TOOLBOX SESIÓN 13

CONGRESO DE MANTENIMIENTO & CONFIABILIDAD COLOMBIA 3ª 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

FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS UNIVERSIDAD DE CHILE

CONGRESO DE MANTENIMIENTO & CONFIABILIDAD COLOMBIA 3ª EDICIÓN

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

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

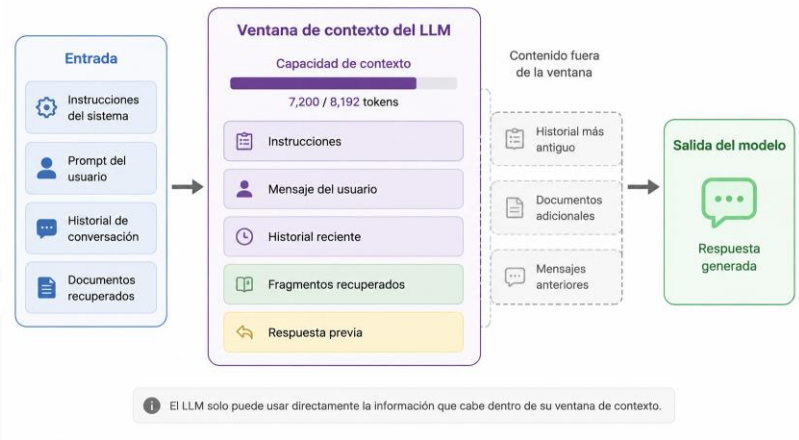
- Cada token se representa como un vector numérico.
- Estos vectores permiten que el modelo capture relaciones entre conceptos técnicos.
- Conceptos relacionados tienden a quedar cerca entre sí:
  - Desgaste
  - Degradación
  - Falla

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

7

Context Window

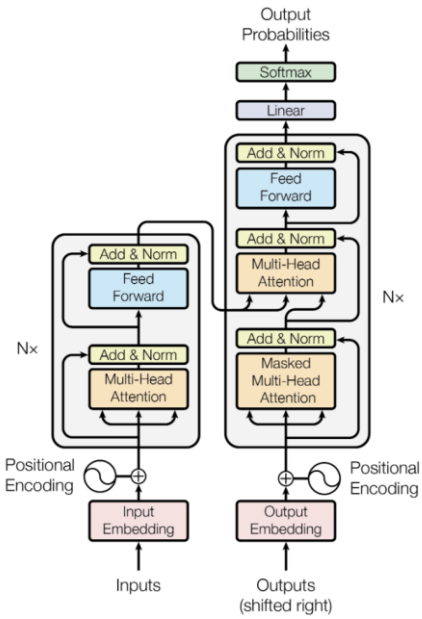
- Es la cantidad de información que el modelo puede considerar en una interacción.



8

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.

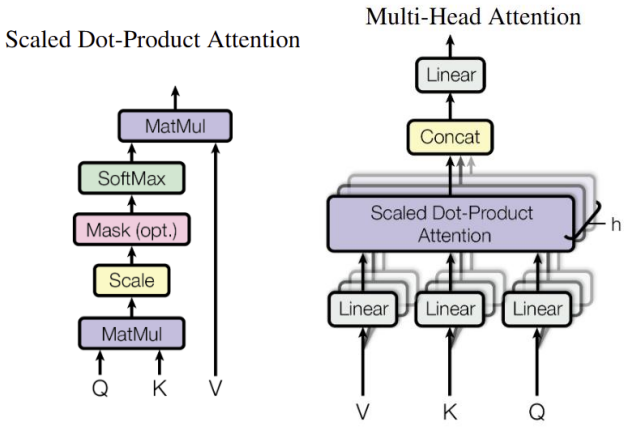


Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention Is All You Need 2017. https://doi.org/10.48550/arXiv.1706.03762.

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



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.

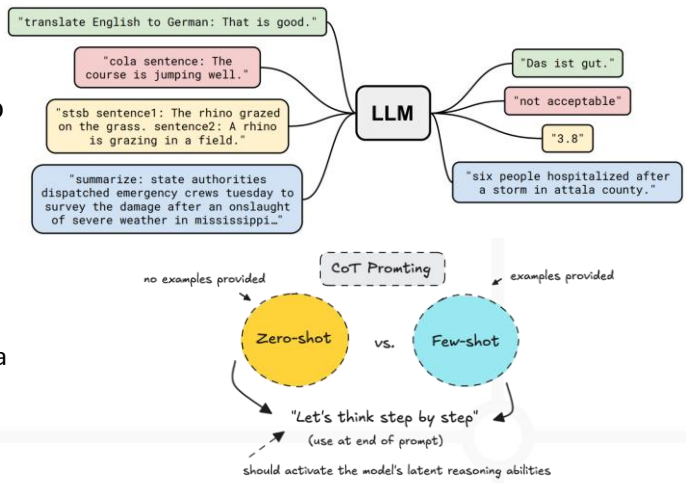
10

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.

