Dear LLM, Why Are They Complaining? On Root Cause Analysis (RCA) In Customer Service Automation Using Large Language Models (LLMs)

Kadhim Hayawi
AICEO
Arjaan Office Tower 302, AL Sufouh Complex, Tecom
Dubai, UAE
kadhim.hayawi@ceoai.ai

Sakib Shahriar
AICEO
Arjaan Office Tower 302, AL Sufouh Complex, Tecom
Dubai, UAE
sakib.shahriar@ceoai.ai

***Abstract*—The goal of Root Cause Analysis (RCA) is to determine the underlying concern from a host of concerns. In customer service, RCA can help organizations identify fundamental issues in customer complaints and dissatisfaction. Traditional RCA methods relied on human analysis, often prone to subjectivity and bias. Moreover, processing a large volume of complaint narratives demands significant resources. With data-driven approaches like machine learning, the efficiency of RCA has improved. However, these approaches require domain expertise and training data for accurate performance. Large Language Models (LLMs) can mitigate this challenge by learning from limited examples through few-shot learning or without any examples through zero-shot learning. This research introduces a dataset for RCA detection collected from real-world customer complaint narratives across diverse issues. Six state-of-the-art LLMs, including OpenAI's o1, GPT-4o, Sonnet, and Gemini, were evaluated on their performance in detecting RCA. Llama performed the best in open-ended RCA (55%) whereas o1 performed the best in multiclass RCA (68%). Our findings highlight the potential of LLMs in automating RCA. The limitations of current approaches, like model interpretability and domain-specific generalization, were discussed along with future research directions to enhance RCA capabilities further using LLMs.**

Keywords—Root Cause Analysis, Large Language Models, Customer Service Automation, Contextual Learning, Few-shot Learning

I. Introduction

Customer satisfaction is foundational for retaining business success and brand reputation. Customers frequently encounter challenges with products and services due to factors like quality, inadequate instructions, and unrealistic expectations. Organizations can analyze customer complaints to discover patterns of underlying issues that can be systematically resolved to improve overall satisfaction. For instance, recurring complaints about difficulties placing orders via a mobile interface, poor user experience with the application, or limited online services point to the need for improving or overhauling the organization's digital platforms. Therefore, understanding the root cause of customer dissatisfaction will enable the organization to focus its resources more efficiently.

Root Cause Analysis (RCA) is a systematic approach to identifying the underlying causes of recurring problems or faults [1]. Using RCA, organizations can implement solutions that prevent recurrence. RCA relies on inductive and deductive reasoning to establish causal relationships ([2], [3]). RCA has been effectively applied across various fields, including healthcare [4], software testing [5], and manufacturing [6]. RCA enables insights to address complex issues at their source rather than merely addressing their symptoms. Traditional methods of RCA include visualizing causal relations (Ishikawa diagrams) [7], iterative interrogation (five whys) [8], and structured problem improvement (A3 problem solving) [9]. These structured frameworks provide a rudimentary approach to identifying root causes by relying on human expertise. Nonetheless, their reliance on manual intervention makes them time-consuming and inefficient. The advent of data-driven approaches, leveraging machine learning and pattern recognition, has improved the scalability and precision of RCA ([10], [11], [12], [13], [14]). However, these methods still face challenges, including difficulty in handling unstructured data and the need for domain expertise in training data annotation and model development.

Large Language Models (LLMs) are a family of deep learning and transformer models that excel in analyzing and generating human-like text [15]. LLMs can process large unstructured data, and can therefore perform RCA and extract actionable insights with minimal human input. LLMs have been effective in various forms of reasoning [16], domain-specific problem-solving [17], and general knowledge [18]. Unlike machine and deep learning, LLMs can learn with few examples (few-shot learning) or with no examples (zero-shot learning) to overcome the need for domain expertise [19]. In this work, we present a dataset collected for RCA detection based on customer complaints. We evaluate six state-of-the-art LLMs, including OpenAI's o1, GPT-4o, Sonnet 3.5, Llama 3.1, and Gemini 1.5 pro. To the best of our knowledge, this is the first study to explore the application of LLMs in RCA detection.

