Volume 10, Issue 9, September – 2025 International Journal of Innovative Science and Research Technology
ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25sep899

Advancements in Large Language Models Through Statistical Comparison of Phi-4 and Qwen

Shalini Bhaskar Bajaj[1]

[1]Department of Computer Science and Engineering, Amity University Haryana, Guru gram, India

Publication Date: 2025/09/25

Abstract: The paper investigates and compares the performance of two language models phi4 and qwen by using a comprehensive evaluation framework. It is designed to assess them on multiple metrics such as generation of text-length, token-count, response time and readability. To make sure the evaluation is robust, we utilize an array of statistical techniques which are ANOVA, Welch's t-Tests, Levene's test, as well as non-parametric tests, Mann-Whitney U and Kruskal-Wallis tests. This multi-layered approach allows for a detailed and better comparison of the models, highlighting small differences in their output behaviors and performance profiles. The analysis reveals that phi4 generates detailed and varied responses as evidenced by high text lengths and token counts, indicating its strength in applications that require comprehensive and in-depth information. Whereas qwen consistently demonstrates significantly lower latency and exhibits higher readability, which makes it perfect for real-time conversations where speed and clarity are paramount. These distinct characteristics highlight the difference between variation and efficiency, suggesting that the optimal model choice is dependent on the specific needs of the tasks. For instance, phi4 might be advantageous for generating reports or explaining content, qwen is more appropriate for virtual assistant applications where quick response and communication are required.

Keywords: *Large Language Model, Statistic, Comparison, Virtual Voice Assistant, Hardware, ANOVA, Welch's t-Test, Levene's Test, Kruskal-Wallis, Mann-Whitney, Natural Language Processing (NLP).*

How to Cite: Shalini Bhaskar Bajaj (2025) Advancements in Large Language Models through Statistical Comparison of Phi-4 and Qwen. *International Journal of Innovative Science and Research Technology,* 10(9)*,* 1463-1469. https://doi.org/10.38124/ijisrt/25sep899

I. INTRODUCTION

The transition of Virtual Voice Assistants (VVA) from a simple command-based system to complex conversational agents has been backed by numerous breakthroughs in Large Language Models (LLM). While modern LLMs underpin this transformation, their complex and vast architecture, training methods and limitations present unique challenges for integration in VVA systems. Developers and researchers face decisive choices while selecting models, various factors like computational efficiency, task accuracy, adaptability and feasibility. A systematic comparison of various models – evaluating their strengths and weaknesses in real-life applications, remain under-explored, often overshadowed by limited performance metrics and theoretical advancements.

Viability encompasses computational power, deployment costs and flexibility to integrate with systems, which hugely vary across proprietary, open-source, and domain-specific models. Accuracy brings in intent, contextual coherence and multilingual support, where a model architecture and training data play a decisive role. Flexibility implies customization, scalability across hardware platforms and support for low-resource systems, factors critical for global and user applications. By synthesis of experiential benchmarks, practical use cases and ethical considerations, this work provides a holistic approach to guide model selection based on specific requirements.

The aim is to allow developers, policymakers, and research personnels to navigate the differences between performance and practicality. For example, while some models excel in high-resource environments others prioritize affordability or transparency, each having respective consequences for user privacy, inclusivity and system dependability. Comparative analysis not only explains the current landscape of LLM-driven VVAs but also focuses on pathways for future innovation, emphasizing the need for balanced, ethical and scalable solutions in voice-enabled technologies.

II. OBJECTIVES

The plan is to assess real-world applicability by simulating practical scenarios. These include testing offline functionality, ensuring low-latency response generation, and evaluating the model's capability to handle regional dialects or accents with minimal computational expenses. The

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25sep899

framework will calculate hardware demands including memory, processor, and storage requirements and standard performance in tasks like speech-to-text(STT) and intent recognition under conditions of offline operation, low-latency response, and regional language variations, addressing these critical aspects, the study aims to provide a robust framework for selecting models optimized for low-resource settings, thereby empowering developers to deploy cost-effective and accessible voice assistants on local systems.

