



TOOLBOX SESIÓN 13



CONGRESO DE
MANTENIMIENTO
& CONFIABILIDAD
COLOMBIA

3^a
EDICIÓN



Large Language Models para Mantenimiento Industrial

Enrique López Droguett

Profesor Titular, Civil and Environmental Engineering
Department, UCLA

Profesor Titular, Mechanical Engineering
Department, Universidad de Chile

1

Contenido



1. Fundamentos del Large Language Models
2. Análisis de aceite con LLMs
3. Modelos multimodales y termografías



CONGRESO DE
MANTENIMIENTO
& CONFIABILIDAD
COLOMBIA

3^a
EDICIÓN



UCLA

2

Fundamentos del Large Language Models



UCLA

3

LLMs y Mantenimiento Industrial



UCLA

4

Qué es un Large Language Model

- Modelo entrenado para **procesar y generar lenguaje** a partir de grandes volúmenes de texto.



Tokens

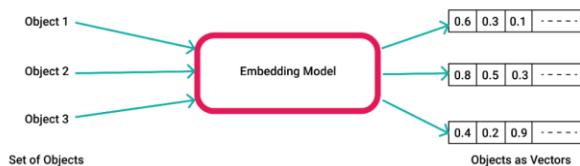
- LLMs no procesan directamente palabras completas
- Procesan unidades llamadas **tokens**.
- Una frase técnica se transforma en **una secuencia de tokens** antes de entrar al modelo:

“Continuar monitoreo tribológico según plan de mantenimiento”

Large Language Models (LLMs), such as GPT-3 and GPT-4, utilize a process called tokenization. Tokenization involves breaking down text into smaller units, known as tokens, which the model can process and understand. These tokens can range from individual characters to entire words or even larger chunks, depending on the model. For GPT-3 and GPT-4, a Byte Pair Encoding (BPE) tokenizer is used. BPE is a subword tokenization technique that allows the model to dynamically build a vocabulary during training, efficiently representing common words and word fragments. Although the core tokenization process remains similar across different versions of these models, the specific implementation can vary based on the model's architecture and training objectives.

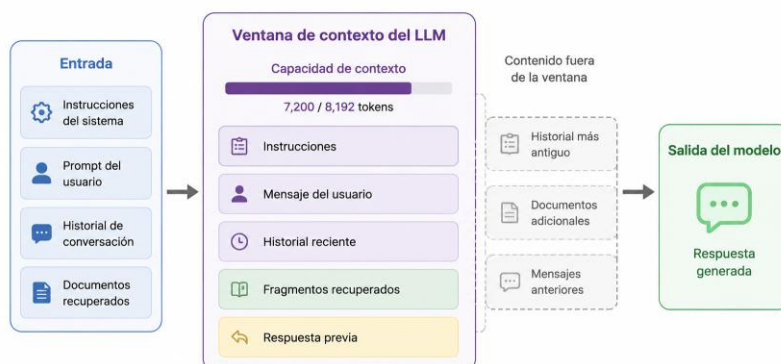
Embeddings

- Cada token se representa como un **vector numérico**.
- Estos vectores permiten que el modelo capture **relaciones** entre conceptos técnicos.
- Conceptos relacionados tienden a quedar cerca entre sí:
 - Desgaste
 - Degradación
 - Falla



Context Window

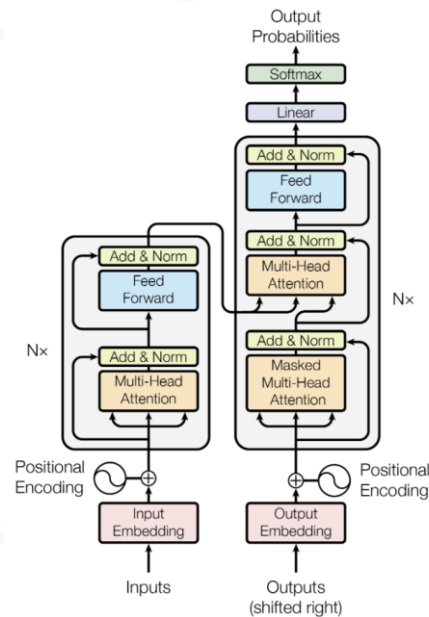
- Es la cantidad de información que el modelo puede considerar en una interacción.



El LLM solo puede usar directamente la información que cabe dentro de su ventana de contexto.

Arquitectura Transformer

- LLMs modernos están basados en la arquitectura **Transformer**.
- Idea central: procesar secuencias de tokens usando mecanismos de atención.
- Esto permite que el modelo relacione partes distintas del contexto antes de generar una respuesta.




Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention Is All You Need 2017. <https://doi.org/10.48550/arXiv.1706.03762>.

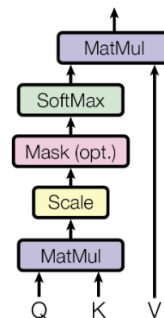


9

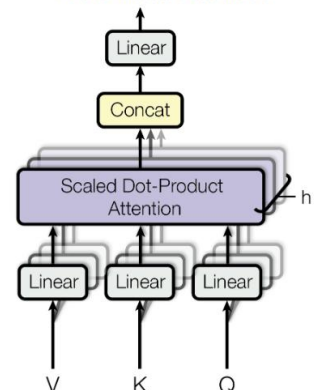
Pre-Training

- Mecanismo de Atención** permite que el modelo asigne distinto peso a diferentes partes del contexto.
- En una tarea de mantenimiento, esto ayuda a relacionar señales como:
 - Metales elevados
 - Contenido de agua
 - Partículas
 - Tipo de activo
 - Nivel de severidad

Scaled Dot-Product Attention



Multi-Head Attention




Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention Is All You Need 2017. <https://doi.org/10.48550/arXiv.1706.03762>.



