Viabilidad de Large Language Models ejecutados en dispositivos on-edge para asistencia administrativa en el sector agroalimentario

Francisco Javier González Ontañón[1], Segundo Autor[2,3], and Tercer Autor[3]

Unizar, Zaragoza, Spain
primer@example.com

Resumen Los Large Language Models (LLMs) han demostrado un rendimiento sobresaliente en tareas de procesamiento del lenguaje natural cuando se ejecutan en infraestructuras cloud. Sin embargo, su dependencia de conectividad permanente, los requisitos de latencia y las preocupaciones relacionadas con la privacidad limitan su adopción en contextos profesionales que operan en entornos con recursos restringidos. En este trabajo se presenta una investigación en curso que analiza la viabilidad de ejecutar LLMs directamente en dispositivos on-edge, como teléfonos móviles, para asistir en tareas administrativas del sector agroalimentario. El artículo se centra en un caso de uso real: el apoyo al rellenado de documentación agrícola oficial en entornos rurales con conectividad limitada. Se discuten las restricciones impuestas por el hardware disponible, así como las limitaciones actuales de las herramientas de ingeniería del software para desplegar y evaluar este tipo de soluciones. Más que proponer una solución cerrada, el trabajo identifica problemas abiertos relacionados con el diseño, la evaluación y la integración de LLMs on-edge en aplicaciones reales, con el objetivo de fomentar la discusión y servir como punto de partida para futuras investigaciones y colaboraciones.

Keywords: LLMs on-edge · ingeniería del software · dispositivos móviles · sector agroalimentario

1. Introducción

Los Large Language Models (LLMs) están impulsando una nueva generación de sistemas capaces de comprender y generar lenguaje natural, pero su despliegue práctico sigue estando condicionado por requisitos computacionales y de memoria elevados. En particular, la ejecución en dispositivos móviles plantea un conjunto de restricciones específicas (memoria disponible, latencia, longitud de entrada y motor de inferencia), cuya caracterización empírica en escenarios reales se ha abordado mediante estudios de medida y análisis comparativo [12,27]. Estos trabajos muestran que, aunque la ejecución local es posible en determinadas condiciones, el rendimiento y la calidad dependen fuertemente del tamaño del modelo y de las optimizaciones aplicadas.

En paralelo, la comunidad está convergiendo hacia el paradigma *on-edge/on-device* como alternativa (o complemento) a arquitecturas basadas en la nube, motivada por la reducción de latencia, la resiliencia ante conectividad intermitente y la necesidad de mantener los datos cerca de su origen. Diversas revisiones recientes sintetizan técnicas y desafíos para habilitar inferencia eficiente en dispositivos con recursos limitados, destacando estrategias basadas en modelos más pequeños, compresión y optimización de inferencia [26,24,6]. No obstante, estas mismas revisiones subrayan problemas abiertos de ingeniería relacionados con evaluación reproducible, integración en aplicaciones reales y consideraciones de seguridad y confiabilidad.

Para aproximar los LLMs a escenarios restringidos se han propuesto familias de técnicas de compresión y ajuste eficiente, incluyendo cuantización, poda, distilación y enfoques de *parameter-efficient fine-tuning* (PEFT), junto con taxonomías sistemáticas que evidencian la falta de marcos unificados que integren estas decisiones de diseño en pipelines de despliegue completos [11]. En el contexto móvil, se exploran tanto arquitecturas sub-billion optimizadas para uso local [15] como técnicas de cuantización específicamente orientadas a hardware móvil [23]. Además, existen implementaciones tempranas que materializan estos enfoques en prototipos de aplicaciones Android on-device, aportando evidencias prácticas sobre ventajas y limitaciones del despliegue local [**?**].

En este contexto, el presente artículo en curso investiga la viabilidad de aplicar LLMs ejecutados completamente en el dispositivo a un escenario real del sector agroalimentario, centrado en asistencia para tareas administrativas y documentación oficial en entornos con conectividad limitada. El objetivo es identificar retos de ingeniería del software (diseño, integración, evaluación y mantenimiento) y delimitar problemas abiertos que orienten futuras colaboraciones y líneas de investigación.

2. Caso de uso y motivación

El sector agroalimentario requiere el cumplimiento de una amplia variedad de obligaciones administrativas, como el registro de labores agrícolas, tratamientos fitosanitarios o inspecciones técnicas. Estas tareas se realizan habitualmente en el propio campo, utilizando dispositivos móviles y, en muchos casos, sin acceso fiable a Internet.

La posibilidad de contar con un asistente basado en lenguaje natural que funcione completamente en el dispositivo podría reducir errores, agilizar la introducción de datos y mejorar el cumplimiento normativo. Sin embargo, las soluciones actuales suelen depender de servicios cloud, lo que limita su aplicabilidad en este contexto y plantea problemas adicionales de privacidad y soberanía del dato.

