.1 focusing on “air conditioning,” “classroom conditions,” and “students’ wellbeing,” while LLaMA 2 includes broader terms such as “schools” and “ministry of education.” However, the relevance scores for keywords remain higher for LLaMA 2, further underscoring the inadequacy of automated scoring in accurately reflecting the quality of Arabic model outputs. This highlights the necessity of manual evaluation to ensure a more reliable assessment of context and alignment with the complaint’s intent.

Table 4.4: Comparison of LLaMA 3.1 and LLaMA 2 Summaries and Keywords for a Complaint Text - Example 3

Complaint Text: مدرسة - اسم مدرسة - الثانوية للبنات جميع مكيفات الصفوف تخرج هواء حار كيف ندرس و نركز في الحصة و المكيفات تخرج هواء حار ! اطالب إعادة صيانة المكيفات في اسرع وقت لانها ليست بيئة تعليم.		
LLaMA	LLaMA 3.1	LLaMA 2
Summary	The complainant is dissatisfied with the extreme heat in their school classrooms due to malfunctioning air conditioning units.	Complaint submitted to the ministry of education regarding poor air conditioning in all girls’ schools, with a focus on addressing the issue as soon as possible.
Keywords	air conditioning, classroom conditions, education, heating, students’ wellbeing	air conditioning, poor, schools, girls, ministry of education, complaint, urgent action
Summary Relevance	0.644093812	0.748834491
Keywords Relevance	0.62213099	0.772108853

By examining complaints with lower relevance values (0.2 and 0.3), it was observed that 0.2 relevance values correspond to cases where complaints primarily include reference numbers or contact information without any substantive text describing the issue. These complaints lack meaningful content for summarization. On the other hand, 0.3 relevance values are associated with complaints written in English. Since this experiment focuses on Arabic-language complaints, it is essential to first filter and split these English complaints from the dataset. This preprocessing step ensures that the relevance evaluation and model performance analysis remain focused on Arabic complaints, aligning with the scope of the experiment.

4.2.2 Classification Performance

The classification process aimed to categorize complaints into predefined categories based on summaries and keywords generated by LLaMA2 and LLaMA3.1. Method 1 (LLaMA2 with LSTM) was the initial approach, but due to its limitations in performance and efficiency, Method 2 (LLaMA3.1 with RAG) was later introduced to address these shortcomings and achieve better results.

Method 1: LLaMA2 with LSTM Classification

Method 1 utilized the LLaMA2 model to extract features, including summaries, keywords, submitter types, and school names, which were then used as input features for a LSTM model to classify complaints into predefined categories. This approach served as an initial attempt to automate complaint categorization using advanced NLP techniques. However, the results revealed several limitations, which ultimately motivated the development of a more advanced methodology.

Feature Extraction Using LLaMA2 The feature extraction process involved generating summaries and keywords for each complaint using LLaMA2. Additionally, the model extracted submitter type (e.g., parent, student, teacher) and school names where mentioned. These features were intended to enhance classification by providing diverse contextual information about each complaint. However, the evaluation (discussed earlier in Section 4.2) revealed that the generated summaries and keywords lacked the necessary specificity to effectively distinguish between complaint categories. For instance:

- Summaries were often generalized, failing to capture nuanced details essential for identifying specific complaint categories.

- Keywords frequently overlapped between different categories, reducing their discriminative power.
- Submitter type and school name, while extracted successfully in many cases, did not significantly improve classification performance due to their limited relevance to most complaint categories.

Classification Performance The extracted features were used as inputs to an LSTM model trained to classify complaints. The model’s performance metrics are summarized below:

- **Overall Accuracy:** The LSTM model achieved an overall accuracy of 49.6%, indicating moderate success but significant room for improvement.
- **Macro F1-Score:** The macro F1-score of 0.34 indicates the model’s difficulty in effectively handling underrepresented categories. This metric calculates the F1-score for each class independently and then takes their unweighted average, so all classes contribute equally regardless of their frequency in the dataset. Consequently, the low macro F1-score highlights the LSTM model’s inability to effectively classify complaints from minority classes (Géron, 2019).
- **Weighted F1-Score:** The weighted F1-score was 0.49, indicating slightly better performance when accounting for class imbalance. This metric also calculates the F1-score for each class but weights each score by the number of samples in that class. As a result, the weighted F1-score is dominated by the model’s performance on the majority classes (Scikit-Learn, 2023).

The discrepancy between the macro and weighted F1-scores demonstrates that the model performed reasonably well for majority classes but struggled significantly with minority classes. This imbalance is a common challenge in classification tasks involving datasets with uneven distributions (Johnson and Khoshgoftaar, 2019).

Challenges with Minority Classes As noted in the data exploration findings (Section 4.1), the dataset exhibited substantial class imbalance, with certain categories containing very few samples. The LSTM model struggled to learn meaningful patterns for these minority classes, as reflected in their near-zero F1-scores. This limitation was compounded by the lack of highly specific and discriminative features generated by LLaMA2.

Impact of Feature Types The inclusion of submitter type and school name as input features did not significantly improve classification performance. These features were often irrelevant to many complaint categories and inconsistently extracted. For instance:

- Submitter type was only relevant for specific complaint types (e.g., student-related issues) but added little value for broader categories.
- School names were inconsistently mentioned in the dataset and often extracted with errors, reducing their utility in improving classification outcomes.

Computational Inefficiency Processing complaints using LLaMA2 and training the LSTM model was computationally expensive and time-consuming. Each batch of 200 records required 1.5–2 hours to process, highlighting the scalability challenges of this approach for large datasets, such as the 45,000 complaints analyzed in this study.

Conclusion for Method 1 While Method 1 represented an initial attempt to integrate LLMs with traditional machine learning models, it demonstrated significant limitations. The combination of low-quality features, poor performance on minority classes, and computational inefficiency underscored the need for a more robust and scalable approach. These challenges motivated the transition to Method 2, which employed LLaMA3.1 and Retrieval-Augmented Generation (RAG) to address these issues and achieve significantly better results, as discussed in the following section.

Method 2: LLaMA3.1 with RAG for Classification

Building upon the lessons learned from Method 1, Method 2 employed LLaMA3.1 combined with RAG to overcome the challenges observed in the initial approach. This method not only retained the extraction of summaries and keywords but also introduced a refined classification mechanism by leveraging contextual embeddings for predefined categories. To evaluate its performance, a sample of 2,000 complaints was analyzed, with human categorizations serving as a benchmark for comparison.

Evaluation of Human Categorization An analysis of human-assigned categories provided valuable insights into the inherent challenges of complaint classification. Out of the 2,000 complaints evaluated:

- **Correctly Categorized Complaints:** Approximately 72% of the complaints were correctly assigned to their respective categories by human evaluators.

- **Incorrectly Categorized Complaints:** Around 23% of the complaints were misclassified due to ambiguities or errors in interpreting the complaint details.
- **Ambiguous Complaints in the “Others” Category:** Roughly 5% of the complaints were placed into the “Others” category due to a lack of alignment with predefined categories. These complaints often involved overlapping themes or insufficient context.

This analysis highlighted the challenges human reviewers faced, particularly with complaints that lacked clear descriptions or fell outside the scope of predefined categories. It also underscored the potential of LLMs to complement human efforts by addressing these ambiguities.

First Run of LLaMA3.1 with RAG In its initial implementation, LLaMA3.1 with RAG demonstrated promising results but also faced notable limitations:

- **Accurately Categorized Complaints:** Around 69% of the complaints were correctly assigned to their intended categories. These included well-defined complaints that matched closely with the predefined category descriptions.
- **Misclassified Complaints:** Approximately 31% of the complaints were misclassified, either due to incorrect matches with predefined categories or overgeneralized interpretations of the complaint text.
- **Complaints Assigned to “Others”:** Complaints categorized as “Others” represented 5% of the total, with an error rate of 69%. These results highlighted the model’s difficulty in handling ambiguous complaints that did not align with any predefined category.