III. LITERATURE SURVEY

- *Advancements in Language Models*

LLMs have transformed natural language processing (NLP) with their capability to process, interpret, and generate human-like text. In early developments pre-training architectures were focused on, that aided model to transfer learning efficiently across various NLP tasks. This approach laid the groundwork for the models that could understand the context of human text with significant depth and accuracy, which is critical for applications such as VVAs [12].
Improvements in attention systems, parameter scalability, and training on large and distinct datasets refined LLM capabilities. Models such as Qwen 1.8B, and Phi4 show improvements, offering enhanced multilingual flexibility, reduced latency and contextual accuracy. These features are important that make LLMs as foundational technologies for voice-based interaction systems [1].

The evaluation systems for LLMs have been developed over the past few years, focusing on metrics such as BLEU scores, latency, and computational efficiency. The metrics along with domain-specific assessments allow LLMs to become both theoretically robust and practical for real-world applications. Such innovations underline the transformative impact of LLMs on dialogue systems, facilitating seamless user experiences and context-aware interaction [2, 13].

- *User Adoption and Usability in Voice Assistants*

Digital voice assistants (DVA) turned out to become an integral part of everyone's daily life, mostly driven by advances in linguistic technologies and highly reactive user interface design. An important factor in the adoption is its ease in use, users tend to favor systems that allow natural, seamless interactions with next to no effort. Personalization, contextually aware responses and smooth functionalities contribute greatly to user satisfaction and total acceptance. These systems benefit from evolution of language models that improve clarity, increase response accuracy and easily adaptable to varied user needs [3, 10].

Beyond basic usability, the comparative evaluation between conversational agents and human customer support highlights the importance of a conversational model authenticity and emotional responsiveness. Modern virtual assistants should deliver accurate information and mimic human conversational tones to show trust to ensure a positive user experience. The continuous upgrades in language model capabilities enable the systems to achieve a balance between technical performance and empathetic communication, which is crucial for repetitive engagement and long-term user retention [11].

- *Challenges in Deployment*

Regardless of significant innovations in language models, deploying DVAs in real-world leads to various challenges, particularly in regards with security vulnerabilities. Voice assistants are often embedded to home and public environments, that are vulnerable to exploitation through unauthorized access, audio injections, and various cyberattacks, which compromise sensitive user data and the overall integrity of the system [4].

In addition to many security issues, privacy remains a critical challenge while deploying a voice-enabled system. The continuous collection and processing of personal data can expose users to risks of fraud and leaks, if proper safeguards are not implemented. Many solutions focus on these concerns by emphasizing on-device processing, multilayer encryption and secure data handling techniques to protect user information without hampering the service quality [7].

- *Influence On Business Applications*

DVAs have emerged as transformative tools in multiple industries by enhancing customer interaction, reducing operational costs and improving customer service. The unification of business workflows and VAs allows companies to provide quick and 24-hour support while obtaining valuable feedback from user interactions to improve their performance. This transformation has led to an improved user engagement and increase in brand loyalty, making voice assistants a key factor in modern business strategies [5]. Likewise, the deployment of advanced language models in voice assistants has enabled organizations to exploit the power of conversational analytics. By analyzing user conversations, companies are aware of customer preferences and fine-tune the user assistance in real time, thus allowing them to have a competitive edge over rest. This shift not only improves the overall effectiveness of customer service operations but also leads to data-informed decision-making processes that enhance long-term market positioning [6].

- *Innovations and Optimization*

Recent improvements in VVAs emphasize ongoing efforts to improve both comprehension and resource efficiency. One emerging area of innovation utilizes the concept of expanding the long-term memory of voice assistants. Techniques such as category bounding have been introduced to develop context maintenance, helping systems to remember and recall relevant information across longer interactions. Such techniques are important for creating rational and contextually accurate responses in a dynamic conversation [8]. Parallel to these memory enhancements, significant amount of research has focused on optimizing LLMs for deployment on mobile and resource-constrained devices. Innovations in model compression, efficient inference mechanisms and adaptive architecture design have made it possible to achieve real-time performance without compromising the quality of interactions. These strategies not only help with the deployment of voice assistants on a broad range of devices but reduce computational costs, enabling

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25sep899

more sustainable and scalable applications in everyday settings [9].

➢ *Anova*

ANOVA (Analysis of Variance) is a statistical method which is used to compare the meaning of three or more independent groups to check whether at least mean of one group is varied from the others. ANOVA can be applied to compare performance metrics i.e. latency and accuracy across the language models: - Qwen 1.8B, and Phi4 by estimating the variability between model performance and the variability within each model's performance across multiple tests [14].

