LLM ejecutados on-edge en el sector agroalimentario

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

Unizar, Zaragoza, Spain
primer@example.com

Resumen Los Large Language Models (LLMs) han mostrado un alto rendimiento en tareas de procesamiento del lenguaje natural en infraestructuras cloud; sin embargo, su dependencia de conectividad permanente, la latencia y las preocupaciones sobre privacidad limitan su adopción en entornos con recursos restringidos. Este trabajo presenta una investigación en curso que analiza la viabilidad de ejecutar LLMs directamente en dispositivos *on-edge*, con especial atención a escenarios móviles.
El artículo se centra en un caso de uso del sector agroalimentario: la asistencia al rellenado de documentación administrativa en entornos rurales con conectividad limitada. A partir de este escenario, se discuten las restricciones del hardware disponible y las limitaciones actuales de las herramientas de ingeniería del software para diseñar y evaluar aplicaciones basadas en LLMs *on-device*. Más que proponer una solución cerrada, el trabajo identifica problemas abiertos que sirven como base para futuras investigaciones.

Keywords: LLMs · on-edge · on-device · agroalimentario

1. Introducción

Los Large Language Models (LLMs) están impulsando una nueva generación de sistemas capaces de comprender y generar lenguaje natural; sin embargo, su despliegue práctico sigue estando condicionado por sus requisitos computacionales y de memoria. La ejecución en dispositivos móviles plantea un conjunto de restricciones específicas —memoria disponible, latencia, longitud de contexto (*context length*) y motor de inferencia— cuya caracterización empírica en escenarios reales se ha abordado mediante estudios de medida y análisis comparativo [4,9]. Estos trabajos muestran que, aunque la ejecución local es posible bajo determinadas condiciones, el rendimiento y la calidad dependen fuertemente del tamaño del modelo y de las optimizaciones aplicadas.

En paralelo, el uso de soluciones *on-edge*/*on-device* como alternativa a arquitecturas basadas en la nube está creciendo, motivada por la reducción de latencia, la resiliencia frente a conectividad intermitente y la necesidad de mantener los datos cerca de su origen. Diversas revisiones recientes coinciden en que la ejecución de LLMs en dispositivos con recursos limitados es técnicamente

viable, pero todavía presenta importantes retos en términos de ingeniería del software, evaluación reproducible y validación en escenarios reales [8,7,2].

Para aproximar los LLMs a escenarios restringidos se han propuesto de técnicas , 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 completos de despliegue [3]. En el contexto móvil, se exploran tanto arquitecturas *sub-billion* optimizadas para uso local [5] como técnicas de cuantización específicamente orientadas a hardware móvil [6]. Además, existen implementaciones tempranas que materializan estos enfoques en prototipos de aplicaciones Android *on-device*, aportando evidencias prácticas sobre las ventajas y limitaciones del despliegue local [1].

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 la asistencia a tareas administrativas y documentación oficial en entornos con conectividad limitada. El objetivo es identificar retos de ingeniería del software —relativos al diseño, la integración, la evaluación y el 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, muchas de ellas reguladas por normativas nacionales y supranacionales. Estas tareas se realizan habitualmente *in situ*, durante la jornada de trabajo en el campo, utilizando dispositivos móviles como herramienta de registro y consulta. En este contexto, la conectividad a Internet es a menudo limitada o inexistente, lo que dificulta el uso de soluciones basadas exclusivamente en servicios en la nube.

Diversos trabajos señalan que los LLMs ejecutados en el propio dispositivo resultan especialmente útiles en entornos con conectividad limitada y fuertes requisitos de privacidad [8,7]. En particular, ejecutar la inferencia directamente en el dispositivo reduce la latencia, mejora la robustez del sistema y evita que datos sensibles salgan del terminal del usuario. Estos aspectos son especialmente relevantes en dominios donde la información gestionada tiene implicaciones legales y requisitos estrictos de trazabilidad y soberanía del dato.

No obstante, la adopción de LLMs en este tipo de escenarios plantea dificultades por las capacidades de los modelos y las restricciones del entorno de ejecución. Diversos estudios señalan que, aunque la ejecución local es técnicamente viable, el rendimiento, la longitud de contexto y la calidad de las respuestas están fuertemente condicionados por las limitaciones de memoria, energía y capacidad de cómputo de los dispositivos móviles [2,9]. Como consecuencia, muchas de las soluciones existentes optan por arquitecturas híbridas o dependen

de infraestructuras cloud, lo que reduce su aplicabilidad en entornos rurales y reintroduce problemas de privacidad.

Este caso de uso permite analizar de manera concreta y aplicada los desafíos identificados en la literatura sobre LLMs *on-edge*. El sector agroalimentario se presenta así como un dominio representativo para estudiar la viabilidad de asistentes basados en lenguaje natural ejecutados completamente en el dispositivo, así como para identificar problemas abiertos de ingeniería del software relacionados con el diseño, la integración y la validación de estas soluciones en condiciones reales de uso.

3. Problemas abiertos y retos de ingeniería

La ejecución de Large Language Models (LLMs) en dispositivos *on-edge* introduce una serie de problemas abiertos que aún no están adecuadamente resueltos por las herramientas y enfoques actuales. Entre los retos más relevantes se encuentran la selección y adaptación de modelos de tamaño reducido, la gestión eficiente de memoria y consumo energético, y la limitación de la longitud de contexto, factores que impactan directamente en la latencia y en la calidad de las respuestas generadas [8,9].