The results from the first run highlighted the need for improvements in both the predefined categories and the LLM prompts. Specifically, ambiguous complaints required more comprehensive category descriptions and fallback mechanisms to ensure accurate classification.

Refinement and Second Run Results To address the challenges identified in the first run, the following refinements were introduced:

- **Enhanced Category Descriptions:** Predefined categories were revised to include more detailed descriptions that captured recurring complaint themes observed in the first run.

- **Introduction of a General “Others” Category:** A clearly defined “Others” category was introduced for complaints that did not fit any predefined category. This provided a structured fallback for ambiguous cases.
- **Improved Prompts for LLM Categorization:** Prompts were refined to emphasize key contextual details, enabling the LLM to better match complaints to relevant categories.

The results of the second run demonstrated significant improvements:

- **Accurately Categorized Complaints:** The proportion of correctly categorized complaints increased to 78%, reflecting the impact of the enhanced category descriptions and refined prompts.
- **Misclassified Complaints:** The proportion of misclassified complaints decreased to 7%, indicating substantial progress in reducing errors.
- **Complaints Assigned to “Others”:** The error rate within the “Others” category dropped to 15%, demonstrating improved handling of ambiguous cases.

Comparison Between Human and LLM Categorization The evaluation of human and LLM categorizations revealed that even human reviewers faced challenges with this dataset, leading to notable errors:

- **Errors in Human Categorization:** Approximately 23% of human categorizations were found to be incorrect upon manual review. These errors were primarily attributed to overlapping themes in complaints, ambiguous descriptions, or insufficient contextual information provided by the submitter.
- **Handling of Ambiguous Complaints:** Both human reviewers and the LLM struggled with complaints that lacked clarity or sufficient detail. The “Others” category, in particular, represented a fallback for complaints that did not align with any predefined category.
- **Improved LLM Performance in the Second Run:** After refinements, the LLM demonstrated greater consistency in handling ambiguous complaints compared to human categorizations. The explicit introduction of a well-defined “Others” category and enhanced category descriptions helped reduce errors in the LLM’s outputs.

- **Clear and Well-Defined Complaints:** Both human reviewers and the LLM performed well on complaints with clear and well-defined descriptions, achieving high levels of accuracy. However, the refined LLM in the second run outperformed humans in terms of consistency and error reduction, particularly for edge cases.

Examples of Misclassification and Contextual Challenges While Method 2 with LLaMA3.1 and RAG demonstrated significant improvements in summarization and categorization accuracy, a few examples illustrate persistent challenges, particularly with nuanced or contextually complex complaints. Below, we present three cases that highlight different types of errors and their implications.

Example 1: Misclassification of Category A complaint regarding reconsideration of grade 3 students’ final exams due to limited capabilities following two years of online learning caused by the COVID-19 pandemic illustrates category misclassification.

- **Original Complaint Text (Arabic):**

السلام عليكم ورحمة الله وبركاته الرجاء إعادة النظر في قرار إعطاء طلاب الصف الثالث الابتدائي امتحانات نهائية هذه المرحلة بدأت الدراسة في جائحة كورونا، وبقي التعليم عن بعد لمدة سنتين، والآن في هذه السنة عاد التدريس الحضوري، هم طلاب الصف الثالث لكن قدراتهم صف الأول الابتدائي، الرجاء النظر في حال الطلاب من هذه الفئة فهم يحتاجون أكثر مراعاة للظروف التي مروا فيها. وشكراً لكم.

- **Correct Category:** “*Tests and Exams*”
- **LLM Summary:**

”The student is requesting reconsideration of the decision to give students in grade 3 final exams due to their limited capabilities after two years of online learning.”

- **LLM Category:** “*Curriculum Remarks*” (Incorrect)
- **Keywords:** *education, exams, online learning, COVID19, students* (Accurate but general)

In this case, the LLM accurately summarized the complaint and generated relevant keywords but misclassified the category as “Curriculum Remarks.” The issue likely stemmed from the broad interpretation of the complaint text, which, while overlapping with curriculum concerns, was primarily focused on final exams.

Example 2: Missing Keywords Despite Correct Categorization Another example involved a complaint about a congested exam schedule for students in a specific academic track, with specific mentions of subjects such as biology, chemistry, and geography. While the LLM correctly categorized the complaint, it failed to capture the key subject-specific details in the extracted keywords.

- **Original Complaint Text (Arabic):**

السلام عليكم ورحمة الله وبركاته معكم طالبو من المرحلة الثاني ثانوي كيم حيا أحب أن أثني على جهودكم الجليلة و لكن لدي طلب وهو إعادة النظر في جدول الامتحانات النهائية للثاني الثانوي كيم حيا للفصل الدراسي الثاني حيث أن الجدول مضغوط و لا نجد الوقت الكافي للمذاكرة و خاصة أن المواد بها كم معرفي كثير و تحتاج للاسترجاع و حل المسائل و خاصة للمواد الكيم و الحيا و الحيو فوفقاً بنا فالرجاء ثم الرجاء بإعادة تنظيم الجدول بما يتناسب مع الكم المعرفي و طول المحتوى الدراسي لكل مادة فيحبذ ترك مجال للدراسة بين مادتي الكيم و الحيا و الحيو و خاصة بأننا نرجع من الامتحان الفترة الثانية و نعاني من الإجهاد فكيف يتسنى لنا المذاكرة للمادة الأخرى في نصف يوم و تقبلوا منا خالص الشكر و التقدير.

- **Correct Category:** “*Tests and Exams*” (Correct)

- **LLM Summary:**

“The student complains that the final exam schedule for Kim Hayy High School is too congested, leaving insufficient time for studying.”

- **Keywords:** *congested schedule, insufficient study time, educational pressure, high school exams, student complaints* (Accurate but missing subject-specific details like biology, chemistry, and geography)

Although the LLM correctly categorized and summarized the complaint, it failed to capture critical details—such as the mentioned subjects—which would have provided a more comprehensive understanding of the issue.

Examples of Misclassification and Contextual Challenges While Method 2 with LLaMA3.1 and RAG demonstrated significant improvements in summarization and categorization accuracy, a few examples illustrate persistent challenges, particularly with nuanced or contextually complex complaints. Below, we present three cases that highlight different types of errors and their implications.

Example 1: Misclassification of Category A complaint regarding reconsideration of grade 3 students’ final exams due to limited capabilities following two years of online learning caused by the COVID-19 pandemic illustrates category misclassification.

- **Original Complaint Text (Arabic):**

السلام عليكم ورحمة الله وبركاته الرجاء إعادة النظر في قرار إعطاء طلاب الصف الثالث الابتدائي امتحانات نهائية هذه المرحلة بدأت الدراسة في جائحة كورونا، وبقي التعليم عن بعد لمدة سنتين، والآن في هذه السنة عاد التدريس الحضوري، هم طلاب الصف الثالث لكن قدراتهم صف الأول الابتدائي، الرجاء النظر في حال الطلاب من هذه الفئة فهم يحتاجون أكثر مراعاة للظروف التي مروا فيها. وشكرًا لكم.

- **Correct Category:** “*Tests and Exams*”
- **LLM Summary:**

”The student is requesting reconsideration of the decision to give students in grade 3 final exams due to their limited capabilities after two years of online learning.”

- **LLM Category:** “*Curriculum Remarks*” (Incorrect)
- **Keywords:** *education, exams, online learning, COVID19, students* (Accurate but general)

In this case, the LLM accurately summarized the complaint and generated relevant keywords but misclassified the category as “Curriculum Remarks.” The issue likely stemmed from the broad interpretation of the complaint text, which, while overlapping with curriculum concerns, was primarily focused on final exams.

