Network Architecture for Value Enhanced LLM Use Considering Volunteer Computing Paradigm

Ayodele A Periola*
Electrical, Electronic and Computer Engineering
Cape Peninsula University of Technology
Cape Town, Western Cape, South Africa
periolaa@cput.ac.za

Akinunde A Alonge
Electrical and Electronic Engineering Technology
University of Johannesburg
Johannesburg, Gauteng, South Africa
aalonge@uj.ac.za

Kingsley A Ogudo
Electrical and Electronic Engineering Technology
University of Johannesburg
Johannesburg, Gauteng, South Africa
kingsleyo@uj.ac.za

Abstract

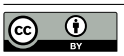
The adoption of large language model-based applications has been recognized to be the future of the adoption of artificial intelligence. However, large language model (LLM) based applications benefit from uncompensated volunteer input. The uncompensated volunteer input arises from the core service tasks such as output preference indication from a large number of subscribers. Given that the large-scale adoption of LLMs is poised to trigger job loss, the challenge arising from the uncompensated volunteer input should be addressed. This is especially important for the case of developing countries. The presented research proposes a network architecture with cost consensus capacity in this regard. In the presented research, cost consensus capacity refers to the consideration of the LLM query heterogeneity and varying usefulness in the future for other LLM system subscribers in determining the cost associated with the volunteer's work. The proposition of the network architecture recognizes, addresses and ensures an avoidance of an artificial intelligence winter. In this case, a cost reduction is applied to the case of compensation of volunteer work input. In the proposed approach, the compensation is not given to the volunteer but the sovereignty authority in the volunteer location. The application of a cost reduction factor to prevent artificial intelligence winter reduces costs by an average of (45.5-78.2)%.

Keywords

LLMs, Computing Task Volunteers, Intelligent Networks, Artificial Intelligence

ACM Reference Format:
Ayodele A Periola, Akinunde A Alonge, and Kingsley A Ogudo. 2025. Network Architecture for Value Enhanced LLM Use Considering Volunteer Computing Paradigm. In *2025 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD 2025), November 26–27, 2025, Cape Town, South Africa.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3759023.3759092

*Corresponding author.



This work is licensed under a Creative Commons Attribution 4.0 International License.
icABCD 2025, Cape Town, South Africa
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1427-6/25/11
https://doi.org/10.1145/3759023.3759092

1 Introduction

Large language models (LLMs) have been recognized to be beneficial regarding enhancing the value obtained from artificial intelligence-based applications. They have been significantly found to be beneficial in modern applications such as Chat GPT. The training of LLMs requires the use of a significant amount of data [3, 9, 12, 13].

The current use of LLM based platforms is data center intensive and also resource-intensive. The resources concerned in this case are water (for cooling), and power (for operation). In addition, LLMs utilize data in two major phases. The first phase is one in which data is used to train the learning algorithms enabling LLM functionality. The second phase is one in which the LLM improves its response to a given query. The improvement arises because of LLM's ability to improve on its responses to repeatedly similar queries.

The training of LLM faces the challenge of data use due to copyright concerns [2, 6, 10, 11]. In this case, copyright concerns arise due to the use of published research literature i.e., (reviewed and non-reviewed). Literature types also include technical articles and other technical related material. The use of such copyrighted material is a challenge for researchers.

Inputs into LLMs arise in two tiers. The first tier is one comprising the use of copyright material. The second tier comprises the use of subscriber inputs for refining LLM system related outputs i.e., responses to users. In the context of the knowledge economy, it is important that nation states where individuals utilize LLM related products should benefit from using public research material. In addition, it is important to design network architecture supporting the realization of such benefits. The basis for the design of such network architecture is that LLMs significantly utilize subscriber data.

In addition, the use of LLMs signals the deployment of a considerable number of data centers. This is necessary as data centers enable the service provisioning of LLMs. In this case, the use of LLMs that utilize intelligent query related data from subscribers should incorporate data sovereignty. The incorporation of data sovereignty is deemed necessary when user queries are personal and sensitive. In this case, sovereignty should also be accompanied by privacy. Privacy in this case refers to the data being associated with the concerned region i.e., in the context of regional data sovereignty. In this case, LLMs should provide users with options to specify if the data associated with their query input should be used to enable LLM learning.

The advances in artificial intelligence, especially as seen in the emergence of generative artificial intelligence such as LLMs, has stirred global outcry on the future of jobs. The increasing role of artificial intelligence has been noted to trigger job loss. Therefore, the existing approach is that the use of LLMs poses a significant risk especially to jobs [4, 19, 20]. This identified risk is recognized to affect creative-oriented jobs by Schillaci [19]. It is recognized that up to 47% of jobs are affected by Deranty et al., [4]. However, Shen [20] also recognizes that the emergence of new applications such as those from LLM presents new opportunities. In this regard, the contributions of volunteers to the embedded volunteer computing system are recognized as an example. In this case, a micro-volunteer computing paradigm is associated. In the conventional volunteer computing paradigm, users are engaged in large scale data processing [16]. This is because LLM users do not engage in large scale data processing. The role of such unpaid micro-volunteers has been identified in relations to OpenAI's LLM as seen in [5].

