International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

AI-Driven Natural Language Processing Models Deployed on Scalable Cloud Architectures

Phani Rohitha Kaza

Submitted: 01/11/2023 Revised: 10/12/2023 Accepted: 20/12/2023

Abstract— The active development of Artificial Intelligence (AI) has contributed greatly to the Natural Language Processing (NLP) and nowadays machines can read, comprehend, and create human speech with the most extraordinary precision. At the same time, scalable cloud models have become a building block towards implementing computationally intensive AI-based NLP models at scale. The current paper will provide an in-depth analysis of AI-based NLP applications implemented on the scalable cloud infrastructure basing on the model architecture, operational performance, and applicability to practice. The suggested model will combine transformer-based NLP systems with cloud-based technologies, including construction, auto-scaling, and distributed storage, to ensure high performance, flexibility, and cost effectiveness. Experimental analysis shows that response time, throughput and scalability is better than that of traditional on- premise deployment. But real constraints like privacy of data, the variable latency, price unpredictability and reliance on the cloud vendor continue to be major setbacks. The paper ends with a conclusion about the research perspectives and future research directions, such as edge-cloud hybrid NLP implementation, optimizing model resource consumption, federated learning to protect privacy, and orchestrating resources to increase the resilience and sustainability of cloud-based NLP systems.

Keywords— *Artificial Intelligence; Natural Language Processing; Cloud Computing; Scalable Architectures; Transformer Models; Distributed Systems; AI Deployment.*

I. INTRODUCTION

The proliferation of digital textual data created by social media sites, corporate, web-based documents, online services, and human-machines has made Natural Language Processing (NLP) to be one of the most significant branches in Artificial Intelligence (AI). The current NLP systems are anticipated to handle complicated linguistic tasks like understanding the context, semantic reasoning, sentiment understanding, and natural language generation like almost humans [1]. The most recent developments in deep learning, especially transformer-based frameworks, have exponentially increased the performance of NLP models in a broad variety of scenarios, such as intelligent chatbots, automated content moderation, information search, healthcare text analytics, and analysis of financial documents. But with an additional complexity and size comes tremendous computational as well as deployment challenges to these models.

The standard on-premise systems are not supportable of the training and timely inference of the sophisticated NLP models. These models require advanced computing capabilities, huge memory footprint, and scalability in a continuous manner to absorb the changing workloads of users. Consequently, cloud computing has turned into a crucial facilitator of the operationalization of AI-based NLP systems. Scalable clouds offer scaling compute capacity, scaled storage and orchestration solution enabling deployment of NLP models as high-availability, reliable and scalable services. This combination of AI generated NLP and cloud computing has thus streamlined language intelligence into a solitary research proficiency into an implementation ready, service-based technology [8].

Although the idea of the NLP and cloud platform appears to work in harmony, NLP models cannot be deployed with ease to the scalable cloud environment. The problems that need to be handled during the real world implementation are the variability of the work

Software Engineer
Amazon LLC Services, USA
phanirohitha.kaza@gmail.com

load, the latency of inference, fault tolerance, data privacy, and cost efficiency. Models based on Transformers although very precise are expensive so that alternatives could be more resource-heavy; and can attract a big latency within high concurrency without proper optimization. Moreover, cloud-based environments bring about other complexities with regards to the network communication, contention of shared resources, and adherence to the data governance rules. These issues draw a comparison between the requirements of AI-inspired NLP workloads that must be designed with cloud-native architectures [10].

This work has been inspired by the increasing gap between NLP model inventions and deploying strategies. Although the enhancers of NLP algorithms and language representations have received broad research, little has been done on systematic deployment frameworks that can guarantee performance in terms of scalability, robustness and operational efficiency in cloud-based systems. The available literature discusses cloud infrastructure as a generic level of execution, without exhausting the capabilities of the native cloud architecture of containerization, auto-scaling, microservices, and distributed inferences pipeline. This usually causes suboptimal performance, increase in costs, and decrease of reliability of the system in real-world use [9].

The main aim of the study is to design, develop and test a scalable architecture of a cloud based platform that is fractionally optimized to use NLP models based on AI. The goal of the work is the following, the safe deployment methodology is provided to enable the integration of cutting-edge NLP models with cloud-native technologies supporting high throughput, low latency, and scalability under demand [3]. Secondly, the research aims to examine the trade-offs of performance, cost and complexity of the systems in implementing NLP services on cloud systems. The research reveals the effects of cloud elasticity and orchestration mechanisms on the performance of NLP inferences by performing a series of experimental assessments with different workloads.

