SMU Data Science Review

Volume 7 | Number 3 Article 3

2023

A Prompt Engineering Approach to Creating Automated Commentary for Microsoft Self-Help Documentation Metric Reports using ChatGPT

Ryan Herrin
Southern Methodist University, rherrin@mail.smu.edu

Luke Stodgel
Southern Methodist University, lstodgel@mail.smu.edu

Brian Raffety
Microsoft Corporation, brian.raffety@microsoft.com

A Prompt Engineering Approach to Creating Automated Commentary for Microsoft Self-Help Documentation Metric Reports using ChatGPT

Ryan Herrin[1], Luke Stodgel[1], Brian Raffety[2]
[1] Master of Science in Data Science, Southern Methodist University, Dallas, TX 75275 USA
[2]Microsoft Corporation, 1 Microsoft Way, Redmond, WA98052 USA
{rherrin, lstodgel}@mail.smu.edu
brian.raffety@microsoft.com

Abstract. Microsoft collects an immense amount of data from the users of their product-self-help documentation. Employees use this data to identify these self-help articles' performance trends and measure their impact on business Key Performance Indicators (KPIs). Microsoft uses various tools like Power BI and Python to analyze this data. The problem is that their analysis and findings are summarized manually. Therefore, this research will improve upon their current analysis methods by applying the latest prompt engineering practices and the power of ChatGPT's large language models (LLMs). Using VBA code, Microsoft Excel, and the ChatGPT API as an Excel add-in, this research will help Microsoft employees more easily identify trends in self-help article metrics, understand the drivers of these trends, and make business decisions that provide the highest return on investment.

1 Introduction

In the beginning, Microsoft built products and services for their customers, and these customers used them. However, some customers had problems, creating a raging river of requests for help. To address this, Microsoft constructed self-help resources—a knowledge dam—to calm the raging river. It provided calm waters for customers to resolve their problems independently. Nonetheless, there were still customers in need of direct assistance from Microsoft. As a solution, Microsoft established an assisted support spillway within the self-help dam, enabling customers to contact Microsoft for help. With each new product and issue, Microsoft continues to fortify the self-help dam, augmenting it with new and enhanced content. To further enhance customer support and optimize their business decisions, Microsoft is now focusing on advancing its methods for analyzing drivers of key performance indicators (KPIs).

Microsoft assesses the impact of self-help content on customer success, support cases created, and cost to identify where it would be best to allocate resources to maximize their return on investment (ROI). However, within their current assessment methods,

there are two places that require improvement: identifying drivers of fluctuations in KPIs and writing summaries of these data insights.

ChatGPT (Chat Generative Pre-Trained Transformer) is an artificial intelligence chatbot released to the public by OpenAI in November 2022. It is a large language model (LLM) that was trained on 154 billion parameters consisting of public books, articles, and webpages. Users can prompt ChatGPT for responses in specific formats, lengths, and language used. Successive prompts and replies are considered at each stage of the conversation as a context. OpenAI's ChatGPT is fine-tuned for conversational applications and was trained using a mix of supervised and reinforcement learning. ChatGPT's ability to generate tailored and context-aware responses makes it the perfect tool for this research, and if applied correctly, it could improve the productivity of Microsoft employees substantially.

Prompt engineering is a concept in artificial intelligence, particularly in natural language processing (NLP), encompassing strategies that enable users to leverage the prompts they create. This approach provides context to pre-trained language models (such as ChatGPT), allowing extracting more accurate and useful responses without the need for retraining the models. One method in prompt engineering is called Chain-of-Thought prompting. This is where you provide the model with logical question-answer examples plus the question you want to answer, all in the same prompt. Wei J. et al. (2023) successfully demonstrated on three large language models that chain-of-thought prompting improved performance on several different arithmetic, commonsense, and symbolic reasoning tasks compared to standard prompting. Past research on prompt engineering will support the research described in this paper.

This research employs VBA code, Microsoft Excel, and the ChatGPT API integrated as an Excel add-in to facilitate Microsoft employees in identifying trends within self-help article metrics, comprehending the underlying factors driving these trends, and making informed business decisions for optimal return on investment. Additionally, the study aims to showcase the capabilities and constraints of the ChatGPT API concerning data ingestion, output, and its interaction with Excel.

2 Literature Review

The literature review focuses on three principal areas: time series data analysis, ChatGPT and large-language models, and prompt engineering.

2.1 Time Series

Numerous research efforts have been dedicated to solving the problem of time series analysis by utilizing contemporary tools like transformer models and neural networks. Notable examples include anomaly detection using decomposition and convolutional neural networks (Gao et al., 2021), probabilistic demand forecasting in retail (Böse et al., 2017), and forecasting, anomaly detection, and classification using Transformers (Wen et al., 2023). These studies demonstrate that modern tools, such as transformer models and neural networks, are being used to push the limits of performance in time series analysis.

This research will use OpenAI's unmodified gpt-3.5-turbo model to conduct commentary on time series web traffic data sourced from Microsoft's product-self-help documentation. Zhou et al. (2023) have conducted relevant research supporting this approach's feasibility. Although their research does not align precisely with the focus of this paper (i.e., this research will not contain fine-tuning of large language models), it still provides valuable support for applying pre-trained transformer models in analyzing time series data.

Zhou et al. (2023) propose a compelling method called "Frozen Pretrained Transformer (FPT)" in their paper titled "Leveraging Pretrained Models for Time Series Analysis" (Zhou et al., 2023). This innovative approach involves utilizing pre-trained models from natural language or image tasks while making minimal alterations to the self-attention and feedforward layers of the residual blocks. The FPT model is subsequently fine-tuned on various time series analysis tasks, including classification, anomaly detection, forecasting, and few-shot learning. Their research findings demonstrate that the FPT model performs on par or better than state-of-the-art approaches across almost all time series tasks (Zhou et al., 2023).

2.2 ChatGPT and Large Language Models

Microsoft employees currently manually write summaries of data insights they extract from self-help content, and they have no cut-and-dry procedure for determining drivers of fluctuations in KPIs. ChatGPT is a Large Language Model OpenAI developed and released to the public in November 2022. Designed to understand and generate human-like responses in natural language conversations, ChatGPT has been trained on a vast amount of text data from the internet, allowing it to have a broad understanding of various topics and generate coherent and contextually relevant responses. To effectively apply ChatGPT in this research, reference is made to several foundational studies that explore its capabilities, limitations, and potential implications (Sun et al., 2023; Qin et al., 2023; Liu et al., 2023; Zhang et al., 2023). These studies highlight ChatGPT's excellence in language understanding, generation, interaction, and reasoning while revealing its potential for search and the challenges it faces in handling specific tasks like sequence tagging. Learning from these studies aims to address limitations and implement prompt engineering methods to enhance ChatGPT's performance in extracting key insights and providing commentary on Microsoft's self-help content data and KPI fluctuations.

One of the objectives of this study revolves around ChatGPT being able to search through data given to extract key insights for providing commentary. According to the research conducted by Sun et al. in "Is ChatGPT Good at Search? Investigating Large Language Models as Re-Ranking Agent" (2023), ChatGPT, and similar language models, excel in language understanding, generation, interaction, and reasoning but do not fully explore the ability to search. The study used ChatGPT turbo-3.5, the model utilized in this research, and found that a custom permutation generation approach enabled ChatGPT to demonstrate search capabilities during passage re-ranking, offering optimism about effectively searching through Microsoft's data.

Another aspect of language models, including ChatGPT, is determining their limits in solving complex questions. This study is from Qin et al. in a paper titled "Is ChatGPT a General-Purpose Natural Language Processing Task Solver?" (Qin et al., 2023). The

researchers explored ChatGPT's proficiency in solving questions through zero-shot learning in various NLP tasks, demonstrating its effectiveness and limitations. Although fine-tuning methods were utilized, the study revealed that ChatGPT was not perfect, but it performed impressively well in reasoning and dialog tasks, solidifying its role as a powerful generalist model. However, the study identified one limitation: ChatGPT's challenges in handling specific tasks, such as sequence tagging. This will be considered when applying prompt engineering strategies to create prompts.