II. Related Works

Machine learning (ML) algorithms marked a shift from traditional, expert-driven approaches to more automated, data-driven strategies. Early efforts focused on addressing the limitations of manual RCA, such as subjectivity, inefficiency, and scalability issues, by leveraging ML algorithms to process large-scale industrial data. To this end, Ma et al. [14] introduced a comprehensive big-data-driven RCA system for manufacturing, which improved the identification of root causes by employing classification models and feature-based descriptions of quality issues. Similarly, Mueller et al. [20] demonstrated the effectiveness of decision tree algorithms in automating RCA for non-conforming production parts. The proposed approach reduces dependency on domain experts and improves cost efficiency. In anomaly detection, the role of ML in identifying and analyzing deviations from expected behavior in industrial settings has been explored in the literature ([11], [21]). Abdelrahman and Keikhosrokiani [11] utilized ML methods such as k-Nearest Neighbors for anomaly detection in assembly processes. The authors used statistical techniques to isolate key failure points after the anomaly detection phase. Interpretability in ML enables developers to pin down factors contributing to the model's decision-making and improves transparency [22]. To this end, Carletti et al. [21] emphasized the importance of interpretability in ML-based anomaly detection for RCA. The authors proposed a feature importance evaluation approach for the Isolation Forest algorithm to obtain insights into model predictions.

Bayesian networks enable modeling complex causal relationships in manufacturing and industrial alarm systems with potential for RCA. Abele et al. [12] developed a Bayesian network-based alarm system for RCA in industrial plants. Their approach reduced alarm floods and enhanced operator efficiency. Similarly, Lokrantz et al. [6] simulated manufacturing processes using Bayesian networks and demonstrated their utility in transferring expert knowledge across factories. Hybrid methodologies that combine knowledge-driven and data-driven approaches have also been explored. For example, Steenwinckel et al. [23] introduced a system that integrates semantic knowledge and ML techniques to enhance RCA in predictive maintenance. This hybrid method bridges the gap between human expertise and automated systems by combining rule-based reasoning with adaptive ML models. Advanced techniques like neural networks have also been applied in specialized domains. Velásquez and Lara [10] proposed a genetic algorithm-tuned neural network for RCA in power transformers. This method obtained high accuracy in fault diagnosis using gas analysis data. In software engineering, Kahles et al. [5] leveraged artificial neural networks for RCA in agile software testing environments. To extract features from code logs, the authors interviewed software testing engineers (manual annotation). Despite the notable performance (88.9%), the main limitation of this approach is the reliance on domain expertise for manual data annotation. Several challenges exist in ML-based RCA detection, including scalability and integrating batch data. Pal et al. [24] proposed a lifelong learning framework that adapts to evolving data patterns. This approach combines historical batch data with real-time streams for incremental RCA in personalized medicine to balance adaptiveness and computational efficiency. For a more comprehensive review, we refer to the survey paper by Papageorgiou et al. [25], who summarized the role of ML in zero-defect manufacturing.

Deep learning models use layers of artificial neural networks and can handle high-dimensional data. More recently, deep learning methods have been applied for RCA. Saleem et al. [26] leveraged graph neural network (GNN) for real-time fault detection in Industrial Internet of Things (IIoT) edge networks. The classification performance of the model was improved by optimizing the aggregation function and employing efficient sampling techniques. Similarly, Liang et al. [27] combined GNN with knowledge graphs for transport network fault analysis, which allowed effective propagation and aggregation of information within fault graphs. This approach reduces the reliance on domain experts. Temporal data analysis has also been a focus of deep learning-based RCA. Choi et al. [28] used long short-term memory (LSTM) networks to analyze microwave network failures. Their approach led to a 95% accuracy rate (surpassing the 80% benchmark of human engineers) by incorporating contextual data like weather and terrain. In industrial applications, Huang et al. [29] proposed a two-stage (hierarchical) neural network for fault detection and RCA in complex systems. The authors generated saliency maps to identify the most likely root causes of faults by integrating time-series anomaly detection and residual regression. The model demonstrated its capability to overcome limitations, such as misattributing symptoms as root causes. Lastly, in mobile networks, Mampaka and Sumbwanyambe [30] proposed a deep neural network to identify the causes of poor throughput. The model provided end-to-end insights by incorporating radio and core network performance indicators.

Machine and deep learning have demonstrated notable performance in RCA across domains. However, these methods require human effort to curate training data, and the complexity of training such models demands specialized expertise. LLMs have introduced the possibility of zero-shot learning in cross-domain applications, and they offer reasoning capabilities [16] that are pivotal for RCA. While the existing literature explores data-driven RCA across various domains, there is a noticeable gap in applying RCA in customer service. To our knowledge, this research is the first to investigate RCA in customer service and the first to apply RCA using LLMs in any domain.

III. Methods and Materials

Figure 1 summarizes the methodology employed in this study. The details are presented in the following subsections.

