International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868
Foundation of Computer Science FCS, New York, USA
Volume 12– No.46, February 2025 – www.ijais.org

Tokens as Currency: A Novel Framework to Sustain AI Adoption and Profitability

Siddharth Nandagopal
Unaffiliated/Independent
Cambridge, Massachusetts. 02139, USA

ABSTRACT
Large Language Models (LLMs) are widely adopted across sectors for tasks like text generation, analytics, and automated support. However, many organizations struggle with token-based cost escalation, hampering long-term sustainability. The objective of this study is to address the urgent demand for an integrated framework that optimizes token usage and preserves financial stability. The research employs a conceptual, theoretical methodology by examining existing literature, synthesizing best practices, and proposing hypothetical use cases that illustrate potential benefits. Results reveal that the proposed Token-Efficient AI Utilization Framework (TEA-UF) can significantly improve token efficiency, enhance operational scalability, energy efficiency, and reduce overall expenditures linked to LLM deployments. Organizations adopting TEA-UF can mitigate vendor lock-in risks, balance on-premises and cloud infrastructures, and forecast costs more accurately. This approach underscores the viability of sustainable AI adoption without sacrificing innovation. In conclusion, the framework holds promise for revolutionizing LLM integration across diverse industries while maintaining economic responsibility. By implementing TEA-UF, businesses can achieve deeper market penetration, faster product evolution, and streamlined budgeting processes. The solution fosters global collaboration and fuels robust AI-driven growth, confirming that token efficiency and resource optimization can serve as cornerstones for successful LLM deployment strategies. Hence, it fosters synergy for enduring, future-proof AI adoption.

General Terms
Algorithms, Econometrics, System Architecture, Scalability, Privacy, Resource Management

Keywords
Token Efficiency, Sustainable AI, LLM Deployment, Cloud-Local Integration, Cost Optimization, Hybrid Strategies

1. INTRODUCTION
Artificial Intelligence (AI) has become a cornerstone for innovation across industries, transforming how organizations operate and compete [30]. In recent years, the adoption of Large Language Models (LLMs) has expanded rapidly, offering unprecedented capabilities in natural language processing, customer engagement, and automation [12]. LLMs, such as GPT-4 and LLaMA, enable tasks like text summarization, content generation, and conversational interfaces, making them indispensable in modern product portfolios [12]. However, building a proprietary LLM from scratch poses substantial challenges, leading many organizations to rely on established LLM service providers [27].

Developing a custom LLM demands extensive resources, including vast datasets, specialized hardware, and expert talent [27]. Training models with billions of parameters requires months of computational effort, often costing millions of dollars [27]. Additionally, the maintenance of such models, including updates and error corrections, adds ongoing expenses. For most organizations, these constraints make the direct adoption of service providers like OpenAI, Google, and Microsoft a more practical choice. These providers offer pre-trained models with ready-to-use APIs, enabling faster deployment and reducing the entry barrier [27].

Despite their advantages, LLM service providers operate on token-based pricing models, where organizations pay based on the number of tokens processed during inference [39]. This consumption-based cost structure can escalate rapidly, mirroring the financial challenges encountered in the early days of cloud computing adoption [37]. Organizations that migrated to the cloud attracted by initial low costs often faced vendor lock-in and surging expenses as usage grew, making transitions back to on-premises systems economically unfeasible [15]. Similar risks now arise with LLMs, where uncontrolled token usage can threaten financial stability and limit scalability [37], [39].

Moreover, token-heavy processing often leads to unnecessary energy consumption, raising questions about sustainability and environmental impact. Lowering token overhead through strategic AI utilization can minimize the carbon footprint of frequent cloud-based inference [2], [6].

The stakes for organizations extend beyond financial implications. Sustaining competitive AI strategies while managing operational costs is critical for long-term profitability [37]. Unchecked expenses in AI utilization can also impact workforce investments, jeopardizing innovation and employee development [37]. To address these challenges, this paper introduces the Token-Efficient AI Utilization Framework (TEA-UF), a novel solution designed to optimize token usage, enhance cost predictability, and enable sustainable LLM adoption. TEA-UF integrates principles of optimization, hybrid deployment, and intelligent design, providing a comprehensive strategy to balance efficiency and scalability.

2. METHODOLOGY
2.1 Appropriateness of the Methods
As depicted in Figure 1, the methodology employed in this research integrates a systematic approach combining literature review, conceptual framework design, and hypothetical applications to develop the TEA-UF [19]. A systematic literature review was conducted to analyze the growing challenges organizations face in managing LLM tokens and the limitations of existing solutions. Recent studies were reviewed to identify the factors contributing to cost escalation, inefficiencies, and vendor lock-in in AI adoption [27], [37],