Liu et al.'s study titled "Summary of ChatGPT/GPT-4 Research and Perspective Towards the Future of Large Language Models" (Liu et al., 2023) delved further into the performance evaluation of ChatGPT and GPT-4, aiming to understand its capabilities, ethical concerns, and potential advancements. The paper thoroughly surveyed existing research on ChatGPT and its applications across various domains. The conclusion highlighted ChatGPT's effectiveness in multiple domains while suggesting that fine model tuning and Reinforcement Learning from Human Feedback (RLHF) could further improve its performance. While fine model tuning is beyond the scope of this research, the paper pointed out some limitations that apply to this study, such as "Insufficient Understanding", where complex domain questions may lead to incorrect answers or raise ethical concerns (ex. giving a wrong answer confidently). This research plans to address these limitations by giving ChatGPT comprehensive data within the prompts to perform its analysis.

Zhang et al.'s research paper "Exploring the MIT Mathematics and EECS Curriculum Using Large Language Models" (Zhang et al., 2023) aimed to prove the effectiveness of language models (LLMs), specifically GPT-3.5 and GPT-4, in fulfilling the graduation requirements for any MIT major in Mathematics and EECS. The study achieved excellent results by embedding questions in a low-dimensional space and utilizing few-shot learning to discover relationships. However, the research encountered a limitation due to the token size of 8k in GPT models, referring to the amount of text processed by the model (both input and response). To address this limitation, this research uses a novel prompt chaining method.

2.3 Prompt Engineering

As the field of prompt engineering continues to evolve alongside the development of large language models like ChatGPT, it becomes evident that achieving consistently accurate answers remains an ongoing challenge. Prompt engineering is a relatively new field that has emerged with the development of large language models and their ability to perform well on various tasks with the right prompts. Several researchers have provided a solid foundation for effectively applying prompt engineering strategies in this research, such as Wei et al. (2023), who show that Chain-of-Thought prompting improves the performance of large language models compared to standard prompting, and Wang et al. (2023) who demonstrate that a new prompt engineering strategy, "self-consistency," boosts the performance of chain-of-thought prompting.

Wei et al.'s (2023) study explored how generating a chain of thought significantly improves the ability of large language models to perform complex reasoning. A chain of thought is a sequence of intermediate reasoning steps added to a prompt. The researchers conducted experiments using different large language models and tested the chain of thought prompts across various arithmetic, commonsense, and symbolic

reasoning tasks. The results showed a significant improvement in the performance of large language models when utilizing chain of thought prompts compared to standard prompting. Notably, the study found the most substantial improvements in the models with the most parameters (Wei et al., 2023).

Wang et al.'s (2023) study introduced a novel decoding strategy called self-consistency and compared its performance against chain-of-thought prompting. They tested their approach on various arithmetic and commonsense reasoning tasks using four large language models of varying parameter sizes. They found that with every language model, self-consistency outperformed chain-of-thought prompting by a striking margin across all tasks (Wang et al., 2023).

Overall, the collective insights from the studies mentioned in this literature review provide a solid foundation for leveraging ChatGPT effectively in this research.

3 Methods

This section will describe the data used in this research, how it was preprocessed, its use in prompt creation, the ChatGPT API for Excel add-in, applied prompt engineering strategies, prompt chaining definitions in the context of this research, and scoring methods. All these topics were necessary to facilitate the objective of this research, creating automated commentary.

3.1 Data Description

The data in this research is sourced directly from the Data Science division of Microsoft's Support Business. This data is comprised of metrics related to user behavior when using Microsoft support documentation online. Due to Microsoft's vast collection of support documents for its extensive range of commercial and consumer products, a substantial amount of data is collected. The key metrics of focus in this research are Deflected Cases, Engagement Rate, Visits, and Helpful Response Rate (HRR).

Deflected Cases is a proprietary metric developed in-house at Microsoft, designed to assess whether a user would have contacted technical support if there were no support articles providing a solution. Engagement Rate measures whether a user spent at least 15 seconds on a page, Visits represents the number of page-visits a document received, and HRR is a percentage that gauges whether a user answered "yes" to a survey at the bottom of the document, indicating whether the document was helpful or not.

Microsoft collects this data in the form of time series data, allowing them to track how each of these metrics have changed compared to the previous month(s) or year(s). Based on these metrics, they can determine which area(s) of their business require improvement or adjustments in resource allocation.

With Microsoft's extensive repository of support articles covering numerous products, this data offers significant insights for analysis.

3.2 Data Preprocessing

In addition to the spreadsheet containing the number of Deflected Cases, Visits, the Engagement Rate, and HRR for each software sector's documentation in the current month and previous months, additional calculations were provided. These calculations included the previous six-month average and the current month's percentage of the six-month average. These metrics proved highly informative and straightforward for assessing documentation performance within a specific product sector for the current month. However, a more comprehensive dataset was required to enable ChatGPT to produce analyses that matched or exceeded those generated by human analysts.

Despite the utility of these metrics, additional data was necessary for ChatGPT to perform at the level of a Microsoft employee. Consequently, z-scores were computed for the Deflected Cases, Engagement Rate, Visits, and HRR of each product's current month in relation to the six-month average. Furthermore, z-scores were calculated to assess the data's skewness for these four metrics over the preceding six months. The z-score for each product offered a measurement of how many standard deviations each metric deviated from the mean of the previous six months, which could also be interpreted as an indicator of current month volatility. By employing these metrics, ChatGPT could be provided with a more comprehensive dataset to integrate into its analyses.

3.3 ChatGPT API for Excel Add-In

For this research, the decision was made to incorporate a ChatGPT API add-in into Microsoft Excel. This integration was chosen to seamlessly introduce the tool into Microsoft's existing workflow. Three ChatGPT for Excel add-ins were evaluated based on criteria such as cost-effectiveness, the supported GPT models, and the ability to accept tables as input. The ChatGPT API add-ins that underwent testing were developed by Apps Do Wonders LLC (2023), Talarian S. a. r. l. (2023), and Deepanshu Bhalla (2023), with Deepanshu Bhalla's add-in emerging as the most suitable choice.

Deepanshu's ChatGPT for Excel demonstrated comprehensive functionality, meeting all essential requirements. It leveraged OpenAI's gpt-3.5-turbo model, allowing for prompts with a maximum token length of 4096. Moreover, gpt-3.5-turbo offered a cost-effective per-API-call pricing structure, facilitating extensive testing without budget constraints. Notably, this add-in also supported the transmission of Excel functions, such as TEXTJOIN(), through its API. This feature enabled the chaining of multiple cells, providing a flexible approach to constructing ChatGPT prompts.

In contrast, the ChatGPT API add-ins developed by Apps Do Wonders LLC and Talarian S. a. r. l. were not utilized primarily due to cost considerations. Apps Do Wonders LLC's add-in employed OpenAI's text-davinci-003 model, which incurred significantly higher costs per API call. Furthermore, text-davinci-003 was constrained by a maximum token limit of 2048, which restricted the length of potential prompts. Talarian S. a. r. l.'s add-in, although appearing to offer comprehensive functionality and more, required a subscription fee of $20 per month for usage, making it impractical for the scope of this research.

This selection process ensured that the chosen ChatGPT add-in for Excel aligns with the research's goals and budgetary constraints while maximizing functionality and versatility within the Microsoft Excel environment.

3.4 Prompt Engineering

In this research, three innovative techniques for harnessing the power of large language models will be applied, and their performances will be compared. These techniques include few-shot prompting, chain-of-thought prompting (as proposed by Wei et al. in 2023), and a novel prompt chaining strategy.

Few-shot prompting is a conventional approach where a large language model is presented with multiple question-and-answer pair examples to understand the task at hand. In contrast, chain-of-thought prompting introduces a novel methodology. It involves providing the model with a prompt consisting of triples: input, chain-of-thought, and output. In this framework, the input represents the initial question or problem statement, the chain-of-thought elucidates the intermediate steps and reasoning required to solve the problem, and the output conveys the desired answer or result.