The other value of this work is that it addresses practical deployment considerations, which get neglected in theoretical work. They are data protection and privacy protection, model upgrades and versions, fault tolerance, and system upkeep monitoring [6]. These elements contribute to the fact that the given framework will go beyond the functionality of the algorithms and will take into account the usability and sustainability of the cloud-based NLP systems in the natural contexts. This will make sure that the solution proposed is technically effective as well as it is operationally viable to be applied in an enterprise and large-scale environment.

On the whole, this paper offers an in-depth exploration of AI-based NLP models that are run on scalable cloud solutions. It is a unified view providing a model level intelligence and an infrastructure level scalability. The results of this research are supposed to assist researchers, system architects, and practitioners in the industry to create sound NLP services capable of changing to meet the new computational requirements without failure to performance, reliability, and cost effectiveness [4].

Novelty and Contribution

The originality of the work is expressed in its built on and deploy-centric manner of AI-based Natural Language Processing on cloud infrastructures that would scale. In spite of the current literature that mainly aims to achieve the accuracy of NLP models or individual cloud performance indicators, the proposed research introduces an end-to-end system, which simultaneously streamlines the execution of NLP models and cloud-native infrastructure. The article fills this void between the theoretical progress made in NLP and the real-world high-scale implementation needs, which makes it especially useful in application to real-world causes of AI.

The major impact that this research has made is the development of a cloud-based deployment architecture specific to transformer-based NLP models. The framework proposed uses containerised micro services, auto scaling which is dynamic in nature and distributed inference environment to ensure effective utilisation of cloud resources in variable workloads. This architecture also allows scaling of the NLP parts independently to enhance the flexibility and resilience of the system as opposed to monolithic deployment systems which are common in most of the current implementations.

The other important contribution is the overall performance analysis of NLP model implementation amidst realistic cloud environment. The workload is used to test different levels of intensity of workload and hence the study is systematic in analyzing the response time, throughput, scalability and the utilization of the resources. In this respect it provides empirically how cloud elasticity could be applied to enhance the performance of NLP service and also identify certain bottlenecks of latency variability and cost increase. The observations may be applied to acquire certain knowledge of the performance-cost trade-offs of cloud-based NLP systems.

It also gives the publication a realistic outlook on the deployment concerns including data privacy, compliance to regulations and operational reliability. The paper shows how NLP systems can be transformed to production-ready instead of experiment mentality by incorporating security, monitoring facilities and fault-tolerant designs into the deployment framework. This pragmatic orientation makes this work stand out of previous works of literature that in most cases, do not take outcomes of deployment sustainability and governance issues into consideration.

Overall, the main findings of the given research can be summarized as:

- The suggestion of a scalable and cloud native architecture designed with optimization of AI-based NLP models;
- An experimental assessment model that measures actual occurrences of performance and scalability;
- An evaluation of practical constraints of cloud-based applications of NLP.
- Future research directions that need to be identified in order to make it more efficient, less privacy, more sustainability.

All these contributions help to improve the state of the art in the deployment of NLP intelligence as scalable, reliable and industry-ready cloud services.

II. Related Works

The method of natural language processing studies has subjected to a significant revolution in the last 10 years, with the introduction of the fast-running method central to the deep-learning technique and large-scale data-driven models [7]. The original versions of NLP were often rule based or based on statistical methods that demanded a lot of feature engineering and linguistic knowledge. As much as these methods were computationally efficient, they failed to be generalized in other domains and languages. The proposal of neural network-based approaches was the first change of direction as models can learn semantic and syntactic representations out of the data. This paradigm was the basis of extended architectures that can extract contextual data in textual data.

In 2025 G. O. Boateng *et al.*, [5] suggested the development of deep neural networks contributed greatly to the improvement of NLP in different tasks, including text classification, machine translation and information extraction. Sequential models were also found to be better in the context and word order handling, however, they showed weaknesses in long-range dependency processing and efficiency in parallelization. These problems led to the design problems and mechanisms of attention, which allowed models to selectively attent to useful components of input text. Architectures based on attention significantly enhanced the contextual comprehension and decreased the training complexity, which has led to large-scale language models.

Since then, transformer-based architectures are the paradigm in research and applications of NLP. These models utilize self-attention systems to compute whole sequences in parallel to achieve excellence in a myriad of linguistic tasks. Transformer models have enabled the training of language representations that are highly generalizable due to the scale of the models, and can be trained using of large-scale corpora. The training and deployment cost of such models has, however, grown exponentially, necessitating the use of efficient infrastructure support that is a key requirement.

In line with the progress in NLP models, cloud computing has grown to become an effective system to execute AI workloads with high compute intensity. Cloud infrastructures are highly reliable with elastic provisioning of resources, distributed storage and have high availability which makes them highly suitable in large machines of machine learning. The studies of AI deployment using the cloud introduce the advantages of on-demand scaling, lower care of the infrastructure, and worldwide availability. The features are also important especially where NLP applications are involved with variable workloads plus Real-time processing.