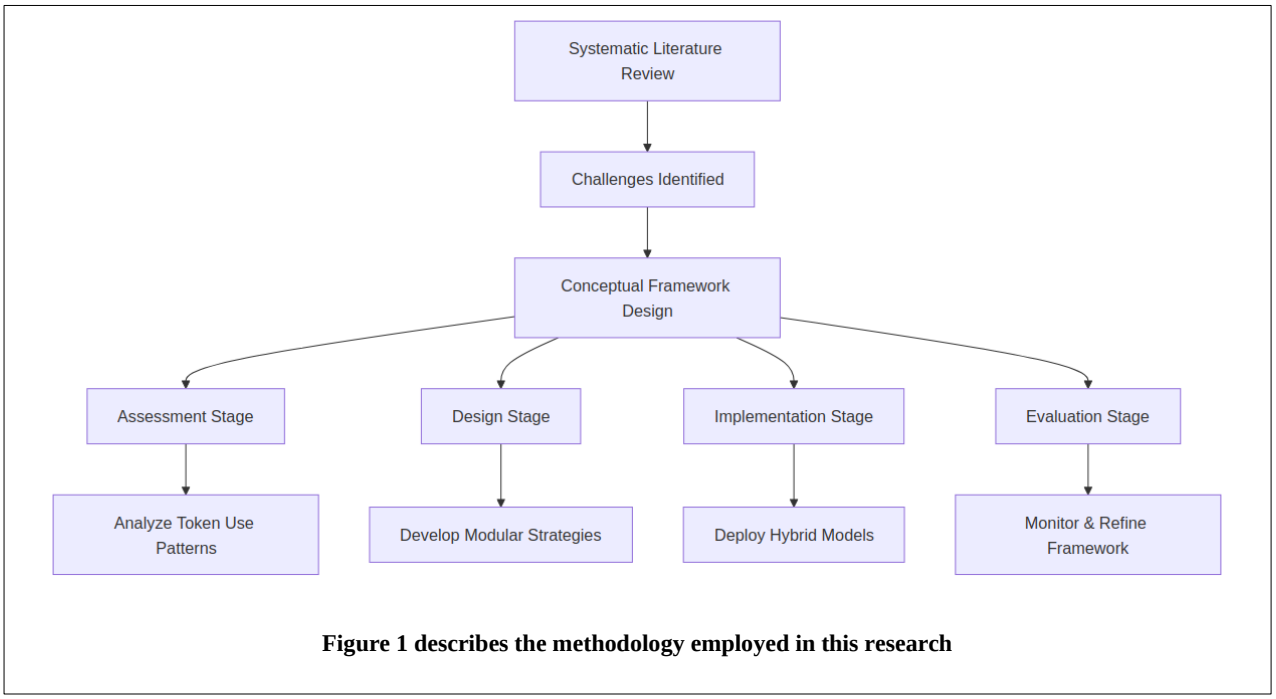


Figure 1 describes the methodology employed in this research

[39]. This review provided a foundation to conceptualize TEA-UF as a sustainable and efficient framework for AI utilization.

The conceptual framework was designed using insights derived from the literature and organizational challenges, focusing on token optimization, hybrid deployment, and secure data preprocessing. This design phase incorporated feedback from case studies of industries reliant on LLMs, such as retail, and healthcare, highlighting the need for adaptable and scalable solutions [7], [27]. Finally, hypothetical applications of TEA-UF were crafted to validate its practical relevance. These scenarios simulated its implementation across diverse sectors, projecting cost savings, enhanced operational efficiency, and improved data security [19].

2.2 Framework Development

The development of TEA-UF followed a structured, multi-stage approach, ensuring its comprehensiveness and adaptability across organizational contexts. The framework was built upon three foundational principles: token optimization, hybrid deployment, and secure preprocessing. Each principle addressed a specific aspect of LLM token management to align with organizational goals [36], [40].

1. Assessment Stage: This stage involved evaluating the AI needs of organizations and analyzing token consumption patterns to identify inefficiencies. Tools like token usage dashboards and predictive analytics were recommended for accurate assessments [1], [28], [35].
2. Design Stage: This phase focused on modularizing AI tasks, enabling organizations to allocate resources based on operational priorities. Strategies included integrating caching mechanisms and summarizing prompts to reduce redundant token consumption [17], [23].
3. Implementation Stage: This stage emphasized deploying hybrid models combining cloud and local infrastructures. Such models offered scalability while ensuring data privacy and compliance with industry regulations [3], [32].
4. Evaluation Stage: Continuous monitoring and iterative refinement were key components of this stage. Real-time dashboards tracked token usage, while feedback mechanisms ensured the framework remained adaptive to changing needs [1], [28], [34], [35].

This methodology establishes TEA-UF as a robust solution to address the complexities of LLM token management, making it highly relevant for adoption across industries.