The work by Wei et al. (2023) demonstrated that chain-of-thought prompting significantly enhances the performance of large language models when compared to the traditional few-shot prompting approach. Their findings indicate that this improvement is especially pronounced in the context of complex reasoning tasks. The use of chain-of-thought prompting essentially equips the model with the ability to perform multi-step reasoning, making it better suited for intricate problem-solving scenarios.

Prompt chaining is a technique developed during this research, in which large language model responses from multiple individual queries are combined into a single prompt, which is then used as input for a follow-up query. This approach is particularly valuable when complex analysis of multiple variables is required.

In this research, the goal is to evaluate the effectiveness of these three prompting techniques within the context of a straightforward data analysis summary writing task. By comparing the performances of few-shot, chain-of-thought, and prompt chaining, valuable insights into which approach yields better results can be obtained. This research not only contributes to the broader discussion on prompt engineering for large language models but also has the potential to enhance our understanding of the underlying mechanisms that influence the model's performance on tasks of varying complexity.

3.5 Prompt Chaining

During this research, two types of prompt chaining were identified: straight chaining and code-mediated dynamic branch chaining. While these strategies may seem specialized, they hold relevance in a wide range of applications. In this research, prompt chaining was indispensable for breaking down multi-step analyses into manageable components, ensuring that the large language model could produce accurate overall product summaries. A modified version of Straight Chaining was implemented for this research.

Straight chaining involves feeding the output from one query directly into a follow-up query. An elementary example of this is receiving an answer to a question in one query and then sending another query to ChatGPT for formatting. The modified Straight Chaining method employed in this research combined and concatenated multiple individual responses into a single string, which was then used as the text for a follow-up ChatGPT query. Figure 1 below shows a visual flowchart of the modified straight chaining process we used.

On the other hand, code-mediated dynamic branch chaining is a more intricate approach that requires significantly more effort to implement. With code-mediated dynamic branch chaining, you send a query, receive a response, and depending on that response, a specific branch is activated, triggering another query.

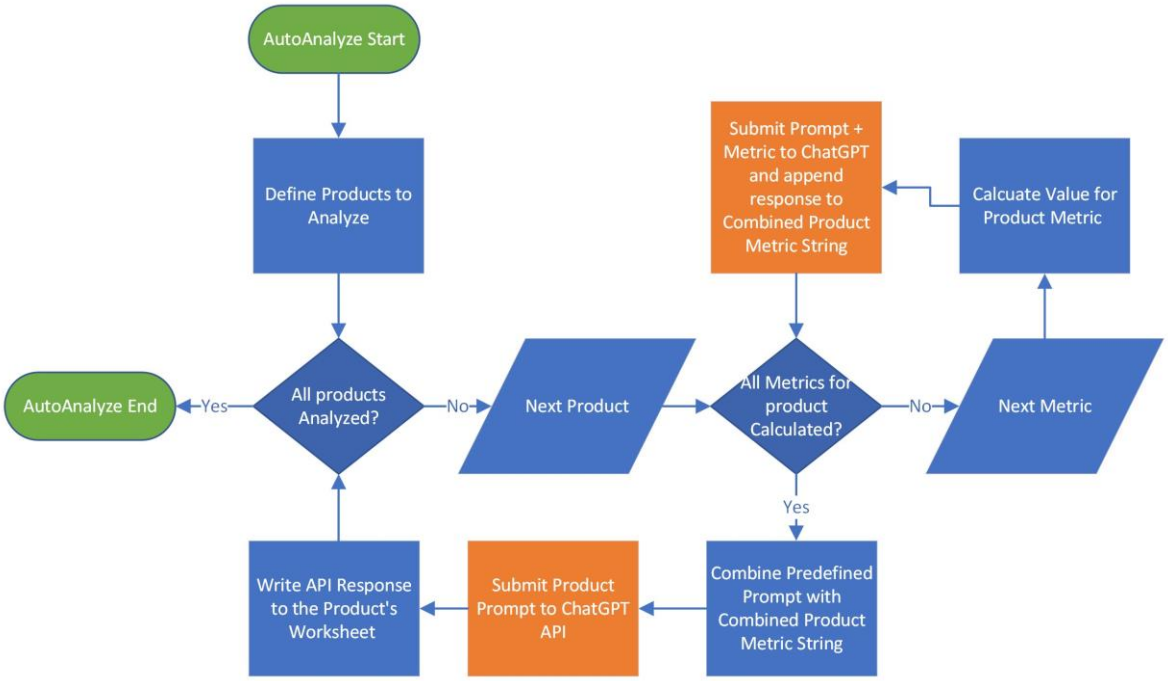

Figure 1. Modified Straight Chaining Method Flowchart

3.6 Prompting Strategy Scoring Method

The determination of analysis accuracy for each prompting strategy tested in this research was based on multiple factors. Each metric, including Deflected Cases, Engagement Rate, Visits, and HRR, had its unique set of criteria: 'Current Month Percent of Six-month Average,' 'Volatility,' and 'Data Distribution' values, for which ChatGPT was instructed to provide specific mentions in its analysis. For example, the 'Current Month Percent of Six-month Average' metric might indicate a value of 80%, signifying a decrease in the current month compared to the previous six months. For an accurate analysis, ChatGPT was required to mention this. Volatility was measured

using a z-score. Depending on the volatility value for Deflected Cases, Engagement Rate, Visits, or HRR, ChatGPT was instructed to indicate whether it was higher or lower than usual for that metric in the current month. Regarding Data Distribution commentary, the z-score of the data from the previous six months for each metric was included. A z-score of one or higher indicated skewed data, suggesting potential unreliability in the analysis, while a lower z-score indicated normal data distribution, making the analysis more dependable. For a visual representation of the metrics and their corresponding variables used in the real prompt tests, please refer to Figure 2 below.

Note: The z-scores for 'Volatility' were converted into one of four categories based on their values, and an absolute value function was applied to facilitate this process. 'Normal' was assigned to z-scores between zero and one, 'Trend' for z-scores above one but below two, 'Changed' for z-scores above two but below three, and 'Warning' for z-scores above three. Furthermore, the z-scores for the data distributions were transformed into 'Normal,' 'Noticeable Departure from Normality,' and 'Substantial Departure from Normality' for z-scores between zero and one, between one and two, and two and above, respectively.

[Metric] = Deflected Cases
[Current Month Percent of Six-month Average] = 97.00%
[Volatility] = Normal
[Distribution] = normal
[Metric] = Engage Rate
[Current Month Percent of Six-month Average] = 103.00%
[Volatility] = Normal
[Distribution] = normal
[Metric] = Visits
[Current Month Percent of Six-month Average] = 109.99%
[Volatility] = Normal
[Distribution] = normal
[Metric] = HRR
[Current Month Percent of Six-month Average] = 85.12%
[Volatility] = Normal
[Distribution] = normal

Figure 2. Example Metrics and Values Used in Prompts

4 Results

The scores for each prompting strategy were determined based on the number of correct analyses relative to the total number of analyses conducted. Specifically, correctness in analysis implied that the product documentation metrics assessment generated by ChatGPT was entirely accurate and encompassed all the information it was instructed to include. Any additional, non-detrimental information did not disqualify an accurate analysis. Conversely, it was considered incorrect if any

information was missing or inaccuracies were present. Each prompting strategy was given a single opportunity to provide analyses for each of the 21 product areas metrics.

The modified straight chaining prompting method exhibited the highest performance, achieving 21 out of 21 accurate analyses. Chain-of-thought prompting ranked as the second best, yielding 10 out of 21 accurate analyses, while the few-shot prompting method performed the least effectively, producing 8 out of 21 accurate analyses. Please refer to the appendix section at the end of this paper for the specific prompts used for each strategy.

For a visual representation of the performance of each prompting strategy, see Figures 3 and 4, below.

