



Desafíos técnicos del uso de LLM on-device en el sector agrícola

F. Javier González-Ontañón, Sergio Martin-Segura, Mario E. Suaza-Medina, Miguel Ángel Latre, F. Javier Zarazaga-Soria

Departamento de Informática e Ingeniería de Sistemas / Instituto de Investigación en Ingeniería de Aragón (I3A)
Universidad de Zaragoza, Mariano Esquillor s/n, 50018, Zaragoza, Spain. · jontanon@unizar.es

01 · CONTEXTO Y CASO DE USO

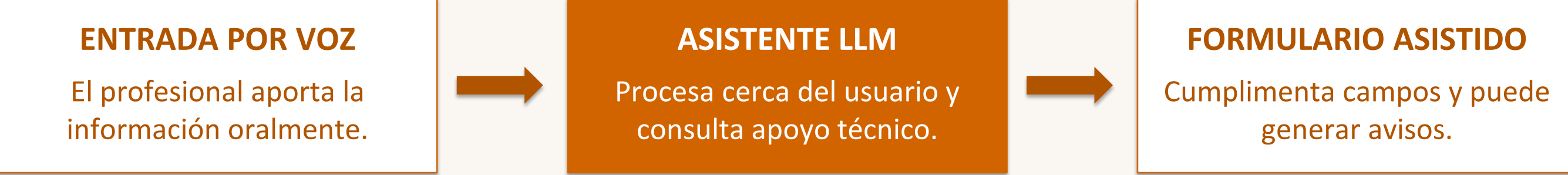
POR QUÉ ESTUDIAR LLM ON-DEVICE PARA USO AGRÍCOLA

Los LLM permiten interactuar en lenguaje natural, pero su ejecución en dispositivos móviles está condicionada por restricciones de memoria, latencia y longitud de contexto. En el sector agrícola, donde la carga administrativa asociada a la digitalización, la trazabilidad y el cumplimiento normativo debe gestionarse a menudo en campo y con conectividad limitada, un asistente local podría facilitar la consulta de documentación técnica y la cumplimentación de formularios, reduciendo la dependencia de la nube y reforzando la privacidad de los datos.

¿En qué condiciones es viable un LLM on-device para documentación agrícola en campo?

DOCUMENTACIÓN AGRÍCOLA EN CAMPO

El profesional aporta información oralmente y el asistente ayuda a trasladarla al formulario. La normativa o la documentación técnica pueden apoyar determinados campos y generar avisos, incluso con conectividad limitada y requisitos estrictos de privacidad y trazabilidad. La entrada por voz añade incertidumbre y debe validarse en condiciones realistas.



Funciones a validar: voz, formularios, avisos, normativa y documentación técnica

02 · ESPACIO DE DISEÑO Y RETOS DE INGENIERÍA

1. ALTERNATIVAS DE DESPLIEGUE

La arquitectura debe equilibrar recursos, conectividad y gestión del dato. Cada ubicación ofrece ventajas y restricciones que deben compararse para el caso agrícola.

NUBE	ON-EDGE	ON-DEVICE
Aporta: mayor capacidad. Limita: conectividad externa y datos procesados fuera.	Aporta: más recursos y procesamiento próximo. Limita: infraestructura cercana.	Aporta: ejecución local, resiliencia y privacidad. Limita: memoria, cómputo, energía y contexto.
El reto es definir el equilibrio entre autonomía local y apoyo on-edge según los requisitos de la tarea, los recursos disponibles y las condiciones de conectividad.		

2. MODELOS Y OPTIMIZACIÓN

Se evaluarán modelos de código abierto con tamaño adecuado para dispositivos móviles. La literatura propone comprimir o adaptarlos eficientemente para conseguir una calidad suficiente en tareas administrativas mediante alguna de las siguientes técnicas .

MODELOS DE CÓDIGO ABIERTO OPTIMIZADOS PARA MÓVILES

CUANTIZACIÓN Reduce la precisión numérica para disminuir memoria y coste de inferencia.	PODA Elimina parámetros o componentes prescindibles para reducir tamaño y coste.
DESTILACIÓN Transfiere capacidades de un modelo mayor a otro pequeño y eficiente.	PEFT Adapta pocos parámetros, por ejemplo con LoRA, con menor coste que un ajuste completo.

También se analizará la factorización de bajo rango para reducir almacenamiento y cómputo.

3. PROBLEMAS ABIERTOS

La viabilidad técnica no basta: el sistema completo debe responder a las necesidades del trabajo en campo.