A. Data Collection

Given the lack of datasets focused on RCA in customer service, we undertook a data collection initiative for this study. Our source was the Consumer Complaint Database provided by the Consumer Financial Protection Bureau (CFPB) of the United States [31]. This publicly available database offers detailed records of consumer complaints about financial products and services like credit cards, loans, mortgages, and bank accounts. Complaints are published after companies respond or after 15 days. Consumers can choose to share narrative descriptions of their complaints publicly. All consumer records are anonymized to prevent re-identification. While the data is unrepresentative of all consumers' experiences, it provides a foundation for identifying potential root causes in customer service interactions. The Consumer Complaint Database offers detailed records of over 7 million complaints. The large scale of the dataset posed computational challenges for our case study. Thus, we performed filtration and sampling to create a manageable (yet representative) dataset. For sampling, we first selected 13 popular product categories out of the 21. From the 178 total issues and sub-issues within these categories, we ensured diversity by randomly sampling 44 distinct issues (Figure 2). This resulted in a refined dataset comprising 420 customer complaints, each labeled with its corresponding issue. The dataset is publicly available to facilitate further research in customer service analytics (https://github.com/sakibsh/RCA).

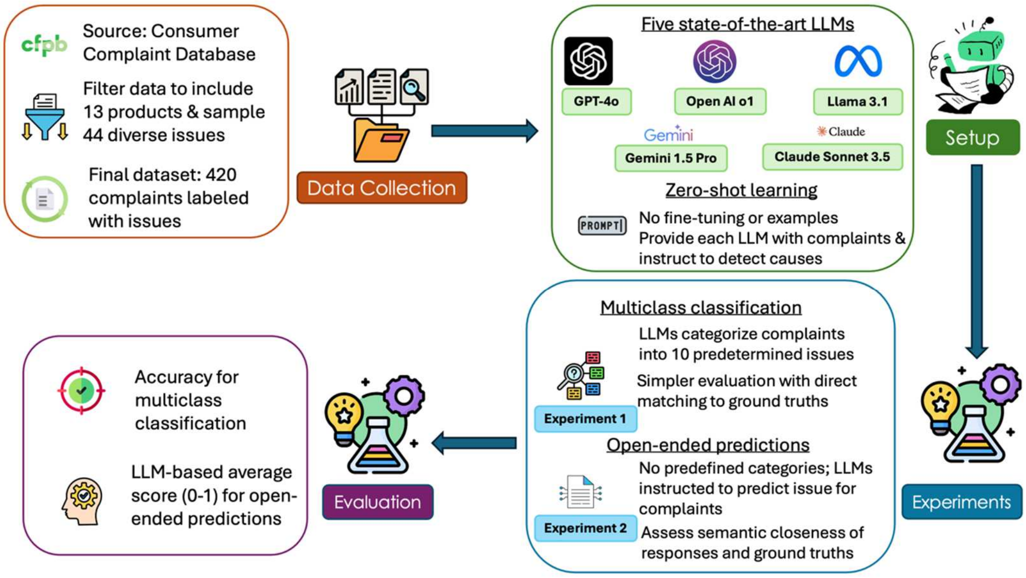

Figure 1 Summary of Methodology

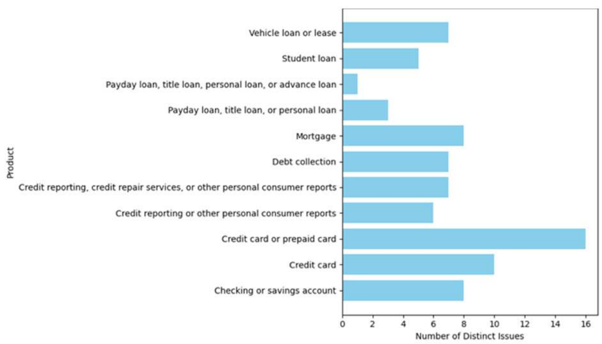

Figure 2. Product by Number of Distinct Issues

B. *Experimental Setup*

The primary objective of our study was to evaluate the effectiveness of LLMs in RCA within the customer service domain. To this end, we selected five state-of-the-art LLMs: GPT-4o, OpenAI O1, Gemini 1.5 Pro, Llama 3.1, and Claude Sonnet 3.5. We focused on zero-shot learning, assuming that the LLMs could detect root causes without requiring fine-tuning or few-shot learning. Each LLM received a customer complaint and was instructed to identify the underlying issue. The primary challenge in this setup was evaluating the correctness of the LLM responses against the ground truth. For example, if the ground truth was "Fraudulent transaction" and the LLM output was "Fraud," the response would still be considered correct, as it indicates the same issue with synonymous phrasing. To address this challenge, our comparative study was segmented into two experiments for appropriate evaluation. In the first experiment, we performed multiclass classification by instructing the LLMs to categorize complaints into 10 predetermined issues. This approach facilitated evaluation since the predicted issues could be directly matched with the ground truth labels without concerns about phrasing variations.