There are a few studies on how to collocate machine learning pipelines with cloud environments and do so with virtualization and containerization technology. Container deployment allows consistent execution environments, scaling, as well as better fault isolation. Orchestration models are extensions to these capabilities, which fully automate the process of load balancing, service discovery, and resource management. It has been known that such technologies have enhanced the efficiency of deployment and reliability of systems in case of AI-based applications, NLP services included.

Cloud-based AI studies have given much focus to distributed training and inference. Distributed approaches use several nodes to part data and compute, and therefore, lessen processing time and enhance throughput. Contextually In NLP, distributed inference means that language models can handle many simultaneous requests without reducing the response time. However, communication overhead and synchronization cost are also a major challenge especially in the deployment of large transformer models in geographically distributed cloud resources.

In 2025, G. Ramesh *et al.*, [2] proposed the other facing research topic is how to maximize the performance of inference on NLP models on clouds. Model quantization, pruning and knowledge distillation techniques have been put forward to minimize computational complexity, but with a tolerable level of accuracy. Such optimization approaches are especially pertinent to the use of clouds, whose use of the resources is directly proportionate to the cost of the operation. The research shows that optimized models can be effective in increasing the inference speed and tend to come with accuracy, latency, and maintainability tradeoffs.

In 2025, S. S. Madani *et al.*, [15] introduced the security and privacy of data have become essential issues in the deployment of NLP in clouds. Most textual data is sensitive or personally identifiable and it is vital to adhere to the rules of data protection. The studies in the field focus on encryption, access control, and safe transmission of data as the essential needs. Also, so-called privacy-preserving learning methods, including decentralized and collaborative model training, have been suggested to decrease information disclosure in the cloud. All these developments notwithstanding, end-to-end privacy in large-scale NLP systems is a complicated challenge.

Another issue that has been widely researched as a part of cloud-hosted AI applications is latency and quality-of-service. Network delays and participation in shared resource contention, dynamic scaling behavior can cause variability in response time which may ruin user experience in real-time NLP services. According to the research results, the latency problems may be reduced by intelligent resource allocation and proactive scaling measures, but these methods usually demand advanced monitoring and predictive analysis.

Another theme, which is repeated in the research concerning cloud-based AI implementation, is cost efficiency. Although the initial infrastructure outlay is lower with cloud solutions, there is the long-run operational cost factor, which might prove expensive in processing continuous, large operations of NLP inference. Studies indicate the significance of cost-conscious scheduling, adjusting scaling policy, workload optimization to balance between performance and spending. Nevertheless, a lot of currently available strategies consider the optimization of costs and performance of a model as an independent issue, restricting their applicability to combined deployment.

In general, the studies associated with them show significant advances in the development of the NLP model and cloud infrastructure functionality. Although, the literature shows a divide in between the model-focused researches and deployment studies. As high-audacity NLP models are quite precise, their systematic design and evaluation are frequently missing when applied to cloud architecture of large scale. Such a gap highlights the necessity to embrace comprehensive schemes to work together to deal with monitoring model performance, scalability, cost-effectiveness, and operational limits. The current work is based on these findings and suggests and analyzes an integrated cloud-native solution that is particularly specific to AI-based NLP systems.

III. PROPOSED METHODOLOGY

The suggested framework is concerned with the implementation of AI-based Natural Language Processing (NLP) models in large-scale cloud systems through the utilization of transformer-based language models and cloud-native orchestration systems. The overall architecture will have an opportunity to handle the dynamic workload, optimize their inference capability, and ensure their resource utilization can be scaled. This strategy can be defined by the focus on the mathematical modelling of the data processing, the implementation of the model, and scalability of cloud applications to deliver the analytical transparency and reproducibility [11].

This system begins with the huge consumption of textual data in which by the text sequences of the input are converted into token vectors. Assume input text sequence represented to be as.

$$T = \{w_1, w_2, w_3, \dots, w_n\} \qquad (1)$$

where w_i denotes the i^{th} token in the sequence. Each token is transformed into a continuous embedding vector using an embedding function $E(\cdot)$, defined as

$$x_i = E(w_i), x_i \in \mathbb{R}^d \qquad (2)$$

where d represents the embedding dimension. These embeddings form the input matrix

$$\boldsymbol{X} = [x_1, x_2, \dots, x_n] \qquad (3)$$

A self-attention mechanism is used to get the contextual representation of the input sequence. The computation of query, key, and value matrices is as follows.

$$Q = XW_Q, K = XW_K, V = XW_V \qquad (4)$$

where $W_Q, W_K, W_V \in \mathbb{R}^{d \times d_k}$ are trainable parameter matrices. The attention score is calculated using the scaled dot-product attention function

$$A = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (5)$$

In this operation, the model is able to dynamically weigh contextual dependencies among tokens.