- Adecuación funcional**
Determinar qué degradación de calidad es aceptable cuando una respuesta incorrecta afecta al formulario.
- Evaluación en condiciones reales**
Combinar eficiencia con entrada por voz, conectividad limitada, normativa extensa y reproducibilidad.
- Decisiones arquitectónicas**
Estudiar cuándo on-device es suficiente y cuándo resulta útil apoyarse en una alternativa on-edge.
- Ingeniería del software**
Diseñar, desplegar y mantener el sistema con decisiones trazables en dispositivos heterogéneos.

03 · METODOLOGÍA, EVALUACIÓN Y PRÓXIMOS PASOS

1. METODOLOGÍA Y ESTADO ACTUAL

El trabajo se encuentra en una fase inicial. La revisión del estado del arte y los cuestionarios a profesionales permitirán definir el prototipo y los criterios de evaluación.

- 1 Mapeo sistemático**
Identificar enfoques, limitaciones y problemas abiertos en on-device y on-edge.
- 2 Captura de requisitos**
Comprender documentación, conectividad habitual y criterios de respuesta útil.
- 3 Prototipo abierto**
Comparar modelos de código abierto, optimizaciones y configuraciones híbridas.
- 4 Validación empírica**
Evaluar implementaciones reales en condiciones controladas y reproducibles.

2. EVALUACIÓN PREVISTA

La evaluación debe aproximarse al uso real en campo, combinando benchmarks del dominio con eficiencia, corrección funcional y comparación arquitectónica.

RENDIMIENTO	¿Responde rápidamente? Latencia de inferencia, memoria y energía.
ADECUACIÓN FUNCIONAL	¿Ayuda a cumplimentar documentación? Calidad y corrección.
CONTEXTO REAL	¿Cómo funciona con voz y conectividad limitada? Reproducibilidad.
ARQUITECTURA	¿Qué compromisos ofrece on-edge frente a on-device ? Latencia, privacidad, eficiencia y conectividad.

Prioridad: condiciones realistas y reproducibles

3. CONCLUSIONES PRELIMINARES

La adopción de LLM on-device y on-edge plantea desafíos de diseño, evaluación y mantenimiento.

La cuestión no es únicamente si el modelo es capaz de funcionar en un móvil, sino que el sistema completo pueda funcionar de forma útil, evaluable y mantenible.

ÚTIL	EVALUABLE	MANTENIBLE
Interpreta voz, asiste formularios y consulta apoyo técnico.	Mide eficiencia y corrección en condiciones representativas.	Permite integrar, desplegar y decidir la arquitectura adecuada.

SIGUIENTE PASO

Sistematizar criterios de evaluación y estudiar un despliegue razonable en dispositivos heterogéneos como base para una validación empírica más completa.

04 · REFERENCIAS Y AGRADECIMIENTOS

REFERENCIAS

- [1] LI, X., LU, Z., CAI, D., MA, X. and XU, M. Large Language Models on Mobile Devices: Measurements, Analysis, and Insights. In: Proceedings of the Workshop on Edge and Mobile Foundation Models (EdgeFM '24). New York: ACM, 2024, pp. 1-6. Available from: doi:10.1145/3662006.3662059.
- [2] YAN, X. and DING, Y. Are We There Yet? A Measurement Study of Efficiency for LLM Applications on Mobile Devices. arXiv preprint arXiv:2504.00002. 2025. Available from: doi:10.48550/arXiv.2504.00002.
- [3] XU, J., LI, Z., CHEN, W., WANG, Q., GAO, X., CAI, Q. and LING, Z. On-Device Language Models: A Comprehensive Review. arXiv preprint arXiv:2409.00088. 2024. Available from: doi:10.48550/arXiv.2409.00088.

- [4] FRIHA, O., FERRAG, M.A., KANTARCI, B., ÇAKMAK, B., OZGUN, A. and GHOUALMI-ZINE, N. LLM-Based Edge Intelligence: A Comprehensive Survey on Architectures, Applications, Security and Trustworthiness. IEEE Open Journal of the Communications Society. 2024, 5, pp. 5799-5856. Available from: doi:10.1109/OJCOMS.2024.3456549.
- [5] WANG, R., GAO, Z., ZHANG, L., YUE, S. and GAO, Z. Empowering large language models to edge intelligence: A survey of edge efficient LLMs and techniques. Computer Science Review. 2025, 57, 100755. Available from: doi:10.1016/j.cosrev.2025.100755.
- [6] KLERKX, L., JAKKU, E. and LABARTHE, P. A review of social science on digital agriculture, smart farming and agriculture 4.0: New contributions and a future research agenda. NJAS - Wageningen Journal of Life Sciences. 2019, 90-91, 100315. Available from: doi:10.1016/j.njas.2019.100315.

AGRADECIMIENTOS

Esta publicación es parte del proyecto de I+D+i T59_23R financiado por el Gobierno de Aragón y del proyecto de I+D+i PID2024-155657OB-C21:AGRARD financiado por el Gobierno de España.