Prompt: Given the complaints below, predict the Issue Type that complaint belongs to. Only respond with the "Issue" type you think it belongs to. Possible Issue Types (choose one for each complaint): Closing an account, Trouble using your card, Problem with a purchase or transfer, Advertising, Fraud or scam, Problem getting a card or closing an account, Communication tactics, Struggling to pay your loan, Received a loan you didn't apply for, Issue with income share agreement

In the second experiment, we did not constrain the LLMs to specific categories. Instead, the models were presented with all complaints and asked to predict the issue for each without predefined categories. To evaluate the responses in this unconstrained setting, we relied on LLM-based scoring to assess the semantic closeness of predictions to the ground truth. This enabled us to account for variations in phrasing and synonyms while still measuring the models' ability to accurately detect root causes.

For the first experiment, where LLMs categorized complaints into ten predetermined issues, we evaluated performance using accuracy. Accuracy was calculated as the ratio of correct predictions to the total number of predictions, as defined by the equation:

$$Accuracy = \frac{N_c}{N_t}$$

where N_c represents the number of correct predictions and N_t represents the total number of samples. In the second experiment, where LLMs were unconstrained by predefined categories, we employed a more flexible evaluation approach. Each LLM was asked to score its predictions on a scale from 0 (totally incorrect) to 1.0 (perfectly correct), based on semantic similarity to the ground truth. The overall performance was then measured by averaging these scores across all predictions. For the LLM-based evaluation, we used the GPT-4o model with the following prompt:

> **Prompt:** Below, First value is Ground Truth of issue, and Second value is predicted issue type by an LLM. I want you to assign a score between 0 and 1 to the prediction based on its accuracy. Wording can be different, as LLM didn't know about the possible issue types, please score based on the prediction.

IV. Results

A. Multiclass RCA

The results from the first experiments are summarized in Table 1, which indicates varying levels of performance across the models in identifying root causes of customer complaints. OpenAI o1 and Gemini 1.5 Pro achieved the highest accuracy at 68%, suggesting they are better equipped to predict categories of causes (Figure 3). GPT-4o follows closely with an accuracy of 66%, showing strong reasoning capabilities. Llama 3.1 and Sonnet, both at 64% accuracy, demonstrate comparable but slightly lower performance. These results show the potential of state-of-the-art LLMs for addressing RCA. Nonetheless, there is room for improvement in minimizing errors that could lead to risks of misidentification. Table 2 presents examples of LLM predictions.

Table 1. LLM Performance in Multiclass RCA

Model	Correct Predictions	Total Predictions	Accuracy
GPT 4o	37	56	0.66
Llama 3.1	36	56	0.64
OpenAI o1	38	56	**0.68**
Gemini 1.5 Pro	38	56	**0.68**
Claude Sonnet	36	56	0.64

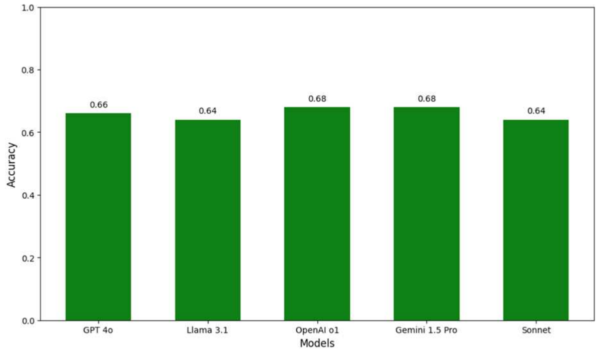

Figure 3. Performance Comparison of LLMs in RCA Classification

Table 2. Sample Complaints and LLM Predictions. Blue shading indicates correct prediction and gray shading indicates incorrect predictions