Example 2: Missing Keywords Despite Correct Categorization Another example involved a complaint about a congested exam schedule for students in a specific academic track, with specific mentions of subjects such as biology, chemistry, and geography. While the LLM correctly categorized the complaint, it failed to capture the key subject-specific details in the extracted keywords.

- **Original Complaint Text (Arabic):**

السلام عليكم ورحمة الله وبركاته معكم طالبو من المرحلة الثاني ثانوي كيم حيا أحب أن أثنى على جهودكم الجليلة و لكن لدي طلب وهو إعادة النظر في جدول الامتحانات النهائية للثاني الثانوي كيم حيا للفصل الدراسي الثاني حيث أن الجدول مضغوط و لا نجد الوقت الكافي للمذاكرة و خاصة أن المواد بها كم معرفي كثير و تحتاج للاسترجاع و حل المسائل و خاصة للمواد الكيم و الحيا و الحيو فرفقاً بنا فالرجاء ثم الرجاء بإعادة تنظيم الجدول بما يتناسب مع الكم المعرفي و طول

المحتوى الدراسي لكل مادة فيجب ترك مجال للدراسة بين مادتي الكيمياء والحيا و الحيو
و خاصة بأننا نرجع من الامتحان الفترة الثانية و نعاني من الإجهاد فكيف يتسنى لنا
المذاكرة للمادة الأخرى في نصف يوم و تقبلوا منا خالص الشكر و التقدير.

- **Correct Category:** “*Tests and Exams*” (Correct)

- **LLM Summary:**

”The student complains that the final exam schedule for Kim Hayy High School is too congested, leaving insufficient time for studying.”

- **Keywords:** *congested schedule, insufficient study time, educational pressure, high school exams, student complaints* (Accurate but missing subject-specific details like biology, chemistry, and geography)

Although the LLM correctly categorized and summarized the complaint, it failed to capture critical details—such as the mentioned subjects—which would have provided a more comprehensive understanding of the issue.

Example 3: Misinterpretation of Contextual References The third example highlights the impact of culturally specific references on LLM performance. In this case, a complaint about the heavy syllabus for third-grade elementary students included a Prophetic Hadith (saying), *أرحموا من في الأرض يرحمكم من في السماء*. This led the LLM to misinterpret the complaint as being related to religious studies rather than general textbooks.

- **Original Complaint Text (Arabic):**

حرام عليكم ارحموا من في الأرض يرحمكم من في السماء صف ثالث ابتدائي من أين
لهم طاقة تستوعب دراسة الكتاب كاملاً كفاية أنهم من أول ابتدائي لي ثاني ابتدائي
عن بعد الدراسة كفاية التقويمات الأسبوعية. الرجاء النظر لموضوع الامتحانات
النهائية والرافة بحال الأطفال وأولياء الأمور.

- **Correct Category:** “*Tests and Exams*” (Correct)

- **LLM Summary:**

“The complainant expresses concern about the ability of third-grade elementary students to study the entire Quran and suggests that their syllabus is too heavy.”

The summary incorrectly inferred that the complaint was about the Quran, likely due to the mention of the Hadith.

- **LLM Category:** *“Other Complaints and Inquiries”* (Incorrect)
- **Keywords:** *elementary education, Quran study, heavy syllabus, student stress* (Inaccurate focus on “Quran study”)

In this case, the inclusion of the Hadith, while unrelated to the core issue, caused the LLM to associate the complaint with religious content rather than exams. This misinterpretation illustrates the need for greater contextual sensitivity in LLMs when handling culturally rich texts.

Insights and Lessons Learned These examples highlight several challenges and opportunities for improvement:

- **Category Refinement:** Categories should be explicitly defined to reduce ambiguity and overlap. For example, the distinction between “Curriculum Remarks” and “Tests and Exams” should be clarified.
- **Keyword Precision:** Improved keyword extraction mechanisms are needed to capture specific details, such as subject names in Example 2, and to avoid misleading keywords, as seen in Example 3.
- **Contextual Understanding:** LLMs must be trained to recognize when cultural or religious references are unrelated to the primary complaint context.
- **Prompt Engineering:** Prompts should explicitly guide the LLM to prioritize the primary focus of the complaint while disregarding extraneous details or expressions.

These cases underscore the iterative nature of improving NLP models for complex datasets, particularly those containing nuanced, culturally specific, or ambiguous text. By refining prompts, categories, and keyword extraction techniques, future iterations can address these challenges more effectively.

Conclusion Method 2 demonstrated substantial advancements in complaint categorization compared to both Method 1 and human categorizations. By leveraging RAG with LLaMA3.1 and introducing iterative refinements, the method achieved higher accuracy and reduced error rates, particularly for ambiguous and undefined complaints. These results highlight the potential of advanced LLMs to complement and enhance traditional approaches to complaint analysis, offering scalable solutions for handling large and diverse datasets.

4.2.3 Entity Recognition and Matching of Schools

Identifying specific school names from complaint texts is a critical aspect of this research. This task, referred to as entity recognition, is essential for determining which institutions are most frequently associated with public concerns and allows targeted action. The process involved using LLaMA 3.1 with tailored prompts to extract school names, followed by matching these names to a predefined list of 276 validated school names to ensure consistency and standardization.

Entity Recognition Process The entity recognition process relied on a prompt designed to extract the name of the school mentioned in the complaint or return “No School” if no specific school was mentioned. This prompt was evaluated on a sample of 200 complaints, demonstrating that LLaMA 3.1 correctly identified schools in **95% of cases** where schools were explicitly mentioned. In complaints without a specific school, the model reliably returned “No School,” ensuring robustness across different scenarios.

However, certain mismatches occurred during extraction due to variations in how school names were presented. A recurring issue involved extracted responses containing additional phrases, such as “School name is: [name],” which caused inconsistencies during subsequent matching.

Matching Extracted School Names After extracting school names, the matching process aimed to align the extracted names with a predefined list of schools. This step was critical for resolving ambiguities and standardizing the data. Three approaches were evaluated:

Approach 1: RAG with Predefined Embeddings The first approach used Retrieval-Augmented Generation (RAG) by embedding the predefined list of school names and querying the extracted names against these embeddings. While conceptually robust, this method struggled with minor variations in naming conventions, such as abbreviations or formatting

differences. This approach achieved an accuracy of (0.07) in a manual evaluation of 100 samples, highlighting its limitations.

Approach 2: Prompt with Full List of Schools In the second approach, the predefined list of schools was embedded into the prompt text, and the model was asked to identify the best match. This method proved ineffective due to token size constraints, as the large predefined list consumed a significant portion of the token limit, reducing the model’s ability to focus on the complaint text itself. As a result, this approach also achieved an accuracy of only (0.11).

Approach 3: String Matching with ‘difflib’ The third approach employed the Python ‘difflib’ library for string similarity matching. This approach significantly outperformed the others, achieving an accuracy of **85%** in manually evaluated samples. String matching proved particularly effective in handling minor spelling variations and formatting inconsistencies, making it the most reliable method for this task.

Challenges in Matching Process The main challenges observed in the entity recognition and matching processes were:

- **Ambiguity in Extracted Names:** In some cases, extracted names included additional phrases like “School name is: [name],” which disrupted matching with predefined names.
- **Variations in Naming Conventions:** Abbreviations, alternate spellings, and dialectical differences posed significant challenges for embedding-based and prompt-based approaches.
- **Token Limitations:** Including the full list of school names in the prompt text overwhelmed the model’s token capacity, reducing the accuracy of matching results.

