Utilising LLMs for Streamlined Analysis of PTA Datasets

Abdullah Arshad
Lahore University of Management Sciences
Lahore, Pakistan
25100281@lums.edu.pk

Danish Athar
Lahore University of Management Sciences
Lahore, Pakistan
25100174@lums.edu.pk

Hamna Shafqat
Lahore University of Management Sciences
Lahore, Pakistan
25100176@lums.edu.pk

Sheikh Hamza Elahi Sodana
Lahore University of Management Sciences
Lahore, Pakistan
25100006@lums.edu.pk

Abstract

PTA releases extensive datasets about the performance of telecom operators in Pakistan but this massive data goes unutilised. To tap into the potential of this data, we make use of large language models (LLMs), which are known to possess amazing potential to derive valuable insights from large amounts of unprocessed data (Chen et al., 2023). We employ LLMs to extract information about telecom operators aiming to investigate disparities in quality of services and assess the spatial and temporal coverage across providers. Our approach involves leveraging LLMs to extract data from PDFs and streamlining the analysis process. Through this method, we can look at the quality and availability of broadband services in Pakistan, comparing provider performance. However, effective prompt engineering is essential to optimize LLM responses for meaningful information retrieval. For this reason, we employ different techniques. First, we observe the sensitivity of LLMs to the original PDFs provided by PTA and compare the LLM's analysis against manual analysis. We find that GPT-4, our LLM of choice (after comparison against other options), struggles if the data is large in size. As a fix, we split the quarterly and yearly data into multiple smaller PDFs according to KPIs and cities. For efficient analysis we extract data as xlsx and JSON files via GPT-4 and then analyze the data. For the analysis, we incorporate various prompt engineering techniques to generate analysis that is most suitable to our goal. These include zero-shot prompting, one-shot prompting, few-shot prompting and various types of chain-of-thought (CoT) prompting and quantify the improvement brought forth by each. Finally, we suggest various kinds of useful analyses to show the kinds of valuable insights we can obtain from this setup.

Keywords: Quality of Service (QoS), Broadband Services, Network Operators, Large Language Models (LLMs), Prompt Engineering, Data Processing, Pakistan Telecommunication Authority (PTA), Policy Decisions

ACM Reference Format:
Abdullah Arshad, Danish Athar, Hamna Shafqat, and Sheikh Hamza Elahi Sodana. 2024. Utilising LLMs for Streamlined Analysis of PTA Datasets. In . ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

The evolution of broadband services has revolutionized communication and connectivity, becoming a cornerstone of modern societies. In Pakistan, the landscape of broadband services is dynamic, with numerous providers striving to offer reliable, high-speed connectivity to consumers. The Pakistan Telecommunication Authority (PTA) plays a crucial role in regulating and monitoring these services to ensure compliance with quality standards and promote healthy competition in the market. To assess the performance and quality of broadband services in Pakistan, the PTA conducts regular Quality of Service (QoS) surveys across various regions. These surveys evaluate key performance indicators (KPIs) related to voice, SMS, and mobile broadband services, including network coverage and speed. The results of these surveys provide valuable insights into the strengths and weaknesses of different service providers, helping to identify areas for improvement and development.

The problem at hand revolves around the assessment of broadband services in Pakistan, focusing on disparities in quality and spatial and temporal coverage across providers. Despite the significant advancements in technology and infrastructure, there are still challenges and discrepancies in the quality of broadband services offered by different

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

providers. These challenges can include inconsistent network coverage, varying data speeds, and differing levels of service reliability. The PTA conducts regular QoS surveys to monitor the performance of broadband services. However, analyzing the vast amount of data collected from these surveys can be complex and time-consuming, requiring sophisticated tools and methodologies. The service providers also need to meet a specific threshold criteria for various KPIs, and this data has the information required to ensure compliance and quality regulations. However, effective regulation requires robust data analytics to identify compliance and the areas for improvement. This can in turn ensure equality of broadband services in different regions in Pakistan.

By leveraging Large Language Models (LLMs) and prompt engineering techniques, we aim to streamline the analysis process and extract meaningful insights from PTA's data. In this process, we ran into several problems including inaccurate readings as well as hallucinations causing incorrect analysis outputs. We explore different methods to counteract these issues like data splitting and prompt engineering to come up with a working LLM-based solution for PTA's datasets.