10

Pre-Training

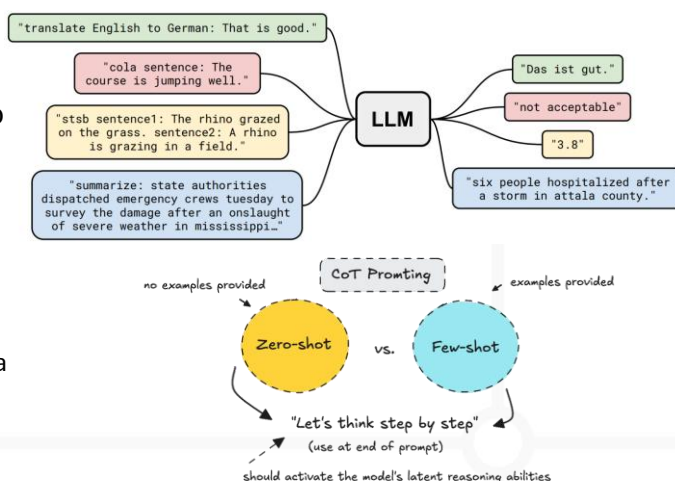
- Antes de ser usados en aplicaciones específicas, los LLMs pasan por una etapa de **pre-training**
 - Aprenden patrones generales de lenguaje, estructura y conocimiento a partir de grandes corpus.



UCLA

Prompting

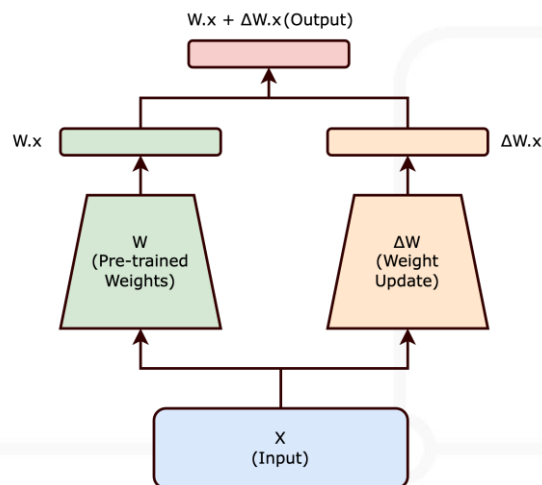
- **Prompting** consiste en guiar al modelo mediante instrucciones, contexto y ejemplos. No modifica los pesos del modelo. Estrategias comunes:
 - **Zero-shot**: instrucción directa sin ejemplos
 - **Few-shot**: instrucción con ejemplos resueltos
 - **Razonamiento estructurado**: respuesta organizada en pasos o campos



UCLA

Fine-Tuning

- **Fine-tuning** adapta un modelo base a una tarea o dominio específico usando datos del problema.
- En modelos grandes, una alternativa eficiente es entrenar adaptadores en vez de modificar todo el modelo.
- **LoRA** permite adaptar un modelo con menor costo computacional.



Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, et al. LoRA: Low-Rank Adaptation of Large Language Models 2021.
<https://doi.org/10.48550/arXiv.2106.09685>.
<https://towardsdatascience.com/understanding-lora-low-rank-adaptation-for-finetuning-large-models-936bce1a07c6/>.

UCLA

13

Prompting x Fine-Tuning

Aspecto	Prompting	Fine-Tuning
Modifica pesos	No	Sí, o adaptadores
Esfuerzo inicial	Bajo	Mayor
Personalización	Limitada	Alta
Control local	Depende del proveedor	Alto con open source



UCLA

14

Modelos Comerciales y Open Source

Existen dos grandes formas de usar LLMs en ingeniería:

1. Modelos Comerciales:

- Alta capacidad general
- Uso inmediato
- Acceso mediante web o API



Modelos Comerciales y Open Source

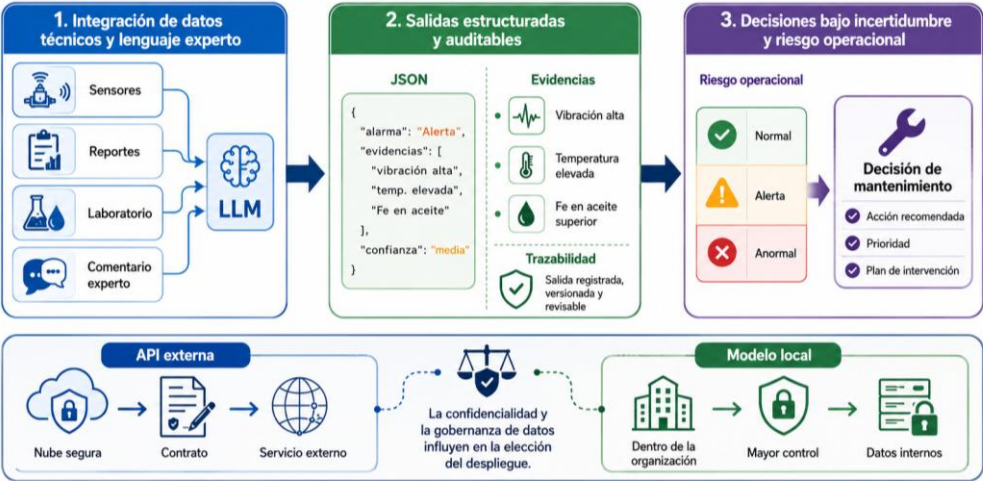
2. Modelos Open Source:

- Ejecución local
- Mayor control del modelo
- Adaptación al dominio
- Mejor compatibilidad con datos sensibles



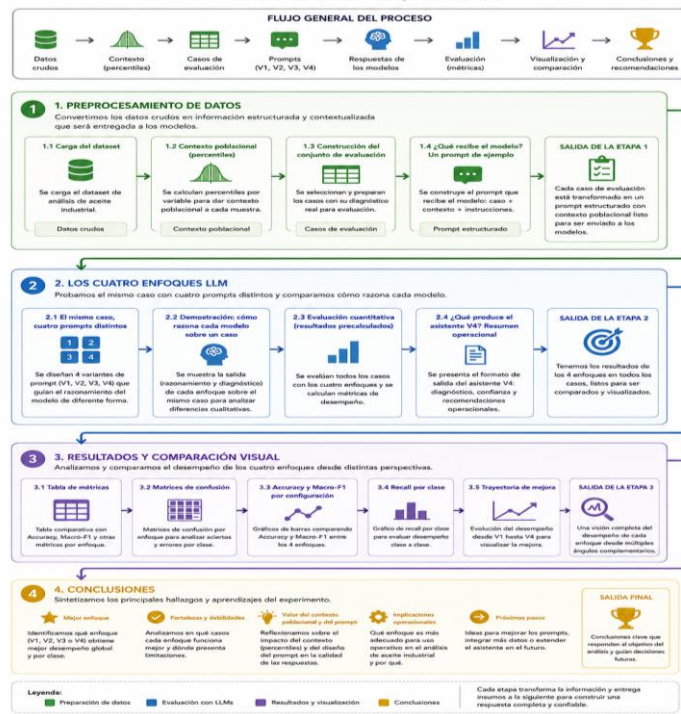
Relevancia para Mantenimiento Industrial

Los LLMs conectan datos, estructuran salidas y apoyan decisiones de mantenimiento.