Evaluation and Insights The evaluation of LLaMA 3.1’s entity recognition and matching capabilities underscores the strengths and limitations of applying LLMs to structured data tasks. While the model demonstrated exceptional performance in extracting school names from complaint texts with an accuracy of 0.95, its performance in structured matching was less effective due to the inherent limitations of LLM-based methods. These findings align with challenges discussed in prior research, offering critical insights into the nuances of LLM performance in such tasks.

Brown et al. (2020) identified token constraints as a significant barrier to LLM performance in tasks involving large or complex inputs. When extensive predefined lists, such as the 276 validated school names used in this study, are included in prompts, they consume a considerable portion of the token limit, leaving less capacity for the model to process the context effectively (Brown et al., 2020). This limitation was evident in the prompt-based matching approach, where including all school names resulted in only 0.11 accuracy. The reduced contextual focus caused by token overflow highlights the need for optimization in handling extensive lists within prompt-based tasks.

Furthermore, Karpukhin et al. (2020) explored how semantic mismatches arise when dense retrieval models are applied to structured tasks requiring exact matches. They demonstrated that while embeddings are well-suited for semantically related queries, they struggle with precise matching due to their reliance on similarity rather than exact string correspondence (Karpukhin et al., 2020). In this study, minor variations in school names, such as “Al Ahad AlZahrer High School” versus “Al Ahad AlZahrer Secondary School,” led to mismatches during embedding-based retrieval, further limiting the effectiveness of this approach.

The challenges of using embeddings for exact matches in domain-specific tasks were also emphasized by Guu et al. (2020). Their research highlighted that embeddings prioritize semantic relationships over lexical precision, leading to inaccuracies when dealing with highly specific data (Guu et al., 2020). This was particularly relevant to this evaluation, where embedding-based matching consistently failed to resolve variations in school names, achieving an accuracy of just (0.07).

These limitations underscore the need for hybrid approaches that combine the strengths of LLMs in entity recognition with traditional algorithms for precise matching. As demonstrated in this study, using the Python ‘difflib’ library for string similarity comparisons significantly improved matching accuracy to (0.85), outperforming all LLM-based methods. This success aligns with Jain and Patel’s (2021) findings, which advocated for combining LLMs with traditional techniques to address structured data challenges (Jain and Patel, 2021).

In summary, the evaluation highlights the following key insights:

- **Strengths of LLMs in Entity Recognition:** LLaMA 3.1 excelled in extracting school names with high accuracy (0.95), demonstrating its capability to handle unstructured complaint texts effectively.
- **Limitations in Structured Matching:** The model struggled with tasks requiring precise matching due to token constraints, semantic mismatches, and embedding limitations, achieving only (0.7) with both RAG and prompt-based approaches.

- **Effectiveness of Hybrid Approaches:** Combining LLM outputs with string similarity algorithms, such as ‘diffib’ addressed these limitations and achieved significantly higher matching accuracy (0.85), illustrating the value of integrating traditional techniques with LLM capabilities.

These findings align closely with prior research, which has consistently highlighted the challenges of using LLMs for structured data tasks. The insights gained from this evaluation provide a strong foundation for future work, emphasizing the importance of hybrid methodologies and tailored solutions for domain-specific applications.

4.2.4 Results of Monthly Summary Analysis

The monthly summary analysis aimed to extract high-level insights, keywords, and recurring trends from complaints submitted to the Ministry of Education in Bahrain. Two iterations of prompt design were used to improve the quality and relevance of outputs while addressing hallucination issues. This section evaluates the results generated from both prompt versions and highlights key improvements and limitations.

The initial prompt, which directed the model to generate summaries, keywords, and trends, provided useful overviews but faced several challenges:

- The generated summaries often included information not explicitly mentioned in the complaints. For instance, in some cases, the model inferred general concerns about education systems in neighboring countries without any such context provided in the complaints.
- Keywords were overly generic or irrelevant, leading to difficulties in identifying specific trends. For example, terms such as “education,” “students,” and “issues” appeared frequently but did not add significant value.
- Trends occasionally included hallucinated information, such as attributing issues to policies of other ministries or countries, which were not supported by the complaints.

These limitations indicated a need for refinement in the prompt design to ensure outputs were both precise and directly tied to the complaint data.

The second prompt iteration addressed these limitations by:

- Increasing the number of keywords extracted to 15–30, focusing on relevance and specificity.

- Introducing stricter instructions to avoid generating information not explicitly mentioned in the complaints. The updated prompt explicitly stated, “Only include information that is explicitly present in the provided complaint text.”
- Providing examples to the model for improved contextual understanding, ensuring the generated outputs aligned more closely with the complaint content.

These adjustments significantly improved the outputs:

- Summaries became more concise and accurate, focusing on the main themes of the complaints without including unrelated information.
- Keywords reflected the core issues more effectively, with terms such as “bullying,” “scholarship delays,” and “school transfers” frequently appearing instead of generic terms.
- Trends highlighted recurring themes directly tied to the complaints, such as dissatisfaction with digital platforms for student registration or concerns about overcrowding in schools.

Examples of Key Findings

- **Meeting Other Countries’ Standards:** In one of the earlier outputs, hallucinated trends suggested that parents were concerned about Bahrain’s education system failing to meet international standards, referencing countries such as Finland and Singapore. After prompt refinement, such unsupported claims were eliminated, and trends focused on actual concerns raised by parents, such as delays in implementing digital services and scholarship processes.
- **Recurring Complaints About Hallucinations:** Initial summaries often inferred dissatisfaction with policies from other ministries or attributed issues to external factors, such as the economic climate, without any textual basis in the complaints. The revised prompts removed such hallucinations, ensuring that summaries and trends were grounded in the dataset.
- **Complaint About Egyptian Education System:** One recurring hallucination involved references to the Egyptian education system, which appeared in summaries despite no such mentions in the complaint text. After the prompt update, the model correctly identified complaints about student transfer issues and online registration systems without introducing unrelated contexts.

Evaluation of Final Outputs and Addressing Hallucinations The refined prompt demonstrated significant improvements in accuracy and reliability across all outputs, including summaries, keywords, and trends. The evaluation highlighted the following improvements:

- **Summaries:** The accuracy of summaries increased to **89%**, as evaluated by human reviewers. This improvement was primarily due to the reduction in hallucinations and the inclusion of concise, relevant information directly tied to the complaints.
- **Keywords:** Keywords achieved an accuracy of **90%**, successfully capturing the main issues while avoiding overly generic terms. For instance, the refined keywords emphasized specific concerns such as “scholarship delays,” “teacher misconduct,” and “overcrowded classrooms.”
- **Trends:** Trends became more actionable and aligned closely with recurring themes in the data, such as dissatisfaction with student registration platforms, delays in financial aid, and concerns about the quality of teaching. Importantly, fabricated trends or irrelevant inferences, such as referencing other countries’ education systems, were eliminated.

One of the significant issues addressed during the refinement process was the hallucination problem, where the model generated information not explicitly mentioned in the complaints. For example, earlier outputs sometimes inferred dissatisfaction with international education systems or attributed complaints to unrelated ministries, despite no textual evidence supporting these claims. These inaccuracies were misleading and could negatively impact decision-making processes.

To address hallucination issues, the refined prompt incorporated explicit instructions, such as: *“Only include information that is explicitly mentioned in the complaints provided.”* This adjustment aligns with findings by Ji et al. (2023), who emphasize that hallucinations in large language models can be mitigated by providing restrictive and unambiguous task descriptions in the prompt design (Ji et al., 2023).

Additionally, the refinement process included examples and clearer task descriptions to guide the model’s outputs, as suggested by Reynolds and McDonell (2021). Their research demonstrates that adding examples and precise instructions reduces the likelihood of hallucinations and ensures the outputs remain grounded in the source data (Reynolds and McDonell, 2021). This approach was instrumental in improving the reliability of trends and summaries, particularly by eliminating unrelated or fabricated content.