3. LITERATURE REVIEW

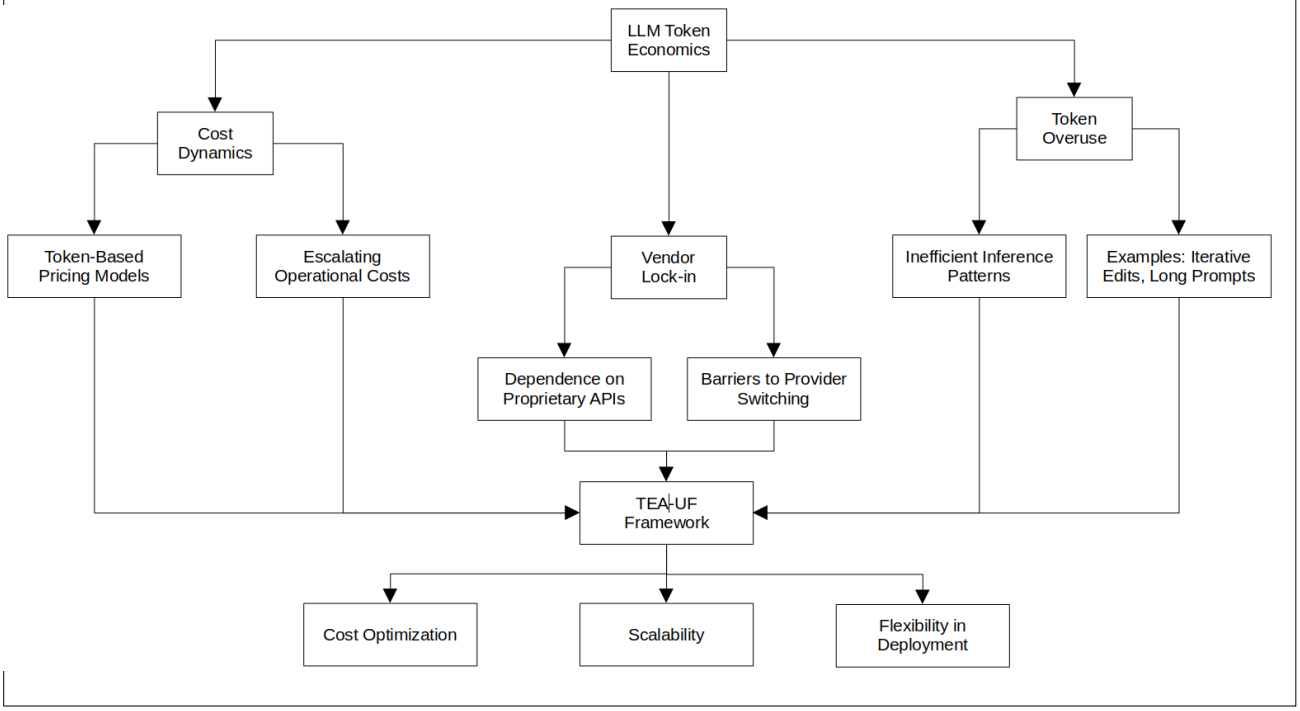
3.1 Comprehensiveness

The landscape of LLM tokenization has evolved significantly in recent years, emphasizing the need to manage token economics effectively. As depicted in Figure 2, Research has highlighted critical challenges such as token-based pricing, escalating costs, vendor lock-in, and inefficiencies in token usage [27], [37], [39]. These studies underline the urgency for frameworks like the TEA-UF.

3.2 The Cost Dynamics of LLM Tokens

Token-based pricing models form the foundation of LLM service offerings, where each token processed incurs a cost. The computational cost of token usage is often underestimated, with expenses accumulating as organizations scale operations. Over time, repetitive processing, redundant token consumption, and large-context prompts exacerbate these costs. Token overuse can result from poor prompt structuring and inefficient utilization of model capabilities, which inflate operational expenses and undermine budget predictability [16], [20], [27], [37], [39].

Figure 2 illustrates the interconnected challenges and their resolution through TEA-UF



3.3 The Vendor Lock-in Dilemma

Dependence on proprietary infrastructures and APIs creates a challenging dynamic for organizations using LLM services. There is difficulty of transitioning to alternative providers due to deeply ingrained dependencies on vendor-specific ecosystems. Furthermore, proprietary models often restrict integration flexibility, forcing organizations to navigate high switching costs and significant operational disruptions. There is need for hybrid solutions to mitigate these barriers, advocating for models that blend cloud-based scalability with local infrastructure autonomy [3], [8], [21], [27], [32], [37], [39].

3.4 Token Overuse in Practical Scenarios

Inefficiencies in token consumption often arise from long-context prompts, iterative edits, and unoptimized tokenization strategies. The excessive token consumption during iterative code editing or document processing inflates costs unnecessarily. Additionally, the lack of caching mechanisms and suboptimal pre-processing techniques compounds the issue. Examples such as managing customer queries or editing large-scale contracts illustrate how improper token management can lead to resource waste [17], [20], [23], [31].

3.5 Relevance to Current Study

The insights from these studies reveal significant gaps in existing token management strategies, emphasizing the need for a structured approach like TEA-UF. While prior research provides a thorough understanding of the challenges, it often falls short in offering comprehensive solutions that combine cost efficiency, scalability, and flexibility. TEA-UF addresses this gap by introducing modular design principles, hybrid deployment strategies, and token optimization techniques. This framework builds upon the lessons from previous studies, providing a practical pathway for organizations to adopt sustainable AI practices.

The need for efficient token management is critical not only for large-scale enterprises but also for small and medium-sized organizations seeking to integrate AI capabilities. As depicted in Figure 2, the proposed framework aligns with the findings of [20] and [21], reinforcing the importance of balancing operational costs with the flexibility to switch providers or adopt open-source alternatives.

4. CONCEPTUALIZING THE TOKEN-EFFICIENT AI UTILIZATION FRAMEWORK (TEA-UF)

4.1 Core Principles of TEA-UF

The TEA-UF incorporates a set of principles designed to manage LLM token usage effectively [8], [36], [40]. Each principle addresses a specific concern that organizations face when handling token-based pricing models and rising operational costs [3], [8], [16], [21], [27], [32], [37], [39]. One core idea involves smart prompts, where shorter and more targeted instructions help reduce token expenses without sacrificing quality. Another important strategy, known as caching, stores previously generated content to avoid regenerating similar text repeatedly. In addition, context summarization condenses lengthy inputs into concise fragments, minimizing token consumption during inference [17], [20], [21], [23], [31].