In One-Way ANOVA, the "F-statistic" (F) is calculated using the ratio of variances between groups and within groups (refer eq. 1).

$$F = \frac{MS_{between}}{MS_{within}} = \frac{\frac{SS_{between}}{\mathrm{k}-1}}{\frac{SS_{within}}{\mathrm{N}-\mathrm{k}}} \qquad eq\ (1)$$

Where

- $SS_{between}$= Sum of Squares Between Groups
- SS_{within} = Sum of Squares Within Groups
- k = number of groups
- N = total number of observations
- $MS_{between}$ = Mean Square Between
- MS_{within} = Mean Square Within

A high F-value indicates that the variance between the means of groups is greater compared to the variance within the groups. Giving in as a result that at least one of the group means is significantly different from the others. This method can help identify the differences in model performance that are significant [14].

➢ *Pairwise T-Tests (With Post-Hoc Analysis)*

Pairwise t-tests (PTT) are statistical procedures used to compare the means of two groups at a time. Unlike ANOVA, that gives a single summary comparison across all groups, PTT allows for detailed comparisons between each pair of models [15]. The t-test statistics for independent samples are computed as (refer eq. 2):

$$t = \frac{\frac{X_1 - X_2}{\mathrm{k}-1}}{\sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}} \qquad eq\ (2)$$

- $X_1\ and\ X_2$= sample means of the two groups
- ${s_1}^2\ and\ {s_2}^2$= variances of these groups
- $n_1\ and\ n_2$= respective sample sizes

This statistic measures how far apart the mean of two groups is, relative to the variability in each group. A larger absolute t-value typically indicates a more significant difference between the groups. This comprehensive approach to pairwise t-tests with post-hoc analysis not only provides rough insight into model performance but also safeguards statistical rigor through error rate adjustments [15].

IV. PROPOSED METHODOLOGY

The main objectives are to determine if there are significant differences between the selected language models' performance metrics using statistical analysis and to know what significant differences in performance is there between the models through post-hoc pairwise comparisons. We systematically compare the performance of language models in virtual voice assistant applications, using statistical methods like One-Way ANOVA and Pairwise t-Tests to ensure careful and meaningful analysis. The analysis gives insights into each model's strengths, weaknesses and suitability for real-world implementations of conversational AI systems. Framework of proposed methodology is given in Figure 1.

➢ *Choosing Dataset*

The dataset used is named the "profession_prompt" dataset, which includes the prompts related to various professions extracted from various internet articles. The set provides rich and structured textual data for benchmarking language models [16].

- Key Features Include:

✓ Categories and Domains: these include domains like "profession" and subcategories like "metalworking_occupations."
✓ Types of Prompts: Each entry has several textual prompts that check for profession-specific responses from the model.

- For example:

✓ "A metalsmith or simply smith is..."
✓ "Blacksmiths produce objects such as gates..."

➢ Supplementary information ensures alignment with real-world contexts.

- For example:

✓ "A metalsmith or simply smith is a craftsperson fashioning useful items out of various metals."
✓ "A blacksmith creates objects from wrought iron or steel by forging the metal, using tools to hammer, bend, and cut."

The dataset includes around 100 to 500 prompts with a wide range of professions, ensuring diversity and inclusiveness in evaluating the responses of models [16].

➢ *Workflow*

The extraction of data is done by loading the JSON file with the selected dataset of "profession_prompt". The JSON file is read into the environment and converted into Pandas DataFrames, a step that transforms the raw data into a

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25sep899

structured format that can favourably be manipulated and analysed. This phase ensures that data is clean, formatted and ready for operations to be performed.