Desde la perspectiva de la ingeniería del software, revisiones recientes señalan una carencia de metodologías estandarizadas para el diseño, despliegue y evaluación de aplicaciones basadas en LLMs ejecutados en el dispositivo. Las revisiones existentes indican que la mayoría de los estudios se centran en métricas aisladas de rendimiento, sin considerar el ciclo de vida completo del sistema ni escenarios de uso continuado en condiciones reales [2]. Esta falta de enfoques integrales dificulta la comparación entre propuestas y limita la transferencia de resultados a aplicaciones prácticas.

Asimismo, integrar LLMs *on-device* en aplicaciones móviles introduce dificultades adicionales, como el mantenimiento del sistema, la actualización de los modelos y la depuración de errores en dispositivos heterogéneos y con recursos limitados. Estudios recientes señalan que estos aspectos suelen resolverse de forma puntual y específica para cada caso, y que aún no existen marcos de ingeniería consolidados que permitan gestionar de manera sistemática los compromisos entre privacidad, rendimiento y fiabilidad. [7].

En conjunto, estos problemas ponen de manifiesto la necesidad de enfoques que tengan en cuenta tanto las restricciones técnicas *on-edge* como los requisitos funcionales y normativos del dominio de aplicación, abriendo un espacio claro para investigaciones orientadas a casos de uso reales y evaluaciones empíricas reproducibles.

4. Metodología y estado actual del trabajo

El trabajo se encuentra en una fase inicial de desarrollo, centrada en la definición del problema y la delimitación del espacio de diseño. En primer lugar, se está llevando a cabo una revisión sistemática del estado del arte sobre Large

Language Models ejecutados en dispositivos *on-edge*, con el objetivo de identificar enfoques existentes, limitaciones técnicas y problemas abiertos relevantes [8,2]. De forma complementaria, se ha iniciado una fase de captura de requisitos mediante cuestionarios dirigidos a profesionales del sector agroalimentario, con el fin de caracterizar escenarios de uso reales, restricciones operativas y necesidades funcionales.

A partir de los requisitos identificados, se prevé el desarrollo de un prototipo experimental que permita evaluar distintas configuraciones de modelos y técnicas de optimización en un entorno controlado. Este enfoque incremental resulta coherente con la literatura reciente, que subraya la necesidad de validar propuestas *on-device* mediante implementaciones reales y no únicamente a través de evaluaciones sintéticas [9]. El prototipo se concibe como una herramienta exploratoria, orientada a estudiar los compromisos entre rendimiento, consumo de recursos y adecuación funcional.

La evaluación se realizará mediante benchmarks específicos del dominio, considerando métricas como latencia de inferencia, uso de memoria y adecuación funcional de las salidas generadas. Siguiendo las recomendaciones de trabajos previos, se priorizará una evaluación empírica reproducible que refleje condiciones de uso realistas, en lugar de optimizar métricas aisladas [7]. El objetivo de esta fase 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 Large Language Models (LLMs) ejecutados en dispositivos *on-edge* en aplicaciones reales plantea desafíos que van más allá del rendimiento del modelo, afectando de forma directa a decisiones de diseño, evaluación y mantenimiento del sistema. La evidencia recogida en la literatura reciente coincide en señalar que, aunque la ejecución local es técnicamente viable, su aplicación práctica se encuentra fuertemente condicionada por restricciones de memoria, consumo energético, longitud de contexto y disponibilidad de herramientas de ingeniería del software adecuadas [8,2,9].

El caso de uso presentado en el sector agroalimentario refuerza estas conclusiones, al situar el problema en un dominio con conectividad limitada y elevados requisitos de privacidad y trazabilidad de los datos. En este contexto, las soluciones basadas exclusivamente en infraestructuras cloud resultan poco adecuadas, mientras que los enfoques *on-device* requieren una cuidadosa adaptación del modelo y del sistema completo para garantizar su viabilidad [7]. Este escenario evidencia la necesidad de enfoques específicos que integren restricciones técnicas y requisitos del dominio desde las fases iniciales de diseño.

Como trabajo en curso, esta investigación no persigue ofrecer una solución cerrada, sino identificar patrones, limitaciones y problemas abiertos que puedan orientar desarrollos futuros. En particular, se abren líneas de investigación relacionadas con la definición de metodologías de evaluación reproducibles, el diseño de prototipos sostenibles a largo plazo y la integración de LLMs *on-edge* en apli-

caciones móviles reales. En este sentido, el trabajo puede servir como punto de partida para reducir la brecha existente entre los avances en modelos de lenguaje y su adopción práctica en entornos con recursos limitados.

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. Bagawan, S., G S, N.: Develop an On - Device LLM Android Application. In: 2024 8th International Conference on Computational System and Information Technology for Sustainable Solutions (CSITSS). pp. 1–5 (Nov 2024). https://doi.org/10.1109/CSITSS64042.2024.10816785, https://ieeexplore.ieee.org/document/10816785/, iSSN: 2767-1097
2. 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/
3. 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
4. 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
5. 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]
6. Tan, F., Lee, R., Dudziak, Ł., 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]
7. 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
8. 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]
9. 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]