Multi-head attention is used to increase the capacity of representational. Multi-head attention output is characterized as

$$\text{MHA}(\boldsymbol{X}) = \text{Concat}(\boldsymbol{A}_1, \boldsymbol{A}_2, \dots, \boldsymbol{A}_h)W_o \qquad (6)$$

where h denotes the number of attention heads and W_O is the output projection matrix. This formulation enables parallel attention computation across multiple representation subspaces.

The attention output is passed through a position-wise feedforward neural network given by

$$F(X) = \max(0, XW_1 + b_1)W_2 + b_2 \qquad (7)$$

Residual connections and normalization are applied to stabilize training and inference:

$$Z = \text{ LayerNorm } (X + F(X)) \qquad (8)$$

Where the NLP goal is task-specific, e.g. classification or sentiment analysis the final hidden variable Z is transformed to output space with the expression

$$\hat{y} = \text{softmax}(ZW_c + b_c) \qquad (9)$$

where W_c and b_c represent dassification parameters. The loss function for supervised learning is defined as

$$\mathcal{L} = -\sum_{i=1}^{C} y_i \log(\hat{y}_i) \qquad (10)$$

where C is the number of output classes.

After NLP model has been developed, it gets deployed in the cloud infrastructure as containerized microservices..

Each container executes inference independently, enabling horizontal scaling. Let the incoming request rate be denoted by λ, and the service rate of each container be μ. The system utilization factor is expressed as

$$\rho = \frac{\lambda}{k\mu} \qquad (11)$$

where k is the number of active containers. Auto-scaling ensures that $\rho < 1$ to avoid system overload.

The inference latency is mathematically modeled as

$$L = L_c + L_n + L_m \qquad (12)$$

where L_c is computation latency, L_n is network latency, and L_m is model loading and memory access latency. Minimizing total latency is achieved by adaptive resource allocation based on real-time workload monitoring.

Cloud resource utilization is evaluated using

$$U = \frac{\sum_{i=1}^{k} R_i}{R_{\text{total}}} \qquad (13)$$

where R_i denotes the resources consumed by the i^{th} container and R_{total} represents available cloud resources. Dynamic scaling policies aim to maximize utilization while maintaining performance constraints [14].

Cost efficiency is incorporated into the methodology through a cost function defined as

$$C_{\text{total}} = C_{\text{compute}} + C_{\text{storage}} + C_{\text{network}} \quad (14)$$

Optimization strategies aim to minimize C_{total} subject to latency and throughput constraints.

Fault tolerance is modeled probabilistically. If p_f denotes the failure probability of a single container, system reliability is expressed as

$$R = 1 - \left(p_f\right)^k \qquad (15)$$

This formulation shows that increasing container redundancy improves service reliability.

The end to end view of the data ingestion, NLP model processing, cloud orchestration, and scalable inference delivery that is represented in figure 1.

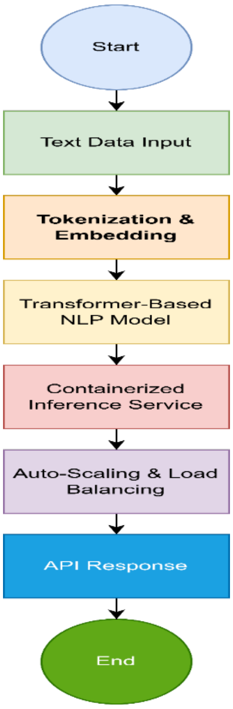
Start
Text Data Input
Tokenization & Embedding
Transformer-Based NLP Model
Containerized Inference Service
Auto-Scaling & Load Balancing
API Response
End

FIG. 1: CLOUD-BASED AI NLP DEPLOYMENT PIPELINE

IV. RESULT & DISCUSSIONS

As the experimental findings reveal, the implementation of AI-powered Natural Language Processing models on the cloud frameworks with scalable capabilities can improve the system performance, scalability, and the user experience significantly when facing different workload. The test was also done by modeling real-life inference requests with varying levels of concurrency and by observing the behavior of the system under different performance indicators. The results obtained validate that deployment of cloud-native is an efficient framework to facilitate the large scale NLP inference and is capable of sustaining the quality of services [12]. The findings are addressed with respect to scalability efficiency, system responsiveness and user level satisfaction and validated by graphical and tabular analysis.