An equally vital principle involves hybrid deployment models, blending cloud services for high-volume workloads with on-premises systems for sensitive data. This approach allows organizations to benefit from the elasticity of external providers while retaining tighter control over critical resources [3], [8], [16], [21], [27], [32], [37], [39]. Secure pre-processing plays a pivotal role by filtering out unnecessary details before sending text to the LLM, which further lessens token demands [20],

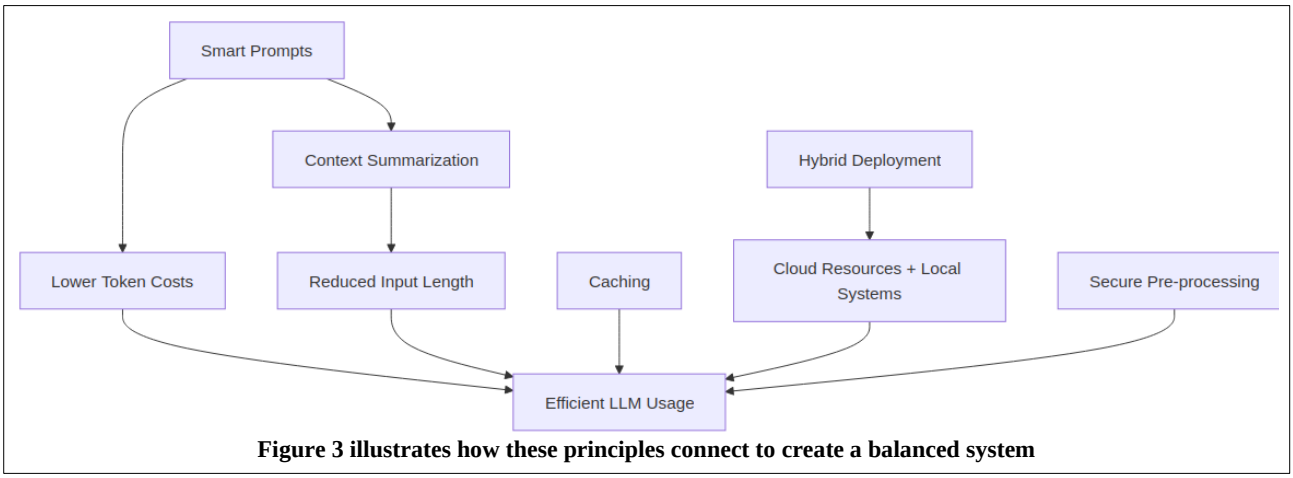


Figure 3 illustrates how these principles connect to create a balanced system

[36], [40]. Maintaining data-locality also bolsters privacy, especially in sectors like healthcare or finance that require strong compliance measures [3], [21], [32]. Together, these principles help users achieve economical token usage and a reliable infrastructure that suits varied operating conditions [20], [31].

Figure 3 demonstrates how each principle feeds into token efficiency, leading to less cost strain on organizations [8], [36], [40]. Smart prompts keep messages relevant, context summarization trims excessive data, caching prevents duplicate outputs, and hybrid deployment merges flexible computation with strict security demands. Meanwhile, secure pre-processing ensures that incoming text is already optimized, further lowering the chance of unnecessary token usage [16], [17], [20], [21], [23], [31].

4.2 The TEA-UF Lifecycle

As depicted in Figure 4, a well-defined lifecycle ensures that TEA-UF remains adaptable as an organization's needs evolve.

Stage 1, Assessment, involves analyzing existing AI requirements, monitoring token consumption patterns to spot inefficiencies and identifying organizational priorities. This stage also considers the sensitivity of data, regulatory factors, and the volume of queries processed daily [1], [8], [16], [21], [28], [35].

Stage 2, Design, uses insights from the assessment to create a modular blueprint for AI operations. Tasks are split into segments that match specific goals, allowing each segment to leverage the correct mix of cloud or local infrastructure. In practice, smaller tasks might be directed to local systems if privacy or bandwidth are priorities, while large-scale text generation might use cloud services for speed [8], [17], [20], [23], [31].

Stage 3, Implementation, translates the design into tangible processes and guidelines. Operators enforce the use of caching protocols, integrate context summarization, and maintain a repository of prompt templates to ensure messages remain concise. Furthermore, administrators configure hybrid deployment to handle sensitive data internally, while harnessing cloud elasticity for non-critical workloads. Implementation also defines cost-monitoring practices, such as real-time dashboards that display token usage for immediate corrective measures [3], [16], [20], [21], [31], [32].

Stage 4, Evaluation, applies continuous monitoring to check if token usage targets and performance benchmarks are being met. This final phase encourages iterative improvements by gathering feedback from users, tracking cost patterns, and revising strategies as needed. When token usage spikes unexpectedly, immediate reviews might highlight poorly structured prompts or unoptimized caching, leading to updated policies or revised frameworks. By viewing TEA-UF as a cycle rather than a static model, organizations retain flexibility in adjusting their AI deployments [1], [8], [16], [20], [21], [28], [34], [35].

4.3 Benefits of TEA-UF

A major advantage of TEA-UF lies in its potential to generate substantial cost savings for organizations aiming to scale AI initiatives. By employing optimized strategies, many institutions can noticeably curtail token-related spending and avoid the unpredictable bills often linked to large-scale LLM usage. This adaptive approach assists management teams in forecasting expenses more reliably, ensuring that growth does not compromise budget stability [8], [16], [20], [27], [31], [37], [39].