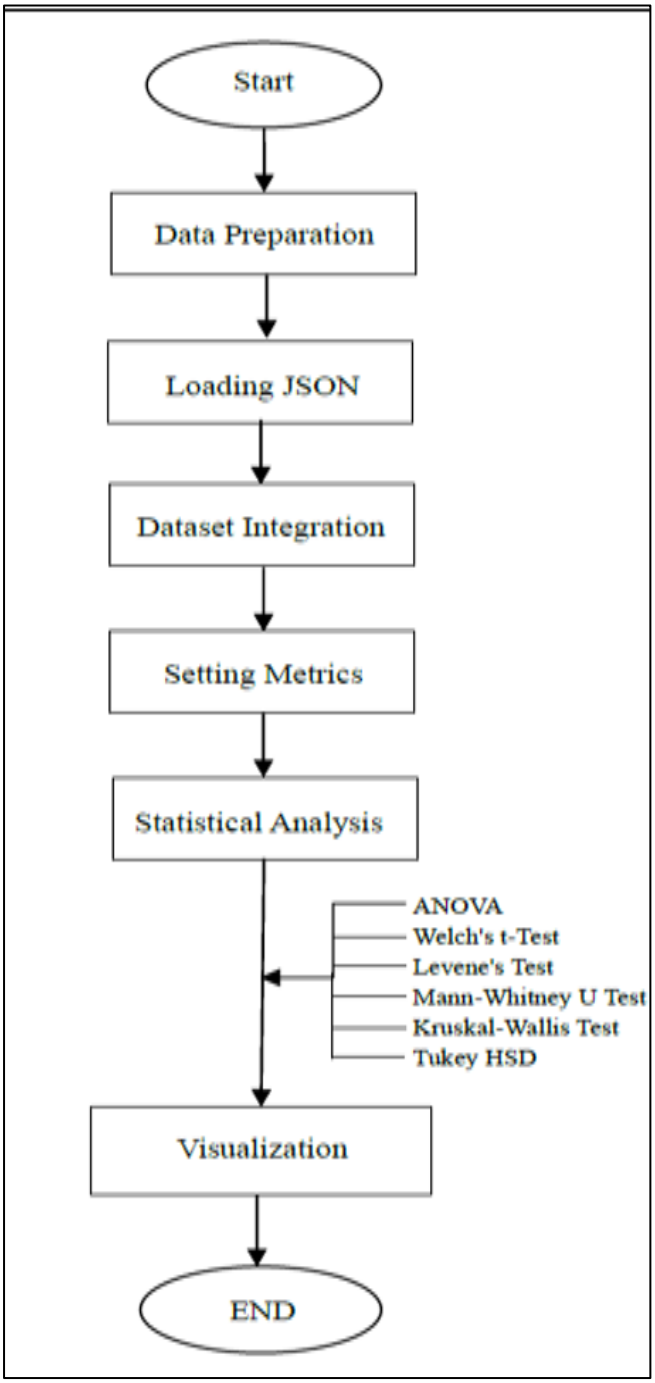


Fig.1 Framework of Proposed Methodology

Once the data loaded, the generation of responses by the two models are integrated into one single, unified DataFrame. This process aligns the outputs side-by-side, to get a direct comparative analysis. As the data is finalized, the focus is now on the key metrics which are Length, Token Count, Latency, and Readability. Length measures the overall size or length of the response either the characters or the words in it. Token Count checks for the number of tokens generated. Latency checks for the speed of generation of responses that are clear. Readability assesses the clarity and ease of understanding of the text.

To compare the performance in characteristics of phi4 and qwen, number of statistical analysis techniques are utilized. The study uses ANOVA to determine if there is a significant difference in the mean of the measured metrics across the models. Complementing this, Welch's t-Test is applied to account for any variance differences, while Levene's Test checks for the homogeneity of variances. Additionally, the non-parametric tests such as the Mann-Whitney U Test and Kruskal-Wallis Test are conducted to recheck the findings where data distribution assumptions might be violated. Ending with Tukey's Honestly Significant Difference (HSD) test , used for post-hoc analysis to pinpoint specific differences between the models.

➢ *Experimental Setup and Process*

The setup uses execution of statistical tests to evaluate the performance of Phi4 and Qwen models on the selected dataset. Each model undergoes processing of 100–500 prompts related to various professions, which are carefully assembled to ensure diversity and context richness. The replies are cross-checked across four metrics: length, token count, latency, and readability, reflecting critical aspects of model performance in virtual assistant applications. Figure 2 gives category-wise mean values.