Análisis de Aceite con LLMs

Procedimiento



Detección Temprana de Fallas en base Análisis de Aceite

- El lubricante transporta señales del estado del activo:

- Desgaste de componentes
- Contaminación externa
- Presencia de agua
- Degradación del lubricante
- Incremento de partículas

Señales del lubricante



Del Laboratorio a la Decisión de Mantenimiento

Del análisis de aceite a la decisión de mantenimiento.



💡 El análisis de aceite convierte una muestra física en una decisión operativa.

Dataset Industrial de Análisis de Aceite

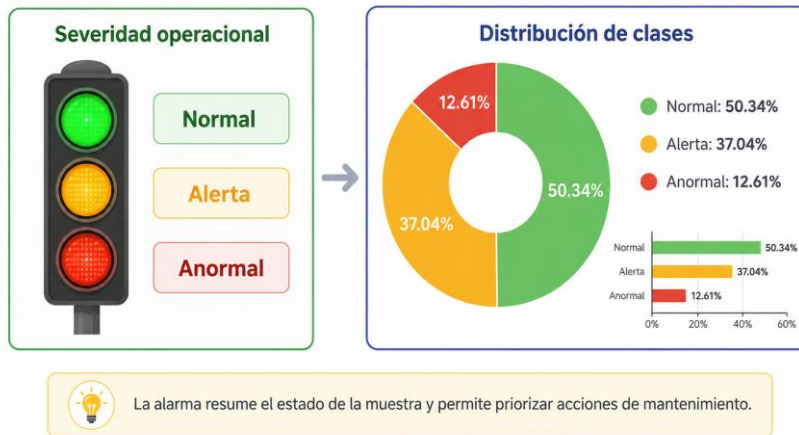
Caso de estudio



💡 El dataset combina mediciones, propiedades del lubricante y juicio experto para el análisis del activo.

Estimar el Astado de Alarma

Objetivo: clasificar la severidad operacional.



UCLA

No todos los Errores Cuestan lo Mismo

Dirección del error en la clasificación de severidad

		Alarma predicha			Leyenda Clasificación correcta Error conservador Error riesgoso Subestimación crítica
Alarma real		Normal	Alerta	Anormal	
	Normal	Correcto	Conservador		
	Alerta	Riesgoso	Correcto		
	Anormal	Error crítico		Correcto	

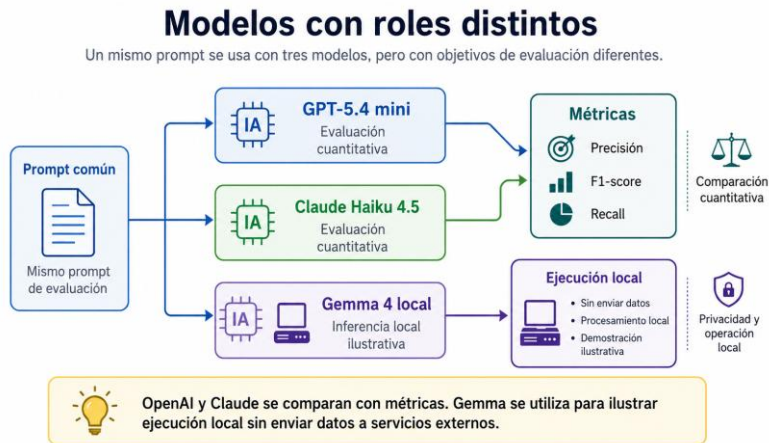
Una subestimación puede ser más grave que una falsa alarma.

UCLA

Tres Modos de Uso del LLM



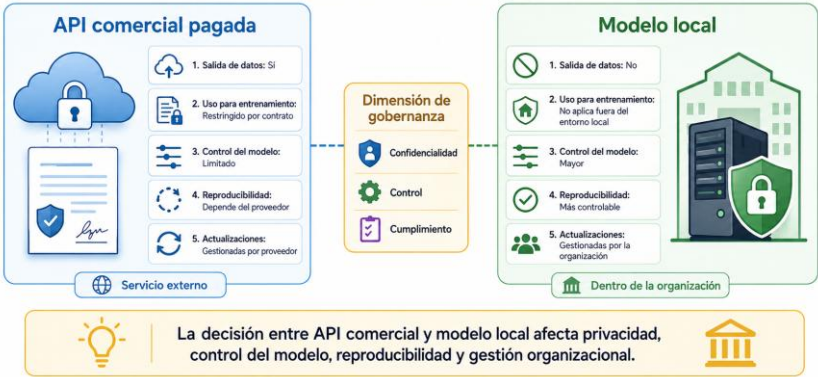
Modelos Usados



Propiedad Intelectual y Confidencialidad

Gobernanza en el uso de modelos

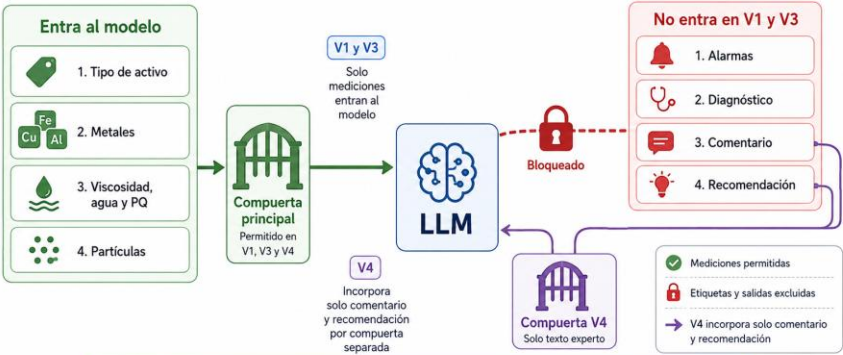
La elección entre API comercial y modelo local no es solo técnica, también implica confidencialidad, control y políticas organizacionales.