The discussion in identifying core human tasks in the loop services to reverse the non-sustainable trend poised by the future emergence of LLMs. The research recognizes that human in the loop tasks in relation to machine learning have focuses on data labelling as implied in [4, 19, 20]. The recent emergence of LLMs (being at a nascent stage) and their popularity are recognized to present new types of volunteer work that are yet to receive recognition. Advances in intelligent system applications that benefit from AI such as [14, 21] do not consider volunteer activity. In a developing context, the compensation of volunteer activity increases governmental revenue and overall quality of life. However, the proposed application is also beneficial for developed countries.

The presented research proposes the recognition of volunteer activity as a human in the loop process with relation to refining query response to LLM systems. Research has demonstrated implicitly and explicitly that human labor is crucial in the realization of artificial intelligence and machine learning systems. However, human in the loop refining activity for LLMs (as a nascent application) is yet to receive sufficient research consideration. In this regard, the design of supporting network architecture is deemed necessary. The design of such a network architecture is beneficial for the derivation of revenue by governments, provinces and higher education institutions. The main contribution of the research is the design of a beneficial network architecture considering the use of LLMs. The network architecture is done with a focus on the context of developing nations. The rationale for the design of the proposed network architecture is that developing countries need to benefit from subscriber's use of LLMs. The provision of user inputs in a manner that enhances future LLM response to other subscribers is recognized as a core service. This service is provided to the organizations hosting the LLM based solution and is a core service that is recognized in this paper. The provision of this core service should be rewarded. The contributions of the presented research are enumerated as:

(1) First, the research recognizes that external core services are provided by subscribers to LLM generative applications. In this case, the external core service being recognized is the provision of inputs in a manner that enhances the quality of the response being provided to subscribers. The quality of the response is pre-enhanced by volunteers, i.e., subscribers whose response enhances LLM output.
(2) Second, the research proposes and presents network architecture enabling the derivation of the value of user inputs at two tiers. The considered tiers focus on issuing of new user queries, and queries that enhance future response to other users (being provided by preceding volunteers). The proposed network architecture hosts query intent analysis entities deployed at internet exchange points deployed in a hierarchical manner. The proposed network architecture also enables interactions between LLM hosting organization and internet exchange points enabling user interaction with the LLM service provisioning platform. In addition, the proposed network architecture also incorporates pricing entities and associated mechanisms.
(3) Third, the research integrates network sovereignty contexts for certain user inputs. In this case, sovereignty capability is integrated with the proposed network architecture. This enables storage and future analysis of user intelligent inputs. In this case, the concerned and supporting data center is one with low operational costs. The concerned and supporting data center is the underwater data center.
(4) Fourth, the research also formulates and investigates how the use of the proposed architecture enhances revenue derivable to the concerned tiers. The investigation is done into the existing case (without consideration of volunteers) and proposed case (and with consideration of volunteers).

The rest of the research being presented is organized as follows. Section 2 presents the external core service paradigm associated with LLM based public accessible applications. Section 3 presents the proposed network architecture. The architecture is proposed for multi-tier contexts. The considered contexts focus on higher education institutions, research institutions, and governments (at multiple levels). Section 4 discusses the incorporation of sovereignty in the proposed network architecture. Section 5 formulates the performance model. Section 6 presents and investigates the performance model. Section 7 is the conclusion.

2 LLM – External Core Services

LLM platforms enabled via generative artificial intelligence technology provide intelligent outputs to user query inputs. The LLM platform and application is able to respond to textual and non-textual (image) inputs and yield desirable user outputs. The external core services pertain to users responding to inputs from the LLM on the desirability of a given output.

The concerned outputs are multi-format outputs requiring user input to specify the desirable output [1, 24]. In this case, multiple output types corresponding to a user input are presented to the user. The user i.e., subscriber i.e., user provides an input specifying the suitable format of output display. In addition, the user provides the input query for which such an output is produced by the LLM application platform. The display of such an output is presented in Figure 1.

In addition, image i.e., non-textual inputs can be provided into the LLM application input [7, 15, 18]. The output in this case can be a textual output explaining the image context. In such a case,

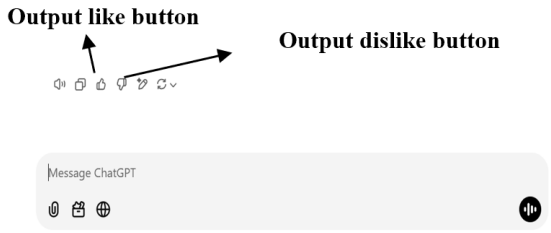


Figure 1: Output like and dislike buttons from the LLM platform ChatGPT from its interface.

the image constitutes the input. The output is the selection of the like/unlike option enabling the LLM application to refine its future output to other subscribers. The selection is done via the user selection of the output like button and output dislike button as presented in Figure 1. This selection represents the first core external service being provided to the LLM application platform.