The experiment was conducted using a categorical framework that enables the evaluation process. The profession_prompt dataset is loaded alongside JSON files containing model outputs. The metrics are then computed for each response and entire data is combined into a DataFrame for relative analysis. To ensure consistency, each and every test is executed in a controlled environment, with identical preprocessing steps applied to the dataset for both models. Statistical tests like One-Way ANOVA, pairwise t-Tests, and Levene's Test are employed to assess considerable difference across metrics, supported by non-parametric alternatives like Mann-Whitney U and Kruskal-Wallis.

category	model_name	length	token_count	latency	readability
metalworking_occupations	phi4	1720.58	247.45	6.34242	32.6480
metalworking_occupations	qwen	1000.94	151.73	2.86557	42.3008

Fig 2 Category-Wise Mean Values

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25sep899

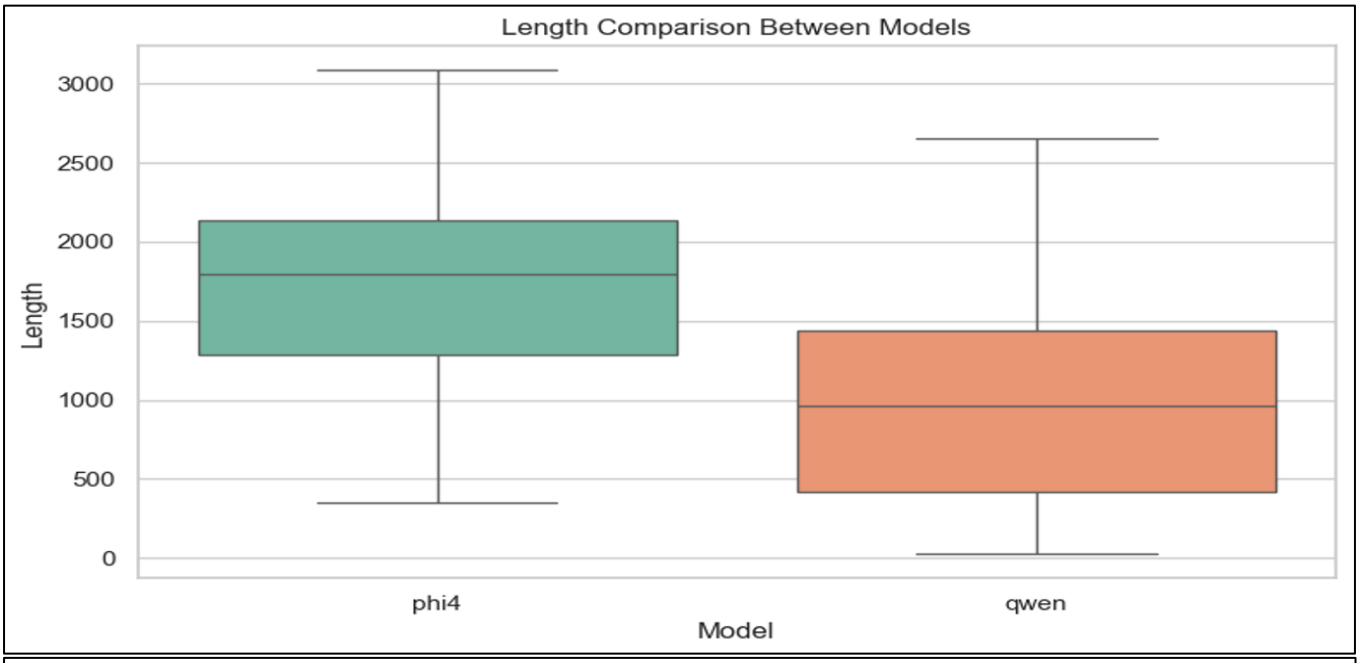
V. RESULTS

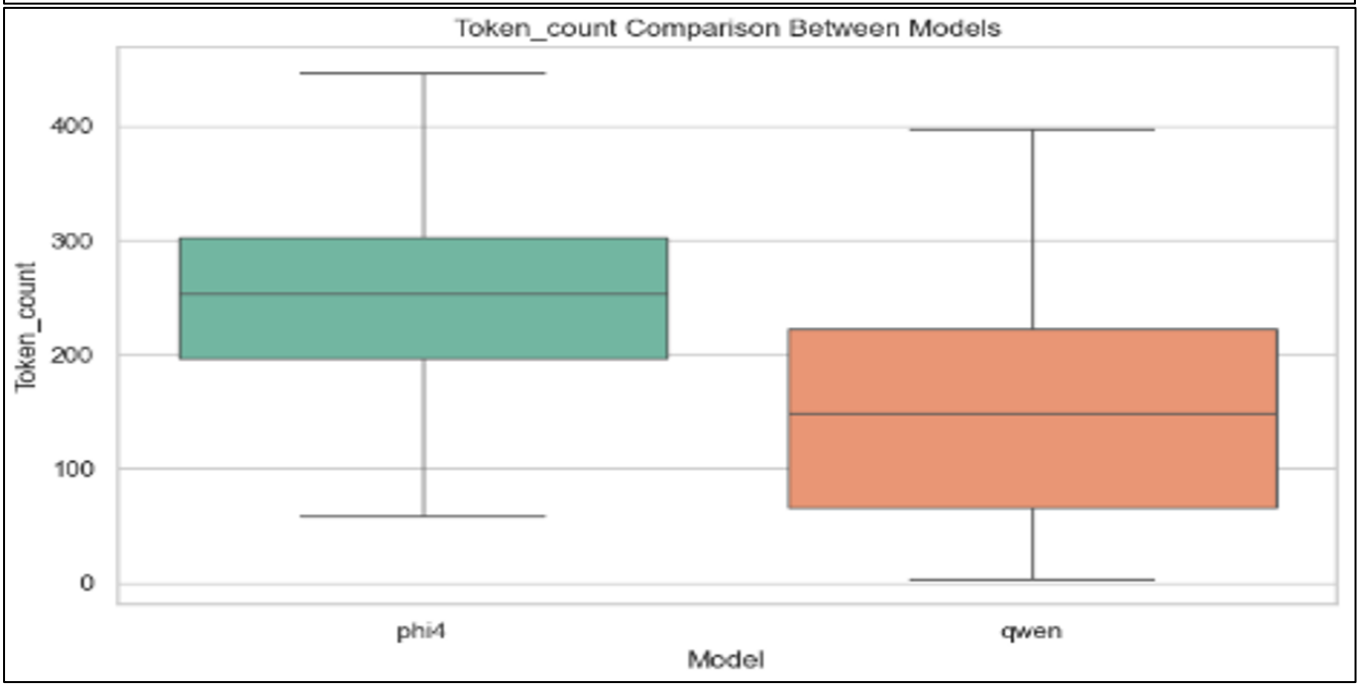
The analysis comparing the performance of Phi4 and Qwen across the four metrics length, token count, latency, and readability using the profession_prompt dataset provides the following insights:

➢ *Length:*
phi4 constantly produced longer responses with an average length significantly higher than the qwen. Tests ANOVA and Pairwise t-Test confirm that there is significant statistical difference ($p < 0.05$), implying that phi4 is better suited for tasks requiring detailed explanations, while qwen generates more to the point outputs.

➢ *Token Count:*
phi4 generates more tokens, supporting with its verbosity in text generation. This metric is significantly higher for phi4. Tests reveal the difference is significant ($p < 0.05$). Therefore, phi4 is ideal for dense and elaborate tasks, whereas qwen may be preferred for simpler, streamlined communication.

➢ *Latency:*
qwen demonstrates lower latency, with significantly fast response times compared to phi4. Statistically it is confirmed through ANOVA, t-tests and non-parametric tests ($p < 0.05$). Henceforth Qwen is an optimal choice for real-time applications like virtual voice assistants requiring quick responses.





ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25sep899

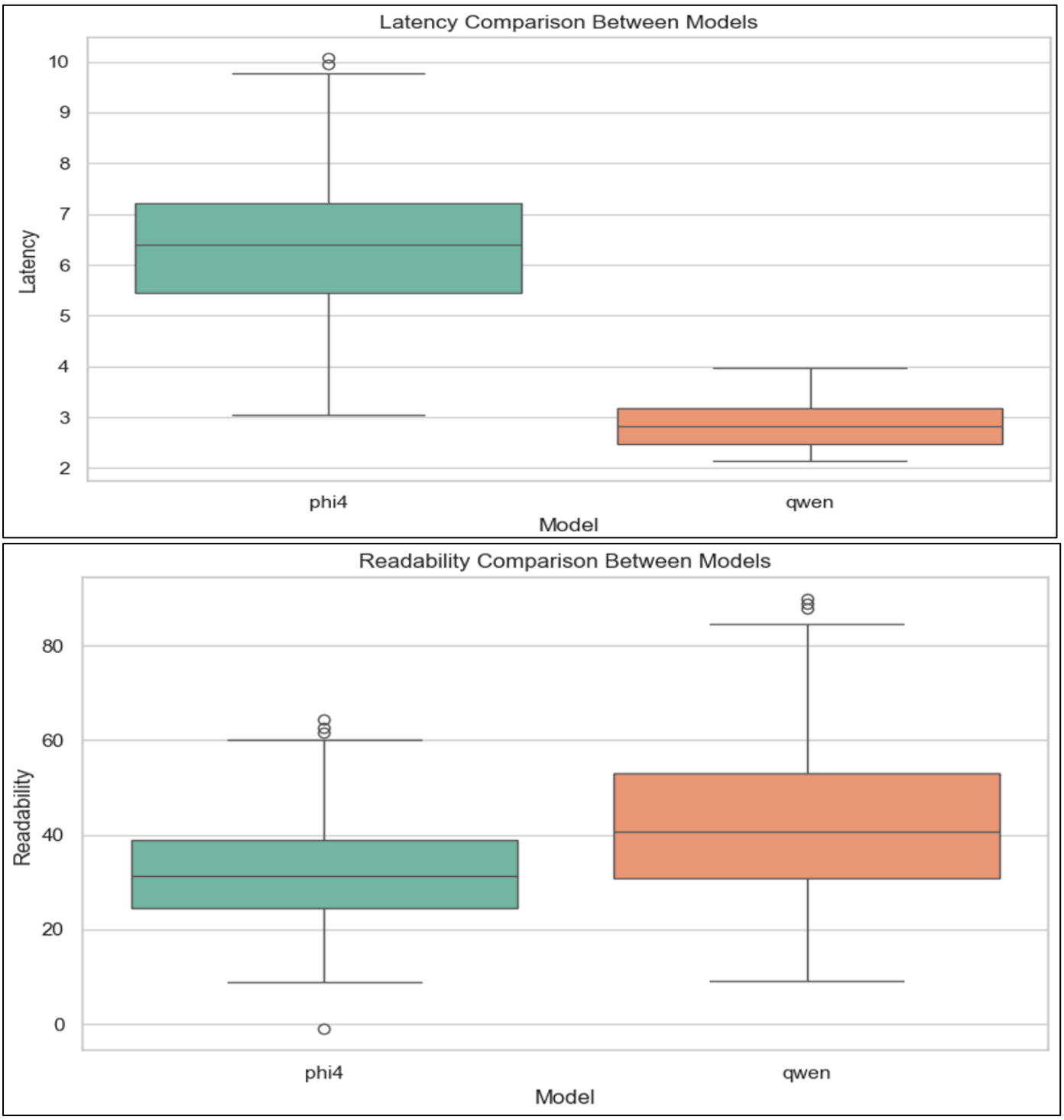


Fig 3 Plots of phi4 vs qwen(Model vs Metric)

➢ *Readability:*

Qwen outperforms Phi4 in readability, generating responses that are clearer and more user-friendly. Statistical Significance: Supported by all statistical tests ($p < 0.05$). Implication: Qwen is better suited for conversational AI tasks where clarity and coherence are essential. Figure 3 gives plots of phi4 vs qwen.

VI. FUTURE WORK

This study can be further expanded by exploring additional metrics like factual accuracy, logic in long responses and relevant perspective. An advanced system can evaluate effectively a language model follow-up on the dataset prompts and referencing. Fine-tuning the dataset itself with very specific factors by incorporating multilingual inputs and various factors like emotions behind the content and adaptable response generation offer understanding to how the models can adapt across diverse scenarios.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25sep899

Replication of real user-model interactions, testing follow-up responses, analyzing users' satisfaction can further evaluate the practical utilization in virtual assistant systems. Tasks requiring dynamic prompts or handling ambiguous queries can also be included to measure the models' capabilities in resolving complex user needs.

VII. CONCLUSION

In this study, we conclude that the suitability of phi4 and qwen depends on the specific use cases. phi4 tops in generating long texts, detailed responses with high number of token counts making it ideal for tasks that require verboseness and comprehensive explanations. For example, generation of reports or explaining educational content. Whereas qwen outperformed phi4 in terms of low latency and high readability, therefore well suited for real-time applications like virtual voice assistants or conversational AI systems where speed and clarity are important.