2 Motivation

We aim to offer a viable solution to the need to identify and address the discrepancies in broadband services in Pakistan, particularly one that utilizes Large Language Models (LLMs). This study is motivated by the potential impact it can have on improving the quality of broadband services in Pakistan. By identifying areas for improvement and highlighting best practices, this research can inform policy decisions and regulatory measures aimed at enhancing the overall broadband experience for consumers. Additionally, by automating the analysis process, this study can provide a more efficient and effective way of evaluating broadband services, benefiting both providers and consumers alike. Despite its extensive nature, the potential of PTA's datasets remains unrealized even after several years of quarterly releases. It could possibly be because of their sheer volume; it is difficult to go over these PDFs spanning several hundred pages manually when a person wants to learn about a broadband performance metric relevant to them. Automating this process can make extracting information from these PDFs more efficient for the parties involved.

3 Overview

We took a multifaceted approach for our solution. Initially, we addressed the challenges posed by the voluminous PDF datasets by segmenting them into smaller, city-specific documents, which significantly improved the focus and accuracy of our analysis. Utilizing LLM-based data extraction techniques, we converted these segmented PDFs into more structured JSON formats, enhancing the usability of the data. To interpret this refined data, we leveraged a variety of prompting techniques with Large Language Models (LLMs), such as zero-shot, few-shot, and chain-of-thought prompting, to generate nuanced insights and reliable recommendations. This methodology not only allowed for a systematic data-based analysis of the performance of different network operators but also supported the generation of actionable recommendations for policy improvements and service enhancements, ultimately aiming to elevate the quality of broadband services across Pakistan.

4 Choice of LLM

After examining the formats of data included in the PTA datasets, we explored several AI based tools to help extract the data in the form of bar graphs, coverage maps, tables, line graphs and data matrices. We learned that texts and tables could be analysed by using GPT 3.5. However, it exhibited limitations in dealing with complex data formats such as graphs and matrices, constraining the ability to generate rich insights and constructing an automated pipeline for analysis. We explored a computer vision based tool called Astica which combined the abilities of GPT 3.5 and OCR to give a detailed image analysis. This tool posed cost considerations. Additionally, Bing and Powerdrill, both tools based on GPT 3.5 and GPT 4 were explored for having abilities in pdf analysis and image analysis. However, powerdrill's inability to handle multiple PDFs at a time would restrain temporal analysis while Bing's query limit and issues with image analysis, including frequent hallucinations weakened the validity of the generated insights. Navigating the limitations of these tools, we concluded that the use of manual splitting coupled with the use of GPT 4 would be most effective. Due to its ability to handle multiple PDFs at the same time for comparison, and a considerable context window, we further explored GPT 4 via different ways of PDF splitting, labelling, and prompt engineering.

5 Splitting

The data PTA provides is in the form of PDFs, divided on a per-quarter, yearly basis. Each PDF spans more than a hundred pages, containing data for many different KPIs within the same PDF. In order to improve the LLM's performance, we split the dataset into smaller, specialised chunks. Initially, we attempted to split the PDFs based on the different KPIs (56 in total), which seemed promising. However, we encountered a significant challenge: the data extracted from these splits was still too voluminous. This posed a twofold issue—it made it complex for GPT to effectively comprehend and interpret the data, and it also led to incomplete results, where some cities were analyzed while others were omitted, resulting in incomplete or insufficient data for analysis. To address this, we pivoted our approach and opted to split the PDFs by cities (regions), which proved to be a more effective strategy.

We went from the original average of 2 PDFs per quarter to about 30 PDFs for each quarter, separating out data related to each city (region) from the original PDF file into a new PDF consisting solely of data related to that single city (region). As a result, we have a more specialised dataset, where we have per-quarter PDFs for each individual city (region), which we show is handled better by GPT-4 as compared to the original larger PDFs containing several KPIs.

The splitting was mostly done manually, but we wrote Python scripts to assist us in the process. We specified the paths to the original PDF files and used the PyPDF2 library to scan the PDFs in those paths for the city (region) we were searching for and extracted the pages on which there was a match into a new PDF. We also printed out the PDF name and the matching page numbers for manual verification of the output against the original files to make sure our script did not miss any pages or include any extra ones. This was done for each of the cities (regions) specified previously.