Key Improvements Through Prompt Refinement The iterative refinement of prompts not only addressed hallucination issues but also enhanced the overall quality of outputs:

- Summaries became more concise and directly reflected the main themes of complaints without including fabricated or irrelevant information.
- Keywords captured the most critical concerns raised in the complaints, ensuring relevance and specificity.
- Trends highlighted recurring issues grounded in the dataset, providing actionable insights for stakeholders.

These findings emphasize the critical role of prompt engineering in improving the performance of LLMs for complex tasks. The success of this approach aligns with research by Goyal et al. (2023), who highlight the importance of tailoring prompts to specific tasks to improve accuracy and reduce hallucinations in LLM-generated outputs.

The evaluation of the refined prompt demonstrates that clear instructions, task-specific examples, and restrictive guidelines are essential to leveraging LLMs effectively for summarization and trend detection. By addressing hallucination issues and enhancing output relevance, the refined prompt significantly improved the accuracy and utility of LLM outputs, providing reliable insights into the complaint data.

Remaining Challenges Despite the improvements, some challenges persisted:

- Summaries occasionally missed nuanced details, such as specific references to subjects like “chemistry” or “biology” in complaints about exam schedules.
- Keywords, while more relevant, sometimes failed to capture specific terminology mentioned in complaints, such as the name of particular educational platforms or initiatives.
- The model still struggled with complaints that included mixed languages (e.g., Arabic and English) or informal phrasing, which sometimes led to less precise outputs.

The iterative refinement of prompts for monthly summary analysis demonstrated the importance of clear instructions and explicit guidelines for LLMs. By addressing hallucination issues and focusing on actionable outputs, the updated prompts significantly improved the relevance and reliability of summaries, keywords, and trends. These findings align with best practices in prompt engineering and highlight the potential of LLMs to provide valuable information on large datasets when carefully tailored to the task.

4.3 Visualization and Analysis

The visualization tools developed for the analysis of complaints provided a comprehensive understanding of data trends and key issues. These visualizations were designed to interactively explore the dataset based on various categories, temporal trends, and geographical distribution. This section highlights the findings from the visual analysis.

4.3.1 Monthly Analysis and Interactive Dashboard

The *Monthly Analysis Dashboard* features an interactive bar chart showcasing the number of complaints received each month, along with a filter for categories. Users can select and deselect categories to focus on specific complaint types. For example:

- In the “Tests and Exams” category, Figure 4.3, the dashboard reveals peak complaints in **May** and **December**, correlating with the timing of final exams in Bahrain (June and January).
- The “Student Transfers” category peaks in **August**,, Figure 4.4 ,aligning with the preparatory period before the new academic year starting in September.

Clicking on a specific month reveals a detailed summary of complaints, recurring trends, and top keywords for that period, as shown in Figure 4.2. This enables policymakers and stakeholders to understand the dynamics of issues during particular periods and align resources accordingly.

4.3.2 Keyword Evolution Over Time

The keyword slider visualization allows users to track the evolution of key terms on a daily basis, providing insights into the changing nature of complaints. For instance:

- Prominent keywords such as *exams*, *bullying*, and *administrative errors* highlight seasonal variations in the nature of issues raised.
- The slider also reveals peaks in keywords like *financial aid* during scholarship announcements or *transportation* issues at the start of the academic year.

This visualization (Figure 4.5) demonstrates the flexibility of the tool in monitoring and responding to trends in real-time.

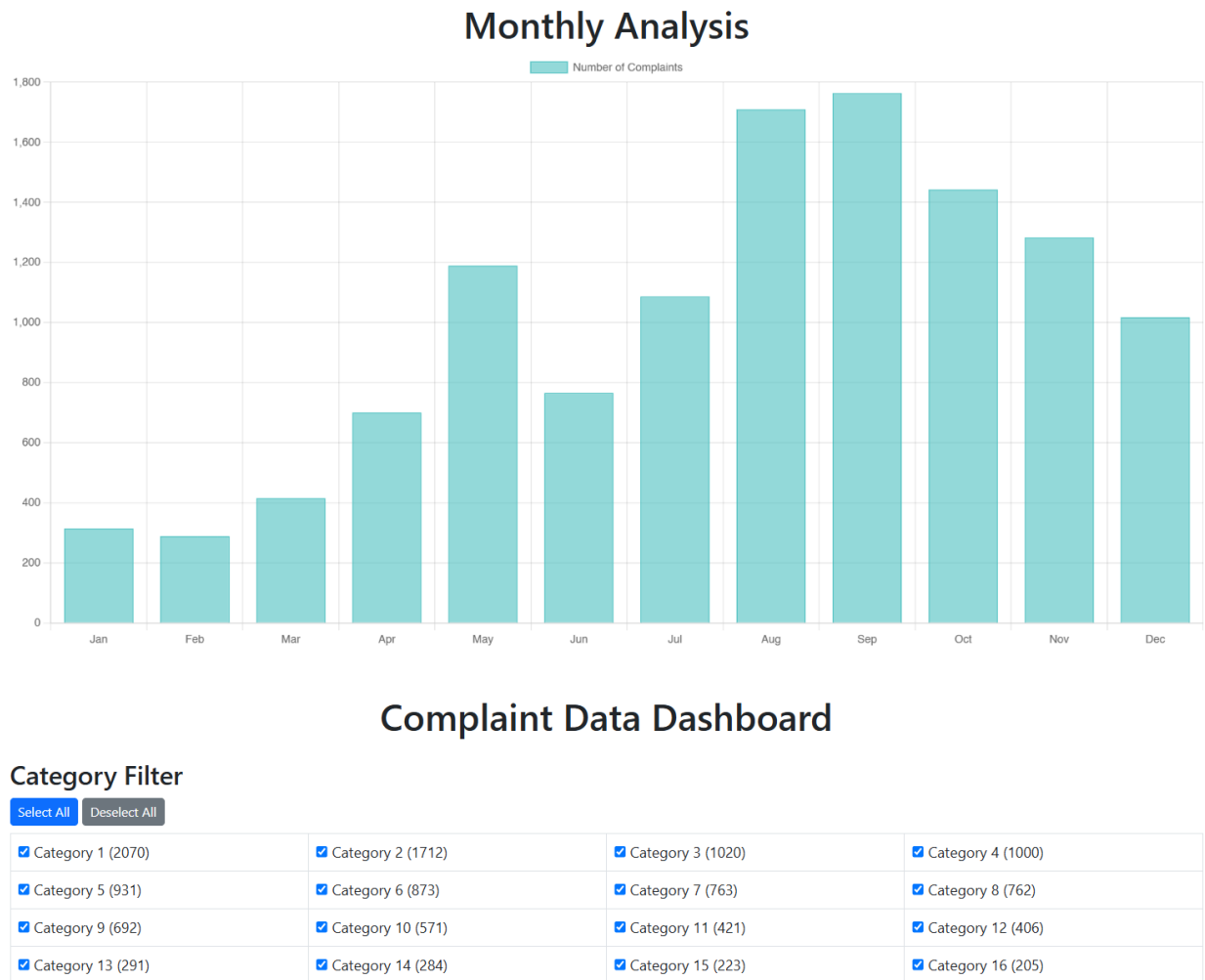


Figure 4.2: Monthly Analysis Dashboard with category filters and trends.

4.3.3 Geographic Heatmap of School Complaints

The *School Complaints Heatmap* (Figure 4.6) provides a geographical representation of complaints across Bahrain. High complaint density is observed in the central area of Isa Town, a known educational hub housing:

- Multiple government and private schools.
- The University of Bahrain and Bahrain Polytechnic.