Complaint Narrative	Actual Issue and Predictions
I had an account in 2010 with fifth third bank. In Now over 10 years began receiving harassing calls from company XXXX stating they just received collection account and submitting lawsuit against me for over {$3000.00}. I was in the XXXX XXXX XXXX XXXX, asked them to stop contacting me as I have never heard of them. They began calling my relatives 3 and 4 times a day, continued contacting me from different numbers. When I asked what the matter was, rep said it was from 2010 account with fifth third bank. I have never received correspondence from anyone stating I owed money. Contacted the branch, they stating they can not do anything as the account is so old they no longer have record of. The account was closed in XX/XX/2010. My credit has never been reported by any company claiming I owed the debt. Again the account was closed. I recall being a class action for the bank at one point as well. They continue to call from XXXX or XXXX area codes or toll free. No prior notice has ever been received previously. As stated above, account was closed over 10 years ago!	Issue (Ground Truth): Closing an account GPT-4o: Communication tactics Llama 3.1: Fraud or scam OpenAI o1: Communication tactics Gemini 1.5 Pro: Struggling to pay your loan Sonnet: Problem with a purchase or transfer
Commenity bank 's website went down sometime in the last two weeks of XX/XX/2022. I was unable to login and retreive account information. About a week later, that issue was resolved and I was able to view most of my account information online. On XX/XX/2022, my first transaction was declined. I figured it was due to their ongoing site/conversion issues and did not use my card for a few days. I have tried to use my card several times up to today 's date - XX/XX/2022 and the card continues to be declined. I do not have a balance on my account and have a credit limit of almost {$18000.00}. I have tried to call in to the customer support number but each time, I get a message that tells me customer support is closed. My calls are during normal business hours. I have tried to call mid-morning and mid-afternoon but get the same message each time. Additionally, I am unable to redeem rewards during this outage. This disruption of business has lasted longer than three weeks with no word from the bank on when it will all be corrected. I have also submitted a message to the bank about the matter via their online messaging system but have not yet received a response.I have several XXXX of dollars in cash back credits that I'd like to use but am unable to do so if I can not use the card to make purchases.	Issue (Ground Truth): Trouble using your card GPT-4o: Trouble using your card Llama 3.1: Trouble using your card OpenAI o1: Trouble using your card Gemini 1.5 Pro: Trouble using your card Sonnet: Trouble using your card
Dear someone who may concern I applied America Express Hilton Honors Surpass Card in XX/XX/2022. According to the advertisement I can earn XXXX Hilton Honors bonus points after I spend {$2000.00} in purchases on the card in the first three months of card membership. My friend referred me and I applied this card online with him together. The entire application process is so smooth and I get my card information directly after file my information. There is no pop up or any notification during entire application. My friend can prove it. Then I get my physical card 5 days later.	Issue (Ground Truth): Advertising GPT-4o: Advertising Llama 3.1: Problem with a purchase or transfer OpenAI o1: Advertising Gemini 1.5 Pro: Advertising Sonnet: Problem with a purchase or transfer

B. *Open-ended Responses*

The ability of LLMs to perform open-ended root cause analysis is necessary because, in real-world scenarios, it is unrealistic to expect pre-determined categories of challenges. The results from the second experiment are summarized in Table 3 and displayed in Figure 4. Llama 3.1 achieved the highest average score of 0.55 (±0.25), and OpenAI o1 closely followed with an average score of 0.53 (±0.24). GPT-4o and Sonnet performed comparably, with average scores of 0.50 and 0.52, respectively, but exhibited higher variability (standard deviation of 0.27), and Gemini 1.5 Pro performed the worst with an average score of 0.48 (±0.24).

Table 3. LLM Performance in Open-ended RCA

Model	Average Score	Standard Deviation
GPT 4o	0.50	0.27
Llama 3.1	**0.55**	0.25
OpenAI o1	**0.53**	0.24
Gemini 1.5 Pro	0.48	0.24
Sonnet	0.52	0.27