The scripting worked out well with a couple of challenges. Since the PDF was continual in nature, some cities (regions) were mentioned on one page but their related tables and graphs were continuing on to the next page which were not found by the city (region) search. Additionally, the same city (region) was mentioned differently across PDFs and within the same PDF as well. For example, in the file "qos survey report cities pak 21062022.pdf", the city (region) 'Gwadar' was mentioned as 'Gwadar' and 'Gawadar' as well, which made the searching script miss out on these pages because of the way string matching works. All such errors were identified manually.

The scripts used for this are available on our GitHub repository.

6 Data Extraction Formats

When the data was split into different regions, GPT-4 struggled to accurately interpret the data and generate high-quality assessments. To mitigate this issue, we performed further data preprocessing. Initially, we converted the split data into an Excel format, which contained the relevant data for the KPIs listed in the prompts. However, this approach led to numerous inaccuracies in extraction, as the values recorded in the Excel file often differed from those stated in the actual PDF. Recognizing the limitations of the Excel format, we explored an alternative approach using JSON objects. Converting the data into JSON format resulted in significantly greater accuracy in the values extracted by GPT-4, as compared to the Excel format. Additionally, the JSON format proved to be more convenient for GPT-4 to understand and interpret. As a result, we decided to use the JSON format for all the PDFs that were split into cities, as it offered improved accuracy and interpretability for GPT-4.

7 Prompting

When pre-processing PDFs to make them suitable for analysis with GPT-4, despite efforts to format the data for easy interpretation, the results obtained were often unsatisfactory. The issues ranged from overly generic responses to inaccuracies in the quoted numerical data when compared with the original PDFs. To address these challenges, several prompting techniques were employed: Zero-Shot Prompting, One-Shot Prompting, Few-Shot Prompting, Zero-Shot Chain-of-Thought(CoT) Prompting, and Few-Shot CoT Prompting.

7.1 Zero-Shot Prompting

```
I have a dataset that provides year-wise Key Performance Indicators (KPIs)
for various network operators across multiple cities. The KPIs are network
accessibility (higher is better), service accessibility (higher is better),
call connection time (lower is better), call completion ratio (higher is
better), mean opinion score (higher is better), ISHO for circuit switch voice
(better if present/'yes'), SMS end-to-end delivery time (lower is better),
SMS success rate (higher is better), 3G signal strength (lower is better),
3G signal strength confidence level (higher is better), 4G signal strength
(lower is better), 4G signal strength confidence level (higher is better),
user data throughput 3G (higher is better), and user data throughput for 4G
(higher is better).

As a new resident of CITYNAME, I am exploring the performance of network
operators in the city to find the most reliable service provider.

Could you assist me in evaluating the performance of these operators in
CITYNAME based on the most recent data available? Please provide the analysis
in the following structured format:

1. A detailed comparison of each KPI for the top-performing operator.
2. A brief summary of the overall performance of each operator.
3. A recommendation of which operator stands out as the best (performs better
   in most KPIs) in CITYNAME.

Keep it concise.
```

Figure 1. Prompt for zero-shot prompt

Zero-shot prompting is a technique where a prompt is given to a language model without any specific examples or training on the topic, relying solely on the model's pre-existing knowledge. In our case, we applied this technique to analyze the broadband data for the region of Nowshera from the 2023 PDF. Despite the potential of this approach, the results we obtained were ultimately unsatisfactory. One of the main issues we encountered was with the accuracy of the KPI (Key Performance Indicator) values quoted in the results. These values often differed significantly from the actual data provided in the JSON file, indicating a lack of precision or understanding on the model's part. Furthermore, the results were overly generic, failing to provide in-depth assessments of the data or offer meaningful suggestions for improvement or comparisons between different network operators. Additionally, we found that the recommendations provided by the model for the best network operator were inconsistent and would vary each time the prompt was rerun, indicating a lack of reliability and consistency in the model's output. The following is an example of the zero shot prompt we used to assess the data:

7.2 One-Shot Prompting

One-shot prompting involves providing a single example or piece of context to a language model along with a prompt, allowing the model to generate a response based on this limited input. We applied this technique to analyze the broadband data for the region of Larkana, using an example analysis of Karachi West from the 2023 PDF. However, the results of this analysis revealed some shortcomings. While the model was able to generate responses based on the provided example, there were inaccuracies noted in the KPI (Key Performance Indicator) values listed in the results. Additionally, the suggestions offered by the model were not very strong, lacking in-depth analysis or meaningful insights. Furthermore, the recommendations provided for the best network operator were still not consistent, indicating a lack of reliability in the model's output.