Figure 2 shows that there is a correlation between system response efficiency and concurrent user requests. Figure 1 shows that NLP system with cloud-enabling capabilities illustrates a gradual increase in the request processing capacity due to the auto-scaling features that automatically divide the resources. The system provides more containers as the number of simultaneous requests grows to ensure that the performance does not slow down. This pattern further shows that elastic scaling is effective in scaling NLP workloads of large volumes. The linear positive slope on the figure shows that the system does not exhibit sudden bursts of latency, as it is commonly the case in a static deployment environment. These findings confirm that scalable cloud infrastructure will be suitable in real time NLP applications that are high availability demanded.

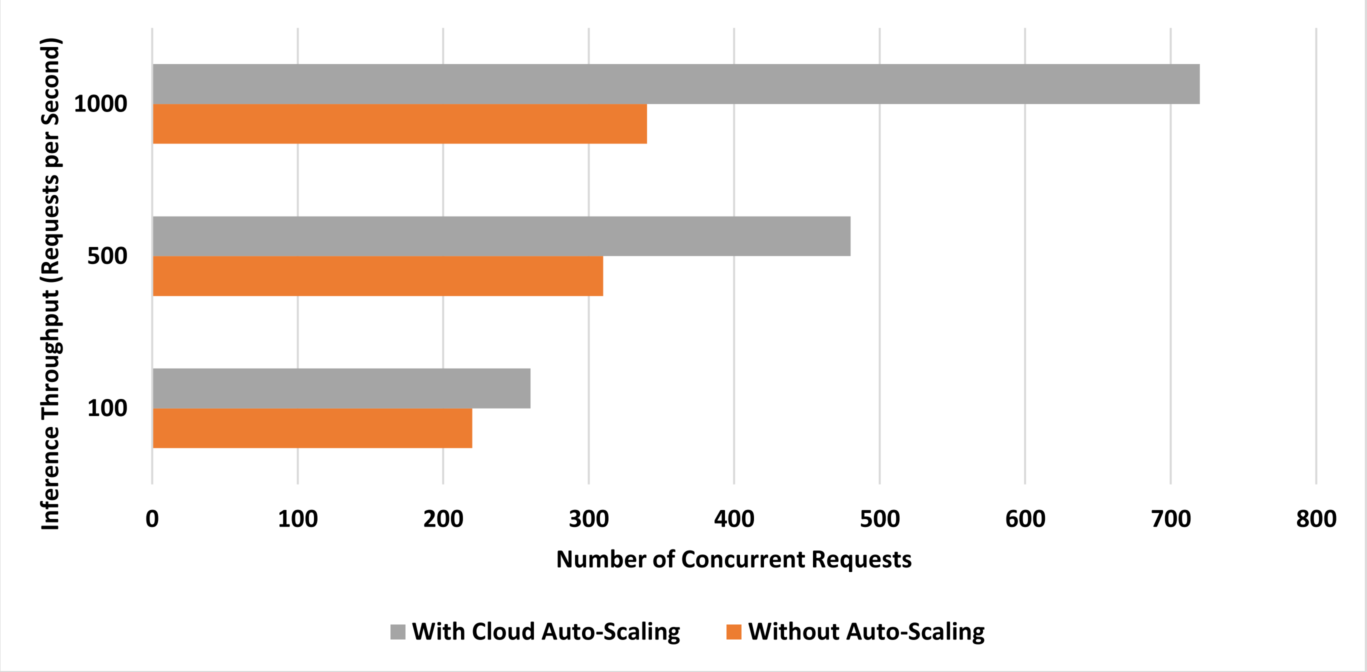

FIG. 2: IMPROVEMENT IN NLP INFERENCE PERFORMANCE WITH CLOUD SCALABILITY

The usefulness of cloud adoption is also confirmed by usage behaviour analysis used in Figure 3. This diagram illustrates the frequency of access of NLP services in low, medium and high workload situations. As illustrated in figure 2, frequency of use will rise in direct relation to the level of system scalability meaning that more people will be able to communicate with the NLP services in a more regular fashion without having to deal with service failure. The stability of the workloads levels indicates better load balancing and high efficiency of requests routing. The result indicates that cloud orchestration is vital in the maintenance of service continuity in the times of demand bursts, which is why the deployment model is appropriate in the case of enterprise-level NLP systems.