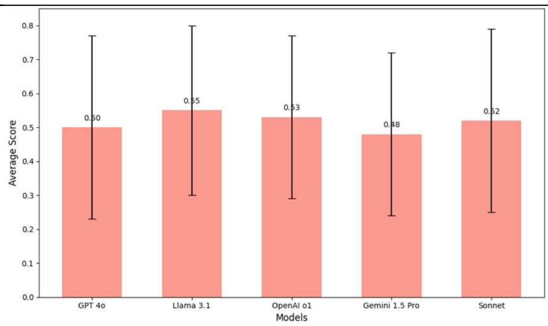

Figure 4. Performance Comparison of LLMs in Open-ended RCA

V. DISCUSSION

A. *Analysis*

The first experiment (with predefined categories of customer complaint issues) allowed LLMs to perform within a constrained framework, which perhaps yielded better performance. This reflects their ability to map inputs to limited and explicit labels. However, such setups are less reflective of real-world scenarios, where categories are often undefined or ambiguous. In contrast, the open-ended second experiment required the models to interpret and reason more flexibly. This led to lower overall scores but arguably offered a more realistic assessment of their RCA capabilities. The results show the trade-off between structured and unstructured tasks. While models excel in tasks with clear boundaries, their performance drops in open-ended contexts due to the lack of fine-tuning or few-shot learning, which could enhance their adaptability. Notably, the variability in performance in the open-ended experiment points to differences in how each model processes semantic nuances and generalizes across diverse scenarios. Perhaps incorporating domain-specific fine-tuning could bridge the gap between the structured and unstructured performance of LLMs.

B. *Implications*

The lack of prior studies on RCA in the customer care domain highlights a critical gap in the literature. We introduce the first dataset curated for RCA in customer care, which will hopefully provide a valuable resource to promote further research in this area. The dataset will facilitate the evaluation of existing methods and serve as a benchmark for developing new approaches. Our findings demonstrate the promise of LLMs in performing zero-shot RCA, enabling them to detect root causes without domain-specific training. However, the results also reveal limitations in their ability to handle open-ended, pointing to the need for few-shot learning to improve their performance. In addition to its technical contributions, this study highlights the importance of using AI to tackle real-world challenges in customer service. These include reducing errors in identifying root causes, speeding up response times, and improving the overall experience for customers.

C. *Limitations*

This study has several limitations that should be addressed in future work. First, the evaluation was limited to only five state-of-the-art LLMs, which may not fully capture the diversity of all models. Second, the study focused exclusively on zero-shot learning without exploring the potential of fine-tuning or few-shot learning in RCA tasks. Third, an ablation study was not conducted to analyze the impact of individual components like prompt designs. Fourth, relying on LLM-based evaluation for the second experiment introduces potential biases, as these models may lack consistency in scoring. The absence of human evaluation further limits the reliability and validity of the findings, particularly in assessing context-dependent outputs. The growing ethical and privacy concerns of AI applications ([32], [33]) also necessitates studying their societal impact in the context of RCA. Lastly, while the introduced dataset is a valuable resource, it is relatively small compared to the scale of real-world customer complaints. The limited size of the dataset may affect the generalizability of the results.

VI. CONCLUSION AND OUTLOOK

Root Cause Analysis (RCA) helps in addressing underlying issues in customer service, yet its application in this domain remains underexplored. In this study, we introduced the first dataset specifically designed for RCA in customer care and evaluated the zero-shot capabilities of five state-of-the-art Large Language Models (LLMs) on structured and open-ended RCA tasks. Our results demonstrate that while LLMs show promise in zero-shot scenarios, their performance is constrained in handling open-ended issues. We found that models performed better when restricted to predefined categories, but open-ended tasks revealed gaps in reasoning and semantic understanding. This highlights the potential for improvement through approaches such as fine-tuning or few-shot learning. In the future, we plan to explore these methods to enhance model adaptability and accuracy. Additionally, we aim to conduct ablation studies to better understand the impact of dataset features and prompt designs on performance. Expanding the dataset to include more diverse and large-scale real-world complaints and incorporating human evaluations for model outputs are also part of our plans. Finally, we also plan to fine-tune the LLMs and compare zero-shot performance with few-shot performance.

References

[1] H. Mohammadnazar, M. Pulkkinen, and H. Ghanbari, "A root cause analysis method for preventing erratic behavior in software development: PEBA," *Reliab. Eng. Syst. Saf.*, vol. 191, p. 106565, Nov. 2019, doi: 10.1016/j.ress.2019.106565.

[2] M. Barsalou, "Explanatory hypothesis generation during a root cause analysis: a multiple case study," *Int. J. Product. Qual. Manag.*, vol. 43, no. 3, pp. 371–384, Jan. 2024, doi: 10.1504/IJPQM.2024.142316.

[3] L. Cuvelier, H. Bencheckroun, and G. Morel, "New vistas on causal-tree methods: from root cause analysis (RCA) to constructive cause analysis (CCA)," *Cogn. Technol. Work*, vol. 19, no. 1, pp. 13–30, Feb. 2017, doi: 10.1007/s10111-017-0404-8.

[4] R. J. Latino, "How is the effectiveness of root cause analysis measured in healthcare?," *J. Healthc. Risk Manag.*, vol. 35, no. 2, pp. 21–30, 2015, doi: 10.1002/jhrm.21198.