7.3 Few-Shot Prompting

Few-shot prompting involves providing a small number of examples or instances to a language model along with a prompt, enabling the model to generate a response based on this limited set of data. In our study, we applied this technique to analyze broadband data for the region of Larkana, providing example analysis of the regions Karachi West, Chakwal, and Multan. The results of this analysis showed significant improvement compared to the one-shot prompting approach. Specifically, the KPI (Key Performance Indicator) values listed in the results were more accurate, aligning closely with the quantitative values from the original data. This suggests that the model was able to better generalize its understanding from the few examples provided, leading to more precise results. Additionally, the recommendations provided for the best performing network operator were more consistent, indicating that the model's assessment was more robust and reliable. Overall, the use of few-shot prompting resulted in more accurate and reliable insights, highlighting its effectiveness in improving the performance of language models for complex data analysis tasks.

7.4 Chain-of-Thought Prompting

Chain-of-thought prompting involves providing a series of prompts to a language model, allowing it to generate a response by considering the reasoning step by step. We applied this technique to analyze broadband data for the region of Larkana, asking the model to think through the reasoning process and explain it at each step.

When we applied the zero-shot approach with CoT Prompting, we noticed significant enhancements over traditional zero-shot prompting. The results showed better accuracies in KPI values, more accurate suggestions, and consistent recommendations for the best network operator. The zero-shot CoT results were more nuanced, offering suggestions for different operators based on individual KPI performance.

Moving to the one-shot approach with CoT Prompting, we observed notable improvements over the zero-shot CoT solution. This step built upon the zero-shot results, further improving KPI accuracy, suggesting operators more accurately, and providing better operator comparisons. The recommendations became more consistent, and the results included suggestions for different operators based on KPI performance.

Finally, when we employed the few-shot approach with CoT Prompting, we witnessed significant improvements over both the zero-shot and one-shot CoT approaches. The few-shot CoT results showed enhanced KPI accuracies, more concise suggestions, and clearer presentations of information.

Overall, the chain-of-thought prompting technique proved to be highly effective in enhancing the performance of language models for complex data analysis tasks, with each step (zero-shot to one-shot to few-shot) showing progressively greater improvements. The exact improvements made with each Prompting technique employed is quantified and mentioned in the evaluation section.

An example of a zero-shot CoT Prompt we used was:

```
I have a dataset that provides year-wise Key Performance Indicators (KPIs) for
various network operators across multiple cities. The KPIs are network accessibility
(higher is better), service accessibility (higher is better), call connection time
(lower is better), call completion ratio (higher is better), mean opinion score
(higher is better), ISHO for circuit switch voice (better if present/'yes'), SMS
end-to-end delivery time (lower is better), SMS success rate (higher is better),
3G signal strength (lower is better), 3G signal strength confidence level (higher
is better), 4G signal strength (lower is better), 4G signal strength confidence
level (higher is better), user data throughput 3G (higher is better), and user
data throughput for 4G (higher is better). As a new resident of CITYNAME, I am
exploring the performance of network operators in the city to find the most reliable
service provider.
Could you assist me in evaluating the performance of these operators in CITYNAME based
on the most recent data available? Please provide the analysis in the following
structured format:
A detailed comparison of each KPI for the top-performing operator, noting the areas
where this operator excels or performs better than others. Use the comparison to
reason through which operator performs the best in most number of KPIs. A brief
summary of the overall performance of each operator. A recommendation of which
operator stands out as the best (performs better in most KPIs) in CITYNAME. Keep
it concise.
Let's think through this problem step-by-step. If we look at the per-KPI data from
the JSON file, we can identify the best performing network for each KPI for a given
year for a given city. The network that is performing best for the most number of KPIs
in that year for that city is the best general performing network in this context.
```

Figure 2. Prompt for zero-shot COT prompt

8 Evaluation

In order to quantify how well our results worked, we decided to manually check for inaccurate in a 3 PDFs. These PDFs were for the first quarters of the years 2021, 2022 and 2023 which had data for the broadband services in cities of Pakistan. This assessment focuses on comparing the accuracy of the model in a series of tests, including direct analysis of PDF documents, handling of split and merged documents, and responses to different structured formats and diverse prompting strategies. We split the PDFs on different basis depending upon the evalutaion given by GPT-4. Furthermore, we leverage different prompting techniques such as zero-shot, few-shot and chain of thought and even amalgamating them with one another to gauge the effectiveness of each model under diverse situations. This allowed to gauge the

strengths and limitations of each model, providing insights into how we can further develop our model for peak performance. We took missing data to be incorrect data for our calculations. We manually checked the values in the output from GPT-4 with the PDFs.