Another benefit involves reducing the risks associated with vendor lock-in, a concern for those who rely heavily on proprietary APIs. TEA-UF's hybrid stance allows seamless pivots between platforms or providers if services become expensive or fail to meet evolving requirements. This agility positions organizations to negotiate better deals and keep pace with new technologies without committing to one exclusive ecosystem [3], [8], [20], [21], [27], [31], [32], [37], [39].

A pivotal aspect of TEA-UF involves curbing redundant computations, which not only reduces expenses but also translates into tangible energy savings. By channeling simpler workloads to local systems, enterprises cut back on persistent cloud resource demands, thereby decreasing overall power usage [2], [6].

Additionally, TEA-UF promotes scalability for businesses of varying sizes, from modest startups handling small user bases to major enterprises processing millions of requests. Teams can configure TEA-UF principles in ways that match their immediate needs, then expand those principles when usage volumes rise. This tailored approach prevents sudden strains on infrastructure and finances by ensuring that token consumption remains controlled [8], [16], [20], [27], [37], [39].

Overall, TEA-UF brings clarity to the complexities of token usage by merging cost optimization, flexibility, and secure deployment model. Many case studies emphasize the difficulty in maintaining AI solutions once costs accelerate, especially

when organizations must update models or scale to new markets. By tackling these concerns at the framework level, TEA-UF bridges the gap between immediate implementation hurdles and long-term growth strategies [3], [8], [20], [21], [27], [31], [32], [37], [39].

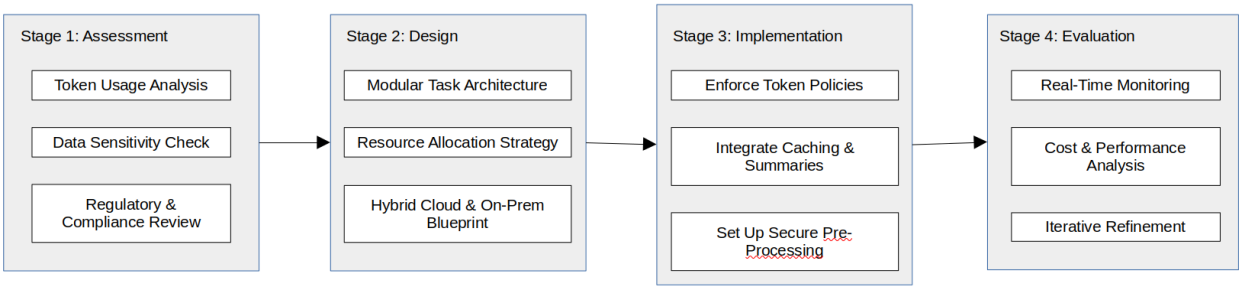


Figure 4 illustrates TEA-UF lifecycle

5. RESULTS AND DISCUSSION

5.1 Interpretation and Discussion

TEA-UF establishes a structured way for organizations to manage LLM usage, particularly when token-based costs threaten budget stability [37], [39]. Each phase of TEA-UF targets common pitfalls in AI adoption, such as excessive token consumption and limited flexibility when vendor lock-in occurs [15]. Many organizations struggle to predict expenses because token surcharges escalate quickly once query volumes rise [8], [16], [20], [27], [31], [37], [39]. TEA-UF minimizes these risks by encouraging prompt caching, modular system design, and hybrid deployments, thereby creating an adaptive foundation for scalable AI solutions [16], [17], [20], [21], [23], [31]. Token budgeting and monitoring, aided by predictive dashboards, forecast usage spikes accurately [1], [8], [16], [20], [21], [28], [34], [35]. Meanwhile, federated learning and edge-based models expand AI reach, lowering cloud reliance [41]. Such strategies future-proof organizations against cost escalations and vendor restrictions, enabling flexible scaling without sacrificing efficiency. These approaches enhance trust in adoption. Such a balanced framework also aligns with the call for adaptive budget controls in mission-critical systems, ensuring that cost is neither a barrier nor an afterthought.

5.2 Hypothetical Applications: TEA-UF in Diverse Contexts

The effectiveness of TEA-UF becomes evident when applying back-of-the-envelope estimates to two key sectors: retail and healthcare [26], [29]. Consider a mid-sized retail operation handling 60,000 daily user queries, each averaging 100 tokens, for a total of 6 million tokens per day. By caching common inquiries and using local inference for 40% of queries, the daily token load on the cloud drops from 6 million to around 3.6 million tokens, potentially cutting monthly token spending from 1.8 million USD to near 1.1 million USD if the base rate is 0.01 USD per token. Furthermore, about 20% of those cached tokens might repeat brand descriptions, sizing charts, and promotional ads, thereby reducing repeated processing costs. This modular approach also deters vendor lock-in, since local systems can handle a significant share of processing.

A similar scenario arises in healthcare, where a hospital receives 25,000 patient record requests daily, each involving 200 tokens for detailed summaries, resulting in 5 million tokens per day. A hybrid strategy might route 50% of queries—especially those carrying sensitive personal details—to local deployments, cutting cloud-based tokens from 5 million to roughly 2.5 million. The local portion might be allocated for shorter summaries, while specialized analyses requiring advanced text generation go to the cloud, thereby lowering monthly token expenditure from an estimated 1.5 million USD to near 750,000 USD if the base rate is 0.01 USD per token. Privacy mandates, such as patient data protected by HIPAA, often involve around 10% of daily tokens, which leads to more secure operations and eases regulatory compliance. This combination illustrates how TEA-UF promotes cost-effective scaling and adherence to confidentiality requirements.