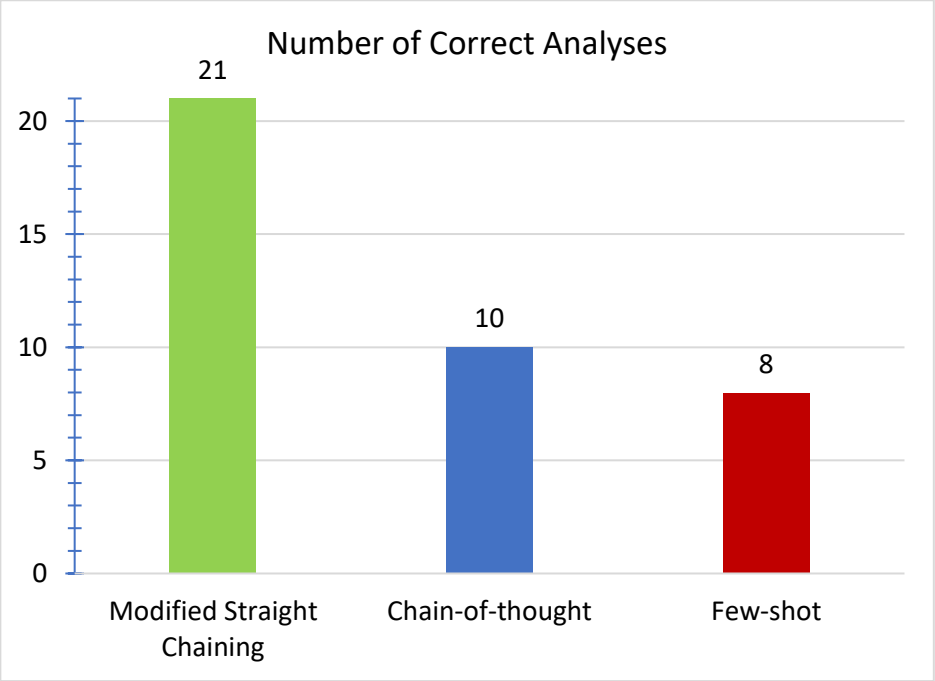

Figure 3. Number of Correct Analyses for Each Prompting Strategy

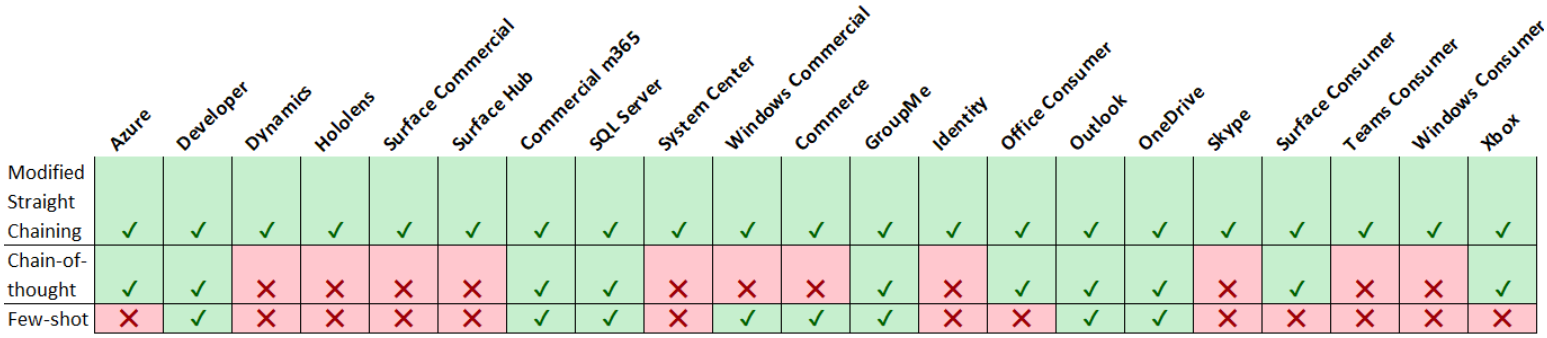

Figure 4. Prompting Strategy Accuracy for Each Product

5 Discussion

5.1 Interpretations

It was initially believed that, with well-crafted prompts, ChatGPT could generate automated commentary using few-shot prompting alone. However, when few-shot prompting product analyses yielded less than 50% accuracy, the chain-of-thought prompting was considered a potential solution to address the errors occurring in ChatGPT's analyses.

Upon observing that chain-of-thought prompting only marginally improved results over the few-shot prompting strategy, it became evident that there might be inherent limitations to the complexity of tasks ChatGPT could handle in a single query using a single prompt. Consequently, the utilization of prompt chaining and the division of analysis tasks into multiple prompt chains emerged as a highly effective solution for applications like the one in this research.

There is also a possibility that the prompts for the few-shot and chain-of-thought prompting strategies employed in this research were not perfect or were too noisy, and those strategies could possibly be successfully applied to this research's task. However, despite the team's strong effort to optimize their effectiveness, success was not achieved.

5.2 Implications

One significant implication of the results is the realization that, despite the application of state-of-the-art, research-backed strategies for crafting robust prompts (Wei et al., 2023), these strategies have consistently improved ChatGPT's response accuracy in various contexts. However, these strategies did not yield the desired results in the context of this research. This research highlights a valuable takeaway: it provides insights into the limitations of ChatGPT in providing accurate responses within specific scopes of information.

5.3 Limitations

One minor limitation involved instructing the program to generate very specific outputs. Responses from large language models consist of predictions based on what the program believes the user wants to hear. Despite the variability in responses, the advantages outweigh the disadvantages in the context of this research. For instance, individuals without programming knowledge but able to formulate effective prompts could successfully operate the tool.

Another limitation was related to the necessity for output to meet the standards of human analysis. The expected format of each product analysis was to first list the product name and then present its associated metric analysis. This format ruled out certain techniques, such as chain-of-thought and self-consistency prompts. In the case of chain-of-thought prompts, the expected output from ChatGPT consists of a series of steps leading to the eventual answer, alongside the answer itself. Regarding self-consistency prompting, you submit the same query multiple times and select the most suitable response, thereby enhancing the overall accuracy of the model's responses. Neither of these methods would have been suitable for the intended product delivery in this research.

The last limitation worth mentioning is related to the use of a ChatGPT API add-in within Microsoft Excel. While the API performed effectively for the research, controlling the API and data through VBA macros in Excel seemed less suitable for collaborative work and data management when compared to Python. For this team specifically, it was the consensus that everything accomplished in this research could be more easily replicated in Python, and using code collaboration tools such as GIT would have been easier as well.

5.4 Ethics

Ethical considerations arise when working with online data, proprietary information, and large language models. This section will address two specific ethical topics: plagiarism and machine learning bias.

Plagiarism is a prominent and complex issue within the AI domain. High-performance large language models are trained on extensive datasets comprising content from various online sources. Consequently, there is ongoing debate about whether the content generated by large language models may be considered plagiarism. This is a crucial aspect to remember, especially when utilizing large-language-model-generated responses in official contexts where ownership and attribution are at stake. Employing such content carries potential legal risks for the organization or individual, as it may lead to claims of copyright infringement.

Machine learning bias is another prominent concern within the field of AI. As AI plays an increasingly substantial role in our daily lives, it is imperative to remain vigilant regarding potential biases in the data used to train AI models and in the output generated by AI systems. Ensuring fairness and non-discrimination is crucial. AI should not exhibit bias based on gender, race, age, or other sensitive attributes.

These ethical considerations are fundamental to the responsible and ethical use of large language models and AI technologies in various applications.

5.5 Future Research

This research has laid the foundation for implementing prompt chaining. Future research could introduce innovative approaches to applying the straight chaining and code-mediated branch chain prompting methods described in section 3.5 of this paper. The potential applications of ChatGPT to automate tasks in the workplace across all industries are extensive.

6 Conclusion

The results of this research signify the achievement of the main research objective: the development of a tool capable of automating commentary for Microsoft self-help documentation metrics using a GPT model.

The prompts to achieve this goal were meticulously crafted to instruct ChatGPT to produce information when specific variables matched certain values for each metric and product. Specifically, ChatGPT was asked to comment on the Current Month Percentage of Six-Month Average, Volatility, and Data distribution for each of the main metrics: Deflected Cases, Engagement Rate, Visits, and Helpful Response Rate (HRR), for each product. The number of product areas the tool provided commentary for was 21.