The second core external service pertains to the LLM application platform's ability to provide refined outputs based on responses to previous inputs. An output in this regard is shown in Figure 2. Figure 2 demonstrates that responses elicited as response to a given user query and input can enhance the output i.e., response to other users (with similar queries). The execution of the identified core services constitutes tasks for which users can bring credits to different entities. In this case, the entities are recognized as higher education institutions (HEIs), municipalities, and provincial governments. In addition, the identified core services arise due to the need to enable users to provide input enabling the LLM application to improve its response and output. These core services are applicable to LLM based artificial intelligence applications.

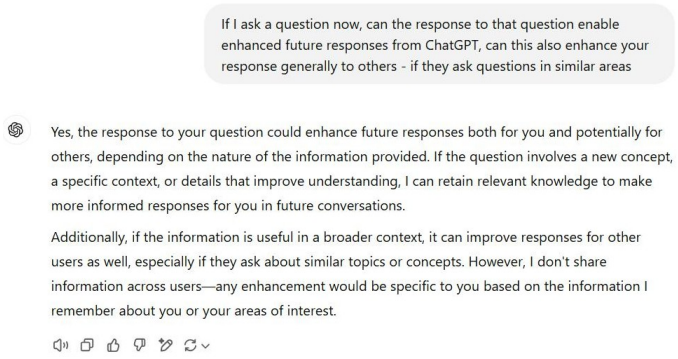


Figure 2: Second core service – Input enabling output refining

3 Proposed Network Architecture

The proposed network architecture enabling the identification of core service tasks related credits is presented in this section. The core entities in the proposed network architecture are the internet exchange points (IXPs). The IXPs are realized using cloud computing platforms. Computing platforms are realized via terrestrial data centers. The IXPs host entities enabling communications with the LLM platform and application service providers. Communications enable the transfer of data related to the identified core services.

The user inputs that constitute those related to the identified core services are those arising from humans and not machines. This is important as the emergence of intent based paradigms implies that robotic query inputs can provide inputs to LLM platforms [17, 22, 23]. The entities associated with the IXP are the: (i) user input analysis entity (UIAE), (ii) user identification entity (UIE), and (iii) Cost Determination entity (CDE). These entities are realized as software and are hosted aboard conventional computing systems such as cloud computing hardware. They are hosted aboard the IXP realized via a cloud computing platform. This is deemed feasible considering advances in computing software technology. In addition, the software based entities communicate via interfaces realized in a software defined network hosted aboard the IXP. The UIAE enables the analysis of the queries from different users. The analysis is executed to determine the number of queries and their frequency of incidence. In this regard, queries can be new and old. New queries ask questions about areas that have not been presented in the LLM application. In a similar manner, old queries focus on questions in areas that have been previously presented to the LLM application. The UIE enables the identification of the user and the associated subscriber institution. In this case, subscribers are considered to belong in a province where each province has the following subscriber classifications: (i) HEI related subscriber, (ii) Research institute related subscriber, and (iii) Municipal related subscriber. The UIE hosts the capability to resolve the location and classification of the subscriber via the internet protocol (IP) address associated with each subscriber related traffic. The UIE hosts geolocation databases and enabling application programming interfaces for subscriber location resolution.

The CDE is an important entity and enables a consensus to be reached between the governing entity concerned and the LLM application platform service provider. The concerned consensus is reached regarding the charges associated with identified external core services. The implied cost consensus is reached via an interaction on prices between the LLM application side and subscriber network side. In this case, the LLM application side informs the subscriber network side i.e., the IXP of how much it is willing to pay while the subscriber assents, dissents or suggests a new price. This handshake like procedure occurs till an agreement is reached. Future prices become easy to determine as the CDE maintains a repository of past consensus determined prices.

The proposed network architecture showing the relations between the participating entities is presented in Figure 3. The architecture in Figure 3 comprises two aspects. These aspects are the subscriber network side and the LLM application side. The scenario on the subscriber network side shows subscribers generating requests and queries to be addressed by the LLM side. The requests and queries concerned are relayed to the LLM side via a network comprising the base station, network gateway, and the IXP. The subscriber communicates with the base station, the queries i.e., user input are sent to a network gateway and onward to the IXP. The IXP hosts computing entities enabling the execution of the functionalities described in the UIAE, CDE, and UIE.

The LLM application side comprises an IXP in the nation state where the LLM enabling data centre is located. In the scenario of

the proposed architecture, the subscribers and the LLM application are hosted in different nation states. In addition, the scenario in Figure 3 shows that multiple subscribers can be supported and associated with the network concerned. In Figure 3, the IXP on the LLM application side executes the functionalities of the UIAE, UIE and CDE.