[5] J. Kahles, J. Törrönen, T. Huuhtanen, and A. Jung, "Automating Root Cause Analysis via Machine Learning in Agile Software Testing Environments," in *2019 12th IEEE Conference on Software Testing, Validation and Verification (ICST)*, Apr. 2019, pp. 379–390. doi: 10.1109/ICST.2019.00047.

[6] A. Lokrantz, E. Gustavsson, and M. Jirstrand, "Root cause analysis of failures and quality deviations in manufacturing using machine learning," *Procedia CIRP*, vol. 72, pp. 1057–1062, Jan. 2018, doi: 10.1016/j.procir.2018.03.229.

[7] M. F. Suárez-Barraza and F. G. Rodríguez-González, "Cornerstone root causes through the analysis of the Ishikawa diagram, is it possible to find them? A first research approach," *Int. J. Qual. Serv. Sci.*, vol. 11, no. 2, pp. 302–316, Dec. 2018, doi: 10.1108/IJQSS-12-2017-0113.

[8] P. Gangidi, "A systematic approach to root cause analysis using 3 × 5 why's technique," *Int. J. Lean Six Sigma*, vol. 10, no. 1, pp. 295–310, Jul. 2018, doi: 10.1108/IJLSS-10-2017-0114.

[9] J. S. Myers, J. M. Kin, J. E. Billi, K. G. Burke, and R. V. Harrison, "Development and validation of an A3 problem-solving assessment tool and self-instructional package for teachers of quality improvement in healthcare," *BMJ Qual. Saf.*, vol. 31, no. 4, pp. 287–296, Apr. 2022, doi: 10.1136/bmjqs-2020-012105.

[10] R. M. Arias Velásquez and J. V. Mejía Lara, "Root cause analysis improved with machine learning for failure analysis in power transformers," *Eng. Fail. Anal.*, vol. 115, p. 104684, Sep. 2020, doi: 10.1016/j.engfailanal.2020.104684.

[11] O. Abdelrahman and P. Keikhosrokiani, "Assembly Line Anomaly Detection and Root Cause Analysis Using Machine Learning," *IEEE Access*, vol. 8, pp. 189661–189672, 2020, doi: 10.1109/ACCESS.2020.3029826.

[12] L. Abele, M. Anic, T. Gutmann, J. Folmer, M. Kleinsteuber, and B. Vogel-Heuser, "Combining Knowledge Modeling and Machine Learning for Alarm Root Cause Analysis," *IFAC Proc. Vol.*, vol. 46, no. 9, pp. 1843–1848, Jan. 2013, doi: 10.3182/20130619-3-RU-3018.00057.

[13] J. M. N. Gonzalez, J. A. Jimenez, J. C. D. Lopez, and H. A. Parada G, "Root Cause Analysis of Network Failures Using Machine Learning and Summarization Techniques," *IEEE Commun. Mag.*, vol. 55, no. 9, pp. 126–131, Sep. 2017, doi: 10.1109/MCOM.2017.1700066.

[14] Q. Ma, H. Li, and A. Thorstenson, "A big data-driven root cause analysis system: Application of Machine Learning in quality problem solving," *Comput. Ind. Eng.*, vol. 160, p. 107580, Oct. 2021, doi: 10.1016/j.cie.2021.107580.

[15] S. Shahriar and K. Hayawi, "Let's Have a Chat! A Conversation with ChatGPT: Technology, Applications, and Limitations," *Artif. Intell. Appl.*, vol. 2, no. 1, Art. no. 1, 2024, doi: 10.47852/bonviewAIA3202939.

[16] S. Shahriar, "A Cross-Domain Performance Report of Open AI ChatGPT o1 Model," Dec. 23, 2024, *Preprints*: 2024121930. doi: 10.20944/preprints202412.1930.v1.

[17] S. Shahriar *et al.*, "Putting GPT-4o to the Sword: A Comprehensive Evaluation of Language, Vision, Speech, and Multimodal Proficiency," *Appl. Sci.*, vol. 14, no. 17, Art. no. 17, Jan. 2024, doi: 10.3390/app14177782.

[18] W. Yu *et al.*, "MM-Vet: Evaluating Large Multimodal Models for Integrated Capabilities," presented at the Forty-first International Conference on Machine Learning, Jun. 2024. Accessed: Jan. 16, 2025. [Online]. Available: https://openreview.net/forum?id=KOTutrSR2y

[19] Y. Ge *et al.*, "OpenAGI: When LLM Meets Domain Experts," *Adv. Neural Inf. Process. Syst.*, vol. 36, pp. 5539–5568, Dec. 2023.