Figure 3. Accuracy difference between Pre-split and Split data

8.1 LLM selection

Gemini and Phind were unable to process PDFs, leading to their exclusion from our study. We tested the capabilities of both Bing and GPT-4 using PDF documents. GPT-4 displayed a superior ability to provide detailed insights. Although it initially mixed up figures across different Key Performance Indicators (KPIs), it offered more accurate data citations than Bing. To quantify the performance of these models, we defined "accuracy" as the proportion of correctly reported values relative to the total values present in the PDF. Surprisingly, Bing demonstrated an accuracy of 0%; it neither improved on repeated prompting nor did it respond to requests for more accurate data. On the other hand, GPT-4 initially achieved an accuracy of 12.01% and this figure increased with further queries. Based on these results, we decided to continue with GPT-4 for further techniques as it seemed to have more potential for efficiency of future data extractions.

8.2 PDF splitting

In this section of our analysis, we delve into the accuracy of data extraction from pre-split versus post-split PDFs. This comparison is crucial for understanding how the structural format of the PDF can affect processing capability of GPT-4. We initially evaluated the accuracy of the model on a pre-split PDF and then on split PDFs extracted from the original pre-split PDFs. We saw that dividing the PDFs into smaller documents, the model gives more accurate results and is able to contextualize the PDFs better.

Again, we define the accuracy as the total correct values per total fields output by GPT-4. We also took missing data to mean incorrect data. This was essential as there were several fields missing from the output from GPT-4 when queried with the pre-split PDF. The difference in evaluations and incorrect fields is as in Figure 3. We can see the huge difference between the pre-split and split data. On average the pre-split accuracy over the three years is 7.84% and for split PDFs the accuracy is 87.4% which is a huge difference.

8.3 Evaluation of JSON and xlsx

This section provides insight into the accuracy difference between the JSON output and the spreadsheet output that we got from GPT-4. This investigation aims to discern which data structuring approach yields more precise results via GPT-4. Comparison of these formats allows us to choose the format that GPT-4 can better extrapolate to bringing us closer to more efficient analysis results.

Again we took the same definition of accuracy as in the previous two sections alongside taking missing values as incorrect values. GPT-4 gave output with approximately the same accuracy as can be seen in Figure 2.

8.4 Evaluation of Prompting techniques

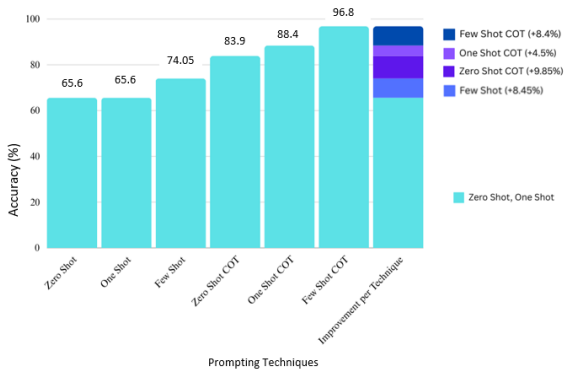

Figure 4. Accuracy difference between different prompting strategies

In this section, we evaluate the accuracy and credibility of the prompting techniques used in our study. We engineered specific prompts tailored to each technique and compared the results against the actual values from the original PDFs. The accuracy scores were calculated based on this comparison and are presented in the ablation study table for prompts as can be seen in Figure 4.

Zero-Shot prompting, despite its ease of use, yielded the lowest accuracy among the techniques employed. The output

generated by this technique did not meet our requirements for accuracy or depth of analysis.

One-Shot prompting, however, presented a unique challenge. We found that the language model tended to memorize the example provided, leading to outputs that closely resembled the example rather than providing a generalized analysis for other cities. This issue became more pronounced when the example output differed from the actual output that should have been returned for a city using that prompt. In such cases, the model would still return results similar to the example, regardless of the differences.