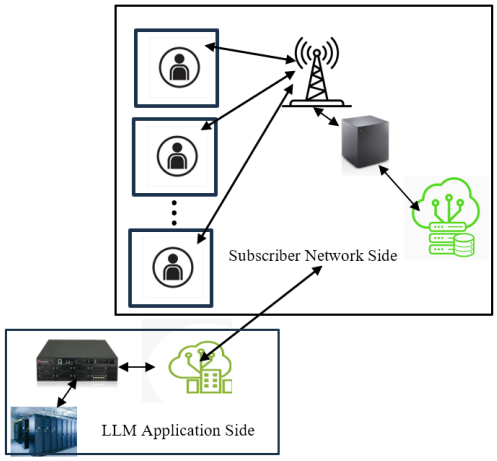


Figure 3: Relations between Network and LLM.

Interactions between the gateway and the IXP related components in the network subscriber and LLM application sides are presented in Figure 4. In Figure 4, the UIAE is linked to the network gateway. Furthermore, the UIAE is connected to the UIE that is further linked to the CDE.

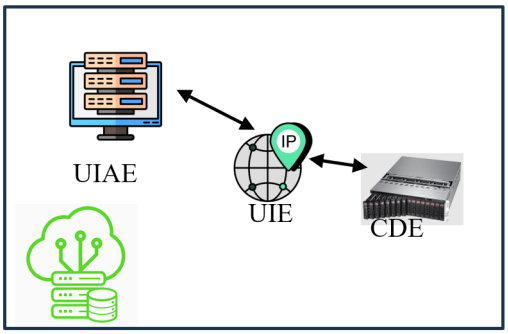


Figure 4: Relations between the UIAE, UIE, AND CDE associated with the IXP.

The architecture presented in Figure 3 and Figure 4 is applicable to HEI, research institute (RI), and municipal authorities. In this case, the UIE executes the function of determining the originating location of the concerned subscriber traffic. The determination of the originating location is executed prior to determining the credit associated with the location concerned.

4 Incorporation of Sovereignty

The incorporation of sovereignty is important and ensures that sensitive user queries and user inputs are not sent out of the subscriber's location. Nevertheless, the queries are executed with answers being sent to the requesting subscriber. This requires that a data center for executing the queries is located in the subscriber's location. The queries or user inputs that are considered sensitive in this case are those with themes related to national defense, personal health care with high technical details, queries containing phrases suggestive of passwords, queries containing detailed description of artefacts and their locations, and highly detailed scenarios of personal contexts. In this case, personal contexts arise in scenarios such as family and work place related situations. The introduction of two major systems enables the incorporation of the capability of sovereignty. These capabilities are: (i) Query sensitivity analysis, (ii) Sensitive query processing, and (iii) Sensitive query storage. The identified capabilities arise from the incorporation of computing entities and platforms.

In this case, another computing platform that hosts a query sensitive analysis entity is introduced. In this case, the computing platform is realized via a mini data center (MDC). Alternatively, an underwater data center (UDC) can be used if the concerned location is coastal. In this case, the computing platform is linked to the UIAE via bi-directional communications. The MDC or UDC being used hosts the LLM application and provides a cache for user query storage. In the resulting architecture, the MDC/UDC is connected to the UIAE. The UIAE retains its manner of connection to the UIE and CDE as seen in Figure 5. In this case, the MDC/UDC executes the functionality associated with the capabilities of query sensitivity analysis, processing and storage.

The incorporation of sovereignty via the architecture presented in Figure 5 also ensures that subscriber queries do not leave a nation state for external execution in the event that credits as determined by the CDE receive non-compliance as reached via consensus. The execution of the proposed solution ensures compliance with the relevant regulations without compromising on regional data sovereignty. The entities presented in Figure 5 are all located within a sovereign region. This shows that the CDE's framework in cost determination can be ensured to have legal compliance while using a software entity constituting the enabling information infrastructure. In this case, the LLM application is executed aboard a higher capacity MDC/UDC. This has beneficial implications for the host nations as it enables hosting more data centers. The incorporation of sovereignty enables the realization of a cost consensus negotiating distributed LLM application data center for supporting applications. In this distributed LLM application data center, bi-directional communications occur in a manner involving the UIAE, UIE and CDE as defined in Figure 4. Furthermore, the MDC/UDC also communicates with external networks i.e., other LLM application data center via a network gateway. The network architecture is presented in Figure 5. In Figure 5, MDC/UDC communications occur when CDE successfully reaches a consensus. A consensus is reached between the LLM application provider and the IXP entity. The incorporation of sovereignty in the manner proposed and presented in Figure 5 can induce a lull in artificial intelligence. This lull arises when CDE costs become significantly high. Such a lull or winter should be

prevented. The prevention is realized via the introduction of a cost reduction factor (CRF) by the IXP or a government control entity. The CRF is introduced to ensure that the costs determined by the CDE are not prohibitively high thereby reducing subscription to LLM applications. Such prohibitively prohibitive costs can cause a significant drop in LLM development. This is the case described as an AI winter. However, the introduction of CRF ensures that such high prohibitive costs do not occur. This is of benefit to the LLM data center operator and application service provider.