[20] T. Mueller, J. Greipel, T. Weber, and R. H. Schmitt, "Automated root cause analysis of non-conformities with machine learning algorithms," *J. Mach. Eng.*, vol. Vol. 18, No. 4, 2018, doi: 10.5604/01.3001.0012.7633.

[21] M. Carletti, C. Masiero, A. Beghi, and G. A. Susto, "Explainable Machine Learning in Industry 4.0: Evaluating Feature Importance in Anomaly Detection to Enable Root Cause Analysis," in *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, Oct. 2019, pp. 21–26. doi: 10.1109/SMC.2019.8913901.

[22] S. Shahriar, S. Allana, S. M. Hazratifard, and R. Dara, "A Survey of Privacy Risks and Mitigation Strategies in the Artificial Intelligence Life Cycle," *IEEE Access*, vol. 11, pp. 61829–61854, 2023, doi: 10.1109/ACCESS.2023.3287195.

[23] "FLAGS: A methodology for adaptive anomaly detection and root cause analysis on sensor data streams by fusing expert knowledge with machine learning," *Future Gener. Comput. Syst.*, vol. 116, pp. 30–48, Mar. 2021, doi: 10.1016/j.future.2020.10.015.

[24] G. Pal, X. Hong, Z. Wang, H. Wu, G. Li, and K. Atkinson, "Lifelong Machine Learning and root cause analysis for large-scale cancer patient data," *J. Big Data*, vol. 6, no. 1, p. 108, Dec. 2019, doi: 10.1186/s40537-019-0261-9.

[25] K. Papageorgiou *et al.*, "A systematic review on machine learning methods for root cause analysis towards zero-defect manufacturing," *Front. Manuf. Technol.*, vol. 2, Oct. 2022, doi: 10.3389/fmtec.2022.972712.

[26] A. Saleem, M. Raeiszadeh, A. Ebrahimzadeh, R. H. Glitho, J. Eker, and R. A. F. Mini, "A Deep Learning Approach for Root Cause Analysis in Real-Time IIoT Edge Networks," in *NOMS 2023-2023 IEEE/IFIP Network Operations and Management Symposium*, May 2023, pp. 1–5. doi: 10.1109/NOMS56928.2023.10154218.

[27] Y. Liang *et al.*, "Network Fault Root Cause Analysis Algorithm Based on Deep Learning," in *Proceedings of the 12th International Conference on Computer Engineering and Networks*, Q. Liu, X. Liu, J. Cheng, T. Shen, and Y. Tian, Eds., Singapore: Springer Nature, 2022, pp. 349–358. doi: 10.1007/978-981-19-6901-0_37.

[28] M. Choi, T. Kim, J. pil Lee, and S. Koh, "An Empirical Study on Root Cause Analysis and Prediction of Network Failure Using Deep Learning," in *2021 International Conference on Information and Communication Technology Convergence (ICTC)*, Oct. 2021, pp. 741–746. doi: 10.1109/ICTC52510.2021.9621097.

[29] H. Huang, T. Shah, J. Karigiannis, and S. Evans, "Deep Root Cause Analysis: Unveiling Anomalies and Enhancing Fault Detection in Industrial Time Series," in *2024 International Joint Conference on Neural Networks (IJCNN)*, Jun. 2024, pp. 1–8. doi: 10.1109/IJCNN60899.2024.10650906.

[30] M. M. Mampaka and M. Sumbwanyambe, "Poor Data Throughput Root Cause Analysis in Mobile Networks using Deep Neural Network," in *2019 IEEE 2nd Wireless Africa Conference (WAC)*, Aug. 2019, pp. 1–6. doi: 10.1109/AFRICA.2019.8843409.

[31] "Search the Consumer Complaint Database," Consumer Financial Protection Bureau. Accessed: Jan. 20, 2025. [Online]. Available: https://www.consumerfinance.gov/data-research/consumer-complaints/search/

[32] S. Shahriar and R. Dara, "Priv-IQ: A Benchmark and Comparative Evaluation of Large Multimodal Models on Privacy Competencies," *AI*, vol. 6, no. 2, Art. no. 2, Feb. 2025, doi: 10.3390/ai6020029.

[33] S. Shahriar, R. Dara, and R. Akalu, "A comprehensive review of current trends, challenges, and opportunities in text data privacy," *Comput. Secur.*, vol. 151, p. 104358, Apr. 2025, doi: 10.1016/j.cose.2025.104358.