Este escenario permite identificar de forma clara las tensiones entre las capacidades de los LLMs y las restricciones reales del entorno de ejecución.

3. Problemas abiertos y retos de ingeniería

La ejecución de LLMs en dispositivos on-edge introduce una serie de problemas abiertos que aún no están adecuadamente resueltos por las herramientas existentes. Entre ellos destacan la selección y adaptación de modelos de tamaño reducido, la gestión eficiente de memoria y energía, y la evaluación de la calidad de las respuestas en condiciones de recursos limitados.

Desde el punto de vista de la ingeniería del software, se observa una falta de metodologías estandarizadas para diseñar, desplegar y evaluar aplicaciones basadas en LLMs on-edge. Los benchmarks disponibles suelen centrarse en métricas aisladas y no reflejan adecuadamente escenarios de uso reales. Asimismo, la integración de estos modelos en aplicaciones móviles plantea desafíos adicionales relacionados con mantenimiento, actualización y depuración.

Estos problemas sugieren la necesidad de enfoques específicos que tengan en cuenta tanto las restricciones técnicas como los requisitos del dominio.

4. Metodología y estado actual del trabajo

El trabajo se encuentra en una fase inicial de desarrollo. Actualmente se está llevando a cabo una revisión del estado del arte y una captura de requisitos mediante cuestionarios dirigidos a profesionales del sector agroalimentario. A partir de estos requisitos se prevé el desarrollo de un prototipo experimental que permita evaluar distintas configuraciones de modelos y técnicas de optimización.

La evaluación se realizará mediante benchmarks específicos del dominio, considerando métricas como latencia, uso de memoria y adecuación funcional. El objetivo no es ofrecer una solución definitiva, sino identificar patrones, limitaciones y oportunidades de mejora que puedan orientar futuras investigaciones.

5. Conclusiones preliminares y líneas futuras

Este trabajo pone de manifiesto que la adopción de LLMs on-edge en aplicaciones reales plantea desafíos que van más allá del rendimiento del modelo, afectando directamente a decisiones de diseño y a la disponibilidad de herramientas de ingeniería adecuadas. El caso de uso presentado evidencia la necesidad de enfoques específicos para dominios con restricciones severas de recursos y alta sensibilidad de los datos.

Como trabajo en curso, este artículo pretende fomentar la discusión en la comunidad de Ingeniería del Software y Bases de Datos, y servir como punto de partida para futuras colaboraciones orientadas a cerrar la brecha entre los avances en modelos de lenguaje y su aplicación práctica en entornos on-edge.

Agradecimientos Agradecimientos

Declaración de intereses El autor declara que no tiene ningún interés financiero ni relación personal que pudiera influir en el trabajo descrito en este artículo.