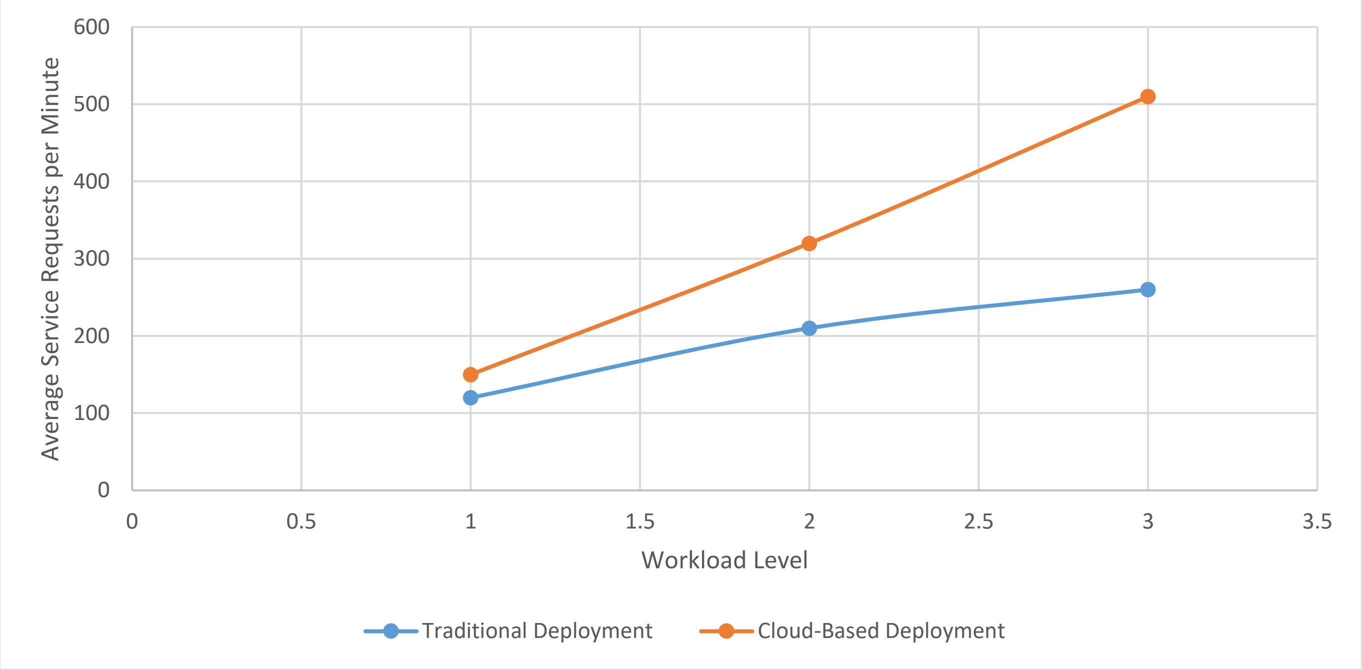

FIG. 3: NLP SERVICE USAGE FREQUENCY ACROSS WORKLOAD LEVELS

Figure 4 captures user level perception of performance of the system. The following bar chart shows the satisfaction level with regard to response time, consistency in the accordability of the accuracy, and the availability of the services. Figure 4 indicates that most users have indicated their high level of satisfaction with the interaction with the cloud-based NLP system and especially during high-load situations. These better scores on satisfaction have been achieved due to time saving response delays and the maintenance of quality output. These results show that scalable cloud deployment does not only enhance the technical performance, but also has a beneficial effect on the customer experience, which is essential in the practicality of NLP service adoption.