The red zones in the heatmap indicate areas requiring further investigation, highlighting potential operational challenges or higher complaint volumes due to larger student and parent populations.

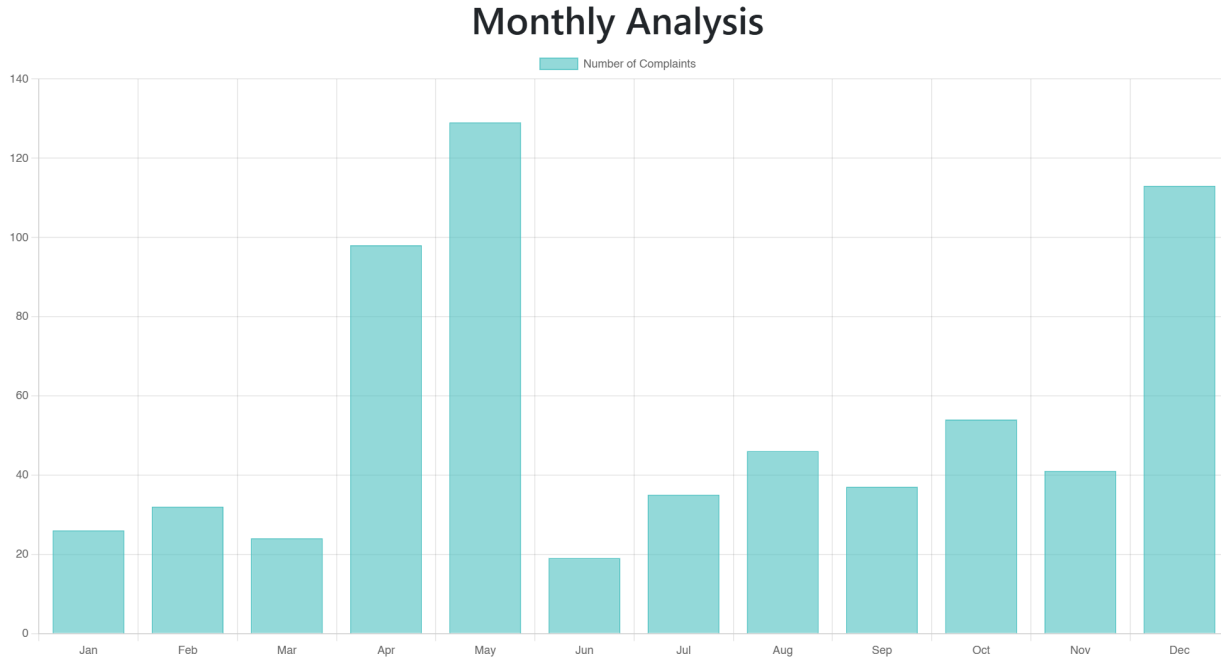


Figure 4.3: Monthly Analysis Dashboard filtered by “Test and Exams Category”

4.3.4 Monthly Summary Example

An example of the monthly summary generated from the dashboard for September 2023 is presented in Figure 4.7. The summary highlights common issues such as inadequate educational materials, administrative inefficiencies, and bullying, along with actionable trends.

4.3.5 Insights and Discussion

Actionable Insights:

- **Seasonal Trends:** The dashboards reveal actionable patterns, such as peaks in complaints related to final exams or student transfers, enabling better resource allocation during critical times.
- **Geographical Focus:** The heatmap identifies hotspots of complaints, providing a foundation for targeted interventions in high-density areas.

Enhanced Decision-Making:

- Policymakers can use the interactive tools to focus on categories or periods of concern, such as addressing high complaint volumes in Isa Town or resolving recurring issues in specific categories like “Administrative Errors.”

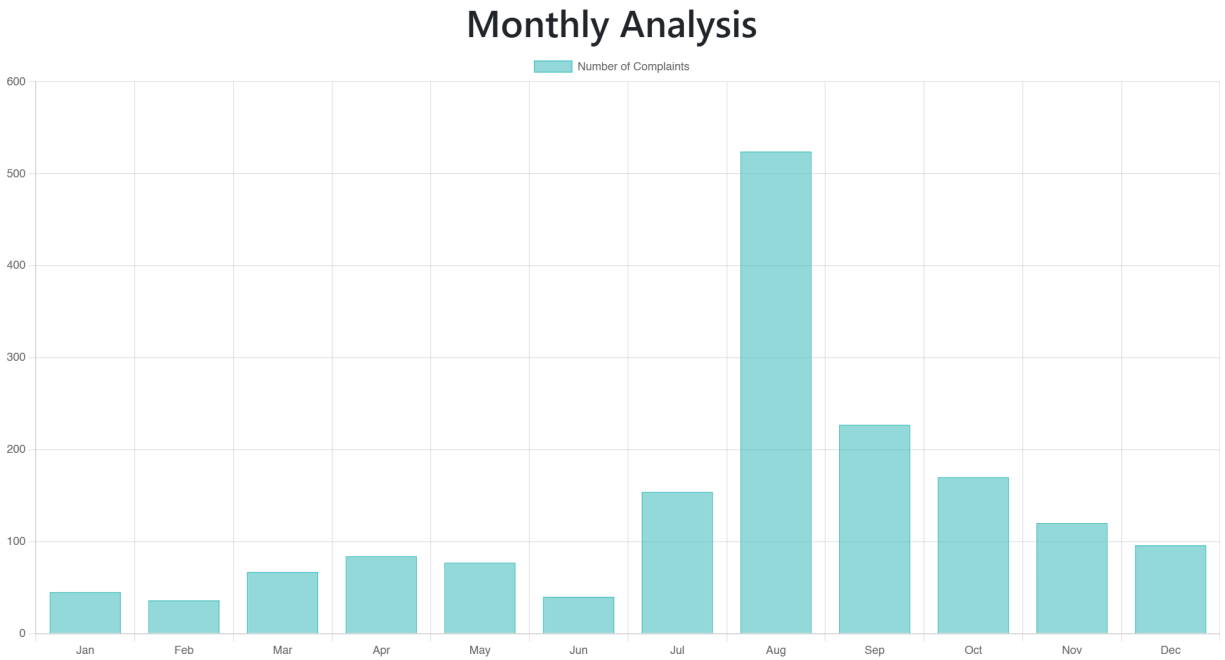


Figure 4.4: Monthly Analysis Dashboard filtered by “Students Transfer” Category

Improving Stakeholder Engagement: The visualizations offer a clear, data-driven method to communicate key findings to stakeholders, making the complaints analysis more transparent and actionable.

[illegible]

Figure 4.5: Keywords per day visualization showing dynamic changes over time.

School Complaints Analysis

Heatmap Description: The heatmap shows the concentration of complaints at school locations across the Kingdom. Areas with more intense color indicate a higher number of complaints, highlighting potential hotspots that may require further investigation.