Three prompting strategies were tested and scored based on accuracy to determine the strongest strategy in the context of this research, and the novel prompt chaining method outperformed both the chain-of-thought and few-shot prompting strategies by a significant margin.

Acknowledgments. Brian Raffety – Capstone Advisor, Jacquelyn Cheun, PhD. – Capstone Professor

References

1. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2023, January 10). *Chain-of-thought prompting elicits reasoning in large language models*. arXiv.org. https://arxiv.org/abs/2201.11903
2. Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., & Zhou, D. (2023, March 7). *Self-consistency improves chain of thought reasoning in language models*. arXiv.org. https://arxiv.org/abs/2203.11171
3. Kamil Malinka, Martin Peresíni, Anton Firc, Ondrej Hujnák, and Filip Janus. 2023. On the Educational Impact of ChatGPT: Is Artificial Intelligence Ready to Obtain a University Degree? In Proceedings of the 2023 Conference on Innovation and Technology in Computer Science Education V. 1 (ITiCSE 2023). Association for Computing Machinery, New York, NY, USA, 47–53. https://doi-org.proxy.libraries.smu.edu/10.1145/3587102.3588827
4. Zhou, T., Niu, P., Wang, X., Sun, L., & Jin, R. (2023, May 25). One Fits All: Power General Time Series Analysis by Pretrained LM. arXiv.org. https://arxiv.org/abs/2302.11939

5. Li, Y., Lu, X., Xiong, H., Tang, J., Su, J., Jin, B., & Dou, D. (2023, January 5). Towards long-term time-series forecasting: Feature, pattern, and distribution. arXiv.org. https://arxiv.org/abs/2301.02068
6. Gao, J., Song, X., Wen, Q., Wang, P., Sun, L., & Xu, H. (2021, September 18). RobustTAD: Robust Time Series Anomaly Detection Via Decomposition and Convolutional Neural Networks. arXiv.org. https://arxiv.org/abs/2002.09545
7. Hyndman, R. J., & Athanasopoulos, G. (2021). Forecasting: Principles and Practice (3rd ed.). OTexts.
8. Böse, J.-H., Flunkert, V., Gasthaus, J., Januschowski, T., Lange, D., Salinas, D., Schelter, S., Seeger, M., and Wang, Y. (2017, August 1). Probabilistic Demand Forecasting at Scale. Proceedings of the VLDB Endowment. https://dl.acm.org/doi/10.14778/3137765.3137775
9. Courty, P. and Li, H. Timing of Seasonal Sales. The Journal of Business. (n.d.). https://www.jstor.org/stable/10.1086/209627
10. Friedman, M. The Interpolation of Time Series by Related Series. (n.d.). https://www.tandfonline.com/doi/abs/10.1080/01621459.1962.10500812
11. Fawaz, H. I., Forestier, G., Weber, J., Idoumghar, L., & Muller, P.-A. (2019, March 2). Deep Learning for Time Series classification: A Review - Data Mining and Knowledge Discovery. https://link.springer.com/article/10.1007/s10618-019-00619-1
12. Wen, Q., Zhou, T., Zhang, C., Chen, W., Ma, Z., Yan, J., & Sun, L. (2023, May 11). Transformers in Time Series: A Survey. arXiv.org. https://arxiv.org/abs/2202.07125
13. Lim, B., Arik, S. O., Loeff, N., & Pfister, T. (2020, September 27). Temporal Fusion Transformers for Interpretable Multi-horizon Time Series Forecasting. arXiv.org. https://arxiv.org/abs/1912.09363
14. Weiwei Sun, Lingyong Yan, Xinyu Ma, Pengjie Ren, Dawei Yin, Zhaochun Ren, (April 19, 2023). Is ChatGPT Good at Search? Investigating Large Language Models as Re-Ranking Agent. https://arxiv.org/abs/2304.09542
15. Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, Diyi Yang (February 15, 2023) Is ChatGPT a General-Purpose Natural Language Processing Task Solver? https://arxiv.org/abs/2302.06476v2
16. Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, Zihao Wu, Dajiang Zhu, Xiang Li, Ning Qiang, Dingang Shen, Tianming Liu, Bao Ge. (May 11, 2023) Summary of ChatGPT/GPT-4 Research and Perspective Towards the Future of Large Language Models. https://arxiv.org/pdf/2304.01852
17. Sarah J. Zhang, Samuel Florin, Ariel N. Lee, Eamon Niknafs, Andrei Marginean, Annie Wang, Keith Tyser, Zad Chin, Yann Hicke, Nikhil Singh, Madeleine Udell, Yoon Kim, Tonio Buonassisi, Armando Solar-Lezama, Iddo Drori. (June 15, 2023) Exploring the MIT Mathematics and EECS Curriculum Using Large Language Models. https://arxiv.org/abs/2306.08997v1
18. Perrigo, B. (2023, April 13). The A to Z of artificial intelligence. Time. https://time.com/6271657/a-to-z-of-artificial-intelligence/
19. Lock, S. (2022, December 5). What Is Ai Chatbot Phenomenon ChatGPT and Could It Replace Humans?. The Guardian. https://www.theguardian.com/technology/2022/dec/05/what-is-ai-chatbot-phenomenon-chatgpt-and-could-it-replace-humans
20. Shekhar, G. (2022, May 26). Causal AI - Enabling Data-Driven Decisions. Medium. https://towardsdatascience.com/causal-ai-enabling-data-driven-decisions-d162f2a2f15e
21. Bhalla, D. (n.d.). *Chatgpt: Best excel plugin for Excel Users*. ListenData. https://www.listendata.com/2023/04/excel-add-in-for-chatgpt.html

Appendix

A.) Modified straight-chaining strategy - first and second chain prompts

First prompt, chain 1

First, say what the [Metric] is. Second, comment on the metric's [Current Month Percent of Six-month Average] and [Volatility]. If [Volatility] = Normal, say '[Metric] is at [Current Month Percent of Six-month Average]% of its six-month average, which is within the normal expected range.' If [Volatility] = Trend, say '[Metric] is at [Current Month Percent of Six-month Average]% of its six-month average, indicating a trend. If this value was displayed on a distribution chart, it would be in the yellow zone.' If [Volatility] = Changed, say '[Metric] is at [Current Month Percent of Six-month Average]% of its six-month average, indicating high volatility and warranting investigation. While there is a less than 5% chance of this value occurring by chance, it is not extreme enough to suggest a data error. If displayed on a distribution chart, [Metric] would be in the red zone.' If [Volatility] = WARNING, say '[Metric] is at [Current Month Percent of Six-month Average]% of its six-month average, indicating extreme volatility. The likelihood of this value occurring by chance alone is less than 0.2%, suggesting a potential data error. If represented on a distribution chart, this metric would fall well into the red zone.' Finally, talk about the [Distribution]. If [Distribution] = normal, say 'The data distribution for [Metric] is normal, supporting the reliability of this analysis.' If [Distribution] = noticeable departure from normality, say 'The data distribution for [Metric] shows a noticeable departure from normality, raising concerns about the validity of its analysis.' If [Distribution] = substantial departure from normality, say 'The data distribution for [Metric] shows a substantial departure from normality, raising concerns about the validity of its analysis.'

Here is the data:

[Metric] = Engage Rate
[Current Month Percent of Six-month Average] = 102.47
[Volatility] = Normal
[Distribution] = normal

Second prompt, chain 2

First say what the product is. Product = Azure
Then, summarize the following text about Deflected Cases, Engage Rate, Visits, and HRR (Helpful Response Rate).
Deflected Cases is at x% of its six-month average, which is within the normal expected range. The data distribution for Deflected Cases is normal, supporting the reliability of this analysis.
Engage Rate is at x% of its six-month average, which is within the normal expected range. The data distribution for Engage Rate is normal, supporting the reliability of this analysis.
Visits is at x% of its six-month average, which is within the normal expected range. The data distribution for Visits is normal, supporting the reliability of this analysis.
HRR is at x% of its six-month average, indicating a trend. If this value was displayed on a distribution chart, it would be in the yellow zone. The data distribution for HRR is normal, supporting the reliability of this analysis.