Separar Mediciones, Alarmas y Texto Experto

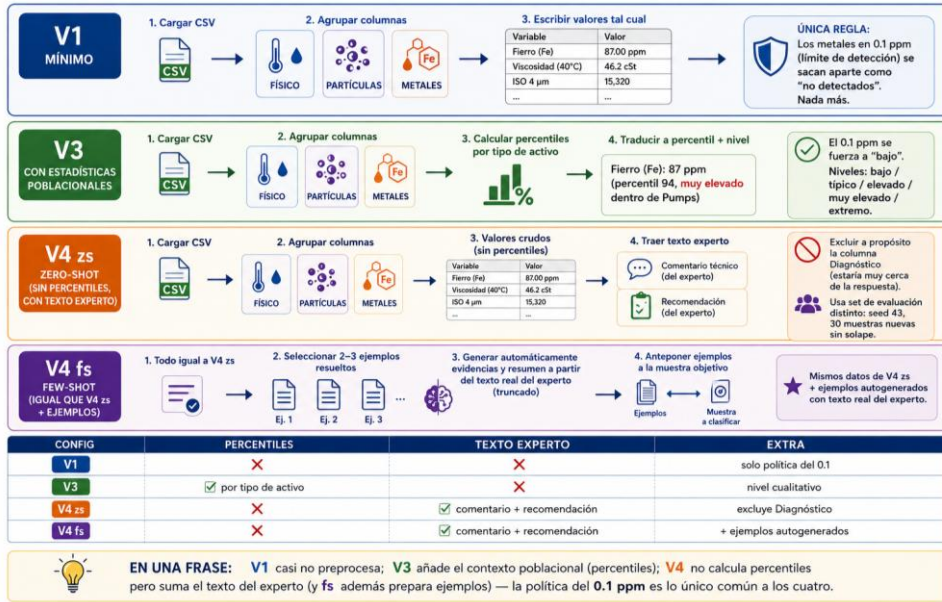
Control de información para evitar leakage

Cada modo controla qué información entra al modelo



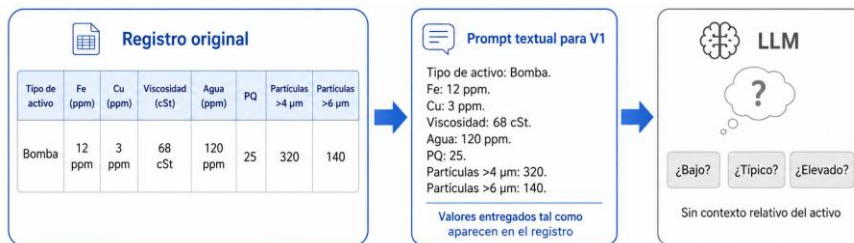
El diseño de entrada controla el leakage. V1 y V3 usan solo mediciones, mientras V4 abre una compuerta separada para comentario y recomendación.

Preprocesamiento de la Data Numérica y Texto



V1: Valores Crudos

- V1 entrega al modelo los valores tal como aparecen en el registro:
 - Nombres de variables
 - Valores numéricos
 - Unidades básicas
 - Sin escala relativa



V1 establece un punto de partida, pero no entrega escala relativa ni juicio operacional.

V3: Criterio Operacional

V3 agrega contexto operacional al dato

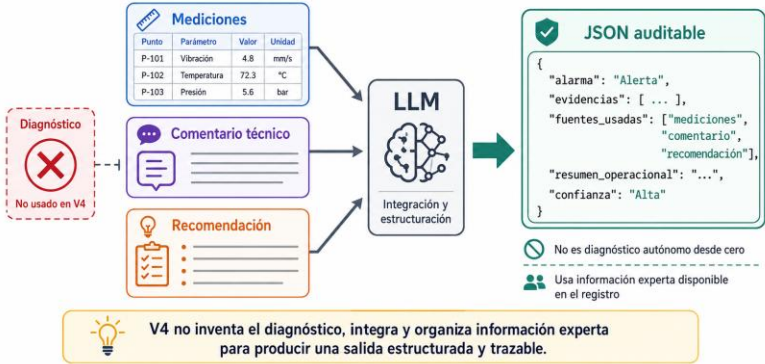
El modelo interpreta señales en conjunto, no como valores aislados.



V4: Mediciones + Texto Experto

V4 integra información experta disponible

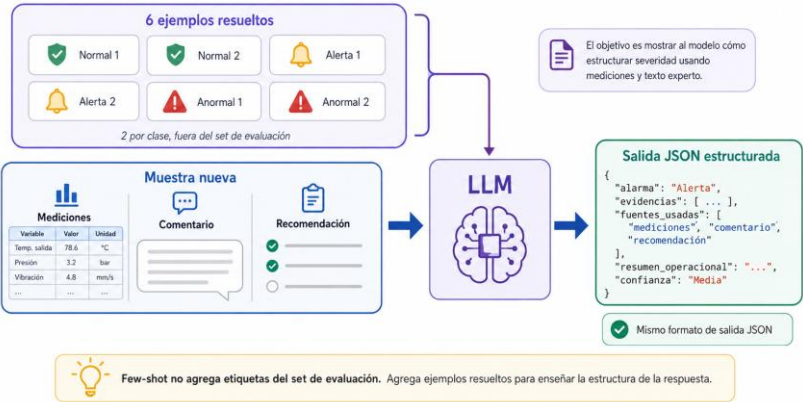
Integra mediciones, comentario técnico y recomendación, sin usar Diagnóstico.



V4 Few-Shot: Ejemplos Resueltos (Observaciones)

V4 Few-shot agrega ejemplos resueltos

El modelo aprende la estructura de severidad a partir de ejemplos resueltos y luego procesa una muestra nueva con el mismo formato JSON.