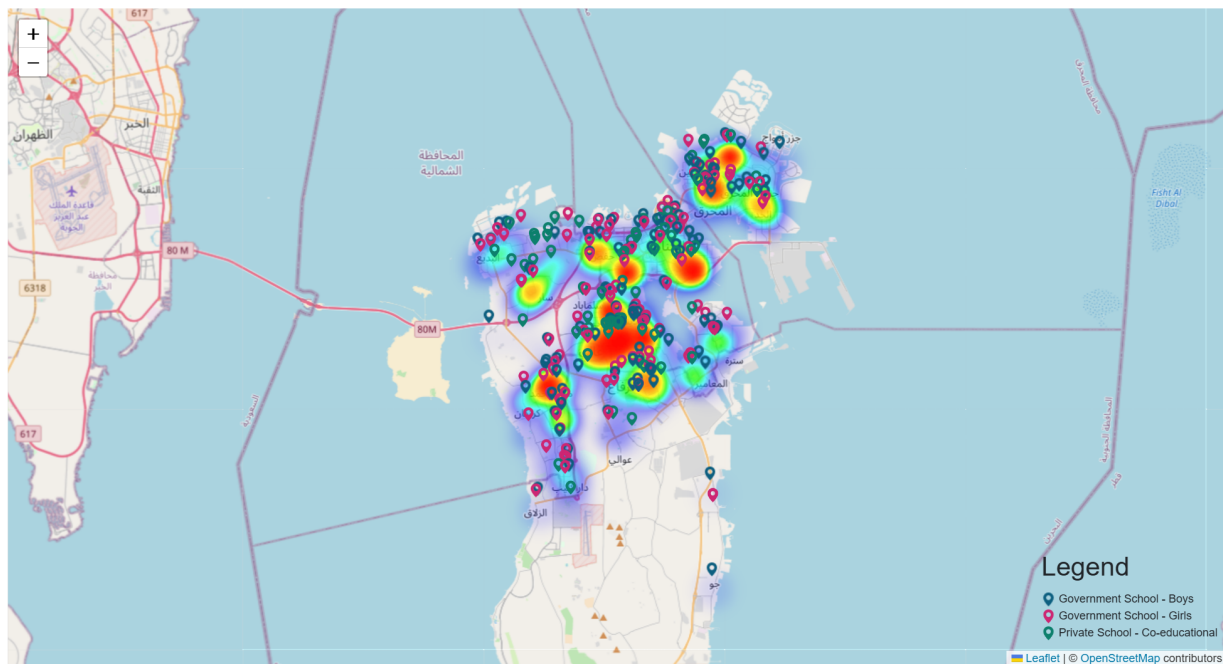


Figure 4.6: School complaints heatmap showing high-density complaint areas in Isa Town.

Monthly Summary

Sep 2023

Summary: Complaints from students and parents highlight various issues in the education system of Bahrain. The most frequently mentioned problems include schoolbook unavailability or delays (e.g., textbooks not delivered, essential books missing), bullying and harassment, educational material scarcity, curriculum-related issues (inconsistent schedules, inadequate completion of elementary-level studies), and administrative errors (mismanaged student transfers, unfair accusations). Some complaints also mention difficulties with contacting authorities or accessing information.

Trends: 1. ****Inadequate Provision of Educational Materials****: Many complaints highlight the scarcity or delayed availability of schoolbooks and essential educational resources. 2. ****Administrative Discrepancies and Inefficiencies****: Several students and parents faced difficulties in managing their education due to administrative errors, including mismanaged student transfers and unfair accusations. 3. ****Inconsistency in Curriculum and Scheduling****: Some schools were criticized for inconsistent schedules and an inadequate completion of elementary-level studies within a standard timeframe. 4. ****Harassment and Bullying****: Several students experienced bullying and harassment within school environments or faced administrative actions unfairly due to these issues. 5. ****Difficulty Accessing Information and Support****: There were instances where parents struggled to contact educational authorities for assistance with their children's education, including inquiries about teaching positions, textbook delivery, and student transfers.

schoolbooks

unavailability

delay

bullying

harassment

education system

curriculum

schedule changes

book delivery

textbook shortages

educational materials

administrative errors

transfer issues

unfair accusations

Figure 4.7: Monthly summary example showing trends, keywords, and summaries for September 2023.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

This research explored the application of LLMs, specifically LLaMA 2 and LLaMA 3.1, to analyze and categorize complaints in the Tawasul system. The goal was not only academic but also practical, to address real-world challenges in large-scale, short-text complaint analysis and support national-level digital governance. Unlike generic natural language processing studies, this work contributes directly to governmental complaint management by proposing a scalable, locally deployable, and privacy-compliant solution that aligns with national data policies.

The experimental results demonstrated substantial performance improvements in key NLP tasks. Summarization and keyword extraction benefited from iterative prompt engineering, achieving accuracies of 0.89 and 0.90, respectively, when benchmarked against human interpretations. Categorization accuracy increased from 0.496 using a traditional LSTM model to 0.78 with the LLaMA 3.1 model integrated with RAG. This improvement brought the model’s performance closer to human-level reliability. Moreover, manual evaluations indicated that, in certain cases, the model handled vague or overlapping complaints with greater consistency than human annotators, particularly after prompt and category refinements.

Unlike approaches dependent on external APIs or cloud services, this system was implemented entirely on secure, locally hosted infrastructure. This design decision was driven by the sensitivity of the dataset and national privacy regulations, ensuring compliance with governmental data protection standards. No cloud-based inference was used at any stage. All model executions and experiments took place on-premise using GPU-equipped compute

nodes under controlled conditions.

From a deployment perspective, the system is technically feasible for real-world governmental use. The model requires high-end GPUs such as NVIDIA A100s with 80 GB of memory. For a single entity or ministry, one such GPU node may be sufficient. However, wider adoption across multiple entities or real-time classification scenarios would necessitate additional compute capacity. The infrastructure requirements include access to high-memory GPUs, quantized LLM versions to reduce computational cost, secure high-performance clusters, and humans to provide quality assurance and dynamic prompt optimization.

The adopted method carries several strengths. The selected LLM supports Arabic dialects and short-text content with minimal preprocessing and generates structured, explainable outputs in the form of summaries, keywords, and predicted categories. The system is adaptable to the addition of new complaint categories and can be trained and enhanced with additional information for retrieval. Moreover, it respects national privacy requirements through full on-premise deployment.

However, the approach also presents certain limitations. It requires advanced hardware. The performance of the RAG pipeline depends heavily on the clarity and completeness of the predefined categories and embedding coverage. Additionally, the current classification system supports only single-label outputs, which limits its ability to represent highly nuanced or overlapping complaints that naturally span multiple domains.

In conclusion, this research demonstrates the viability of combining open-source LLMs, prompt engineering, and retrieval-based techniques to improve governmental complaint analysis. The system offers a practical, secure, and adaptable framework that can enhance public service responsiveness, reduce manual workload, and provide deeper insights into citizen concerns. With further investment and policy alignment, the approach has strong potential for scaling across government institutions and becoming a core component of modern digital governance infrastructure.

5.2 Future Work

Building on this research, future work should focus on applying the developed methods to analyze all complaints submitted by government entities. Expanding the scope will help generalize the findings and further validate the methodology’s effectiveness. Additionally, efforts should include developing approaches for multi-class categorization, allowing complaints to belong to multiple categories, thus capturing the complexity of overlapping or

multi-dimensional issues. Advanced clustering techniques, such as Latent Dirichlet Allocation (LDA) or K-means, could also be explored to uncover hidden patterns and improve thematic grouping.

Another important direction is evaluating the quality of responses provided to complaints. Future work should involve leveraging general-purpose large language models (gLLMs) to assess response quality based on predefined metrics such as relevance, clarity, and resolution effectiveness. This evaluation can guide improvements in government communication and response strategies.

Entity recognition and structured matching should also be refined. Fine-tuning LLMs on domain-specific datasets will likely improve their ability to identify structured entities like school names and extract additional contextual details. Combining LLM-based approaches with traditional similarity measures, such as those provided by difflib or embedding-based techniques, can enhance matching accuracy and reduce dependency on manual corrections.

Lastly, expanding visualization capabilities to incorporate geospatial tools, resolution times, and sentiment analysis will provide deeper insights into complaint trends and stakeholder satisfaction. Integrating external datasets, such as socioeconomic indicators, can further contextualize complaints and uncover correlations across domains. By addressing these areas, the full potential of LLMs can be harnessed to enhance public services, improve citizen satisfaction, and establish robust methodologies for complaint analysis across government entities.