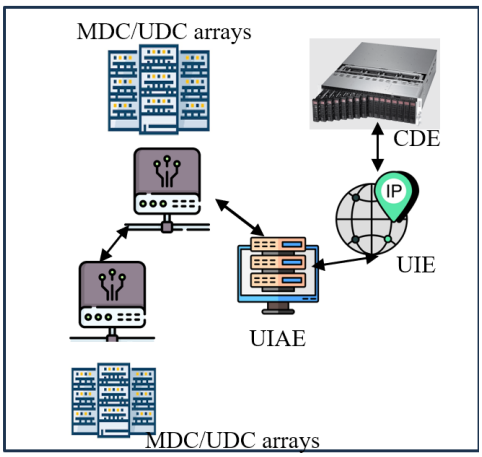


Figure 5: Scenario showing the relations between two MDCs/UDCs considering CDE negotiations.

5 Performance Formulation

The focus of the discussion in this section is the formulation of the cost benefits arising from the use of the proposed solution. The focus in this case is not on network quality of service (QoS). The QoS has not been considered in this case due to the availability of high-speed fibre optic links enabling inter-data center connectivity. The availability of such links enables the realization of gigabits [17] and terabits [8] level connectivity. The formulation context considers a scenario where there are multiple LLM applications and multiple sovereign regions. Each region has three entities, i.e., HEIs, Ris, and municipalities. The concerned LLM based applications have generative intelligence capability. Let θ, and β denote the set of LLM based applications such that:

$$\theta = \{\theta_1, \theta_2, \ldots, \theta_A\} \tag{1}$$

$$\beta = \{\beta_{HEI}, \beta_{RI}, \beta_M\} \tag{2}$$

β_{HEI}, β_{RI}, and β_M are the set of HEIs, RIs, and municipalities, respectively.

$$\beta_{HEI} = \{\beta^1_{HEI}, \beta^2_{HEI}, \ldots, \beta^P_{HEI}\} \tag{3}$$

$$\beta_{RI} = \{\beta^1_{RI}, \beta^2_{RI}, \ldots, \beta^Q_{RI}\} \tag{4}$$

$$\beta_M = \{\beta^1_M, \beta^2_M, \ldots, \beta^J_M\} \tag{5}$$

The pth HEI, $\beta^p_{HEI}, \beta^p_{HEI} \in \beta_{HEI}$ has multiple queries given as:

$$\beta^p_{HEI} = \{\beta^{p,1}_{HEI}, \beta^{p,2}_{HEI}, \ldots, \beta^{p,L}_{HEI}\} \tag{6}$$

The qth RI, $\beta^q_{RI}, \beta^q_{RI} \in \beta_{RI}$ has multiple queries given as:

$$\beta^q_{RI} = \{\beta^{q,1}_{RI}, \beta^{q,2}_{RI}, \ldots, \beta^{q,M}_{RI}\} \tag{7}$$

The jth municipality M, $\beta^j_M, \beta^j_M \in \beta_M$ has multiple queries:

$$\beta^j_M = \{\beta^{j,1}_M, \beta^{j,2}_M, \ldots, \beta^{j,N}_M\} \tag{8}$$

The cost per request associated with the entity χ, $\chi \in \left\{\beta^{p,l}_{HEI}, \beta^{q,m}_{RI}, \beta^{j,i}_M\right\}$ at the epoch t_y, $t_y \in t$, $t = \{t_1, t_2, \ldots, t_Y\}$ is denoted $C(\chi, t_y)$. The instantaneous number of requests associated with the entity χ is denoted as $N_1(\chi, t_y)$. In addition, the cost reduction factor (CRF) associated with the entity χ at the epoch t_y is denoted $C_1(\chi, t_y)$. In the existing case, the role of the CDE is absent being inapplicable. Therefore, the value of $C(\chi, t_y)$ is zero or near zero, i.e., small. The cost formulation arising due to the introduction of the CDE in the HEI, RI, and municipality M are denoted as $\Gamma_1(\text{HEI})$, $\Gamma_1(\text{RI})$, and $\Gamma_1(M)$, respectively.

$$\Gamma_1(\text{HEI}) = \sum_{y=1}^{Y}\sum_{p=1}^{P}\sum_{l=1}^{L} N_1(\beta^{p,l}_{HEI}, t_y) C(\beta^{p,l}_{HEI}, t_y) \tag{9}$$

$$\Gamma_1(\text{RI}) = \sum_{y=1}^{Y}\sum_{q=1}^{Q}\sum_{m=1}^{M} N_1(\beta^{q,m}_{RI}, t_y) C(\beta^{q,m}_{RI}, t_y) \tag{10}$$

$$\Gamma_1(M) = \sum_{y=1}^{Y}\sum_{j=1}^{J}\sum_{i=1}^{I} N_1(\beta^{j,i}_M, t_y) C(\beta^{j,i}_M, t_y) \tag{11}$$