Construcción de los Prompts

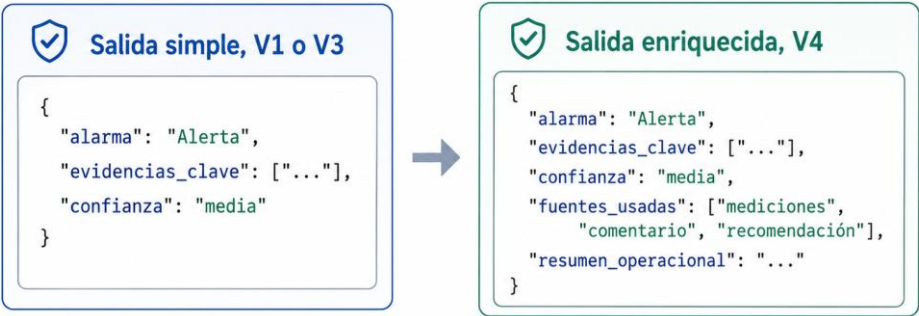
LOS 4 PROMPTS DE ENTRADA

V1 — CLASIFICADOR BÁSICO	V3 — CLASIFICADOR CON CONTEXTO + REGLA DE DECISIÓN	V4 ZERO-SHOT — SINTETIZADOR CON TEXTO EXPERTO	V4 FEW-SHOT — IGUAL QUE V4 ZS + EJEMPLOS
Solo las mediciones crudas (ej. Hierro: 87.00 ppm), sin percentiles ni contexto poblacional. Rol de analista tribológico + esquema JSON. El modelo clasifica la alarma únicamente con los números tal cual.	Mediciones anotadas con su percentil y nivel por tipo de activo (Hierro: 87 ppm (percentil 94, muy elevado)) + contexto de dominio (qué es aditivo/desgaste/contaminación) + un criterio explícito: no marcar Alerta/Anormal por un valor aislado si el resto es normal.	Cambia la tarea: mediciones crudas + Comentario técnico + Recomendación del experto. Ya no diagnostica desde cero, sino que integra esas fuentes. JSON con campos extra (fuentes_usadas, resumen_operacional).	Mismo prompt de V4, pero antepone 2–3 ejemplos ya resueltos (con su Comentario + Recomendación + JSON esperado) antes de la muestra a clasificar. El modelo ve ejemplos y aprende el patrón esperado.
CONFIG	DATOS QUE ENTRAN	RASGO CLAVE	
V1	mediciones crudas	sin contexto, línea base	
V3	mediciones + percentiles	contexto poblacional + criterio de decisión	
V4 zs	mediciones crudas + comentario + recomendación	sintetiza texto experto	
V4 fs	= V4 zs	+ ejemplos resueltos	

EN UNA FRASE: V1 clasifica con números pelados, V3 les añade contexto y una regla más estricta, y V4 deja de clasificar para resumir la severidad usando lo que el experto ya escribió — fs solo agrega ejemplos.

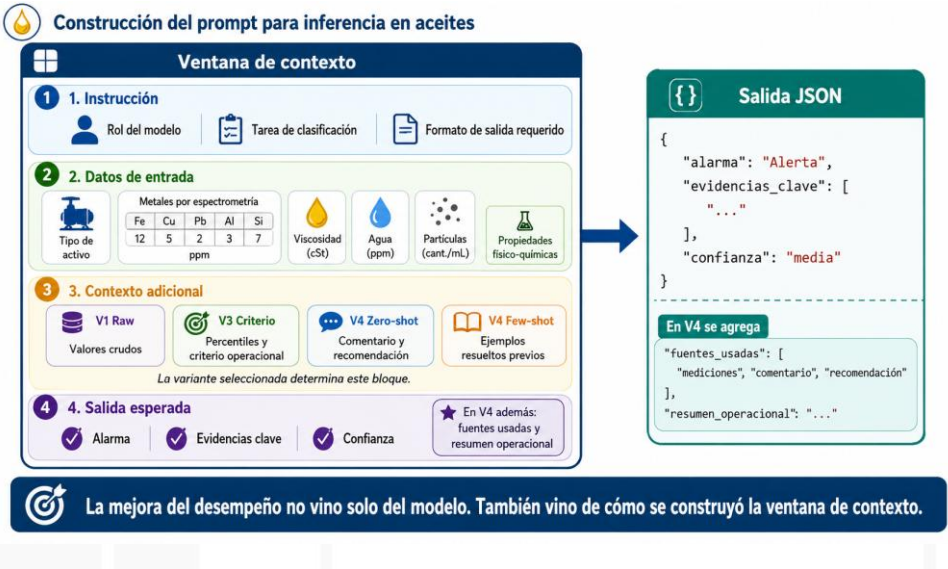
Salidas Estructuradas y Auditables

Todas las configuraciones responden en formato JSON. V4 agrega mayor trazabilidad y contexto operacional.



💡 Ambas salidas son auditables. V4 agrega fuentes usadas y resumen operacional para enriquecer la trazabilidad.

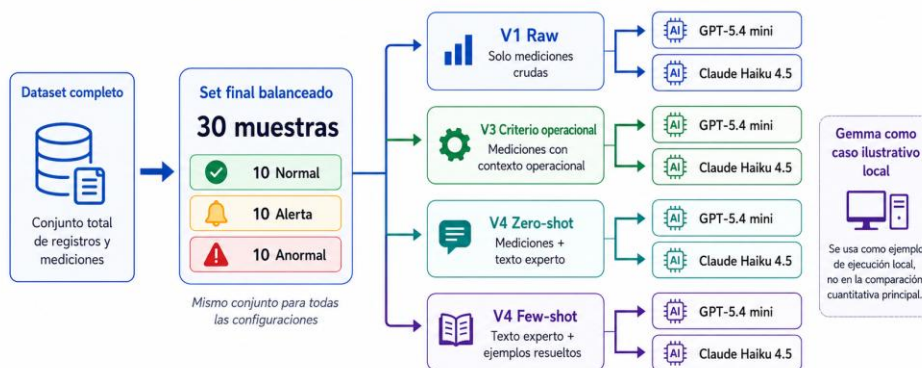
Context Window



Resumen: LLMs en Mantenimiento Predictivo

Comparación final de cuatro configuraciones

Las mismas 30 muestras balanceadas se evalúan en cuatro configuraciones paralelas.



El experimento compara V1, V3, V4 Zero-shot y V4 Few-shot sobre las mismas 30 muestras balanceadas, usando GPT-5.4 mini y Claude Haiku 4.5.