In both the retail and healthcare scenarios, local handling of recurring tasks and selective cloud usage reduce monthly token fees, diminish redundant data processing, and help maintain energy-aware operations, aligning with industry standards for sustainable AI adoption [2], [6].

By showcasing flexible deployments across retail and healthcare, TEA-UF underscores its capacity to unify cost savings, operational agility, and data protection in a single coherent framework. Leaders in different industries can customize token allocation or caching policies, encouraging further research into how TEA-UF might be enhanced for emerging AI use cases.

5.3 Policy Implication at Organization Level and National Level

A policy-driven strategy prompts teams to adopt standardized practices for token control, enhancing clarity and accountability within AI operations. Many organizations set local inference thresholds for sensitive workloads, limiting the tokens transmitted to external providers when privacy rules demand caution. On a broader scale, governments could require AI vendors to reveal pricing structures more thoroughly, empowering businesses to secure fair deals or migrate to new providers without extensive reintegration challenges. Enforcing interoperability across AI platforms further elevates operational efficiency and nurtures competitive markets. Standardized protocols for token usage, data formats, and deployment models allow seamless transitions between systems without significant overhead. At the same time, national authorities could promote shared technical standards that minimize vendor lock-in, enhance cost visibility, and ensure a level playing field. This strategy fuels innovation by letting firms explore multiple tools while contributing to sustainable economic growth and open industry dynamics. As a result, providers must compete on cost-effectiveness and technical quality, creating a healthier environment for all stakeholders. Regulatory bodies also benefit from clearer usage metrics, guiding them toward policies that protect the interests of both enterprises and end consumers.

6. CONCLUSION AND RECOMMENDATIONS

A well-designed approach to LLM adoption can transform organizational workflows, yet the TEA-UF remains conceptual and relies on theoretical modeling that awaits validation in real-world contexts [18], [33]. Each principle of TEA-UF draws upon back-of-the-envelope approximations rather than detailed empirical studies, which underscores the need for pragmatic evaluations in specific sectors such as healthcare, finance, or retail [26], [29]. These industries often involve unique data volumes, compliance mandates, and usage spikes that can influence how token usage patterns unfold in practice. Practical trials would clarify how TEA-UF performs under varied operational pressures, guiding practitioners on ways to balance privacy needs with cost objectives. Further research might also examine how unexpected factors, like sudden market changes or regulatory revisions, affect the viability of TEA-UF in live settings [18], [26], [29], [33].

Additionally, the reduction in token overhead yields notable environmental advantages. Enterprises that optimize workloads using TEA-UF find that fewer superfluous computations not only safeguard budgets but also foster energy efficiency, strengthening the case for immediate and widespread adoption [2], [6].

Future empirical studies should systematically gauge how TEA-UF handles token fluctuations, vendor lock-in avoidance, and policy compliance across different scales of enterprise deployment [24]. Comparative research might analyze how organizations fare before and after adopting TEA-UF, highlighting shifts in overall token expenses and the extent of resource reallocation [13]. Cross-industry adaptations could explore whether certain design elements, like hybrid deployments or caching strategies, yield greater benefits in manufacturing versus customer-facing services [5], [9], [10]. Explore whether this adaptation affects growth processes in small and medium-sized enterprises (SMEs) [14]. Another fruitful avenue involves refining token pricing models, possibly by collaborating with LLM service providers to propose tiered rates or specialized packages that incentivize efficient usage [4]. In-depth simulations and pilot programs can capture nuances in data sensitivity, infrastructure constraints, and regulatory demands to sharpen best practices for TEA-UF [38].

Organizations worldwide should consider TEA-UF a guiding framework when planning AI initiatives, especially to avoid sudden budget hikes and operational bottlenecks [37], [39]. By implementing token usage dashboards, subdividing tasks for local or cloud inference, and adopting advanced summarization approaches, companies stand to save both funds and valuable staff time [1], [8], [16], [20], [21], [28], [34], [35]. Decision-makers might also align these policies with national regulations, ensuring that token consumption is transparent and easily audited [25]. At a policy level, government bodies could promote interoperability standards among LLM vendors, encouraging fair pricing and spurring healthy competition that benefits end-users [11], [22]. This macro-level encouragement of uniform protocols can establish stable AI ecosystems that support innovation, lower financial risk, and maintain robust data security.

In summary, TEA-UF offers a holistic framework for sustainable AI adoption, balancing agile deployments with managed expenses and clear protocols. By bridging cost forecasting, privacy mandates, and scalable architectures, TEA-UF stands ready to facilitate global AI-driven growth [1], [8], [16], [20], [21], [28], [34], [35]. A collaborative effort among corporations, policymakers, and researchers would help refine the framework's components, improve token pricing approaches, and deepen trust in AI solutions. Through ongoing refinements and policy support, TEA-UF can significantly impact organizations, leading them toward long-term efficiency, stronger market performance, and ethical stewardship of AI technology.

7. DECLARATION OF INTERESTS

The author declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

8. DECLARATION OF FUNDING

The author declare that they did not receive any funding for this paper.

9. STATEMENT GENERATIVE AI

During the development of this manuscript the author used AI tools in order to assist with:

1. Structuring ideas and thoughts more coherently.
2. Checking for grammar, spelling, and sentence clarity.
3. Rephrasing to improve readability and flow.

While these tools helped enhance clarity and structure, all intellectual contributions, theoretical frameworks, arguments, and analyses are entirely by the author. All sources are properly cited. After using the AI tools, the author reviewed and edited the content as needed and takes full responsibility for the content of the published article.