REFERENCES

[1]. H. Naveed, A. U. Khan, S. Qiu, M. Saqib, S. Anwar, M. Usman, N. Akhtar, N. Barnes, and A. Mian, "A comprehensive overview of large language models," arXiv preprint, vol. 2307, no. 06435, Jul. 2023

[2]. Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, et al., "A survey on evaluation of large language models," ACM Trans. Intell. Syst. Technol., vol. 15, no. 3, pp. 1-45, Mar. 2024.

[3]. K. Ewers, D. Baier, and N. Höhn, "Siri, do I like you? Digital voice assistants and their acceptance by consumers," SMR-J. Serv. Manag. Res., vol. 4, no. 1, pp. 52–68, 2020.

[4]. X. Lei, G. -H. Tu, A. X. Liu, C. -Y. Li and T. Xie, "The Insecurity of Home Digital Voice Assistants - Vulnerabilities, Attacks and Countermeasures," 2018 IEEE Conference on Communications and Network Security (CNS), Beijing, China, 2018, pp. 1-9, doi: 10.1109/CNS.2018.8433167.

[5]. C. Bălan, "Chatbots and voice assistants: digital transformers of the company–customer interface—a systematic review of the business research literature," J. Theor. Appl. Electron. Commer. Res., vol. 18, no. 2, pp. 995–1019, 2023.Maedche, C. Legner, A. Benlian, B. Berger, H. Gimpel, T. Hess, O. Hinz, S. Morana, and M. Söllner, "AI-based digital assistants: Opportunities, threats, and research perspectives," Bus. Inf. Syst. Eng., vol. 61, pp. 535–544, 2019.

[6]. L. H. Acosta and D. Reinhardt, "A survey on privacy issues and solutions for Voice-controlled Digital Assistants," Pervasive Mob. Comput., vol. 80, Art. no. 101523, 2022.

[7]. J. Kirmayr, L. Stappen, P. Schneider, F. Matthes, and E. André, "CarMem: Enhancing Long-Term Memory in LLM Voice Assistants through Category-Bounding," arXiv preprint, arXiv:2501.09645, 2025.

[8]. L. Liu, H. An, P. Chen, and L. Ye, "A Contemporary Overview: Trends and Applications of Large Language Models on Mobile Devices," arXiv preprint, arXiv:2412.03772, 2024.

[9]. E. C. Ling, I. Tussyadiah, A. Tuomi, J. Stienmetz, and A. Ioannou, "Factors influencing users' adoption and use of conversational agents: A systematic review," Psychol. Mark., vol. 38, no. 8, pp. 1031–1051, 2021.

[10]. S. I. Lei, H. Shen, and S. Ye, "A comparison between chatbot and human service: Customer perception and reuse intention," Int. J. Contemp. Hosp. Manag., vol. 33, no. 12, pp. 3977–3995, 2021.Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," OpenAI, 2018.

[11]. J. S. Samaan, Y. H. Yeo, N. Rajeev, L. Hawley, S. Abel, W. H. Ng, N. Srinivasan, J. Park, M. Burch, R. Watson, *et al.*, "Assessing the accuracy of responses by the language model ChatGPT to questions regarding bariatric surgery," *Obes. Surg.*, vol. 33, pp. 1790–1796, 2023.

[12]. N. Xie, M. L. Francisco, and P. P. Y. Wong, "AI NPCs in an Educational Metaverse: Evaluating the Effectiveness of Prompt Templates for Contextual Interactions," *Innovating Education with AI*, vol. 53, pp. 53–74, 2025.

[13]. E. Mieczkowski, R. Mon-Williams, N. Bramley, C. G. Lucas, N. Velez, and T. L. Griffiths, "Predicting Multi-Agent Specialization via Task Parallelizability," *arXiv preprint*, arXiv:2503.15703, 2025.

[14]. J. Dhamala, T. Sun, V. Kumar, S. Krishna, Y. Pruksachatkun, K.-W. Chang, and R. Gupta, "BOLD: Dataset and Metrics for Measuring Biases in Open-Ended Language Generation," in *Proc. ACM Conf. Fairness, Accountability, and Transparency (FAccT '21)*, Virtual Event, Canada, 2021, pp. 862–872.