UCLA

Demo en Vivo

1. Desarrollar las soluciones predictivas

2. Interpretar los resultados

3. Esto se hace cargando el notebook:

“02_oil_analysis_llm_demo.ipynb”

1. El notebook y sus librerías se encuentran disponibles en:

“CMC_CO_2026_Toolbox_Code.zip”

UCLA

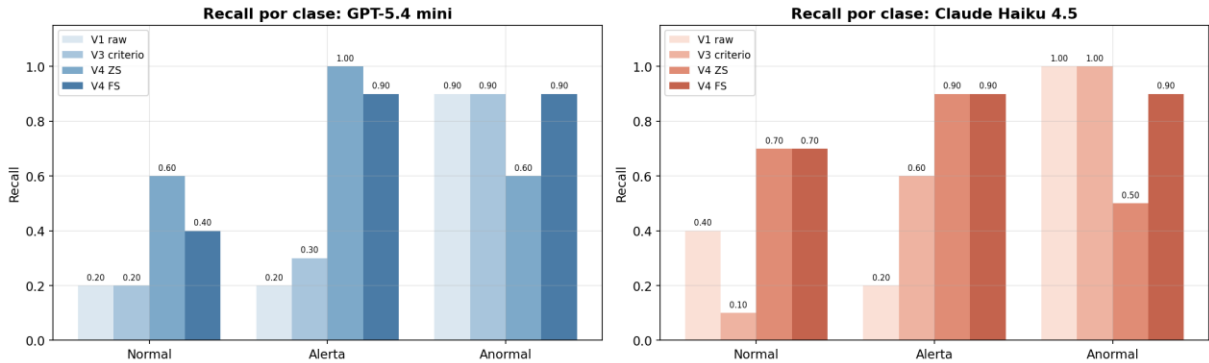
Resultados: Matrices de Confusión

Matrices de confusión: comparación de las cuatro configuraciones (% por clase real)



Resultados: Recall por Clase (Alarma)

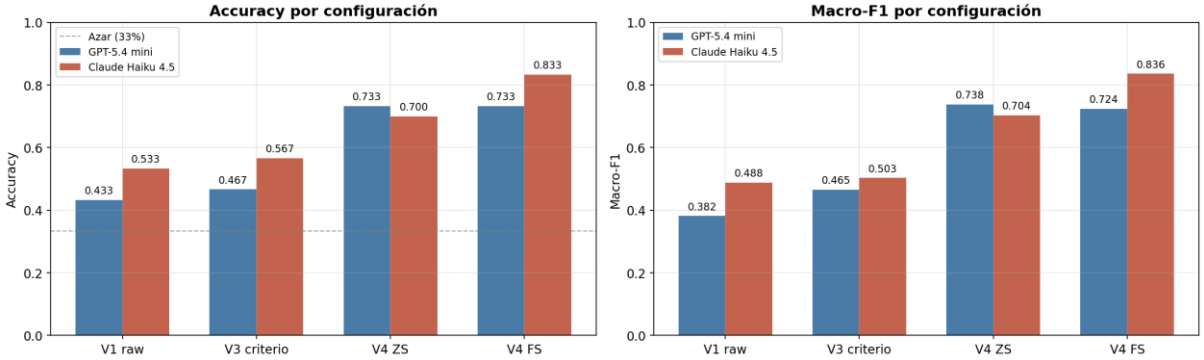
Recall por clase: evolución a través de las cuatro configuraciones



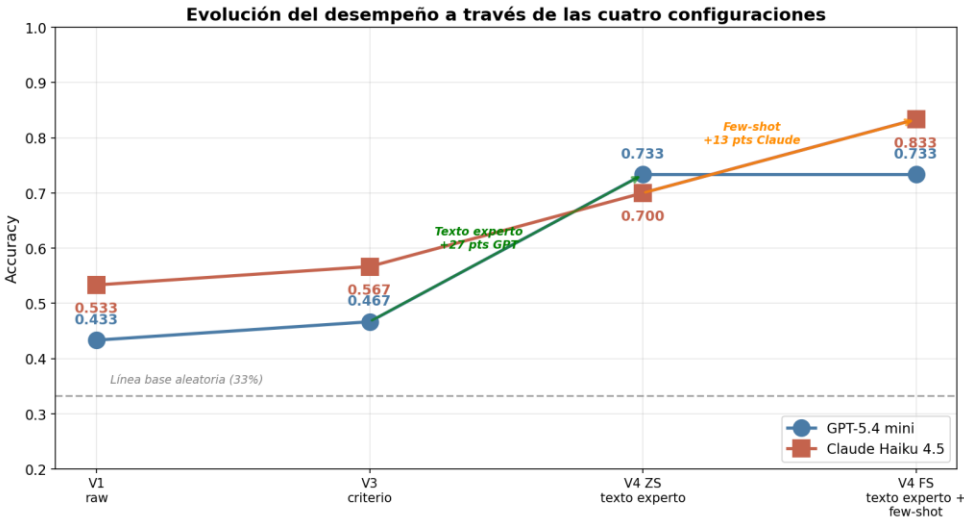
Resultados: Métricas de Performance



Comparación de las cuatro configuraciones sobre las mismas 30 muestras independientes



Resultados: Evolución de las Distintas LLMs Predictivas



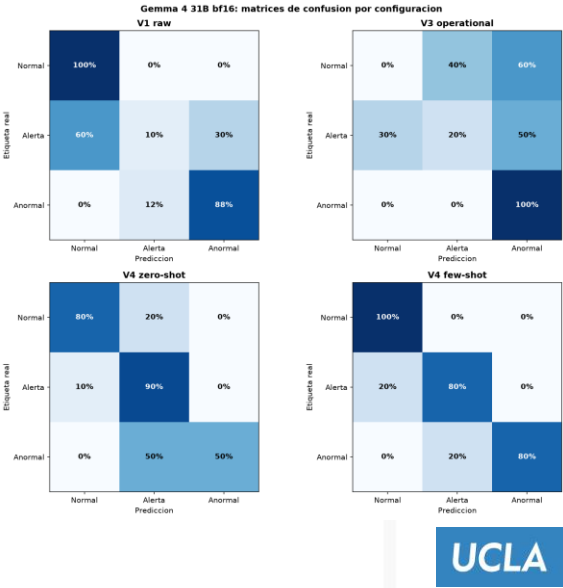
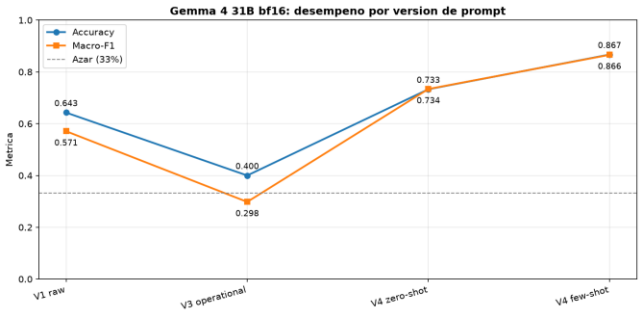
Mejor LLM Predictiva Desarrollada