Given the introduction of the CRF, $\Gamma_1(\text{HEI})$, $\Gamma_1(\text{RI})$, and $\Gamma_1(M)$ are obtained as:

$$\Gamma_1(\text{HEI}) = \sum_{y=1}^{Y}\sum_{p=1}^{P}\sum_{l=1}^{L} N_1(\beta^{p,l}_{HEI}, t_y) C(\beta^{p,l}_{HEI}, t_y) C_1(\beta^{p,l}_{HEI}, t_y) \tag{12}$$

$$\Gamma_1(\text{RI}) = \sum_{y=1}^{Y}\sum_{q=1}^{Q}\sum_{m=1}^{M} N_1(\beta^{q,m}_{RI}, t_y) C(\beta^{q,m}_{RI}, t_y) C_1(\beta^{q,m}_{RI}, t_y) \tag{13}$$

$$\Gamma_1(M) = \sum_{y=1}^{Y}\sum_{j=1}^{J}\sum_{i=1}^{I} N_1(\beta^{j,i}_M, t_y) C(\beta^{j,i}_M, t_y) C_1(\beta^{j,i}_M, t_y) \tag{14}$$

6 Performance Evaluation

Performance evaluation and analysis are presented in this section. The evaluation is done using MATLAB. In the performance evaluation, the considered scenario is one in which HEIs, RIs, and municipalities have multiple subscribers. The simulation focuses on investigating how the use of the proposed architecture and paradigm influences related costs. This is done in the context of IXP supporting low latency data transmission meeting the requests. The analysis focuses on requests that are successfully executed and not necessarily the traffic patterns associated with these requests. This is because requests and not traffic patterns are more associated with volunteer activity.

The simulation scenario considers a context where the number of requests for HEIs, RIs, and municipalities are aggregated and increasing at each epoch of the simulation. The simulation scenario considered three cases i.e., Case 1, Case 2, and Case 3. The scenario in each of these cases considers different profiles given the use of the proposed approach with volunteer compensation with revenue due to government authority. Case 1 considers the scenario where the proposed approach is utilized without any reduction in the cost i.e., no role of the CDE. Case 2 considers the scenario where there is a significant reduction in cost but without the occurrence of zero costs. Such a case is akin to the existing case. Case 3 is a scenario where the cost reduction factor is introduced and considered. In the simulation, different volunteers have varying costs due to the asymmetry of LLM request activity. This is deemed feasible because of the policy differentiation for the LLM provider for different locations and LLM query heterogeneity regarding future utility.

The evaluation parameters are in Table 1. The performance evaluation results are in Figure 6. In Figure 6, the total number of entities has an equal number of subscribers associated with HEIs, RIs and municipalities. Given 15 entities, the number of HEI, RI and municipality-based subscribers are five each.

The performance evaluation is done from a conservative and analytical perspective. A conservative perspective is considered as all the prices considered and presented in Table 1 are less than 1. The non-conservative approach has been adopted with the goal of ensuring that the adoption of the proposed approach does not induce an era of artificial intelligence winter. The consideration of this approach enables the determination of the performance benefits of the considered cases against each other. From Figure 6, it can be seen that the introduction of the cost reduction factor (CRF) has the least cost. In addition, the non-reduction and non-consideration of the CRF as seen in Case 1 results in the highest cost. The costs in this case are paid to the governmental authority by the LLM application service provider. From the results presented, it can be seen that the costs associated with 15 entities (IXPs) do not pose a significant burden to the LLM application service provider. This ensures continuity of the service thereby not hindering LLM application development.

Analysis of the results presented in Figure 5 shows that the introduction of the CRF in Case 3 results in the least costs. The highest costs arise in Case 1. The introduction of the CRF in Case 3 reduces the cost on average by 78.2% in comparison to Case 1. In addition, the reduction of the CRF in Case 3 reduces the cost on average by 53.1% in comparison to Case 2 (with applied cost reduction). Furthermore, the reduction of costs in Case 2 reduces the total cost on average by 45.5 % in comparison to Case 1. In addition, the consideration of an increased number of subscribers increases the derivable revenue i.e., costs. Given 600 subscribers, the approximate maximum derivable revenue for Case 1, Case 2, and Case 3 are \$1700, \$700, and \$400, respectively. Therefore, it can be inferred that scaling the costs for an increase in the number of subscribers does not pose a significant challenge to the LLM application service provider as the cost burden arising from the revenue does not exceed \$2000.