Referencias

1. Alipio, M., Bures, M.: The Role of Large Language Models in Designing Reliable Networks for Internet of Things: A Short Review of Most Recent Developments. IEEE Access **13**, 168527–168545 (2025). https://doi.org/10.1109/ACCESS.2025.3614246, https://ieeexplore.ieee.org/document/11179972/
2. Amini, H., Mia, M.J., Saadati, Y., Imteaj, A., Nabavirazavi, S., Thakker, U., Hossain, M.Z., Fime, A.A., Iyengar, S.S.: Distributed LLMs and Multimodal Large Language Models: A Survey on Advances, Challenges, and Future Directions (Mar 2025). https://doi.org/10.48550/arXiv.2503.16585, http://arxiv.org/abs/2503.16585, arXiv:2503.16585 [cs]
3. Baccour, E., Erbad, A., Mohamed, A., Hamdi, M., Guizani, M.: Active Prompt Caching in Edge Networks for Generative AI and LLMs: An RL-Based Approach. In: 2025 IEEE Wireless Communications and Networking Conference (WCNC). pp. 01–07. IEEE, Milan, Italy (Mar 2025). https://doi.org/10.1109/WCNC61545.2025.10978306, https://ieeexplore.ieee.org/document/10978306/
4. Chai, Y., Kwen, M., Brooks, D., Wei, G.Y.: FlexQuant: Elastic Quantization Framework for Locally Hosted LLM on Edge Devices (Jan 2025). https://doi.org/10.48550/arXiv.2501.07139, http://arxiv.org/abs/2501.07139, arXiv:2501.07139 [cs]
5. Fareed, M., Fatima, M., Uddin, J., Ahmed, A., Sattar, M.A.: A systematic review of ethical considerations of large language models in healthcare and medicine. Frontiers in Digital Health **7**, 1653631 (Sep 2025). https://doi.org/10.3389/fdgth.2025.1653631, https://www.frontiersin.org/articles/10.3389/fdgth.2025.1653631/full
6. Friha, O., Amine Ferrag, M., Kantarci, B., Cakmak, B., Ozgun, A., Ghoualmi-Zine, N.: LLM-Based Edge Intelligence: A Comprehensive Survey on Architectures, Applications, Security and Trustworthiness. IEEE Open Journal of the Communications Society **5**, 5799–5856 (2024). https://doi.org/10.1109/OJCOMS.2024.3456549, https://ieeexplore.ieee.org/document/10669603/
7. Golec, M., Khamayseh, Y., Melhem, S.B., Alwarafy, A.: LLM-Driven APT Detection for 6G Wireless Networks: A Systematic Review and Taxonomy. IEEE Access **13**, 145271–145288 (2025). https://doi.org/10.1109/ACCESS.2025.3595665, https://ieeexplore.ieee.org/document/11112774/
8. Hadish, S., Bojković, V., Aloqaily, M., Guizani, M.: Language Models at the Edge: A Survey on Techniques, Challenges, and Applications. In: 2024 2nd International Conference on Foundation and Large Language Models (FLLM). pp. 262–271. IEEE, Dubai, United Arab Emirates (Nov 2024). https://doi.org/10.1109/FLLM63129.2024.10852473, https://ieeexplore.ieee.org/document/10852473/
9. Husom, E.J., Goknil, A., Astekin, M., Shar, L.K., Kåsen, A., Sen, S., Mithassel, B.A., Soylu, A.: Sustainable LLM Inference for Edge AI: Evaluating Quantized LLMs for Energy Efficiency, Output Accuracy, and Inference Latency (Apr 2025). https://doi.org/10.48550/arXiv.2504.03360, http://arxiv.org/abs/2504.03360, arXiv:2504.03360 [cs]
10. Khatiwada, K., Hopper, J., Cheatham, J., Joshi, A., Baidya, S.: Large Language Models in the IoT Ecosystem – A Survey on Security Challenges and Applications (May 2025). https://doi.org/10.48550/arXiv.2505.17586, http://arxiv.org/abs/2505.17586, arXiv:2505.17586 [cs]

11. Kim, G.I., Hwang, S., Jang, B.: Efficient Compressing and Tuning Methods for Large Language Models: A Systematic Literature Review. ACM Comput. Surv. **57**(10), 253:1–253:39 (May 2025). https://doi.org/10.1145/3728636, https://dl.acm.org/doi/10.1145/3728636
12. Li, X., Lu, Z., Cai, D., Ma, X., Xu, M.: Large Language Models on Mobile Devices: Measurements, Analysis, and Insights. In: Proceedings of the Workshop on Edge and Mobile Foundation Models. pp. 1–6. ACM, Minato-ku Tokyo Japan (Jun 2024). https://doi.org/10.1145/3662006.3662059, https://dl.acm.org/doi/10.1145/3662006.3662059
13. Lin, Z., Qu, G., Chen, Q., Chen, X., Chen, Z., Huang, K.: Pushing Large Language Models to the 6G Edge: Vision, Challenges, and Opportunities (Jun 2025). https://doi.org/10.48550/arXiv.2309.16739, http://arxiv.org/abs/2309.16739, arXiv:2309.16739 [cs]
14. Liu, H.I., Galindo, M., Xie, H., Wong, L.K., Shuai, H.H., Li, Y.H., Cheng, W.H.: Lightweight Deep Learning for Resource-Constrained Environments: A Survey. ACM Comput. Surv. **56**(10), 267:1–267:42 (Jun 2024). https://doi.org/10.1145/3657282, https://dl.acm.org/doi/10.1145/3657282
15. Liu, Z., Zhao, C., Iandola, F., Lai, C., Tian, Y., Fedorov, I., Xiong, Y., Chang, E., Shi, Y., Krishnamoorthi, R., Lai, L., Chandra, V.: MobileLLM: Optimizing Sub-billion Parameter Language Models for On-Device Use Cases (Jun 2024). https://doi.org/10.48550/arXiv.2402.14905, http://arxiv.org/abs/2402.14905, arXiv:2402.14905 [cs]
16. Luo, H., Liu, Y., Zhang, R., Wang, J., Sun, G., Niyato, D., Yu, H., Xiong, Z., Wang, X., Shen, X.: Toward Edge General Intelligence with Multiple-Large Language Model (Multi-LLM): Architecture, Trust, and Orchestration (Jul 2025). https://doi.org/10.48550/arXiv.2507.00672, http://arxiv.org/abs/2507.00672, arXiv:2507.00672 [cs]
17. Northeastern University, Boston, MA, Kompally, V.S.: A Review of Large Language Models in Edge Computing: Applications, Challenges, Benefits, and Deployment Strategies. International journal of data science and machine learning **05**(01), 300–322 (Jun 2025). https://doi.org/10.55640/ijdsml-05-01-25, https://www.academicpublishers.org/journals/index.php/ijdsml/article/view/5302/6234
18. Picano, B., Hoang, D.T., Nguyen, D.N.: A Matching Game for LLM Layer Deployment in Heterogeneous Edge Networks. IEEE Open Journal of the Communications Society **6**, 3795–3805 (2025). https://doi.org/10.1109/OJCOMS.2025.3561605, https://ieeexplore.ieee.org/document/10966456/
19. Pozi, M.S.M., Sato, Y.: A data-augmented model routing framework for efficient LLM deployment in edge–cloud environments. The Journal of Supercomputing **81**(17), 1573 (Nov 2025). https://doi.org/10.1007/s11227-025-08034-8, https://link.springer.com/10.1007/s11227-025-08034-8
20. Qin, R., Liu, D., Xu, C., Yan, Z., Tan, Z., Jia, Z., Nassereldine, A., Li, J., Jiang, M., Abbasi, A., Xiong, J., Shi, Y.: Empirical Guidelines for Deploying LLMs onto Resource-constrained Edge Devices (Oct 2024). https://doi.org/10.48550/arXiv.2406.03777, http://arxiv.org/abs/2406.03777, arXiv:2406.03777 [cs]
21. Sammangi, H.: Harnessing Generative AI and Large Language Models for Revolutionizing Cybersecurity in the Internet of Things: Ethical and Privacy Implications
22. Sun, K., Wang, X., Miao, X., Zhao, Q.: A review of AI edge devices and lightweight CNN and LLM deployment. Neurocomputing **614**, 128791 (Jan 2025). https://doi.org/10.1016/j.neucom.2024.128791, https://linkinghub.elsevier.com/retrieve/pii/S0925231224015625