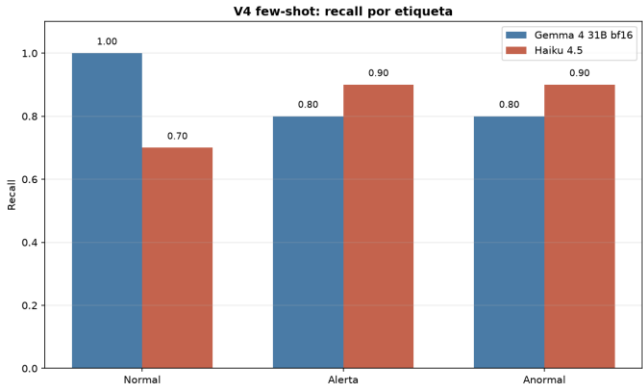
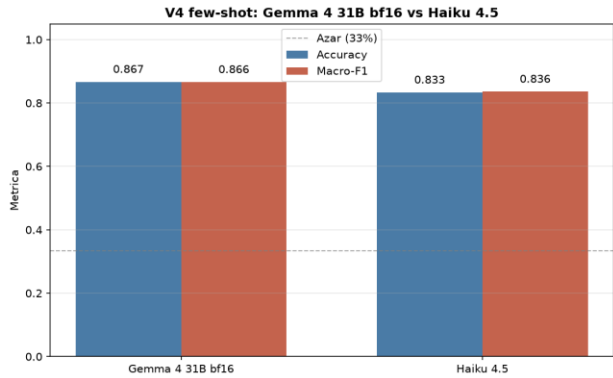
Riesgo Operacional: Claude Haiku 4.5



Qué Pasa Si Hacemos el Mismo Ejercicio con un LLM Local?



Resultados Comparables al Mejor Modelo Comercial



Qué Aprendimos de los Cuatro Modos de LLMs

Aprendizajes por configuración

Cada configuración aporta una mejora metodológica distinta en el flujo de información.



Qué Significa Esto para Mantenimiento Predictiva

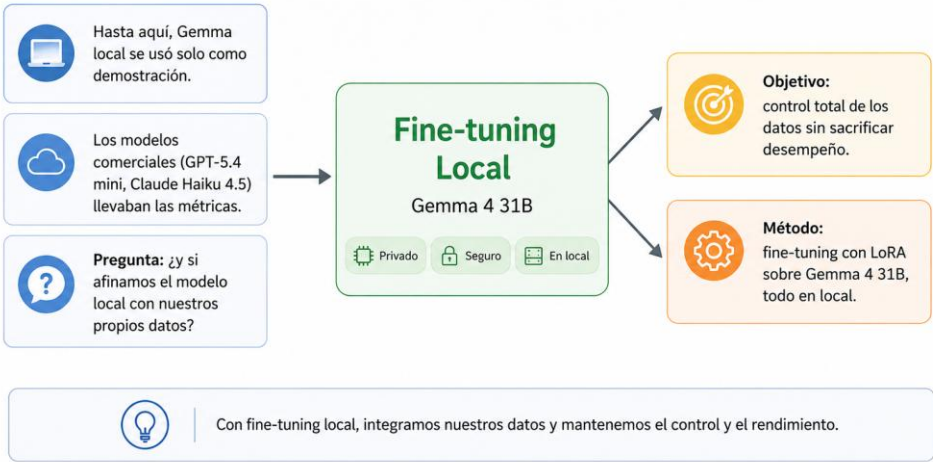


El Experto Sigue en el Sistema

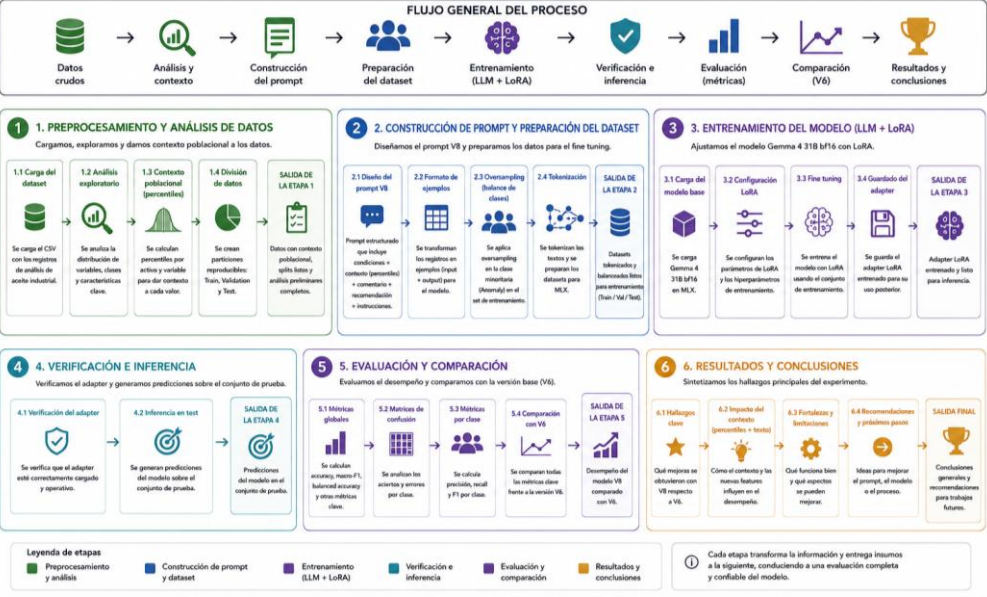


Inteligencia Artificial Local (Local AI)