To address this memorization issue, we adopted a different approach in Few-Shot Prompting. Instead of relying on a single example, we included three different examples for each prompt. This change led to a notable improvement in the generated results. By using multiple examples, we mitigated the problem of memorization, resulting in more diverse and accurate recommendations.

Conversely, Few-Shot CoT Prompting produced the highest accuracy. This technique not only provided accurate results but also offered insightful recommendations supported by robust reasoning.

9 Analysis Options

Using the extracted data, our goal was to leverage GPT-4 to provide analysis on various aspects.

Firstly, there's the Recommendations for the Best Performing Network Provider, which involves utilizing extracted data to suggest the optimal network provider in a specific region. This recommendation can be based on diverse criteria such as overall performance across various Key Performance Indicators (KPIs), user preferences for specific KPIs like speed, reliability, and coverage, as well as historical performance data.

Secondly, the analysis includes Trends assessment, which focuses on identifying and quantitatively evaluating trends in KPIs over time. This historical analysis helps in recognizing patterns and developments in network provider performance, aiding in understanding market evolution and potentially predicting future trends.

Thirdly, Benchmarking is another crucial aspect, which entails comparing the performance of different network operators against the top performer in each region. This comparative analysis can reveal areas where other operators are lagging behind the top performer, highlighting potential areas for improvement. Moreover, it can offer insights into effective practices that can be adopted to enhance overall performance.

Lastly, there's Policy Making which involves leveraging extracted data to assess areas where particular KPIs fall short or could be improved. This information is vital for policymakers as it helps in formulating policies that incentivize network providers to enhance their performance in deficient areas. Aligning policy objectives with identified KPI shortcomings can lead to substantial improvements in broadband service quality and delivery.

10 Future Works

Our work offers a great base for future developments in this context. Our simple python scripts allow for an easy process of converting PTA's PDFs into LLM-friendly JSON files so all of this can be extended to any PDFs PTA may release. As shown in our limited evaluation, the results are very promising with GPT-4 returning accurate analyses that capture the crux of what data from a 100-page PDF (or multiple PDFs) is about, in accordance with the requirements specified in our prompts.

We believe that such quick and flexible, data-driven analysis can be integrated in the future into LLM-based dashboards that would allow users to make informed decisions, telecom operators to identify areas of improvement, and policymakers to develop effective strategies. All stakeholders would benefit greatly from such applications.

Finally, an exploration of this study's methodology in terms of generalization (for example, for datasets from different companies) is an important part that we leave out for a later time.

11 Related Works

For our initial exploration of this domain, we explored several related works about LLMs' usage for data processing and prompt engineering.

LLMs for Data Processing. Chen et al. introduced Data Juicer, an LLM-based solution to the processing of massive heterogeneous data. It offers a solid vantage point on the potential of LLMs in this domain. We have set up a base for a similar solution to the problem of PTA's unused datasets.

Prompt Engineering. Many recent works (Nori et al., 2023) have shown the impact of prompt engineering on LLM performance. We have utilised many of these prompting ideas and explored their effects on the performance we observed.

12 Conclusion

In this paper, we detail various techniques we employed to enhance the accuracy of our analysis of PDFs using GPT-4. These techniques include splitting PDFs into manageable segments, advanced prompt engineering, and the merging of JSON files. By integrating these approaches, we were able to effectively query GPT-4 to analyze the data from different years, as captured in the respective JSON files. This methodological advancement led to a dramatic improvement in accuracy, increasing from 7.84% to an impressive 96.8%. Our results affirm the efficacy of these automated PDF analysis methods. This research lays the groundwork for further developments, such as the integration of this system into a

dashboard and the exploration of additional analytical techniques to enhance our understanding of connectivity trends in Pakistan.

References

[1] D. Chen et al., "Data-Juicer: A one-stop data processing system for large language models," 2023. [Online]. Available: https://arxiv.org/abs/2309.02033. [Accessed: May 11, 2024].

[2] H. Nori et al., "Can Generalist Foundation models outcompete special-purpose tuning? case study in medicine," 2023. [Online]. Available: https://arxiv.org/abs/2311.16452. [Accessed: May 11, 2024].

[3] B. Chen et al., "Unleashing the potential of prompt engineering in large language models: A comprehensive review," 2023. [Online]. Available: https://arxiv.org/abs/2310.14735. [Accessed: May 11, 2024].