B.) Chain-of-Thought Prompt

Q: Product = Jeans
[Metric] = Deflected Cases
[Current Month Percent of Six-month Average] = 97.00%
[Volatility] = Normal
[Distribution] = normal
[Metric] = Engage Rate
[Current Month Percent of Six-month Average] = 103.00%
[Volatility] = Normal
[Distribution] = normal
[Metric] = Visits
[Current Month Percent of Six-month Average] = 109.99%
[Volatility] = Normal
[Distribution] = normal
[Metric] = HRR
[Current Month Percent of Six-month Average] = 85.12%
[Volatility] = Normal
[Distribution] = normal

A: The Product = Jeans, add that to the beginning of the final summary. Next, talk about the metrics' [Current Month Percent of Six-month Average] and [volatility]. Pair together metrics based on [volatility] and append the analysis to the final summary. Because all metrics have [volatility] = normal, omit mentioning where they would land on a distribution chart. For example: 'The Deflected Cases, Engage Rate, Visits, and HRR for this month are at 97.00%, 103.00%, 109.99%, and 85.12% of their respective six-month averages, all falling within normal ranges.' Next comment on the [Distribution] of each metric, pairing metrics that have similar values. In this case, append a sentence like this to the final summary: 'The data distribution of Deflected Cases, Visits, and HRR were normal, reinforcing the reliability of their analyses. The data distribution of Engage Rate on the other hand, showed a noticeable departure from normality, raising concerns about the validity of its analysis.' The answer is 'Product = Je
ans. The Deflected Cases, Engage Rate, Visits, and HRR for this month are at 97.00%, 103.00%, 109.99%, and 85.12% of their respective six-month all falling within normal ranges. Additionally, the data distributions of Deflected Cases, Visits, and HRR were normal, reinforcing the reliability of their analyses. The data distribution of Engage Rate on the other hand, showed a noticeable departure from normality, raising concerns about the validity of its analysis.

Q: Product = Sofas
[Metric] = Deflected Cases
[Current Month Percent of Six-month Average] = 130.46%
[Volatility] = WARNING
[Distribution] = noticeable departure from normality.
[Metric] = Engage Rate
[Current Month Percent of Six-month Average] = 65.34%
[Volatility] = WARNING
[Distribution] = normal
[Metric] = Visits
[Current Month Percent of Six-month Average] = 70.27%
[Volatility] = WARNING
[Distribution] = normal
[Metric] = HRR
[Current Month Percent of Six-month Average] = 105.90%
[Volatility] = Changed
[Distribution] = normal

A: The Product = Sofas, add that to the beginning of our final summary. Next, talk about the metrics' [Current Month Percent of Six-month Average] and [Volatility]. Pair together metrics based on [Volatility] and append the analysis to the final summary. Since Deflected Cases, Engage Rate, and Visits have [Volatility] = WARNING, they exhibit extreme volatility, low probability of occurring, and if presented on a distribution

chart they would be well into the red. HRR has [Volatility] = Changed, therefore showing high volatility, low probability of occurring and if presented on a distribution chart, would be in the red. Here is an example of how to summarize these: 'Deflected Cases, Engage Rate, and Visits are exhibiting extreme volatility this month, standing at 130.46%, 65.34%, and 70.27% of their respective six-month averages. The likelihood of these values occurring by chance alone is less than 0.2%, suggesting a potential data error. If represented on a distribution chart, these metrics would fall well into the red zone. HRR for this month is at 105.90% of its six-month average, indicating high volatility and warranting investigation. While there is a less than 5% chance of this value occurring by chance, it is not extreme enough to suggest a data error. If displayed on a distribution chart, HRR would also be in the red zone.' Next comment on the [Distribution] of each metric, pairing metrics that have similar values. In this case, Deflected cases has a [Distribution] value of 'noticeable departure from normality', and the three other metrics have normal data distributions. In this case we would append a sentence like this to the final summary: 'The data distribution of Deflected Cases showed a noticeable departure from normality, raising concerns about the validity of its analysis. The data distribution of Engage Rate, Visits, and HRR were normal, reinforcing the reliability of their analyses.' The answer is 'Product = Sofas. Deflected Cases, Engage Rate, and Visits are exhibiting extreme volatility this month, standing at 130.46%, 65.34%, and 70.27% of their respective six-month averages. The likelihood of these values occurring by chance alone is less than 0.2%, suggesting a potential data error. If represented on a distribution chart, these metrics would fall well into the red zone. HRR for this month is at 105.90% of its six-month average, indicating high volatility and warranting investigation. While there is a less than 5% chance of this value occurring by chance, it is not extreme enough to suggest a data error. If displayed on a distribution chart, HRR would also be in the red zone. Additionally, the data distributions of Deflected Cases showed a noticeable departure from normality, raising concerns about the validity of its analysis. The data distribution of Engage Rate, Visits, and HRR were normal, reinforcing the reliability of their analyses.'

Q: Product = Beds
[Metric] = Deflected Cases
[Current Month Percent of Six-month Average] = 108.46%
[Volatility] = Changed
[Distribution] = normal
[Metric] = Engage Rate
[Current Month Percent of Six-month Average] = 83.34%
[Volatility] = Changed
[Distribution] = noticeable departure from normality
[Metric] = Visits
[Current Month Percent of Six-month Average] = 95.27%
[Volatility] = Normal
[Distribution] = normal
[Metric] = HRR
[Current Month Percent of Six-month Average] = 145.90%
[Volatility] = WARNING
[Distribution] = normal

A: The Product = Beds, add that to the beginning of our final summary. Next, talk about the metrics' [Current Month Percent of Six-month Average] and [Volatility]. Pair together metrics based on [Volatility] and append the analysis to the final summary. Because Deflected Cases and Engage Rate both have [Volatility] = Changed, they exhibit high volatility, low probability of occurring, and if presented on a distribution chart they would be in the red. So we append this to the summary: 'Deflected Cases and Engage Rate for this month are at 108.46% and 83.34% of their respective six-month averages, indicating high volatility and warranting investigation. While there is a less than 5% chance of these values occurring by chance, they are not extreme enough to suggest data errors. If displayed on a distribution chart, they would be in the red zone. Visits has normal [Volatility] so we append this to the final summary: 'Visits is at 95.27% of its six-month average, remaining within the normal expected range.' Because HRR has [Volatility] = WARNING, it exhibits extreme volatility, low probability of occurring, and if presented on a distribution chart they would be well into the red. So we would append something like this to the final summary: 'HRR is showing extreme volatility, standing at 145.90% of its six-month average. The likelihood of this value occurring by chance alone is less than 0.2%, suggesting a potential data error. This metric would fall significantly into the red zone.' Next comment on the [Distribution] of each metric, pairing metrics that have similar values. In this case, Deflected cases, Visits, and HRR all have a [Distribution] = normal, so we would append this to the summary: 'The data distribution of Engage Rate, Visits, and HRR were normal, reinforcing the reliability of

their analyses.' Engage Rate has a [Distribution] = noticeable departure from normality, so we would append this to the summary: 'In contrast, the data distribution of Engage Rate showed a noticeable departure from normality, raising concerns about the validity of its analysis.' The answer is 'Product = Beds. Deflected Cases and Engage Rate for this month are at 108.46% and 83.34% of their respective six-month averages, indicating high volatility and warranting investigation. While there is a less than 5% chance of these values occurring by chance, they are not extreme enough to suggest data errors. If displayed on a distribution chart, they would be in the red zone. Visits is at 95.27% of its six-month average, remaining within its normal expected range. HRR is showing extreme volatility, standing at 145.90% of its six-month average. The likelihood of this value occurring by chance alone is less than 0.2%, suggesting a potential data error. This metric would fall significantly into the red zone. The data distributions of Engage Rate, Visits, and HRR were normal, reinforcing the reliability of their analyses. In contrast, the data distribution of Engage Rate showed a noticeable departure from normality, raising concerns about the
validity of its analysis.'