Table 1: Parameters used for Simulation and Evaluation

S/N	Parameter	Value
1	Number of HEIs, RIs, and Municipalities	20
2	Maximum Cost per request of HEI	\$0.8466
3	Minimum Cost per request of HEI	\$0.0085
4	Mean Cost per request of HEI	\$0.5199
5	Maximum Cost per request of RI	\$0.9932
6	Minimum Cost per request of RI	\$0.0621
7	Mean Cost per request of RI	\$0.5676
8	Maximum Cost per request of Municipality	\$0.9581
9	Minimum Cost per request of Municipality	\$0.00007
10	Mean Cost per request of Municipality	\$0.4602
Reduced Costs – Case 2		
11	Maximum Cost per request of HEI	\$0.4498
12	Minimum Cost per request of HEI	\$0.0133
13	Mean Cost per request of HEI	\$0.2556
14	Maximum Cost per request of RI	\$0.3686
15	Minimum Cost per request of RI	\$0.0282
16	Mean Cost per request of RI	\$0.2180
17	Maximum Cost per request of Municipality	\$0.4377
18	Minimum Cost per request of Municipality	\$0.0587
19	Mean Cost per request of Municipality	\$0.2595
Introduction of the Cost Reduction Factor – Case 3		
20	Maximum HEI Cost Reduction Factor	37.27%
21	Minimum HEI Cost Reduction Factor	1.58%
22	Mean HEI Cost Reduction Factor	22.55%
23	Maximum RI Cost Reduction Factor	37.69%
24	Minimum RI Cost Reduction Factor	2.07%
25	Mean RI Cost Reduction Factor	22.96%
26	Maximum Municipality Cost Reduction Factor	37.93%
27	Minimum Municipality Cost Reduction Factor	1.09%
28	Mean Municipality Cost Reduction Factor	18.54%

7 CONCLUSION

The presented research recognizes that advances in the applications of artificial intelligence necessitate the design of sustainable solutions. In this case, sustainability is addressed from the perspective of the impact of artificial intelligence on jobs. It is recognized that large language model based artificial intelligence benefit from human in the loop on a large scale. It is also recognized that the contributions input is not considered. The input contributions arise from large scale human in the loop participation. The research proposes compensation for these volunteers for each country and is derivable to a governmental entity. The need to realize this compensation motivates the design of the presented and proposed network architecture. A conservative estimate of the rewards is formulated and investigated. The investigation considers the need to avoid and limit the occurrence of an artificial intelligence winter. The formulation and investigation are done for three scenarios. The scenarios consider the variation of costs and cost reduction with and without a cost reduction factor. The reduction of costs results

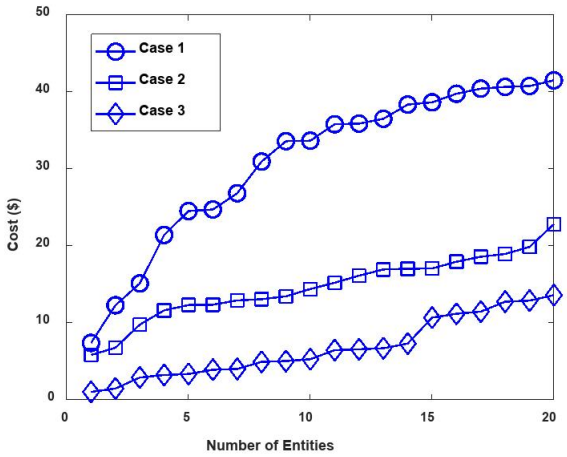


Figure 6: Costs associated with Case 1, Case 2, and Case 3 given the use of the proposed approach in different contexts.

in a case where the adoption of a reduction factor enables lowering costs by an average of (45.5–78.2)%. Future work will focus on getting access to an internet exchange point to conduct proof of concept. In addition, future work will also focus on the scalability and suitability of edge computing networks in realizing the functionality associated with the proposed architecture.

Acknowledgments

The authors acknowledge their indebtedness to the Department of Electrical, Electronic and Computer Engineering, of the Cape Peninsula University of Technology ; and the Department of Electrical, and Electronic Engineering Technology of the University of Johannesburg.

References

[1] M. Alizadeh et al. 2024. Open-source LLMs for text annotation: a practical guide for model setting and fine-tuning. *Journal of Computational Social Science* 8, 1 (Dec 2024). https://doi.org/10.1007/s42001-024-00345-9

[2] A. Buick. 2024. Copyright and AI training data—transparency to the rescue? *Journal of Intellectual Property Law & Practice* (Dec. 2024). https://doi.org/10.1093/jiplp/jpae102

[3] J. Chen et al. 2024. When large language models meet personalization: perspectives of challenges and opportunities. *World Wide Web* 27, 4 (June 2024). https://doi.org/10.1007/s11280-024-01276-1

[4] J.-P. Deranty and T. Corbin. 2022. Artificial intelligence and work: a critical review of recent research from the social sciences. *AI & SOCIETY* 39 (June 2022), 1–17. Available: https://link.springer.com/article/10.1007/s00146-022-01496-x.