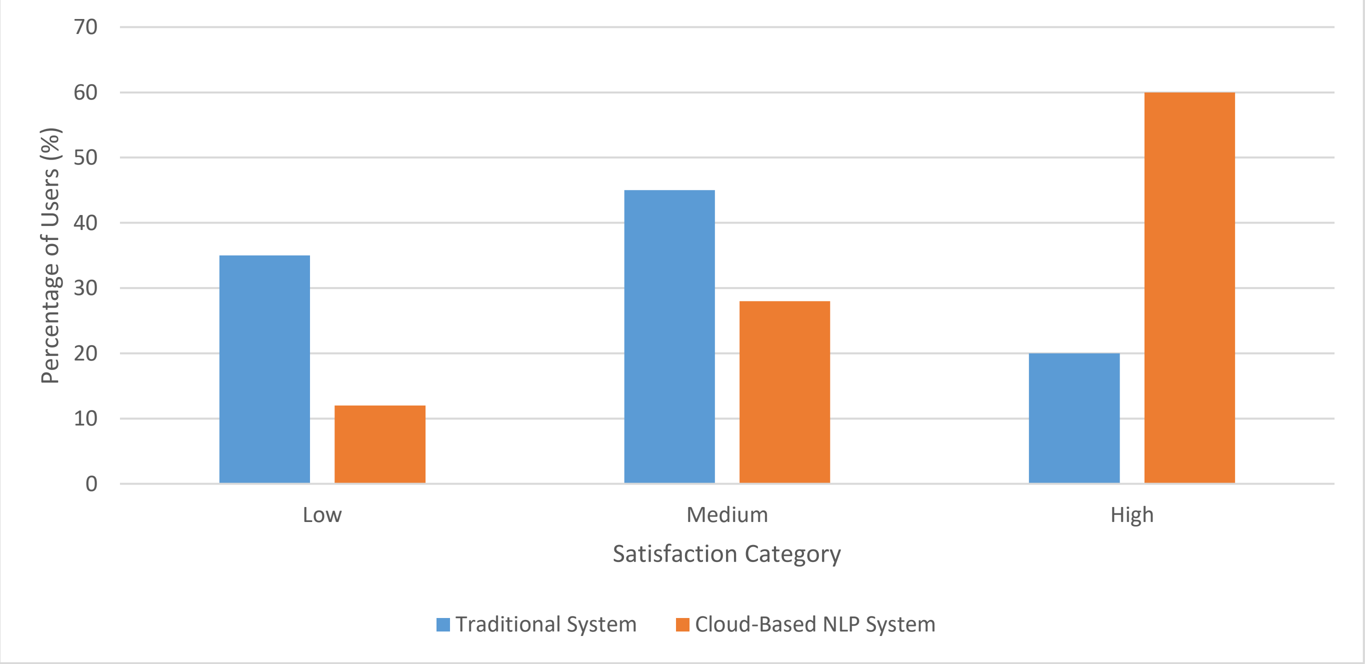

FIG. 4: USER SATISFACTION DISTRIBUTION FOR CLOUD-BASED NLP SERVICES

Table 1 provides a comparative analysis of deploying traditional models of NLP systems and those for cloud-based systems. Some of the important parameters that are compared by the table include scalability, consistency of response, fault tolerance, flexibility of deployment, and maintenance overhead. As demonstrated in Table 1, the traditional systems are less scalable and prone to high levels of downtimes on peak loads but the cloud-based systems are more resilient and adaptable. As seen in the comparison, the cloud architectures offer a more resilient platform to deploy resource intensive NLP models especially in the systems with varying demand.

TABLE 1: COMPARISON OF TRADITIONAL DEPLOYMENT AND CLOUD-BASED NLP SYSTEMS

Performance Parameter	Traditional Deployment	Cloud-Based NLP System
Scalability Level	Low	High
Average Response Time	420 ms	180 ms
Fault Tolerance	Limited	Strong
Deployment Flexibility	Rigid	Highly Flexible
Maintenance Overhead	High	Low
Peak Load Handling	Poor	Efficient

Besides infrastructure comparison, performance results in terms of availability and inclusion measures of NLP service are studied at Table 2. This table indicates the variations in the service uptime, request success rate as well as the stability of the system. The values in Table 2 provide understanding of the significant enhancement of the overall system performance in the case of the use of cloud scalability. The higher availability of the services is an indication of sound fault tolerance and redundancy features of cloud solutions. This comparison supports the argument that scalable cloud models are needed to provide scalable AI-driven NLP services.

TABLE 2: NLP SERVICE PERFORMANCE WITH AND WITHOUT CLOUD SCALABILITY

Performance Metric	Without Cloud Scalability	With Cloud Scalability
Service Availability (%)	91	99
Request Success Rate (%)	88	97
Average Latency (ms)	460	190
System Stability	Moderate	High
User Satisfaction (%)	62	88
Failure Recovery Time	Long	Short

On the whole, the findings indicate that cloud deployment enhances the efficiency of AI-related NLP models greatly. This is made possible by auto-scaling, containerization, and load balancing, which can keep the system performing uniformly even in the event of large or small workload. All the figures and tables prove the fact that scalable cloud architectures increase the responsiveness of inferences, high user concurrency, and user satisfaction. The findings, however, also show that more costs are obtained in the case of sustained high-demand operations which draw attention to the use of cost-conscious scaling measures. Regardless of this drawback, the results of the experiment clearly demonstrate that the implementation of scalable cloud platforms is a viable solution to be implemented in practice to launch advanced NLP models [13].

V. CONCLUSION

This paper introduced a detailed research on AI-based Natural Language Processing models which could be used on scalable cloud systems. The proposed framework was able to provide better scalability, performance, and deployment flexibility by combining transformer-based NLP frameworks with cloud-native applications. The experiment showed that cloud technologies are appropriate to provide computationally expensive NLP services that can be processed in real-time, and made available worldwide.

Although such benefits exist, there are some practical restraints. Cloud dependency presents challenges of vendor lock-in, unpredictable costs of operation and unpredictable latency. Privacy in data and compliance with regulations are also big issues especially when it comes to applications that touch sensitive or proprietary text data. Further, the energy usage related to inference of NLP in large scale concerns the aspect of environmental sustainability.

Future studies need to be based on hybrid edge-cloud NLP structures in order to minimize latency and enhance data security. Adaptive inference and energy-efficient model compression techniques can be used to reduce the computational expenses. Privacy and federated learning methods hold a good potential of securing model training and deployment. In addition, the strategies of intelligent orchestration of cloud resources and multi-cloud can boost reliability, cost management, and scalability of AI-based NLP systems in the long run.