References

- Abdul Aziz, A. and Starkey, A. (2020). Predicting supervise machine learning performances for sentiment analysis using contextual-based approaches. *IEEE Access*, 8:17722–17733.
- Ainslie, J., Lee-Thorp, J., de Jong, M., Zemlyanskiy, Y., Lebrón, F., and Sanghai, S. (2023). GQA: Training generalized multi-query transformer models from multi-head checkpoints. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4895–4901, Singapore. Association for Computational Linguistics.
- Al Sulaimani, S. and Starkey, A. (2023). Real-time event detection using self-evolving contextual analysis (seca) approach. *IEEE Access*, 11:127011–127034.
- Alzanin, S. M., Azmi, A. M., and Aboalsamh, H. A. (2022). Short text classification for arabic social media tweets. *Journal of King Saud University - Computer and Information Sciences*, 34(9):6595–6604.
- Amazon Web Services (2024). Amazon mechanical turk. <https://www.mturk.com/>.
- Amur, Z. H., Hooi, Y. K., Bhanbhro, H., Dahri, K., and Soomro, G. M. (2023). Short-text semantic similarity (stss): Techniques, challenges and future perspectives. *Applied Sciences*, 13(6):3911.
- Antypas, D., Ushio, A., Camacho-Collados, J., Neves, L., Silva, V., and Barbieri, F. (2022). Twitter topic classification. *arXiv preprint arXiv:2209.09824*.
- Brown, T., Mann, B., Ryder, N., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Chen, J., Hu, Y., Liu, J., Xiao, Y., and Jiang, H. (2019). Deep short text classification with knowledge powered attention.

- Diab, M., Habash, N., and Zitouni, I. (2017). Nlp for arabic and related languages. *TAL*, 58(3):9–13. The George Washington University, USA, mtdiab@gwu.edu; New York University Abu Dhabi, UAE, nizar.habash@nyu.edu; Microsoft Research, USA, iz-itouni@microsoft.com.
- Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al. (2024). The llama 3 herd of models. *Meta AI Technical Report*. Submitted on 31 Jul 2024, last revised 15 Aug 2024. A detailed contributor list is available in the appendix.
- Feng, F., Yang, Y., Cer, D., Arivazhagan, N., and Wang, W. (2020). Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*.
- Gerganov, G. (2023). Ctransformers: A fast c++ library for transformer models. *CTransformers Documentation*. Accessed: 2024-11-11.
- Goel, A., Gueta, A., Gilon, O., Liu, C., Erell, S., Nguyen, L. H., Hao, X., Jaber, B., Reddy, S., Kartha, R., Steiner, J., Laish, I., and Feder, A. (2023). Llms accelerate annotation for medical information extraction. In Hagselmann, S., Parziale, A., Shanmugam, D., Tang, S., Asiedu, M. N., Chang, S., Hartvigsen, T., and Singh, H., editors, *Proceedings of the 3rd Machine Learning for Health Symposium*, volume 225 of *Proceedings of Machine Learning Research*, pages 82–100. PMLR.
- Goker, S. and Canbuglalilar, B. (2018). Neural text normalization for turkish social media.
- Goyal, T., Raj, A., and Gupta, S. (2023). Tailoring prompts for task-specific large language model outputs. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Guu, K., Lee, K., Tung, Z., et al. (2020). Realm: Retrieval-augmented language model pre-training. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 3929–3938.
- Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. O’Reilly Media, Inc.
- Hu, Y., Ding, J., Dou, Z., and Chang, H. (2022). Short-text classification detector: A bert-based mental approach. *Computational Intelligence and Neuroscience*, 2022:8660828.
- IBM (2024). Gguf versus ggml. <https://www.ibm.com/think/topics/gguf-versus-ggml>.

- Information and eGovernment Authority (2017). Tawasul system. <https://www.iga.gov.bh/en/article/tawasul-system>. Accessed: 2024-11-13.
- Jain, H. and Patel, M. (2021). Combining string similarity and semantic matching for entity recognition in domain-specific texts. *AI & ML Journal*.
- Ji, Z., Lee, N., Fries, J., Yu, T., Wimala, E., Ghosh, S., and Sachan, D. (2023). Survey of hallucination in natural language generation. *arXiv preprint arXiv:2303.07762*.
- Johnson, J. M. and Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1):27.
- Karl, F. and Scherp, A. (2023). Transformers are short text classifiers: A study of inductive short text classifiers on benchmarks and real-world datasets. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction (CD-MAKE 2023)*, volume 14285 of *Lecture Notes in Computer Science*, pages 110–127. Springer.
- Karpukhin, V., Oguz, B., Min, S., et al. (2020). Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781.
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *arXiv preprint arXiv:2005.11401*.
- Li, T., Chen, X., Dong, Z., Yu, W., Yan, Y., Keutzer, K., and Zhang, S. (2022). Domain-adaptive text classification with structured knowledge from unlabeled data. *arXiv preprint arXiv:2206.09591*.
- Liu, Y., Ouyang, X., Bai, R., Zhu, X., Zhou, Y., and Zhang, Y. (2024). A framework for categorizing complaint text via large language model. In *2024 7th International Conference on Advanced Algorithms and Control Engineering (ICAACE)*, pages 1–8. IEEE.
- Ollama AI (2024). Ollama: Optimized framework for large language models. <https://www.ollama.ai>.
- Reynolds, L. and McDonell, K. (2021). Prompt programming for large language models: Beyond the few-shot paradigm. *arXiv preprint arXiv:2102.07350*.

- Saeed, A. M., Hussein, S. R., Ali, C. M., and Rashid, T. A. (2022). Medical dataset classification for kurdish short text over social media. *Computer Science Department, University of Halabja, KRG, Halabja, Kurdistan, Iraq and Computer Science and Engineering Department, University of Kurdistan-Hawler, KRG, Erbil, Kurdistan, Iraq*.
- Scikit-Learn (2023). Classification metrics documentation. https://scikit-learn.org/stable/modules/model_evaluation.html#classification-report.
- Sun, X., Liu, Z., and Huo, X. (2023). Six-granularity based chinese short text classification. *IEEE Access*, 11:35841–35852.
- Tan, Z., Beigi, A., Wang, S., Guo, R., Bhattacharjee, A., Jiang, B., Karami, M., Li, J., Cheng, L., and Liu, H. (2024). Large language models for data annotation: A survey. *arXiv preprint arXiv:2402.13446*.
- Tang, Z., Chen, Z., Dai, D., and Chen, T. (2022). Short text classification combining keywords and knowledge. In *2nd International Conference on Consumer Electronics and Computer Engineering (ICCECE)*, pages 662–665.
- Touvron, H., Lavril, T., Izacard, G., et al. (2023a). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Touvron, H., Martin, L., Stone, K., et al. (2023b). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Wan, Y., Yang, B., Wong, D. F., Chao, L. S., Yao, L., Zhang, H., and Chen, B. (2022). Challenges of neural machine translation for short texts. *Journal of Computational Linguistics*.
- Wang, H. et al. (2023). Langchain: A framework for building applications with large language models. <https://langchain.com/>. *LangChain Documentation*. Accessed: 2024-11-11.
- Yong, L. P. X., Tung, J. Y. M., Lee, Z. Y., Kuan, W. S., and Chua, M. T. (2024). Performance of large language models in patient complaint resolution: A single-blind comparative evaluation. *Journal of Medical Internet Research*. Preprint.
- Zhao, S., Yang, Y., Wang, Z., He, Z., Qiu, L. K., and Qiu, L. (2024). Retrieval-augmented generation (rag) and beyond: A comprehensive survey on how to make your llms use external data more wisely. *arXiv preprint arXiv:2409.14924*.