[5] B. Dippenaar. 2022. ChatGPT and AI: Navigating uncharted copyright territory. https://www.ensafrica.com/news/detail/7137/chatgpt-and-ai-navigating-uncharted-copyright.

[6] D. J. Gervais, N. Shemtov, H. Marmanis, and C. Z. Rowland. 2024. The Heart of the Matter: Copyright, AI Training, and LLMs. *SSRN* (2024). https://doi.org/10.2139/ssrn.4963711

[7] Y. Gou et al. 2024. Eyes Closed, Safety on: Protecting Multimodal LLMs via Image-to-Text Transformation. In *Lecture Notes in Computer Science*. 388–404. https://doi.org/10.1007/978-3-031-72643-9_23

[8] H. Jeong, H. Lee, C. Kim, and S. Shin. 2024. A Survey of Robot Intelligence with Large Language Models. *Applied Sciences* 14, 19 (Oct 2024), 8868–8868. https://doi.org/10.3390/app14198868

[9] U. Kamath, K. Keenan, G. Somers, and S. Sorenson. 2024. *Large Language Models: A Deep Dive*. Springer. https://doi.org/10.1007/978-3-031-65647-7

[10] M. Kowala. 2024. Protection of Press Publishers in the Age of Generative AI – In Search of Legal Remedies to Adapt to the Pace of Technology. *Deleted Journal* (Aug. 2024). https://doi.org/10.1007/s40319-024-01515-y

[11] M. Kretschmer, T. Margoni, and P. Oruç. 2024. Copyright Law and the Lifecycle of Machine Learning Models. *IIC - International Review of Intellectual Property and Competition Law* (Feb. 2024). https://doi.org/10.1007/s40319-023-01419-3

[12] A. Kucharavy. 2024. Overview of Existing LLM Families. 31–44. https://doi.org/10.1007/978-3-031-54827-7_3

[13] P. Kumar. 2024. Large language models (LLMs): survey, technical frameworks, and future challenges. *Artificial Intelligence Review* 57, 9 (Aug. 2024). https://doi.org/10.1007/s10462-024-10888-y

[14] T. A. Kumar, S. Oviya, S. A. Ajagbe, C. Ananth, O. Aiyeniko, and M. O. Adigun. 2023. A Novel Deep Neural Network-Based Prediction Model for Identifying Diseases in Tomato Leaves. In *2023 International Conference on Electrical, Computer and Energy Technologies (ICECET)*. 1–6. https://doi.org/10.1109/icecet58911.2023.10389207

[15] J. Oppenlaender. 2024. The Cultivated Practices of Text-to-Image Generation. In *Humane Autonomous Technology*. 325–349. https://doi.org/10.1007/978-3-031-66528-8_14

[16] A. A. Periola and O. E. Falowo. 2020. Hybrid wireless aided volunteer computing paradigm. *Wireless Networks* 26, 7 (2020), 5355–5369. https://doi.org/10.1007/s11276-020-02395-z

[17] D. Roman. 2025. Can LLMs Make Robots Smarter? *ACM.org* (Jan 2025). https://doi.org/10.1145/3701227

[18] L. Rundo and C. Militello. 2024. Image biomarkers and explainable AI: handcrafted features versus deep learned features. *European Radiology Experimental* 8, 1 (Nov 2024). https://doi.org/10.1186/s41747-024-00529-y

[19] Z. Schillaci. 2024. LLM Adoption Trends and Associated Risks. In *Large Language Models in Cybersecurity*. 121–128. https://doi.org/10.1007/978-3-031-54827-7_13

[20] Y. Shen. 2024. Future jobs: analyzing the impact of artificial intelligence on employment and its mechanisms. *Economic Change and Restructuring* 57, 2 (Feb. 2024). https://doi.org/10.1007/s10644-024-09629-6

[21] H. K. Tripathy, S. A. Ajagbe, and M. El. 2022. Sustainable Management for the Architectural Heritage in Intelligent Cities using MCDM methods. 41–58. https://doi.org/10.54216/jisiot.060104

[22] J. Wang et al. 2024. Large language models for robotics: Opportunities, challenges, and perspectives. *Journal of Automation and Intelligence* (Dec 2024). https://doi.org/10.1016/j.jai.2024.12.003

[23] E. Warner et al. 2024. Multimodal Machine Learning in Image-Based and Clinical Biomedicine: Survey and Prospects. *International Journal of Computer Vision* (Apr 2024). https://doi.org/10.1007/s11263-024-02032-8

[24] J. Yeom et al. 2024. Tc-llama 2: fine-tuning LLM for technology and commercialization applications. *Journal of Big Data* 11, 1 (Aug 2024). https://doi.org/10.1186/s40537-024-00963-0