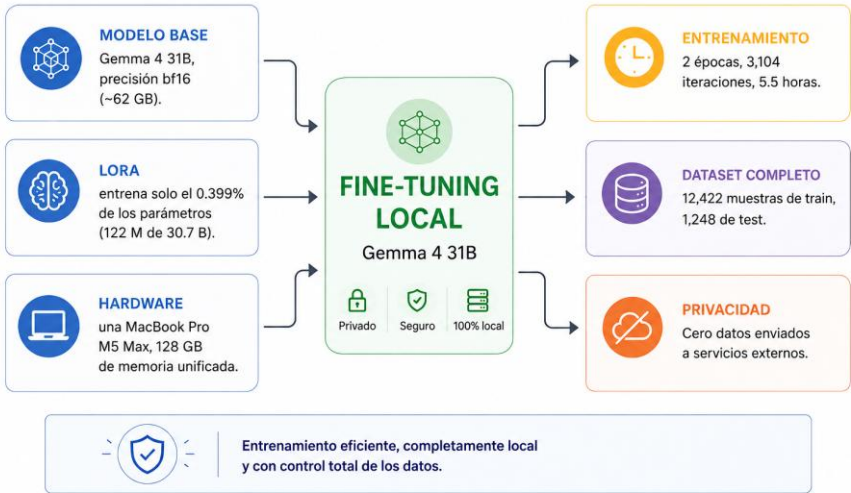
De Ilustrativo a Competitivo: Gemma Local Afinado



Procedimiento del Fine Tuning



Cómo se Afinó: LoRA Local sobre Gemma 4-31B

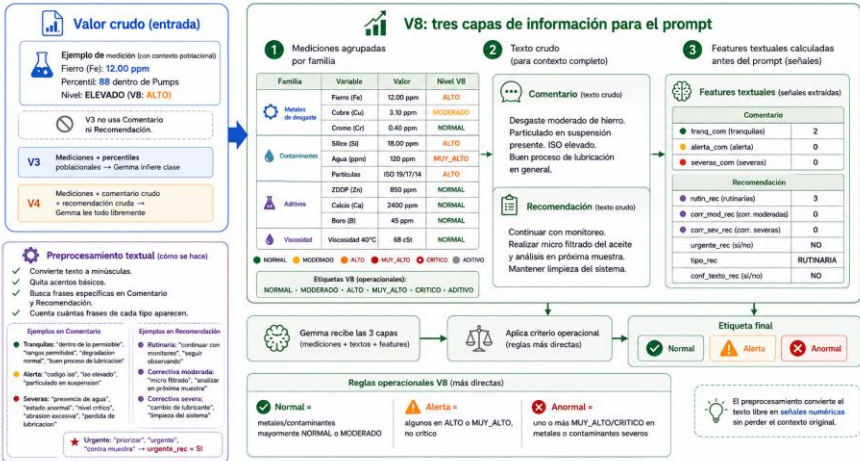


53

Importancia del Pre-Procesamiento

V8 agrega señales textuales estructuradas al prompt

El modelo interpreta señales en conjunto, no como valores aislados.



54

V8 no sustituye el texto: lo entiende y lo resume en señales que ayudan a Gemma a clasificar mejor, con reglas más claras y consistentes.

Tipo y Flujo de la Información

Aprendizajes por configuración

Cada configuración aporta una mejora metodológica distinta en el flujo de información.



Modelo Local Supera a los Comerciales

Resultados de evaluación

Comparación de desempeño en el test completo

Enfoque	Accuracy	Recall Anormal
Red neuronal (baseline del advisor)	85.00%	90.6%
Mejor comercial: Claude Haiku 4.5 (V4 Few-shot)	83.3%	90%
Gemma 4 31B local, afinado (V8)	91.03%	96.2%
Gemma local alcanza 91.03% de accuracy sobre el test completo.		
Supera al mejor comercial por 7.7 puntos de accuracy.		
En la clase crítica (Anormal), 96.2% de recall: 6.2 puntos por encima.		
Balanced Accuracy: 91.63%. F1 macro: 0.8819.		
Nota de evaluación: el comercial se midió sobre 30 muestras balanceadas; Gemma sobre las 1,248 del test real desbalanceado.		
Gemma 4 31B afinado localmente ofrece el mejor desempeño global, con mayor precisión y mejor detección de anomalías.		

Dónde Acierta y Dónde Falla

RENDIMIENTO DEL MODELO: MATRIZ DE CONFUSIÓN

Filas = clase real | Columnas = predicción



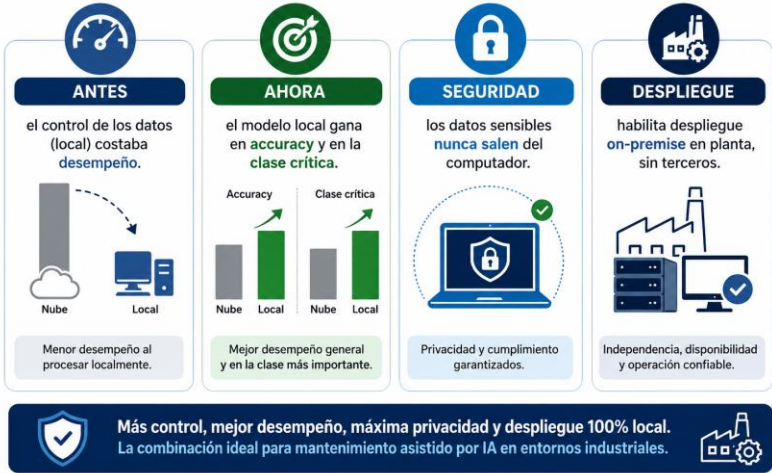
UCLA

57



Privacidad y Desempeño: Ya No Hay que Elegir

¿QUÉ GANAMOS CON MODELOS LOCALES?



UCLA

58



Demo en Vivo: Fine-Tuning Gemma 4

1. Desarrollar la solución predictiva en base a Gemma 4
2. Interpretar los resultados
3. Esto se hace cargando el notebook:
"03_oil_analysis_llm_gemma4_demo.ipynb"
1. El notebook y sus librerías se encuentran disponibles en:
"CMC_CO_2026_Toolbox_Code.zip"

Conclusiones

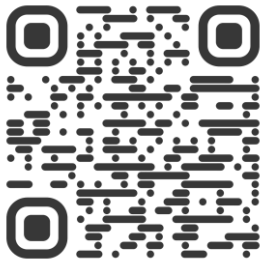
Qué Demostramos

- Los casos de aceite y termografía muestran que las LLMs ya pueden apoyar el mantenimiento industrial:



Por Qué Aprender Esto Ahora ?



SESIÓN 13**ESCANEA EL
CÓDIGO QR****RESPONDE UNA
BREVE ENCUESTA****CONGRESO DE
MANTENIMIENTO
& CONFIABILIDAD
COLOMBIA****3^a
EDICIÓN**

¡GRACIAS!

Enrique López Droguett

eald@ucla.edu