10. REFERENCES

[1] Alarcia, R. M. G., & Golkar, A. 2024. Optimizing token usage on large language model conversations using the design structure matrix. arXiv, 2410.00749. https://doi.org/10.35199/dsm2024.08

[2] Argerich, M. F., & Patiño-Martínez, M. 2024. Measuring and improving the energy efficiency of large language models inference. *IEEE Access*, *12*, 80194–80207. https://doi.org/10.1109/ACCESS.2024.3409745

[3] Asimiyu, Z. 2023. Scalable Inference Systems for Real-Time LLM Integration. Retrieved January 5, 2025. https://www.researchgate.net/profile/Zainab-Asimiyu/publication/387223134_Scalable_Inference_Systems_for_Real-Time_LLM_Integration/links/6764805ae74ca64e1f1ebc27/Scalable-Inference-Systems-for-Real-Time-LLM-Integration.pdf

[4] Aurangzeb, K., Aslam, S., Mohsin, S. M., & Alhussein, M. 2021. A fair pricing mechanism in smart grids for low energy consumption users. IEEE Access, 9, 22035-22044. https://doi.org/10.1109/ACCESS.2021.3056035

[5] Bader, K. 2013. How to benefit from cross-industry innovation? A best practice case. International journal of innovation management, 17(06), 1340018. https://doi.org/10.1142/S1363919613400185

[6] Bai, G., Chai, Z., Ling, C., Wang, S., Lu, J., Zhang, N., Shi, T., Yu, Z., Zhu, M., Zhang, Y., Song, X., Yang, C.,

Cheng, Y., & Zhao, L. 2024. *Beyond efficiency: A systematic survey of resource-efficient large language models* Version4. arXiv:2401.00625 cs.LGcs.LGcs.LG. https://doi.org/10.48550/arXiv.2401.00625

[7] Balasubramaniam, S., Kadry, S., Prasanth, A., & Dhanaraj, R. K. (Eds.). 2024. Generative AI and LLMs: Natural language processing and generative adversarial networks. De Gruyter. https://doi.org/10.1515/9783111425078

[8] Barkan, G. 2024. The Emerging Economy of LLMs—Part 2. Wix Engineering.

[9] Behne, A., Heinrich Beinke, J., & Teuteberg, F. 2021. A framework for cross-industry innovation: Transferring technologies between industries. International Journal of Innovation and Technology Management, 18(03), 2150011. https://doi.org/10.1142/S0219877021500115

[10] Carmona-Lavado, A., Gimenez-Fernandez, E. M., Vlaisavljevic, V., & Cabello-Medina, C. 2023. Cross-industry innovation: A systematic literature review. Technovation, 124, 102743. https://doi.org/10.1016/j.technovation.2023.102743

[11] Cerf, V. G. 2024. Thoughts on AI interoperability. Communications of the ACM, 67(4), 5. https://doi.org/10.1145/3649475

[12] Chui, M., Hazan, E., Roberts, R., Singla, A., Smaje, K., Sukharevsky, A., Yee, L., & Zemmel, R. 2023. The economic potential of generative AI: The next productivity frontier. McKinsey & Company. Retrieved January 5, 2025. https://www.mckinsey.de/~/media/mckinsey/locations/europe%20and%20middle%20east/deutschland/news/presse/2023/2023-06-14%20mgi%20genai%20report%2023/the-economic-potential-of-generative-ai-the-next-productivity-frontier-vf.pdf

[13] Esser, F., & Vliegenthart, R. 2017. Comparative research methods. The international encyclopedia of communication research methods, 1-22. https://doi.org/10.1002/9781118901731.iecrm0035

[14] Faeroevik, K. H., & Maehle, N. 2022. The outcomes of cross-industry innovation for small and medium sized enterprises. Journal of Small Business & Entrepreneurship, 36(4), 675–704. https://doi.org/10.1080/08276331.2022.2070711

[15] George, A. S. 2024. The cloud comedown: Understanding the emerging trend of cloud exit strategies. Partners Universal International Innovation Journal, 2(5), 1–32. https://doi.org/10.5281/zenodo.13993933

[16] Ghosh, S. 2024. Unlocking the Mystery of Tokens in Large Language Models (LLMs). Medium.

[17] Gim, I., Chen, G., Lee, S., Sarda, N., Khandelwal, A., & Zhong, L. 2024. Prompt cache: Modular attention reuse for low-latency inference. Proceedings of Machine Learning and Systems 6 (MLSys 2024) Conference.

[18] Graebner, C. 2018. How to relate models to reality? An epistemological framework for the validation and verification of computational models. Journal of Artificial Societies and Social Simulation, 21(3), 8. http://dx.doi.org/10.18564/jasss.3772

[19] Heyvaert, M., Maes, B., & Onghena, P. 2013. Mixed methods research synthesis: Definition, framework, and potential. Quality & Quantity, 47(2), 659–676. https://doi.org/10.1007/s11135-011-9538-6

[20] Kuzminykh, N. 2024. Calculating LLM Token Counts: A Practical Guide. Winder AI.

[21] Lawlor, A. 2024. Is Building Your Own LLM Worth It? Probably Not. Aptaria.