23. Tan, F., Lee, R., Dudziak, , Hu, S.X., Bhattacharya, S., Hospedales, T., Tzimiropoulos, G., Martinez, B.: MobileQuant: Mobile-friendly Quantization for On-device Language Models (Oct 2024). https://doi.org/10.48550/arXiv.2408.13933, http://arxiv.org/abs/2408.13933, arXiv:2408.13933 [cs]
24. Wang, R., Gao, Z., Zhang, L., Yue, S., Gao, Z.: Empowering large language models to edge intelligence: A survey of edge efficient LLMs and techniques. Computer Science Review **57**, 100755 (Aug 2025). https://doi.org/10.1016/j.cosrev.2025.100755, https://linkinghub.elsevier.com/retrieve/pii/S1574013725000310
25. Wang, X., Xu, Z., Sui, X.: Intelligent data analysis in edge computing with large language models: applications, challenges, and future directions. Frontiers in Computer Science **7**, 1538277 (May 2025). https://doi.org/10.3389/fcomp.2025.1538277, https://www.frontiersin.org/articles/10.3389/fcomp.2025.1538277/full
26. Xu, J., Li, Z., Chen, W., Wang, Q., Gao, X., Cai, Q., Ling, Z.: On-Device Language Models: A Comprehensive Review (Sep 2024). https://doi.org/10.48550/arXiv.2409.00088, http://arxiv.org/abs/2409.00088, arXiv:2409.00088 [cs]
27. Yan, X., Ding, Y.: Are We There Yet? A Measurement Study of Efficiency for LLM Applications on Mobile Devices (Mar 2025). https://doi.org/10.48550/arXiv.2504.00002, http://arxiv.org/abs/2504.00002, arXiv:2504.00002 [cs]
28. Yuan, X., Li, H.: LLM-Driven Offloading Decisions for Edge Object Detection in Smart City Deployments. Smart Cities **8**(5), 169 (Oct 2025). https://doi.org/10.3390/smartcities8050169, https://www.mdpi.com/2624-6511/8/5/169
29. Zhang, M., Shen, X., Cao, J., Cui, Z., Jiang, S.: EdgeShard: Efficient LLM Inference via Collaborative Edge Computing. IEEE Internet of Things Journal **12**(10), 13119–13131 (May 2025). https://doi.org/10.1109/JIOT.2024.3524255, https://ieeexplore.ieee.org/document/10818760/
30. Zhang, R., He, J., Luo, X., Niyato, D., Kang, J., Xiong, Z., Li, Y., Sikdar, B.: Toward Democratized Generative AI in Next-Generation Mobile Edge Networks (Nov 2024). https://doi.org/10.48550/arXiv.2411.09148, http://arxiv.org/abs/2411.09148, arXiv:2411.09148 [cs]
31. Zheng, Y., Chen, Y., Qian, B., Shi, X., Shu, Y., Chen, J.: A Review on Edge Large Language Models: Design, Execution, and Applications (Feb 2025). https://doi.org/10.48550/arXiv.2410.11845, http://arxiv.org/abs/2410.11845, arXiv:2410.11845 [cs]