Q: Product = Tires
[Metric] = Deflected Cases
[Current Month Percent of Six-month Average] = 105.76%
[Volatility] = Trend
[Distribution] = normal
[Metric] = Engage Rate
[Current Month Percent of Six-month Average] = 98.89%
[Volatility] = Normal
[Distribution] = normal
[Metric] = Visits
[Current Month Percent of Six-month Average] = 89.32%
[Volatility] = Changed
[Distribution] = normal
[Metric] = HRR
[Current Month Percent of Six-month Average] = 131.17%
[Volatility] = WARNING
[Distribution] = normal

A: The Product = Tires, add that to the beginning of our final summary. Next, talk about the metrics' [Current Month Percent of Six-month Average] and [Volatility]. Pair together metrics based on [Volatility] and append the analysis to the final summary. Since Deflected Cases has [Volatility] = trend, mention the direction it is trending and also that if it were displayed on a distribution chart, it would land in the yellow zone. For example: 'This month, Deflected Cases is trending upward at 105.76% of its six-month average. If displayed on a distribution chart, it would be in the yellow zone.' Engage Rate has [Volatility] = normal, so all you have to mention is it's Current Month Percent of Six-month Average and nothing about a distribution chart. For example: 'On the other hand, Engage rate is normal, standing at 98.89% of its six-month average.' Visits has [Volatility] = Changed, indicating high volatility, low probability of occurring and if presented on a distribution chart, it would be in the red. So, append this to the final summary: "Visits for this month is at 89.32% of its six-month average, indicating high volatility and warranting investigation. While there is a less than 5% chance of this value occurring by chance, it is not extreme enough to suggest a data error. If displayed on a distribution chart, Visits would be in the red zone.' HRR has [Volatility] = WARNING, meaning it is exhibiting extreme volatility, low probability of occurrence, and if presented on a distribution chart it would be well into the red. Here is an example of how to summarize this: 'HRR is exhibiting extreme volatility, standing at 131.17% of its six-month average. The likelihood of this value occurring by chance alone is less than 0.2%, suggesting a potential data error. If represented on a distribution chart, it would fall well into the red zone.' Next comment on the Distribution of each metric, pairing metrics that have similar values. In this case, Deflected cases has a [Distribution] = noticeable departure from normality, and the three other metrics have normal data distributions. In this case we would append something like this to the final summary: 'Additionally, the data distribution of Deflected Cases showed a noticeable departure from normality, raising concerns about the validity of its analysis. The data distributions of Engage Rate, Visits, and HRR were normal, reinforcing the reliability of their analyses.' So, the answer is 'Product = Tires. This month, Deflected Cases is trending upward at 105.76% of its six-month average. If displayed on a distribution chart, it would be in the yellow zone. On the other hand, Engage rate is normal, standing at 98.89% of its six-month average. Visits for this month is at 89.32% of its six-month average, indicating high volatility and warranting investigation. While there is a less than 5% chance of this value occurring by chance, it is not extreme enough

to suggest a data error. If displayed on a distribution chart, HRR would also be in the red zone. HRR is exhibiting extreme volatility, standing at 131.17% of its six-month average. The likelihood of these values occurring by chance alone is less than 0.2%, suggesting a potential data error. If represented on a distribution chart, it would fall well into the red zone. Additionally, the data distribution of Deflected Cases showed a noticeable departure from normality, raising concerns about the validity of its analysis. The data distributions of Engage Rate, Visits, and HRR were normal, reinforcing the reliability of their analyses.

Q: Product = Azure
[Metric] = Deflected Cases
[Current Month Percent of Six-month Average] = x
[Volatility] = x
[Distribution] = x
[Metric] = Engage Rate
[Current Month Percent of Six-month Average] = x
[Volatility] = x
[Distribution] = x
[Metric] = x
[Current Month Percent of Six-month Average] = x
[Volatility] = x
[Distribution] = x
[Metric] = HRR
[Current Month Percent of Six-month Average] = x
[Volatility] = x
[Distribution] = x

A:

C.) Few-shot prompt

Q: Product = Jeans
[Metric] = Deflected Cases
[Current Month Percent of Six-month Average] = 97.00%
[Volatility] = Normal
[Distribution] = normal
[Metric] = Engage Rate
[Current Month Percent of Six-month Average] = 103.00%
[Volatility] = Normal
[Distribution] = normal
[Metric] = Visits
[Current Month Percent of Six-month Average] = 109.99%
[Volatility] = Normal
[Distribution] = normal
[Metric] = HRR
[Current Month Percent of Six-month Average] = 85.12%
[Volatility] = Normal
[Distribution] = normal

A: Product = Jeans
Deflected Cases, Engage Rate, Visits, and HRR currently stand at 97.00%, 103.00%, 109.99%, and 85.12% of their respective six-month averages, all falling within the normal range.
Additionally, the data distribution for all four metrics is normal, reinforcing the reliability of this analysis.

Q: Product = Refrigerators
[Metric] = Deflected Cases
[Current Month Percent of Six-month Average] = 105.46%
[Volatility] = Normal
[Distribution] = normal
[Metric] = Engage Rate

[Current Month Percent of Six-month Average] = 96.34%
[Volatility] = Normal
[Distribution] = normal
[Metric] = Visits
[Current Month Percent of Six-month Average] = 94.27%
[Volatility] = Normal
[Distribution] = normal
[Metric] = HRR
[Current Month Percent of Six-month Average] = 115.90%
[Volatility] = Normal
[Distribution] = normal

A: Product = Refrigerators
Deflected Cases, Engage Rate, Visits, and HRR currently stand at 105.46%, 96.34%, 94.27%, and 115.90% of their respective six-month averages, all falling within the normal range.
Additionally, the data distribution for all four metrics is normal, reinforcing the reliability of this analysis.

Q: Product = Skincare products
[Metric] = Deflected Cases
[Current Month Percent of Six-month Average] = 107.46%
[Volatility] = Changed
[Distribution] = normal
[Metric] = Engage Rate
[Current Month Percent of Six-month Average] = 94.91%
[Volatility] = Trend
[Distribution] = normal
[Metric] = Visits
[Current Month Percent of Six-month Average] = 96.27%
[Volatility] = Trend
[Distribution] = normal
[Metric] = HRR
[Current Month Percent of Six-month Average] = 106.90%
[Volatility] = Trend
[Distribution] = normal

A: Product = Skincare products
This month, Deflected Cases is at 107.46% of its six-month average, indicating very high volatility and warranting investigation. While there is a less than 5% chance of this value occurring by chance, it is not extreme enough to suggest a data error. If displayed on a distribution chart, it would be in the red zone. In contrast, Engage Rate and Visits are trending downward, while HRR is trending upward, registering at 94.91%, 96.27%, and 106.90%, respectively. If displayed on a distribution chart, these metrics would fall within the yellow zone.
Additionally, the data distribution for Deflected Cases, Engage Rate, and Visits remain normal, reinforcing the reliability of the analysis for these metrics. However, the data distribution for HRR exhibits a noticeable departure from normality, prompting concerns about the validity of the analysis for this metric.

Q: Product = Sofas
[Metric] = Deflected Cases
[Current Month Percent of Six-month Average] = 130.46%
[Volatility] = WARNING
[Distribution] = noticeable departure from normality
[Metric] = Engage Rate
[Current Month Percent of Six-month Average] = 65.34%
[Volatility] = WARNING
[Distribution] = normal
[Metric] = Visits
[Current Month Percent of Six-month Average] = 70.27%
[Volatility] = WARNING
[Distribution] = normal

[Metric] = HRR
[Current Month Percent of Six-month Average] = 105.90%
[Volatility] = Changed
[Distribution] = normal

A: Product = Sofas
Deflected Cases, Engage Rate, and Visits exhibit extreme volatility, standing at 130.46%, 65.34%, and 70.27% of their respective six-month averages. The likelihood of these values occurring by chance alone is less than 0.2%, suggesting a potential data error. If represented on a distribution chart, these metrics would fall well into the red zone. HRR for this month registers at 105.90% of its six-month average, indicating high volatility and warranting investigation. While there is a less than 5% chance of this value occurring by chance, it is not extreme enough to suggest a data error. If displayed on a distribution chart, HRR would also be in the red zone.
Additionally, the data distribution for Deflected Cases shows a noticeable departure from normality, raising concerns about the validity of the analysis for this metric. In contrast, the data distribution for Engage Rate, Visits, and HRR remain normal, reinforcing the reliability of the analysis for these metrics.