[22] Lehmann, R. 2024. Towards interoperability of APIs - an LLM-based approach. Proceedings of the 25th International Middleware Conference: Demos, Posters and Doctoral Symposium, 29-30. Association for Computing Machinery. https://doi.org/10.1145/3704440.3704788

[23] Liu, S., Biswal, A., Cheng, A., Mo, X., Cao, S., Gonzalez, J. E., Stoica, I., & Zaharia, M. 2024. Optimizing LLM queries in relational workloads. arXiv, 2403.05821. https://doi.org/10.48550/arXiv.2403.05821

[24] MacDonell, S., Shepperd, M., Kitchenham, B., & Mendes, E. 2010. How reliable are systematic reviews in empirical software engineering? IEEE Transactions on Software Engineering, 36(5), 676-687. https://doi.org/10.1109/TSE.2010.28

[25] Mökander, J., Schuett, J., Kirk, H. R., & Floridi, L. 2023. Auditing large language models: a three-layered approach. AI and Ethics, 1-31. https://doi.org/10.1007/s43681-023-00289-2

[26] Murillo-Gonzalez, G., & Burkholder, E. W. 2022. What decisions do experts make when doing back-of-the-envelope calculations? Physical Review Physics Education Research, 18(1), 010125. https://doi.org/10.1103/PhysRevPhysEducRes.18.010125

[27] Nagarajan, R., Kondo, M., Salas, F., Sezgin, E., Yao, Y., Klotzman, V., Godambe, S. A., Khan, N., Limon, A., Stephenson, G., Taraman, S., Walton, N., Ehwerhemuepha, L., Pandit, J., Pandita, D., Weiss, M., Golden, C., Gold, A., Henderson, J., Shippy, A., Celi, L. A., Hogan, W. R., Oermann, E. K., Sanger, T., & Martel, S. 2024. Economics and equity of large language models: Health care perspective. Journal of Medical Internet Research, 26, e64226. https://doi.org/10.2196/64226

[28] Nobre, R., Roberts, J., Donovan, L., Callahan, T., & Sinclair, G. 2024. Optimizing token context utilization for efficient inference in large language models. Authorea. https://doi.org/10.22541/au.172953932.23206891/v1

[29] Paritosh, P. K., & Forbus, K. D. 2003. Qualitative modeling and similarity in back of the envelope reasoning. Proceedings of the 26th Annual Conference of the Cognitive Science Society, Boston. Retrieved January 5, 2025. https://www.qrg.northwestern.edu/papers/Files/BotEProposalCogSci03.pdf

[30] Rashid, A. B., & Kausik, M. A. K. 2024. AI revolutionizing industries worldwide: A comprehensive overview of its diverse applications. Hybrid Advances, 7, 100277. https://doi.org/10.1016/j.hybadv.2024.100277

[31] Rottmann, K. 2024. GenAI Disenchanted: Tokenization the Currency of LLMs. LinkedIn.

[32] Saputra, A., Suryani, E., & Rakhmawati, N. A. 2024. Secure and scalable LLM-based recommendation systems: An MLOps and security by design. Proceedings of the 2024 IEEE International Symposium on Consumer Technology (ISCT), 623-629. https://doi.org/10.1109/ISCT62336.2024.10791207

[33] Sargent, R.G. 1979. Validation of Simulation Models, Proceedings of the 1979 Winter Simulation Conference, edited by Highland, H.J., et al., San Diego, California, pp. 497-503.

[34] Shang, Y., Cai, M., Xu, B., Lee, Y. J., & Yan, Y. 2024. LLaVA-PruMerge: Adaptive token reduction for efficient large multimodal models. arXiv, 2403.15388. https://doi.org/10.48550/arXiv.2403.15388

[35] Shekhar, S., Dubey, T., Mukherjee, K., Saxena, A., Tyagi, A., & Kotla, N. 2024. Towards optimizing the costs of LLM usage. arXiv, 2402.01742. https://doi.org/10.48550/arXiv.2402.01742

[36] Sivakumar, S. 2024. Performance optimization of large language models (LLMs) in web applications. International Journal of Trend in Scientific Research and Development, 8(1), 1077–1096.

[37] Subramanian, S. 2024. Large Language Model-Based Solutions: How to Deliver Value with Cost-Effective Generative AI Applications. John Wiley & Sons.

[38] Walsh-Bailey, C., Palazzo, L. G., Jones, S. M., Mettert, K. D., Powell, B. J., Wiltsey Stirman, S., Lyon, A. R., Rohde, P., & Lewis, C. C. 2021. A pilot study comparing tools for tracking implementation strategies and treatment adaptations. Implementation Research and Practice, 2. https://doi.org/10.1177/26334895211016028

[39] Wang, J., Jain, S., Zhang, D., Ray, B., Kumar, V., & Athiwaratkun, B. 2024. Reasoning in token economies: Budget-aware evaluation of LLM reasoning strategies. arXiv, 2406.06461v3. https://doi.org/10.48550/arXiv.2406.06461

[40] Zhou, H., Hu, C., Yuan, Y., Cui, Y., Jin, Y., Chen, C., Wu, H., Yuan, D., Jiang, L., Wu, D., Liu, X., Zhang, C., Wang, X., & Liu, J. 2024. Large language model (LLM) for telecommunications: A comprehensive survey on principles, key techniques, and opportunities. arXiv, 2405.10825v2. https://doi.org/10.48550/arXiv.2405.10825

[41] Zhou, J., Pal, S., Dong, C., & Wang, K. 2024. Enhancing quality of service through federated learning in edge-cloud architecture. Ad Hoc Networks, 156, 103430. https://doi.org/10.1016/j.adhoc.2024.103430