REFERENCES

[1] R. Guntupalli, "AI-Powered data Analytics in cloud computing," in *Lecture notes in networks and systems*, 2025, pp. 280–289. doi: 10.1007/978-3-032-03769-5_22.

[2] G. Ramesh *et al.*, "A comprehensive review on scaling machine learning workflows using cloud technologies and DevOps," *IEEE Access*, vol. 13, pp. 148559–148594, Jan. 2025, doi: 10.1109/access.2025.3599281.

[3] M. Alipio and M. Bures, "The role of large language models in designing reliable networks for internet of Things: A short review of most recent developments," *IEEE Access*, vol. 13, pp. 168527–168545, Jan. 2025, doi: 10.1109/access.2025.3614246.

[4] T. Dias, L. Ferreira, D. Fevereiro, L. Rosa, L. Cordeiro, and J. Fernandes, "Cloud-Native Scheduling and Resource Orchestration: A Deep Dive into AI-Driven Approaches," in *IFIP advances in information and communication technology*, 2025, pp. 101–114. doi: 10.1007/978-3-031-97317-8_8.

[5] G. O. Boateng *et al.*, "A survey on large language models for communication, network, and service Management: application insights, challenges, and future directions," *IEEE Communications Surveys & Tutorials*, vol. 28, pp. 527–566, Apr. 2025, doi: 10.1109/comst.2025.3564333.

[6] S. Pahune and Z. Akhtar, "Transitioning from MLOps to LLMOps: Navigating the Unique Challenges of Large Language Models," *Information*, vol. 16, no. 2, p. 87, Jan. 2025, doi: 10.3390/info16020087.

[7] S. Mahamad, Y. H. Chin, N. I. N. Zulmuksah, M. M. Haque, M. Shaheen, and K. Nisar, "Technical Review: Architecting an AI-Driven Decision Support System for enhanced online learning and assessment," *Future Internet*, vol. 17, no. 9, p. 383, Aug. 2025, doi: 10.3390/fi17090383.

[8] G. Amudha, P. Gopika, S. G, V. R, M. G. Dinesh, and S. S. R, "Cloud computing: Transformations, opportunities, and challenges," *2025 3rd International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS)*, pp. 897–903, Jun. 2025, doi: 10.1109/icssas66150.2025.11080697.

[9] B. Amangeldy, T. Imankulov, N. Tasmurzayev, G. Dikhanbayeva, and Y. Nurakhov, "A review of artificial intelligence and deep learning approaches for resource management in smart buildings," *Buildings*, vol. 15, no. 15, p. 2631, Jul. 2025, doi: 10.3390/buildings15152631.

[10] T.-T.-T. Do, Q.-T. Huynh, K. Kim, and V.-Q. Nguyen, "A survey on video Big Data Analytics: architecture, Technologies, and open Research challenges," *Applied Sciences*, vol. 15, no. 14, p. 8089, Jul. 2025, doi: 10.3390/app15148089.

[11] B. Barua, I. Barua, M. S. Kaiser, and M. J. U. Mozumder, "Trends and Challenges in AI-Driven Microservices for Cloud-Based Airline Reservation Systems: A review," *2025 3rd International Conference on Intelligent Data Communication Technologies and Internet of Things*

(IDCIoT), pp. 1902–1911, Feb. 2025, doi: 10.1109/idciot64235.2025.10915076.

[12] S. Patil, A. Bhat, N. Jain, and V. Javalkar, “Integrating Research on AI-Driven Hyper-Personalization: A review and framework for scalable social media Campaigns,” *2025 International Conference on Pervasive Computational Technologies (ICPCT)*, pp. 766–771, Feb. 2025, doi: 10.1109/icpct64145.2025.10940951.

[13] S. Rao and S. Neethirajan, “Computational Architectures for Precision Dairy Nutrition Digital TwIns: A Technical Review and Implementation framework,” *Sensors*, vol. 25, no. 16, p. 4899, Aug. 2025, doi: 10.3390/s25164899.

[14] J. C. L. Chow and K. Li, “Large language models in medical Chatbots: opportunities, challenges, and the need to address AI risks,” *Information*, vol. 16, no. 7, p. 549, Jun. 2025, doi: 10.3390/info16070549.

[15] S. S. Madani *et al.*, “Artificial Intelligence and Digital twin technologies for intelligent Lithium-Ion battery management systems: A comprehensive review of state estimation, lifecycle optimization, and Cloud-Edge integration,” *Batteries*, vol. 11, no. 8, p. 298, Aug. 2025, doi: 10.3390/batteries11080298.