Q: Product = Beds
[Metric] = Deflected Cases
[Current Month Percent of Six-month Average] = 108.46%
[Volatility] = Changed
[Distribution] = normal
[Metric] = Engage Rate
[Current Month Percent of Six-month Average] = 83.34%
[Volatility] = Changed
[Distribution] = noticeable departure from normality
[Metric] = Visits
[Current Month Percent of Six-month Average] = 95.27%
[Volatility] = Normal
[Distribution] = normal
[Metric] = HRR
[Current Month Percent of Six-month Average] = 145.90%
[Volatility] = WARNING
[Distribution] = normal

A: Product = Beds
Deflected Cases and Engage Rate for this month are at 108.46% and 83.34% of their respective six-month averages, indicating high volatility and warranting investigation. While there is a less than 5% chance of these values occurring by chance, they are not extreme enough to suggest data errors. If displayed on a distribution chart, they would be in the red zone. Visits is at 95.27% of its six-month average, remaining within the normal expected range. In contrast, HRR is showing extreme volatility, standing at 145.90% of its six-month average. The likelihood of this value occurring by chance alone is less than 0.2%, suggesting a potential data error. This metric would fall significantly into the red zone.
Additionally, the data distribution for Deflected Cases, Visits, and HRR are normal, reinforcing the reliability of the analysis for these metrics. In contrast, the data distribution for Engage Rate showed a noticeable departure from normality, raising concerns about the validity of the analysis for this metric.

Q: Product = Motorcycles
[Metric] = Deflected Cases
[Current Month Percent of Six-month Average] = 101.33%
[Volatility] = Normal
[Distribution] = normal
[Metric] = Engage Rate
[Current Month Percent of Six-month Average] = 96.12%
[Volatility] = Trend
[Distribution] = normal
[Metric] = Visits
[Current Month Percent of Six-month Average] = 91.34%
[Volatility] = Changed

[Distribution] = normal
[Metric] = HRR
[Current Month Percent of Six-month Average] = 155.55%
[Volatility] = WARNING
[Distribution] = normal

A: Product = Motorcycles
Deflected cases is normal this month. Engage Rate is trending downward, at 96.12% of its six-month average. If displayed on a distribution chart, it would be in the yellow zone. Visits showed very high volatility at 91.34% of its six-month average, warranting investigation. While there is a less than 5% chance of this value occurring by chance, it is not extreme enough to suggest a data error. If displayed on a distribution chart, Visits would be in the red zone. HRR showed extreme volatility, standing at 155.55% of its six-month average. The likelihood of these values occurring by chance alone is less than 0.2%, suggesting a potential data error. If displayed on a distribution chart, HRR would also be in the red zone.
Additionally, the data distribution for all four metrics is normal, reinforcing the reliability of this analysis.

Q: Product = Tires
[Metric] = Deflected Cases
[Current Month Percent of Six-month Average] = 105.76%
[Volatility] = Trend
[Distribution] = normal
[Metric] = Engage Rate
[Current Month Percent of Six-month Average] = 98.89%
[Volatility] = Normal
[Distribution] = normal
[Metric] = Visits
[Current Month Percent of Six-month Average] = 89.32%
[Volatility] = Changed
[Distribution] = normal
[Metric] = HRR
[Current Month Percent of Six-month Average] = 131.17%
[Volatility] = WARNING
[Distribution] = normal

A: Product = Tires
This month, Deflected Cases is trending upward at 105.76% of its six-month average. If displayed on a distribution chart, it would be in the yellow zone. Engage Rate is normal this month. Visits for this month is at 89.32% of its six-month average, indicating high volatility and warranting investigation. While there is a less than 5% chance of this value occurring by chance, it is not extreme enough to suggest a data error. If displayed on a distribution chart, it would be in the red zone. HRR is exhibiting extreme volatility, standing at 131.17% of its six-month average. The likelihood of these values occurring by chance alone is less than 0.2%, suggesting a potential data error. If represented on a distribution chart, it would fall well into the red zone.
Additionally, the data distribution for all four metrics is normal, reinforcing the reliability of this analysis.

Q: Product = Cookware
[Metric] = Deflected Cases
[Current Month Percent of Six-month Average] = 114.21%
[Volatility] = Changed
[Distribution] = normal
[Metric] = Engage Rate
[Current Month Percent of Six-month Average] = 196.29%
[Volatility] = WARNING
[Distribution] = normal
[Metric] = Visits
[Current Month Percent of Six-month Average] = 95.38%
[Volatility] = Normal
[Distribution] = normal
[Metric] = HRR

[Current Month Percent of Six-month Average] = 112.47%
[Volatility] = Trend
[Distribution] = normal

A: Product = Cookware
This month, Deflected Cases registers at 114.21% of its six-month average, indicating high volatility and warranting investigation. While there is a less than 5% chance of this value occurring by chance, it is not extreme enough to suggest a data error. If displayed on a distribution chart, it would be in the red zone. Engage Rate exhibits extreme volatility, standing at 196.29% of its six-month average. The likelihood of this value occurring by chance alone is less than 0.2%, suggesting a potential data error. If represented on a distribution chart, Engage Rate would fall well into the red zone. Visits is at 95.38% of its six-month average, remaining within the normal expected range. HRR is trending upward, at 112.47% of its six-month average. If displayed on a distribution chart, this metric would fall into the yellow zone.
Additionally, the data distribution for all four metrics is normal, reinforcing the reliability of this analysis.

Q: Product = Accessories
[Metric] = Deflected Cases
[Current Month Percent of Six-month Average] = 133.65%
[Volatility] = WARNING
[Distribution] = noticeable departure from normality
[Metric] = Engage Rate
[Current Month Percent of Six-month Average] = 96.98%
[Volatility] = Trend
[Distribution] = substantial departure from normality
[Metric] = Visits
[Current Month Percent of Six-month Average] = 92.11%
[Volatility] = Changed
[Distribution] = normal
[Metric] = HRR
[Current Month Percent of Six-month Average] = 100.41%
[Volatility] = Normal
[Distribution] = noticeable departure from normality

A: Product = Accessories
This month, Deflected cases showed extreme volatility, standing at 133.65% of its six-month average. The likelihood of this value occurring by chance alone is less than 0.2%, suggesting a potential data error. If represented on a distribution chart, Deflected Cases would fall well into the red zone. Engage Rate is trending downward at 96.98% of its six-month average. If displayed on a distribution chart, this metric be in the yellow zone. Visits is at 92.11% of its six-month average, indicating high volatility and warranting investigation. While there is a less than 5% chance of this value occurring by chance, it is not extreme enough to suggest a data error. If displayed on a distribution chart, Visits would be in the red zone. HRR is normal this month.
Additionally, the data distribution for Deflected Cases and HRR show a noticeable departure from normality, and Engage Rate a substantial departure from normality, raising concerns about the validity of the analysis for these metrics. The data distribution for Visits is normal, supporting the reliability of the analysis for this metric.

Q: Product = Azure
[Metric] = Deflected Cases
[Current Month Percent of Six-month Average] = x
[Volatility] = x
[Distribution] = x
[Metric] = Engage Rate
[Current Month Percent of Six-month Average] = x
[Volatility] = x
[Distribution] = x
[Metric] = Visits
[Current Month Percent of Six-month Average] = x
[Volatility] = x
[Distribution] = x

[Metric] = HRR
[Current Month Percent of Six-month Average] = x
[Volatility] = x
[Distribution] = x

A: