TOOLBOX SESIÓN 20

CONGRESO DE MANTENIMIENTO & CONFIABILIDAD MÉXICO 19ª EDICIÓN

Large Language Models para Mantenimiento Industrial

Enrique López Droguett

Profesor Titular, Civil and Environmental Engineering Department, UCLA

1

Contenido

1. Fundamentos del Large Language Models (LLMs).

2. Predictivo en base a Análisis de Aceite con:
   a. IA en la Nube: LLMs Comerciales,
   b. IA Local: Gemma 4.

FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS UNIVERSIDAD DE CHILE

UCLA

2

Fundamentos del Large Language Models

3

LLMs y Mantenimiento Industrial

4

Qué es un Large Language Model

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

5

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"

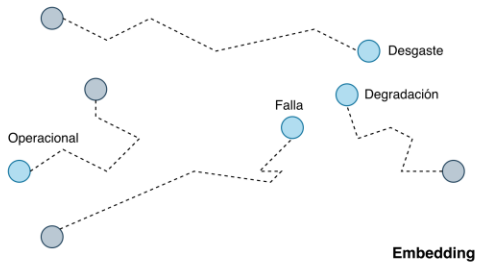
```
Large Language Models (LLMs), such as GPT-3 and GPT-4, util
ize a process called tokenization. Tokenization involves br
eaking down text into smaller units, known as tokens, which
the model can process and understand. These tokens can rang
e from individual characters to entire words or even larger
chunks, depending on the model. For GPT-3 and GPT-4, a Byte
Pair Encoding (BPE) tokenizer is used. BPE is a subword tok
enization technique that allows the model to dynamically bu
ild a vocabulary during training, efficiently representing
common words and word fragments. Although the core tokeniza
tion process remains similar across different versions of t
hese models, the specific implementation can vary based on
the model's architecture and training objectives.
```

https://ai.plainenglish.io/why-future-ais-wont-use-tokens-at-all-8bf5cdcd6fa6

6

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.

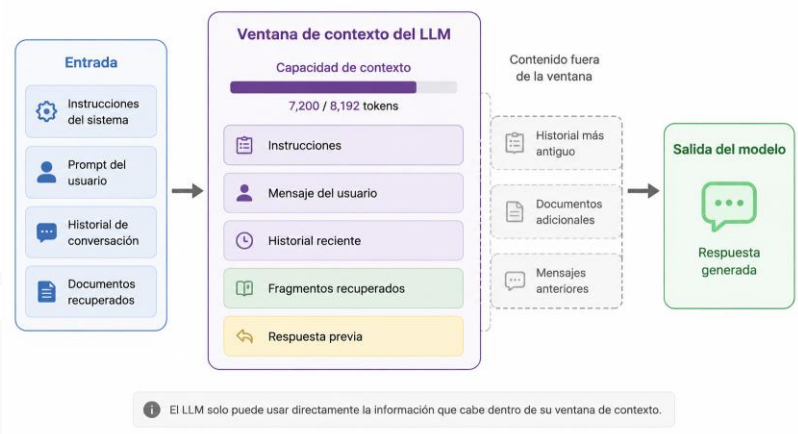


https://www.linkedin.com/pulse/how-do-embeddings-work-large-language-model-llm-onfinanceofficial-54awc/

7

Ventana de Contexto (Context Window)

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



8

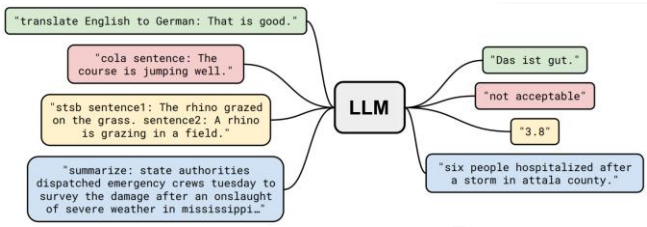
Pre-Training

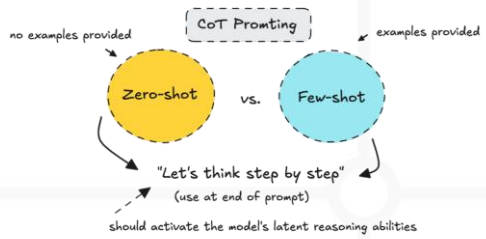
- Antes de ser usados en aplicaciones específicas, los LLMs pasan por una etapa de pre-training:
  - Aprenden patrones generales de lenguaje, estructura y conocimiento a partir de grandes corpus.

9

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 etapas.



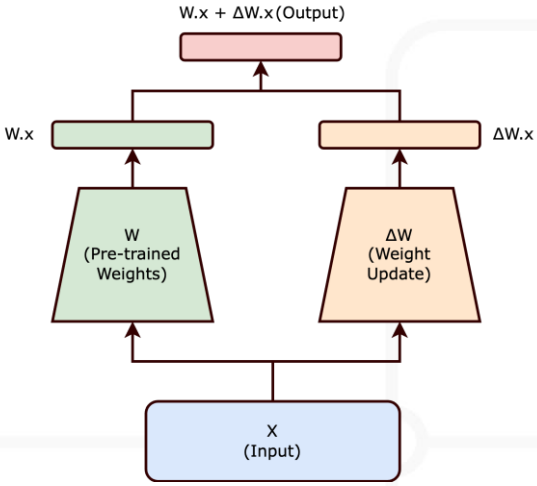


https://www.linkedin.com/pulse/chain-of-thought-prompting-few-shot-learning-nikitha-r-3stvf/
https://cameronrwolfe.substack.com/p/modern-advances-in-prompt-engineering

10

Fine-Tuning

- Adapta un modelo base a una tarea o dominio específico usando datos del problema.
- Con LLMs, 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/.

11

Prompting x Fine-Tuning

Aspecto	Prompting	Fine-Tuning
Modifica Pesos	No	Sí
Esfuerzo Inicial	Bajo	Mayor
Personalización	Limitada	Alta
Control Local	Bajo o Ninguno (Depende del Proveedor)	Alto (con Open Source)

12

Modelos Comerciales y Open Source

Existen dos grandes formas de usar LLMs en ingeniería:

1. **Modelos Comerciales:**
   - Ejecución en la nube,
   - Alta capacidad general,
   - Uso inmediato,
   - Acceso mediante web o API.

13

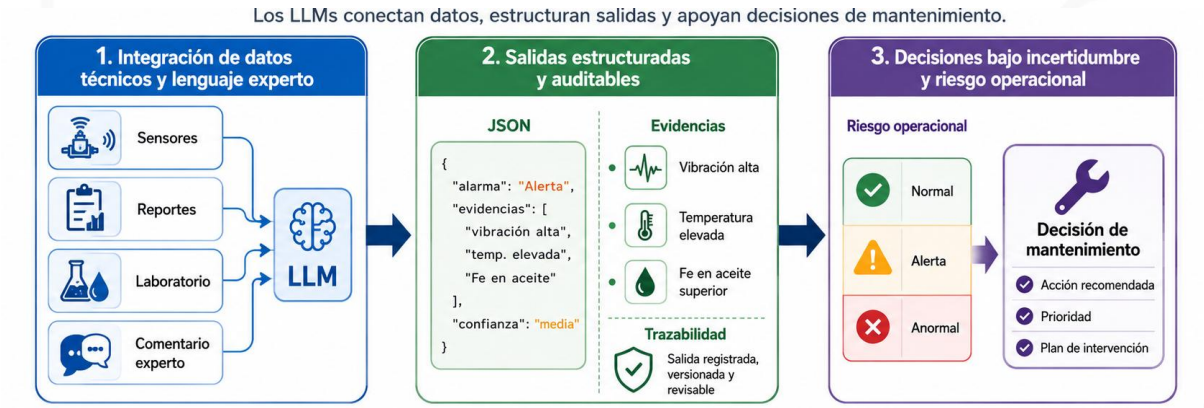
Modelos Comerciales y Open Source

2. **Modelos Open Source:**
   - Ejecución local,
   - Mayor control del modelo,
   - Adaptación al dominio,
   - Mejor compatibilidad con datos sensibles.

https://law.co/blog/fine-tuning-open-source-llms-for-case-specific-legal-ai-agents

14

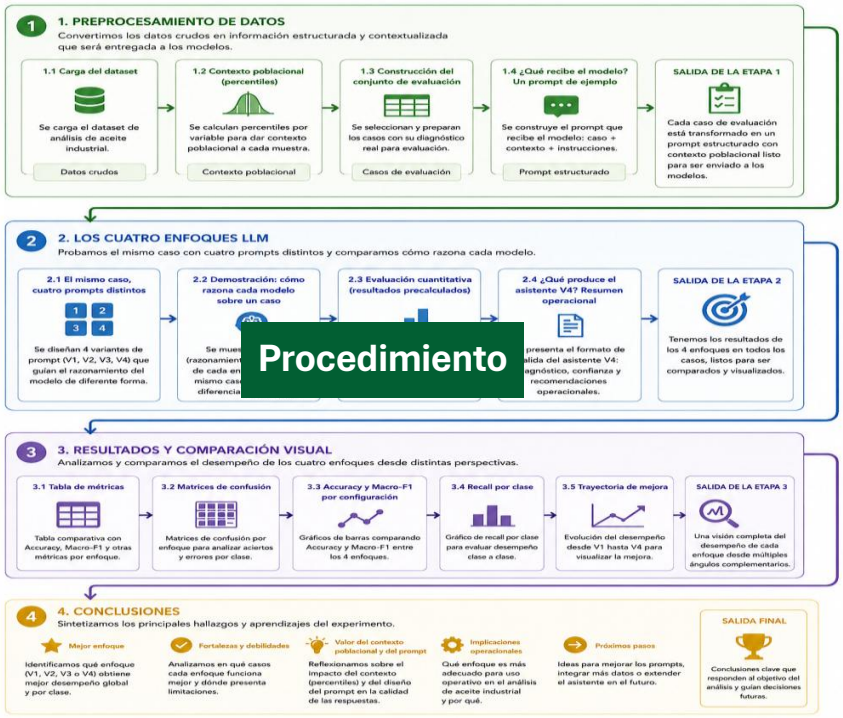
Relevancia para Mantenimiento Industrial



15

IA en la Nube:
LLMs Comerciales

16



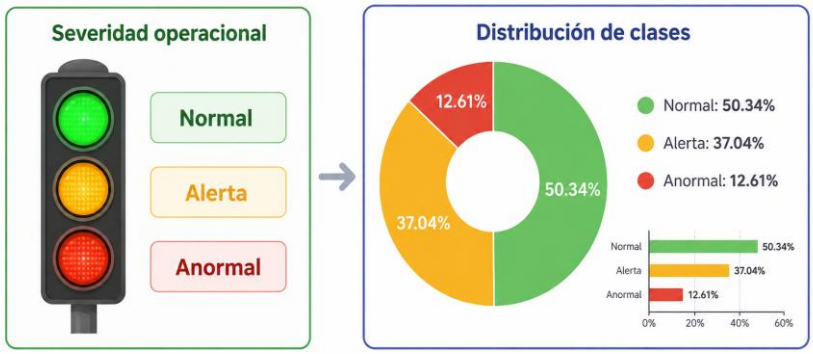
17

Dataset Industrial de Análisis de Aceite

20

Estimar el Estado de Alarma

Objetivo: clasificar la severidad operacional.



La alarma resume el estado de la muestra y permite priorizar acciones de mantenimiento.

21

No Todos los Errores Cuestan lo Mismo

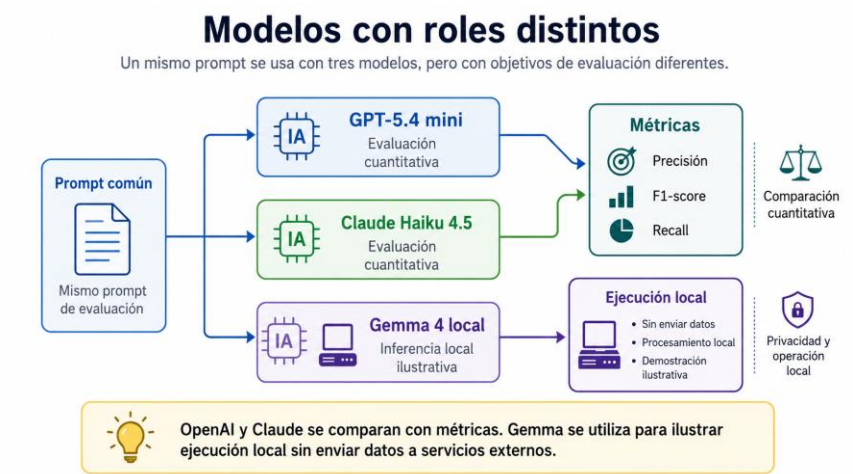
Una subestimación puede ser más grave que una falsa alarma.

22

Tres Modos de Uso del LLM

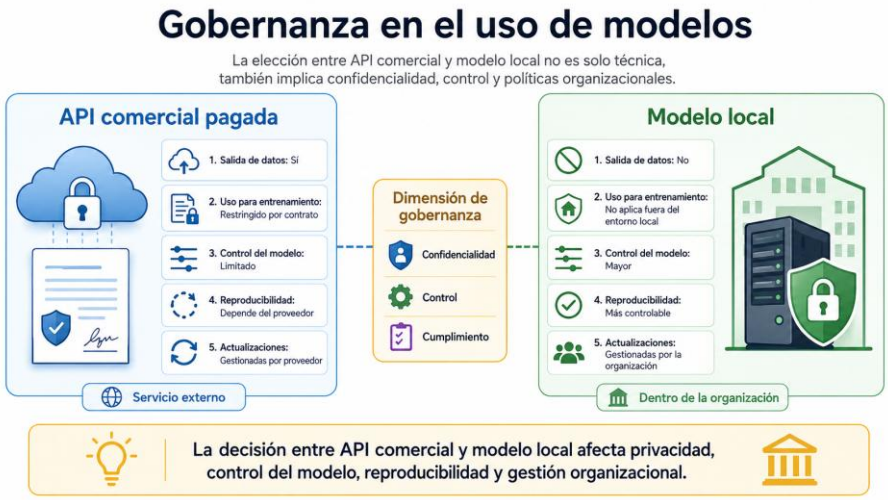
23

Modelos Usados



24

Propiedad Intelectual y Confidencialidad

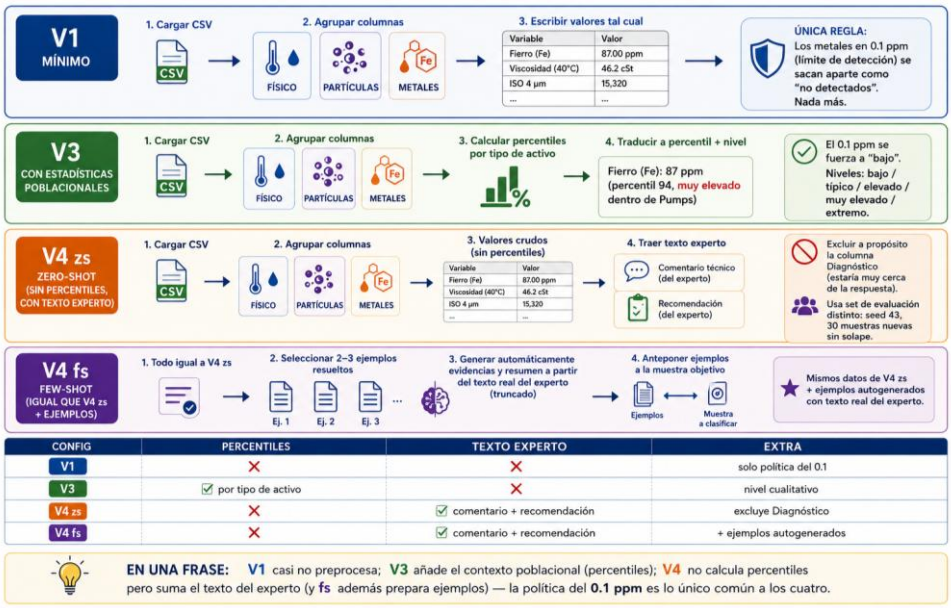


25

Separar Mediciones, Alarmas y Texto Experto

26

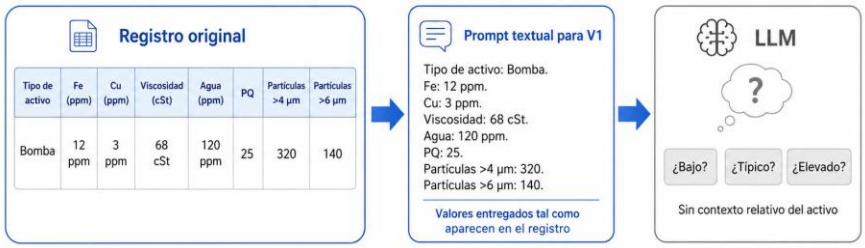
Preprocesamiento de la Data Numérica y Texto



27

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.



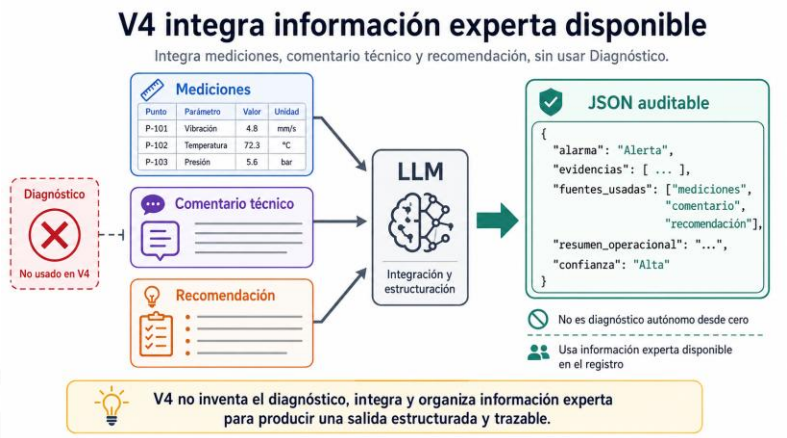
V1 establece un punto de partida, pero no entrega escala relativa ni juicio operacional.

28

V3: Criterio Operacional

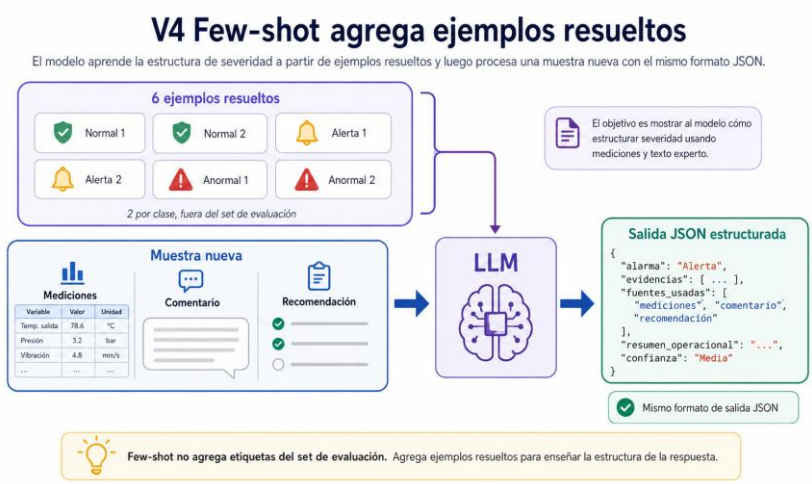
29

V4 Zero-Shot: Mediciones + Texto Experto



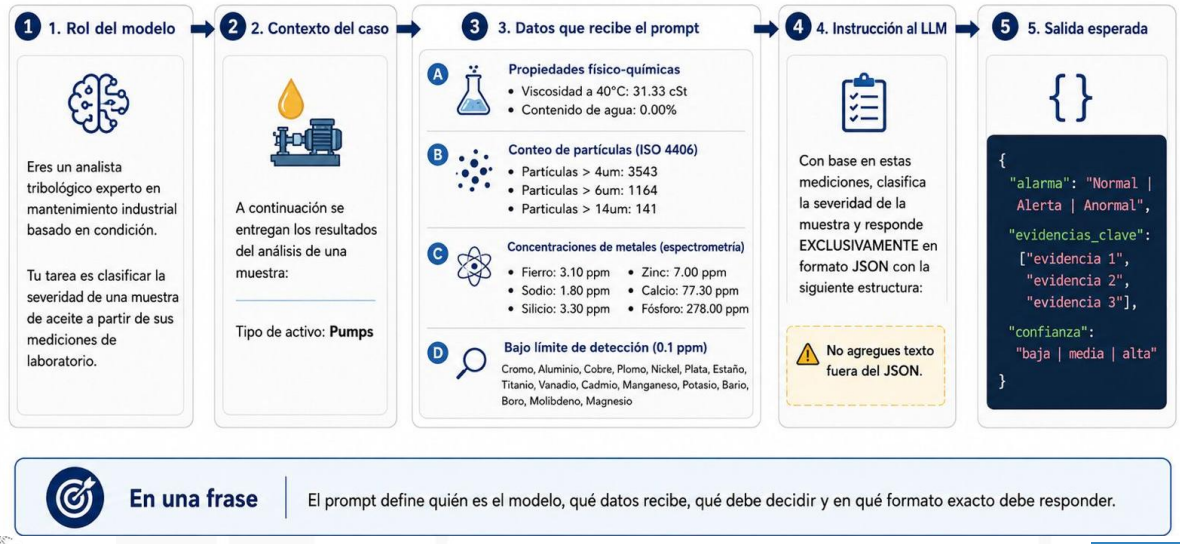
30

V4 Few-Shot: Ejemplos Resueltos (Observaciones)



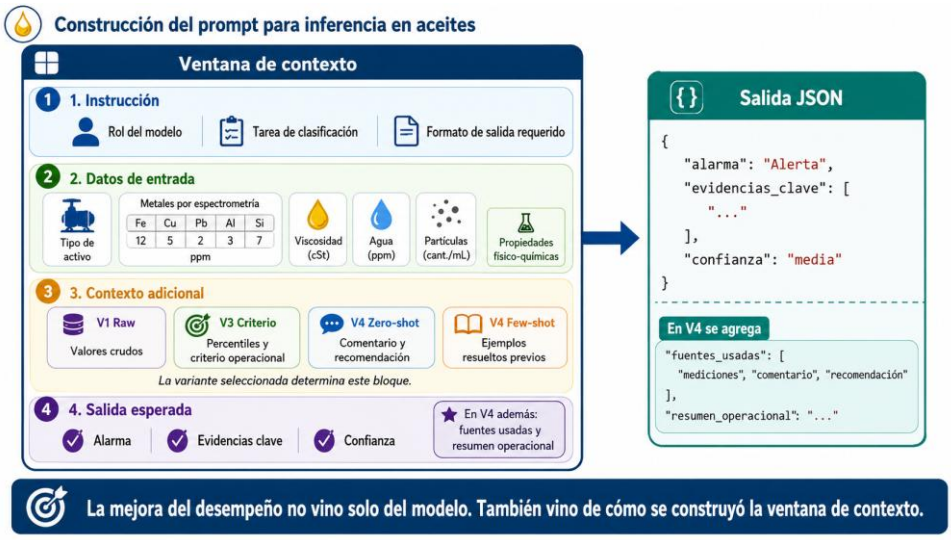
31

Prompt de Entrada para LLMs



33

Context Window



FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS UNIVERSIDAD DE CHILE

UCLA

34

Salidas Estructuradas y Auditables

FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS UNIVERSIDAD DE CHILE

UCLA

35

Demo en Vivo

1. Desarrollar las soluciones predictivas

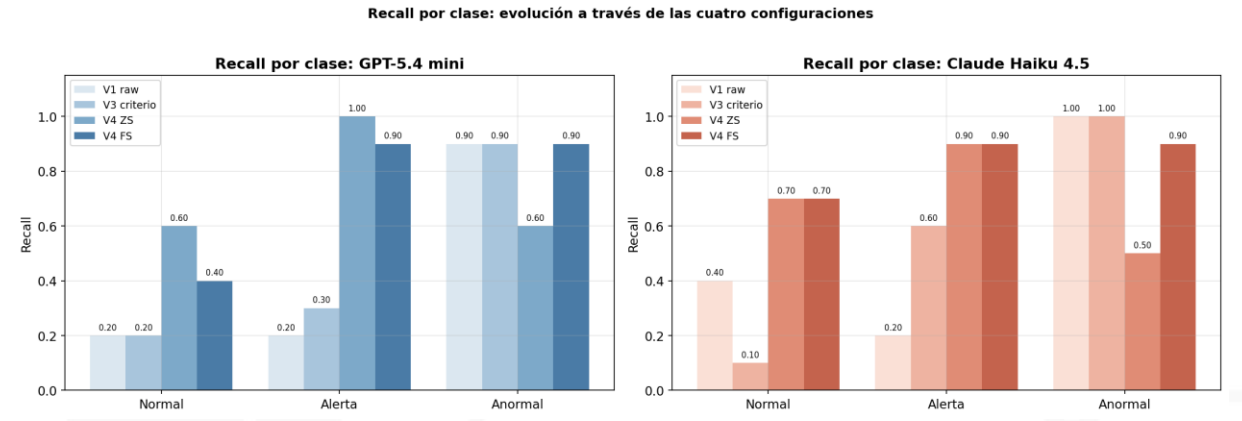
2. Interpretar los resultados

3. Esto se hace cargando el notebook:
   “02_oil_analysis_llm_demo.ipynb”

1. El notebook y sus librerías se encuentran disponibles en:
   “CMC_MX_2026_Toolbox_Code_Nube.zip”

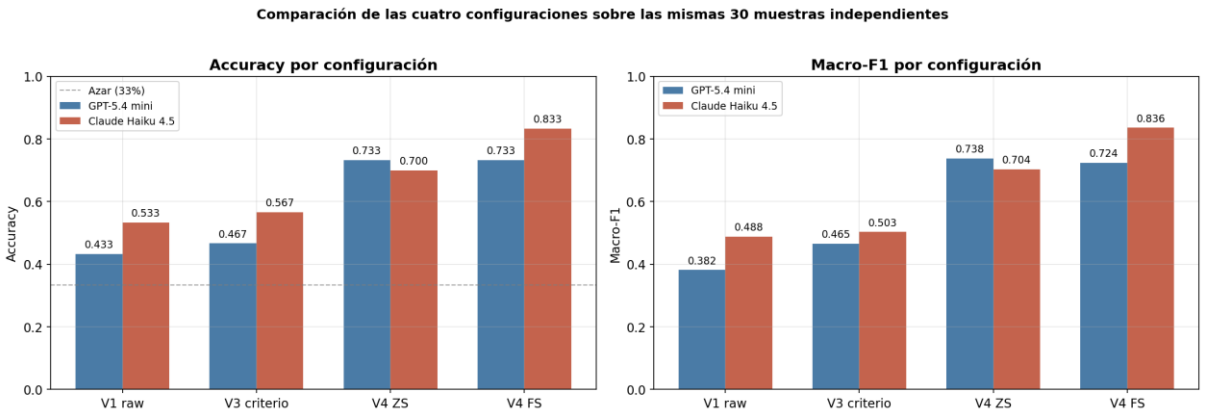
36

Resultados: Recall por Clase (Alarma)



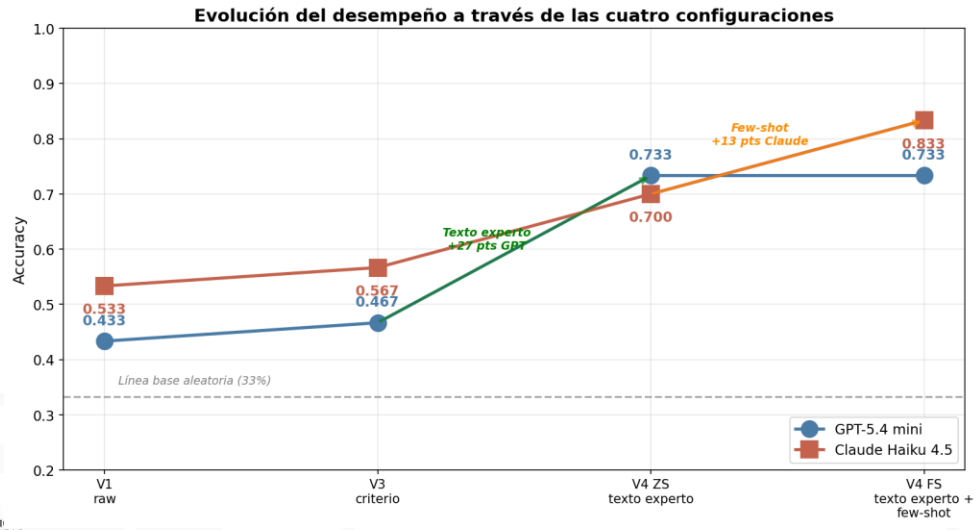
37

Resultados: Métricas de Performance



38

Resultados: Evolución de las Distintas LLMs Predictivas



39

Mejor LLM Predictiva Desarrollada

40

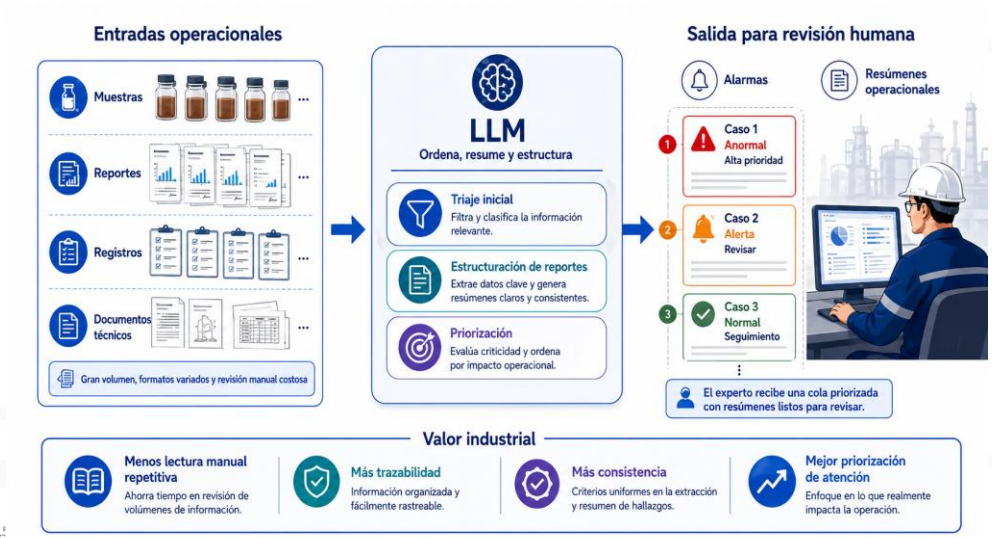
Riesgo Operacional: Claude Haiku 4.5

41

Qué Aprendimos de los Cuatro Modos de LLMs

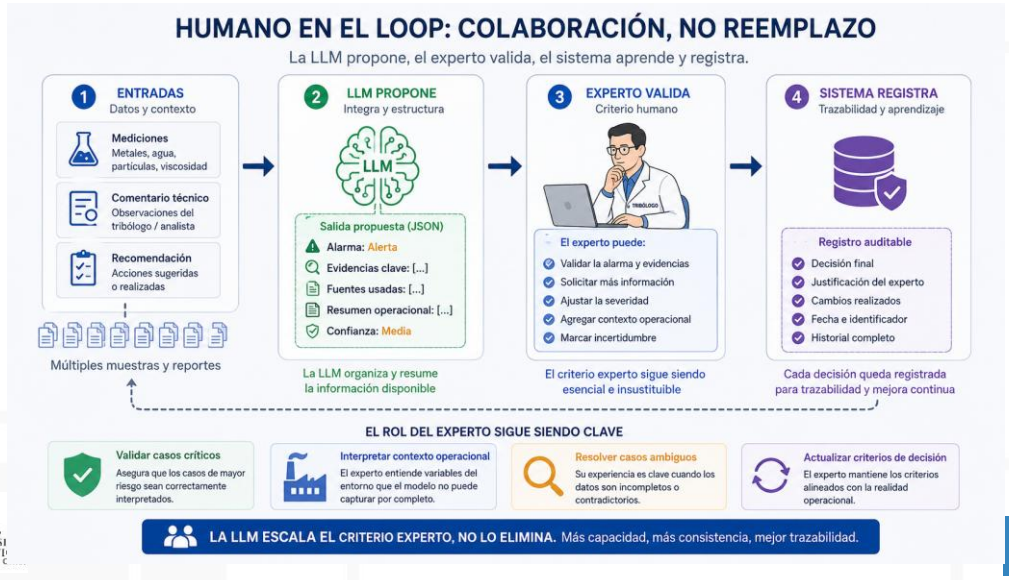
42

Qué Significa Esto para Mantención



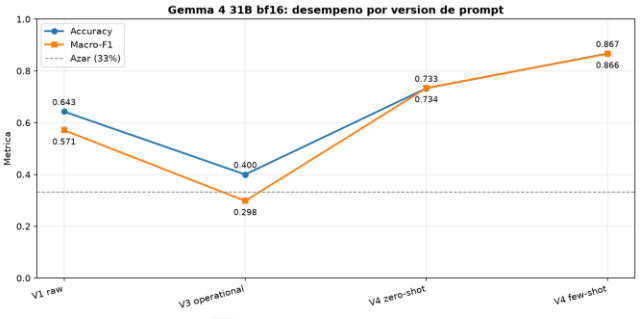
43

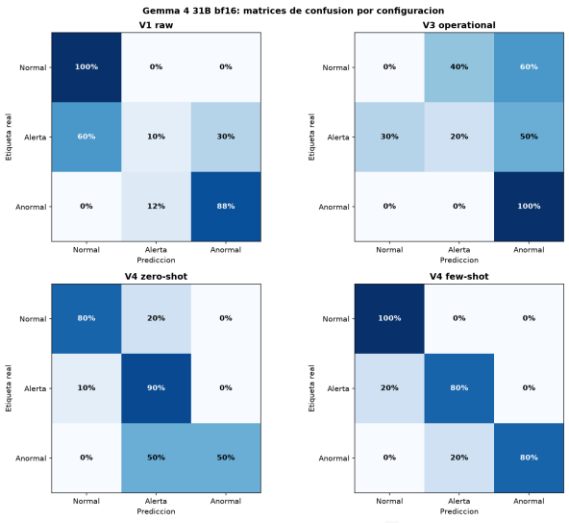
El Experto Sigue en el Sistema



44

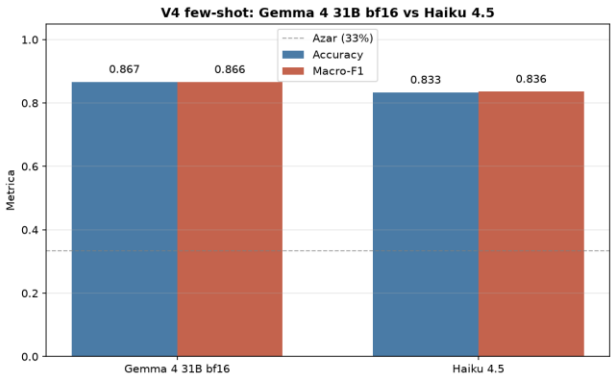
Qué Pasa Si Hacemos lo Mismo con un LLM Local?

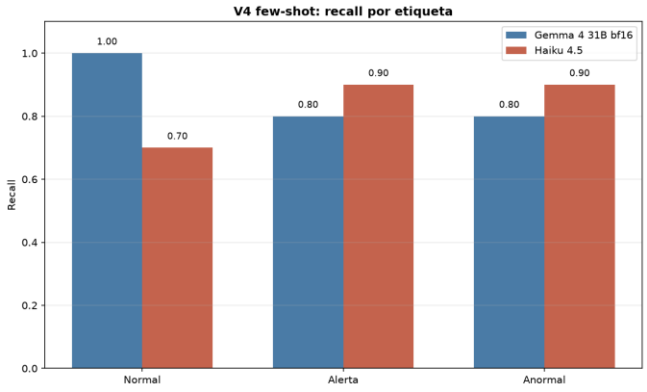




45

Resultados Comparables al Mejor Modelo Comercial



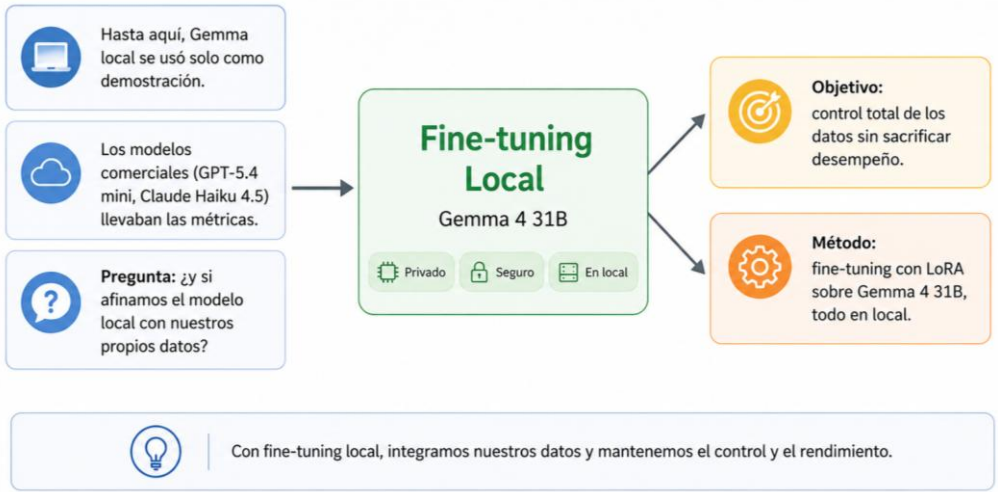


46

IA Local: Gemma 4

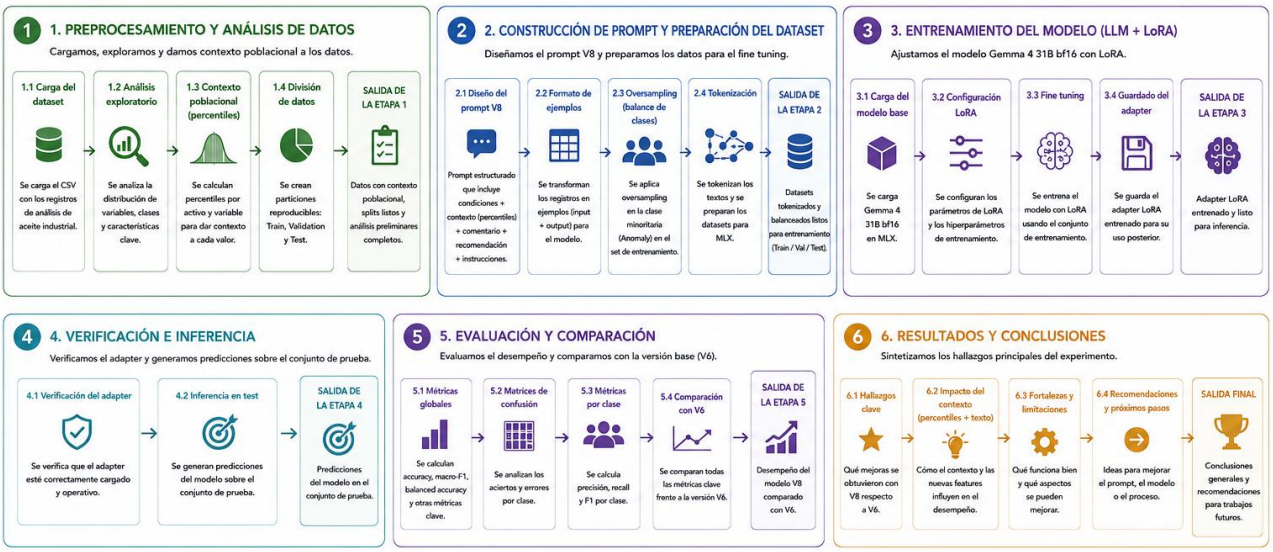
47

De Ilustrativo a Competitivo: Gemma Local Afinado



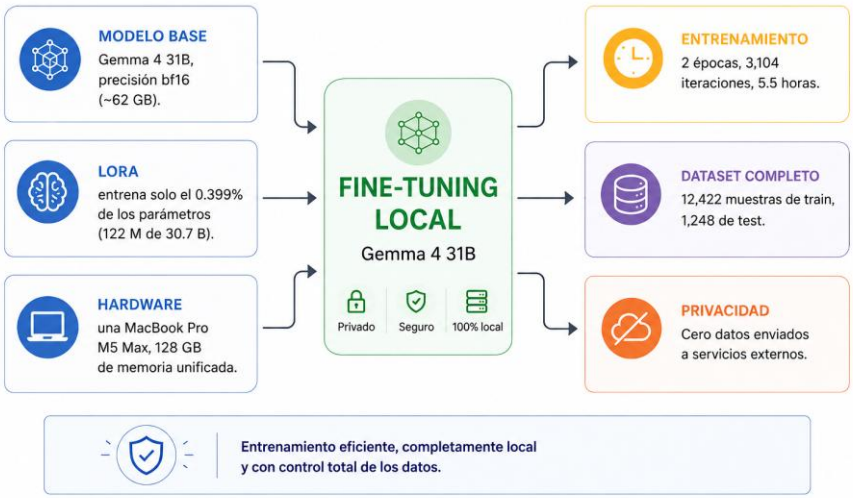
48

Procedimiento del Fine Tuning



49

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

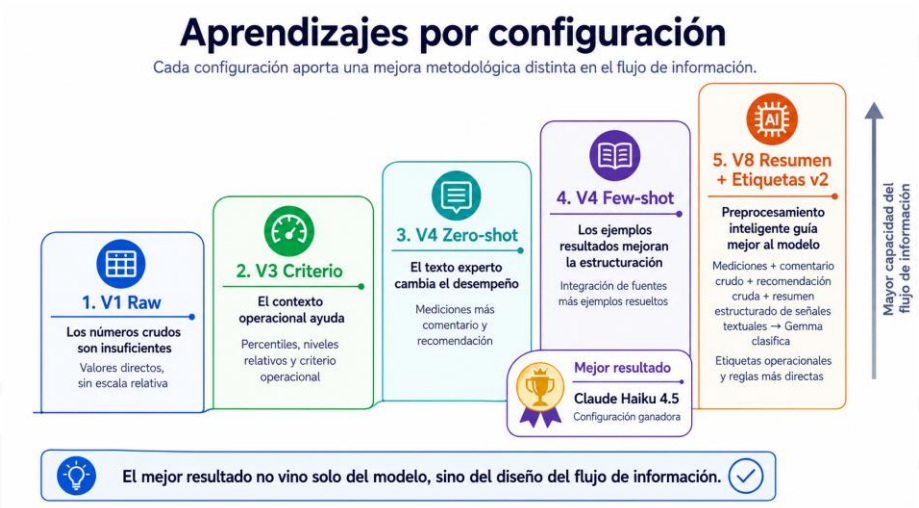


50

Importancia del Pre-Procesamiento

51

Tipo y Flujo de la Información



52

Demo en Vivo: Fine-Tuning Gemma 4

1. Desarrollar la solución predictiva
2. Interpretar los resultados
3. Esto se hace cargando el notebook:
   “09_v8_finetuning_demo.ipynb”

1. El notebook y sus librerías se encuentran disponibles en:
   “CMC_MX_2026_Toolbox_Code_AI_Local.zip”

53

Solución en base a AI 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.

54

Dónde Acierta y Dónde Falla

RENDIMIENTO DEL MODELO: MATRIZ DE CONFUSIÓN

Filas = clase real | Columnas = predicción

MATRIZ DE CONFUSIÓN

Real \ Predicho	Normal	Alerta	Anormal
Normal (627)	599 (95.5%)	28 (4.5%)	0 (0.0%)
Alerta (464)	10 (2.2%)	386 (83.2%)	68 (14.7%)
Anormal (157)	0 (0.0%)	6 (3.8%)	151 (96.2%)

HALLAZGOS CLAVE

- **Anormal a Normal** — 0 casos. Cero subestimaciones críticas.
- **Normal a Anormal** — 0 casos. Cero falsas alarmas extremas.
- **Detección de Anormales** — De 157 Anormales reales, detecta 151; pierde solo 6.
- **Errores residuales** — Los errores residuales son conservadores, no peligrosos.

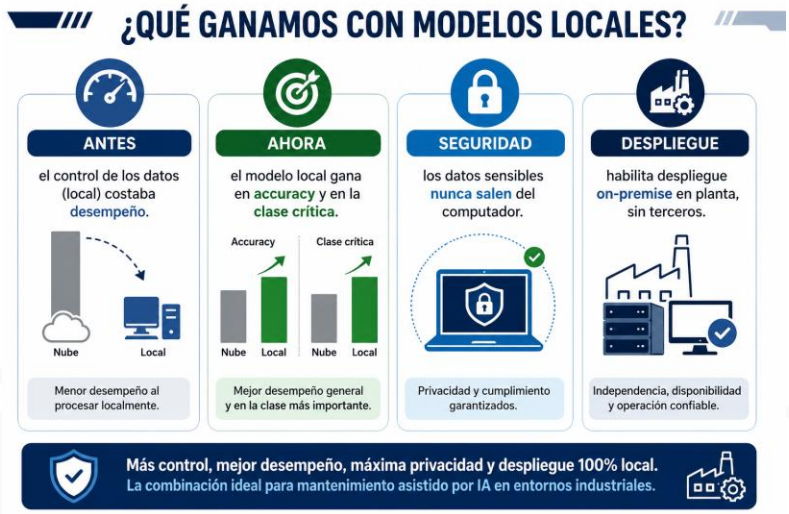
DE LA INFORMACIÓN A LA DECISIÓN

1. **Comprender** — Analizar la matriz e identificar patrones clave.
2. **Evaluar** — Validar que los errores sean conservadores.
3. **Confiar** — Tomar decisiones con base en resultados confiables.
4. **Mejorar** — Aprender y optimizar para seguir elevando el rendimiento.

Los resultados muestran un modelo confiable y seguro: cero subestimaciones críticas y cero falsas alarmas extremas.

55

Privacidad y Desempeño: Ya No Hay que Elegir



56

Conclusiones

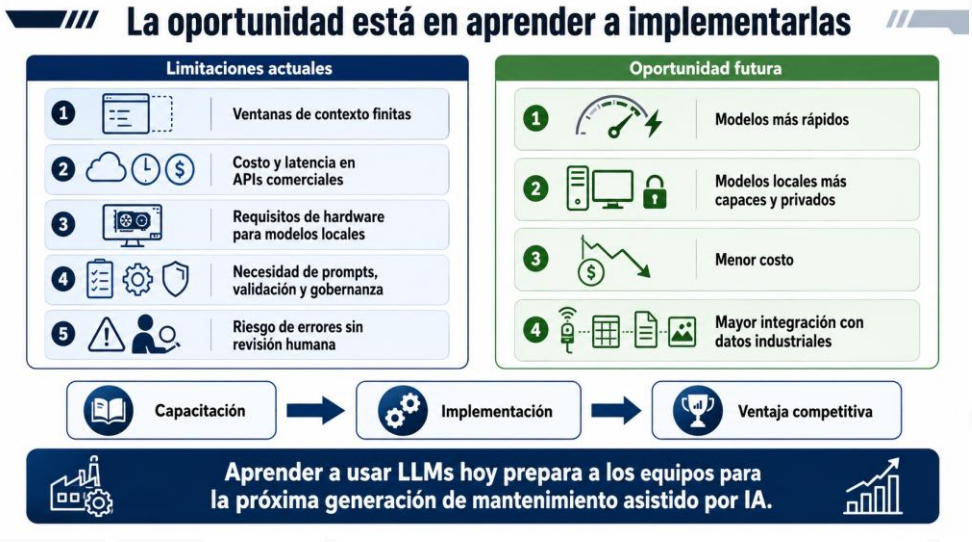
57

Qué Demostramos

- LLMs ya pueden apoyar el mantenimiento industrial:

58

Por Qué Aprender Esto Ahora ?



59


SESIÓN 20
ESCANEA EL CÓDIGO QR
RESPONDE UNA BREVE ENCUESTA
CONGRESO DE MANTENIMIENTO & CONFIABILIDAD MÉXICO
19ª EDICIÓN
¡GRACIAS!
Enrique López Droguett
eald@ucla.edu

60