Prompting

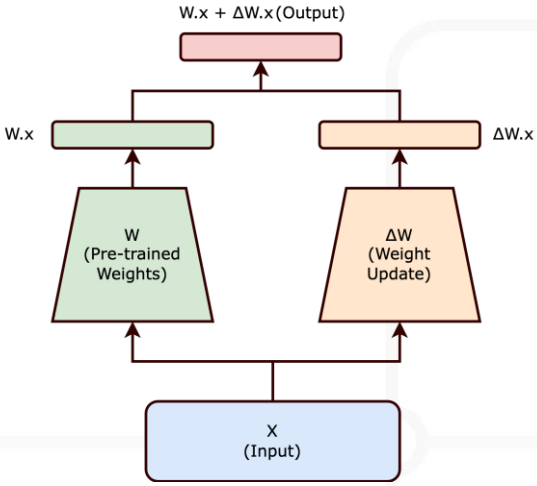
- Prompting consiste en guiar al modelo mediante instrucciones, contexto y ejemplos. No modifica lo pesos del modelo. Estrategias comunes:
  - Zero-shot: instrucción directa sin ejemplos
  - Few-shot: instrucción con ejemplos resueltos
  - Razonamiento estructurado: respuesta organizada en pasos o campos



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

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/.

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

14

Modelos Comerciales y Open Source

Existen dos grandes formas de usar LLMs en ingeniería:

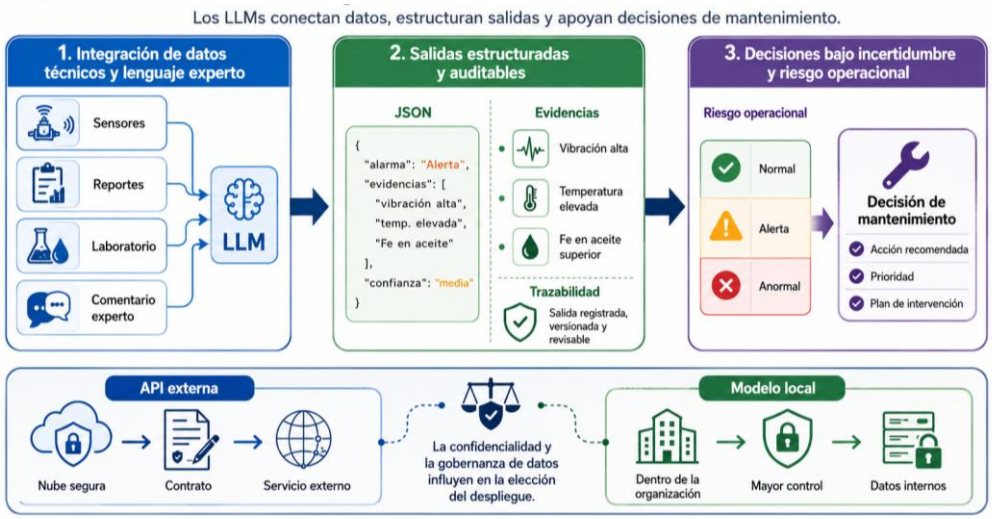
1. **Modelos Comerciales:**
   - Alta capacidad general
   - Uso inmediato
   - Acceso mediante web o API

Modelos Comerciales y Open Source

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

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

Relevancia para Mantenimiento Industrial

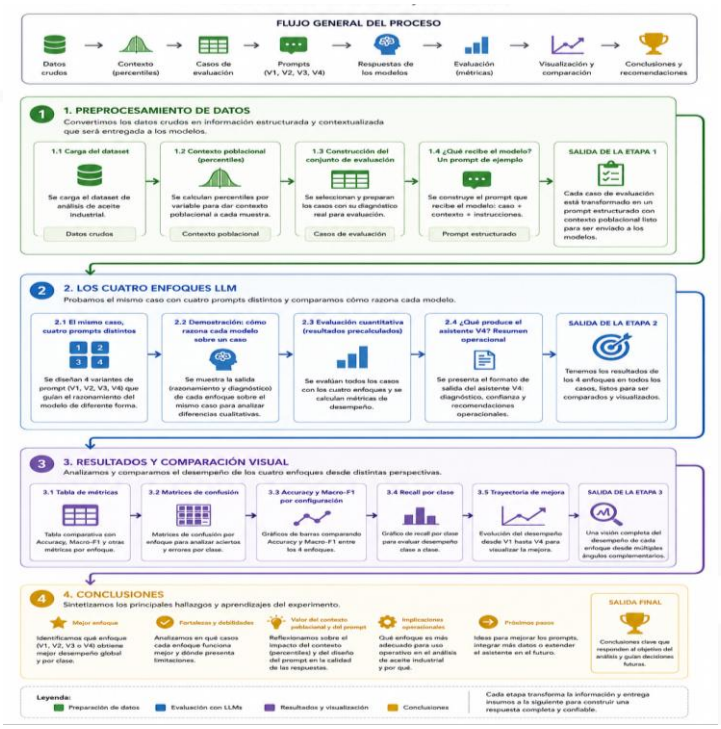


17

Análisis de Aceite con LLMs

18

Procedimiento



19

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

20

Del Laboratorio a la Decisión de Mantenimiento

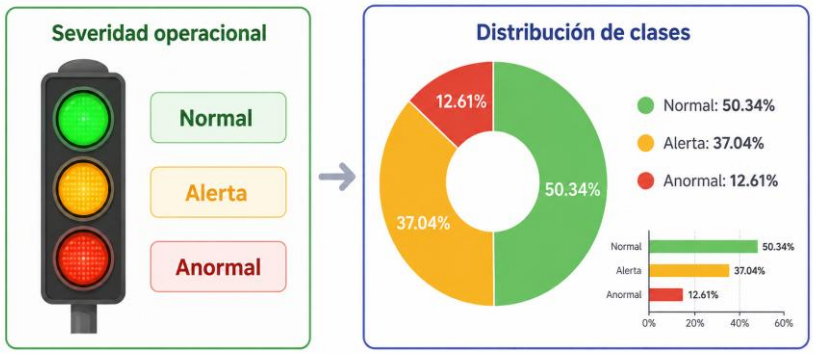
21

Dataset Industrial de Análisis de Aceite

22

Estimar el Astado de Alarma

Objetivo: clasificar la severidad operacional.



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

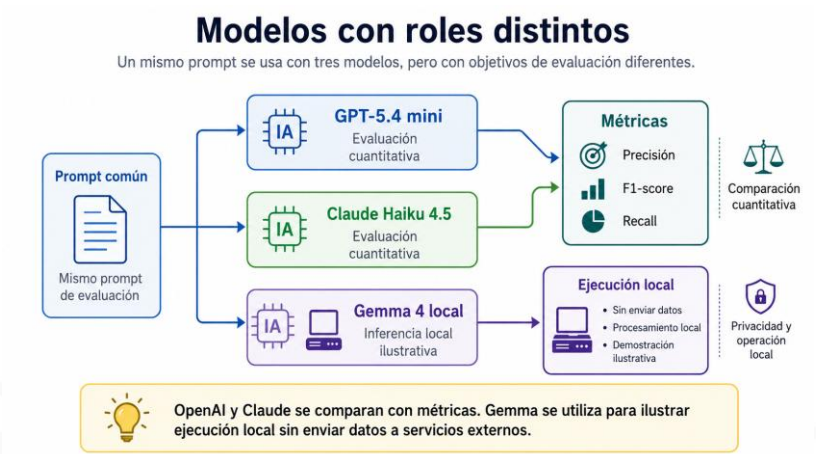
No todos los Errores Cuestan lo Mismo

Una subestimación puede ser más grave que una falsa alarma.

Tres Modos de Uso del LLM

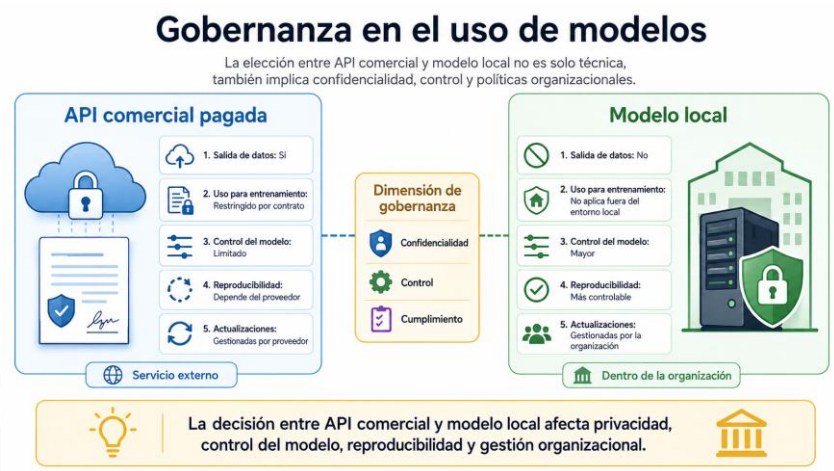
25

Modelos Usados



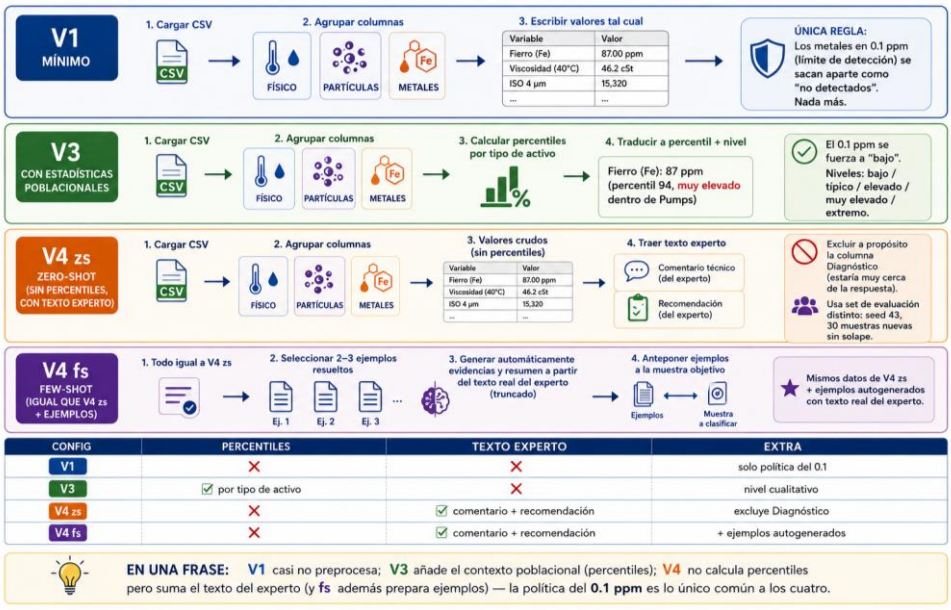
26

Propiedad Intelectual y Confidencialidad



Separar Mediciones, Alarmas y Texto Experto

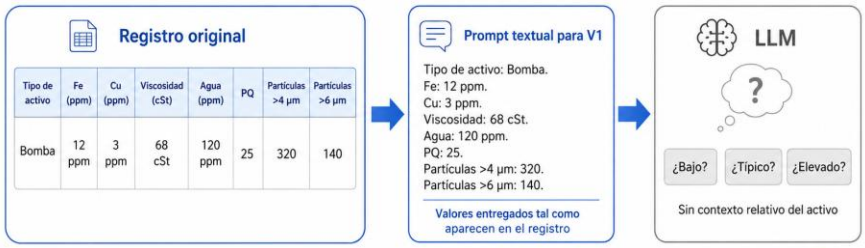
Preprocesamiento de la Data Numérica y Texto



29

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



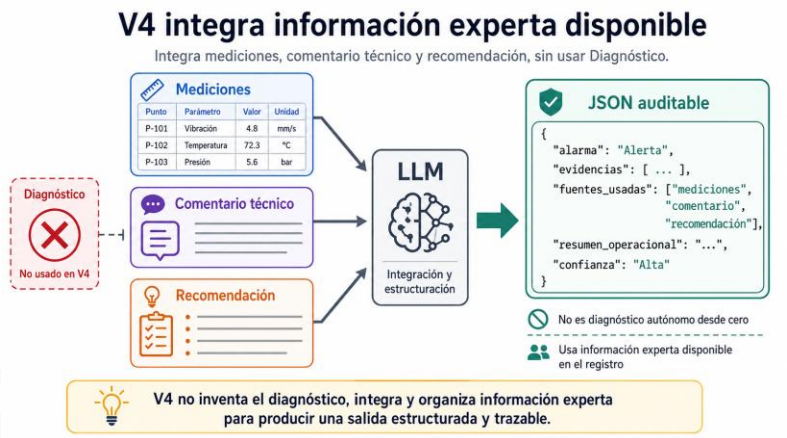
V1 establece un punto de partida, pero no entrega escala relativa ni juicio operacional.

30

V3: Criterio Operacional

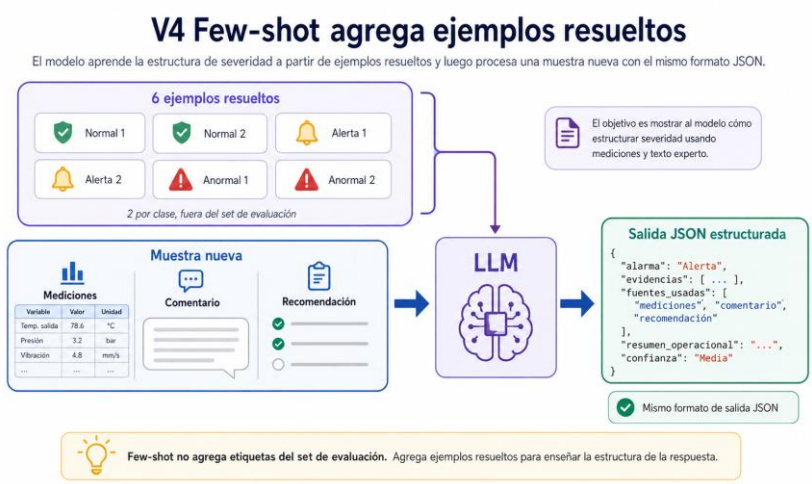
31

V4: Mediciones + Texto Experto



32

V4 Few-Shot: Ejemplos Resueltos (Observaciones)



33

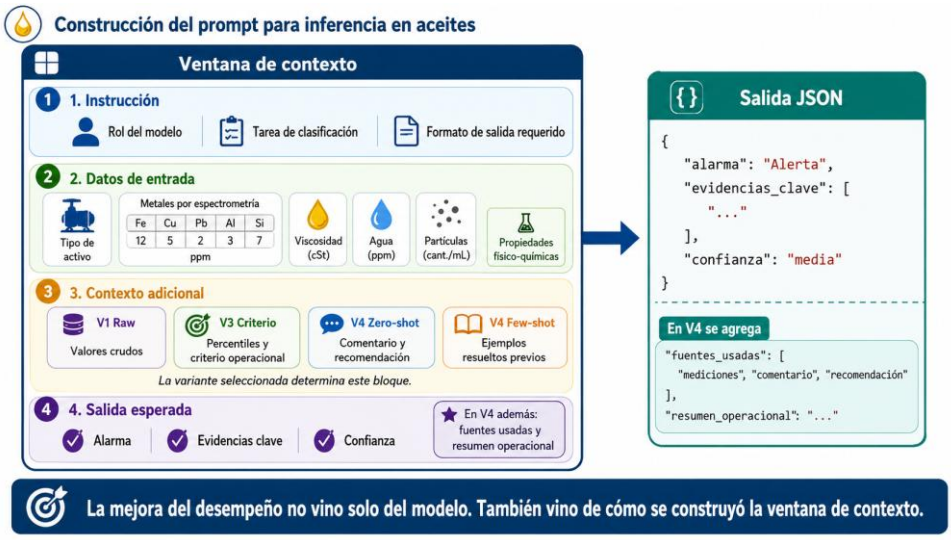
Construcción de los Prompts

34

Salidas Estructuradas y Auditables

35

Context Window



36

Resumen: LLMs en Mantenimiento Predictivo

37

Demo en Vivo

1. Desarrollar las soluciones predictivas

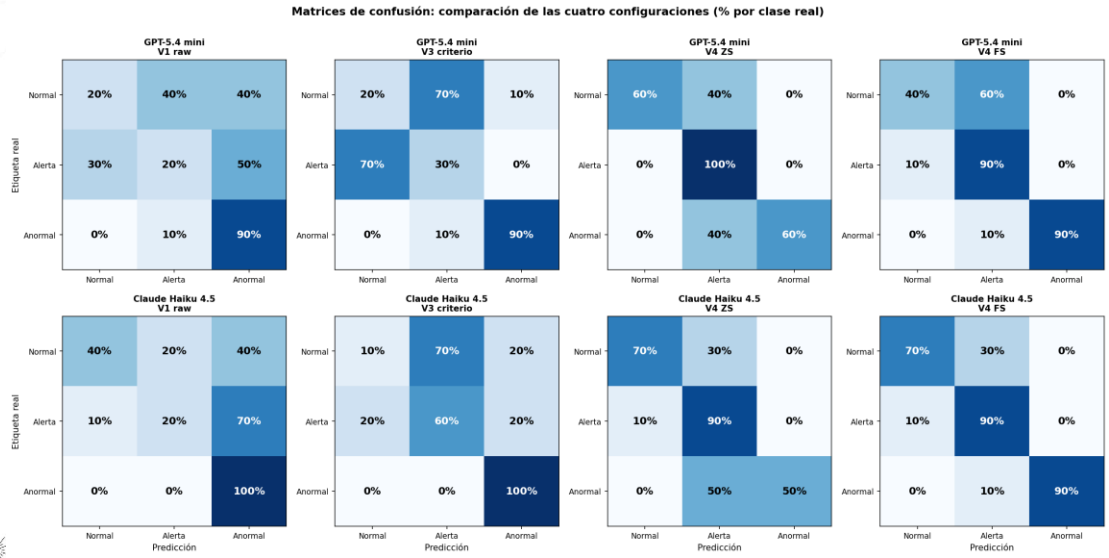
2. Interpretar los resultados

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

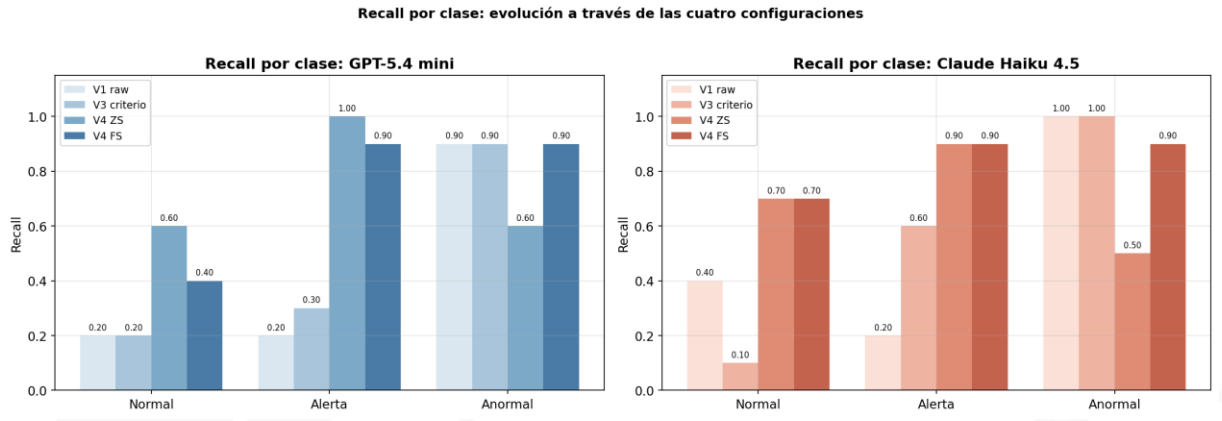
1. El notebook y sus librerías se encuentran disponibles en:
   “CMC_CO_2026_Toolbox_Code.zip”

38

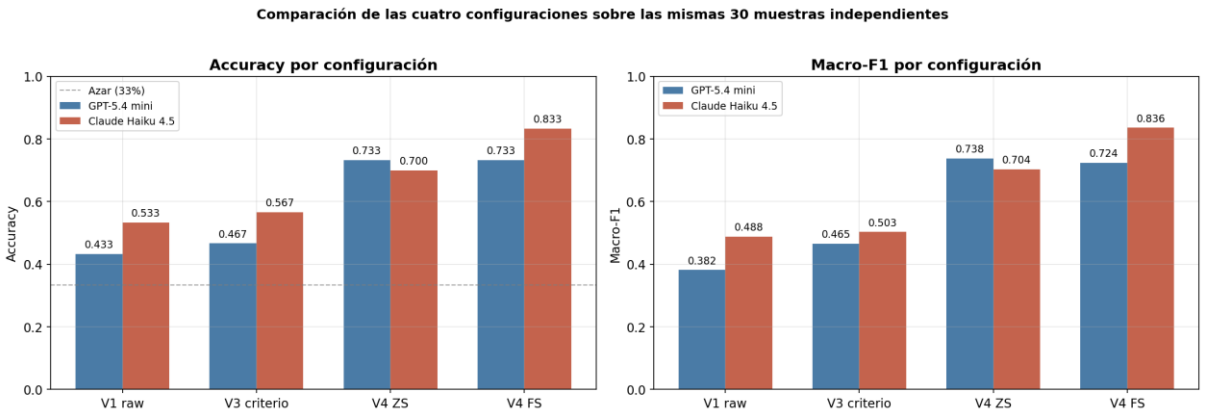
Resultados: Matrices de Confusión



Resultados: Recall por Clase (Alarma)

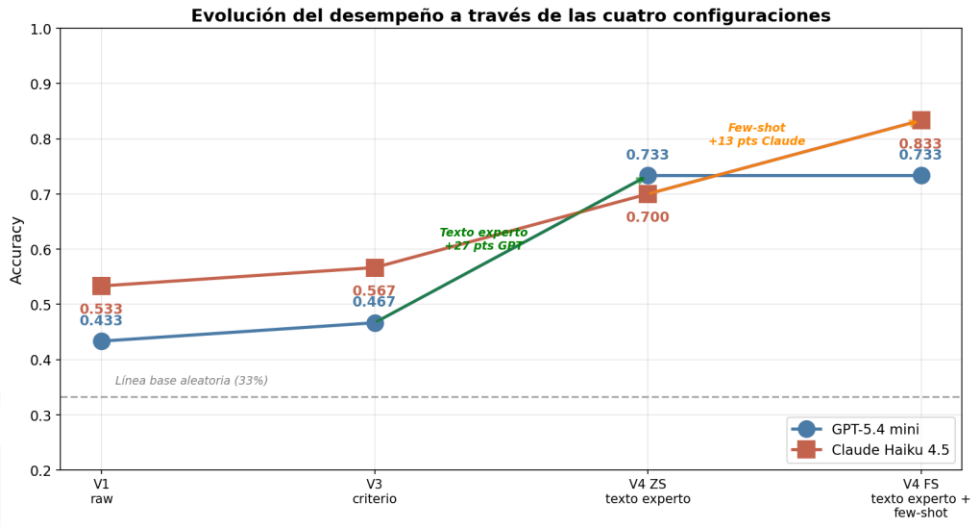


Resultados: Métricas de Performance



41

Resultados: Evolución de las Distintas LLMs Predictivas

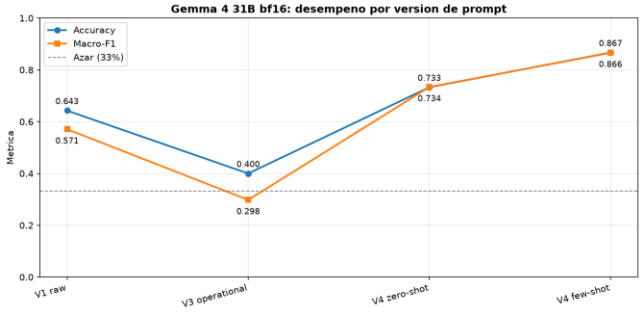


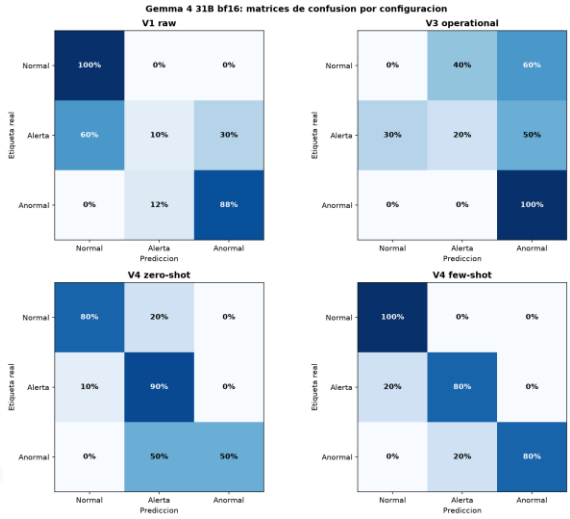
42

Mejor LLM Predictiva Desarrollada

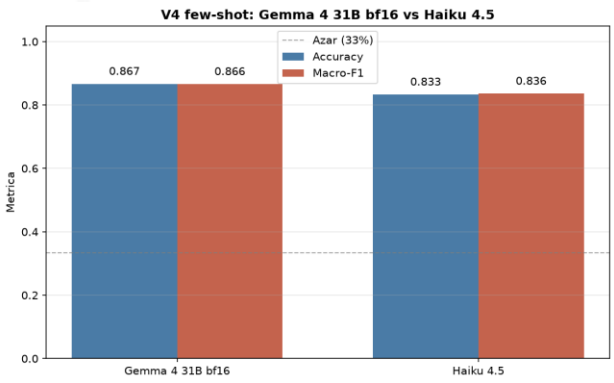
Riesgo Operacional: Claude Haiku 4.5

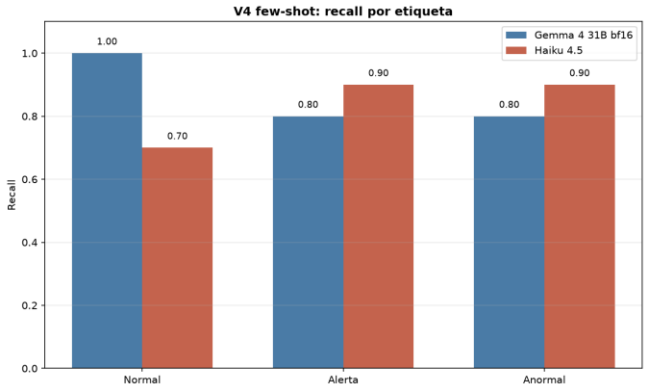
Qué Pasa Si Hacemos el Mismo Ejercicio con un LLM Local?





Resultados Comparables al Mejor Modelo Comercial





Qué Aprendimos de los Cuatro Modos de LLMs

47

Qué Significa Esto para Mantención Predictiva

48

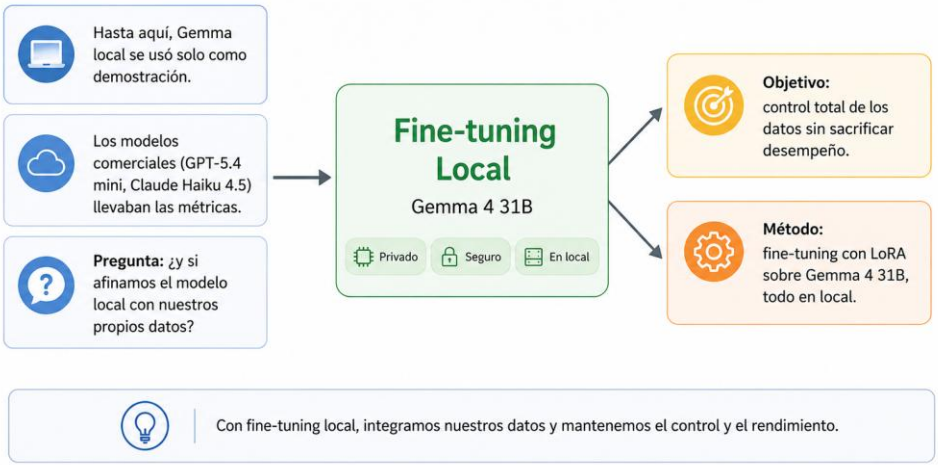
El Experto Sigue en el Sistema

49

Inteligencia Artificial Local (Local AI)

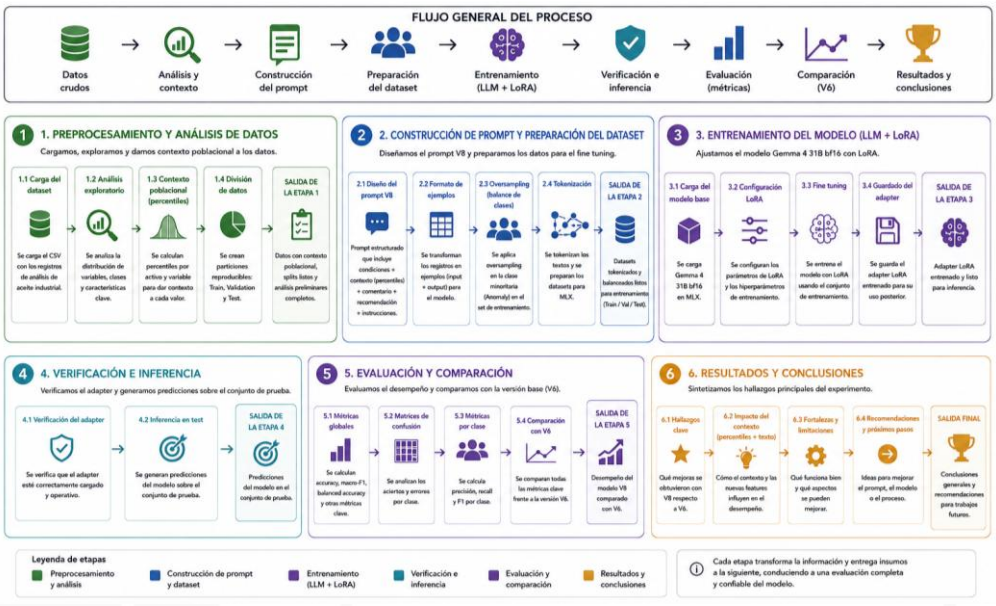
50

De Ilustrativo a Competitivo: Gemma Local Afinado



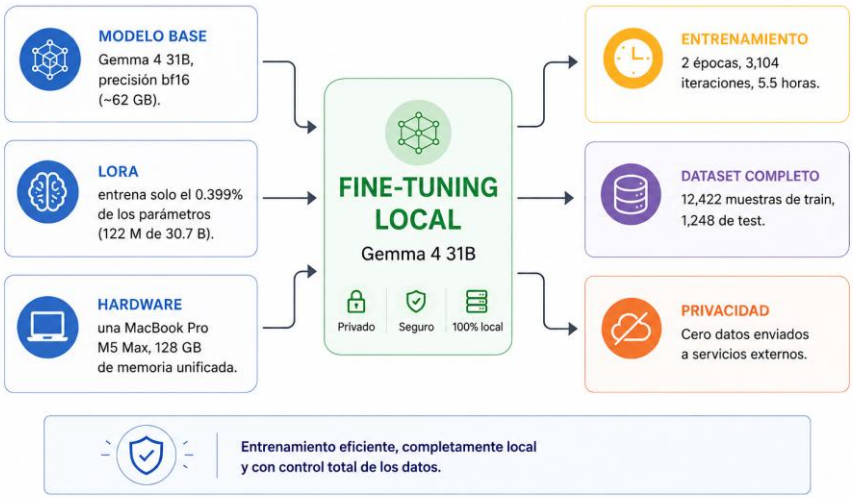
51

Procedimiento del Fine Tuning



52

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



53

Importancia del Pre-Procesamiento

54

Tipo y Flujo de la Información

55

Modelo Local Supera a los Comerciales

Resultados de evaluación

Comparación de desempeño en el test completo

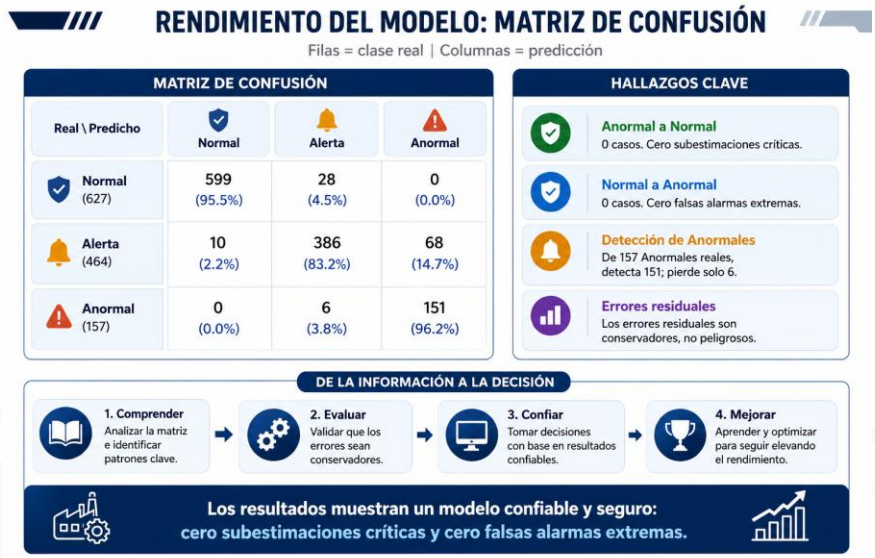
Enfoque	Accuracy	Recall Anormal
Red neuronal (baseline del advisor)	85.00%	90.6%
Mejor comercial: Claude Haiku 4.5 (V4 Few-shot)	83.3%	90%
Gemma 4 31B local, afinado (V8)	91.03%	96.2%

- **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.

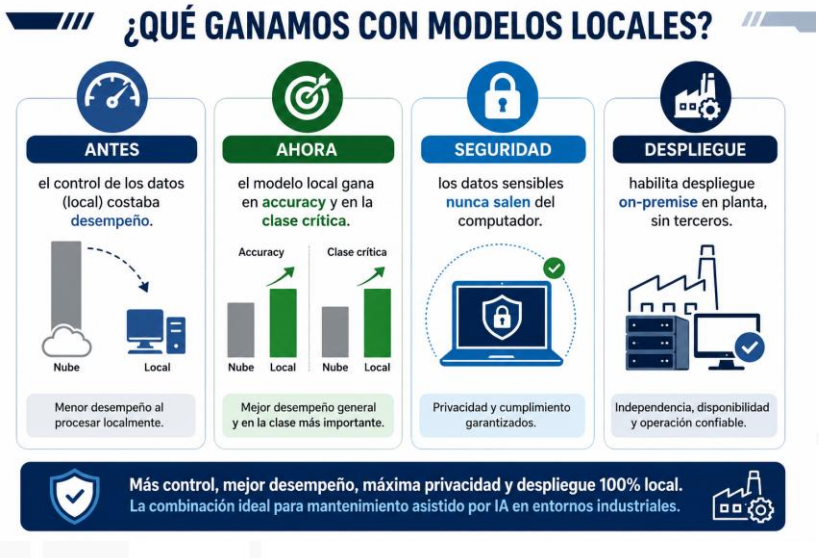
56

Dónde Acierta y Dónde Falla



57

Privacidad y Desempeño: Ya No Hay que Elegir



58

Demo en Vivo: Fine-Tuning Gemma 4

1. Desarrollar la solucione 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”

59

Conclusiones

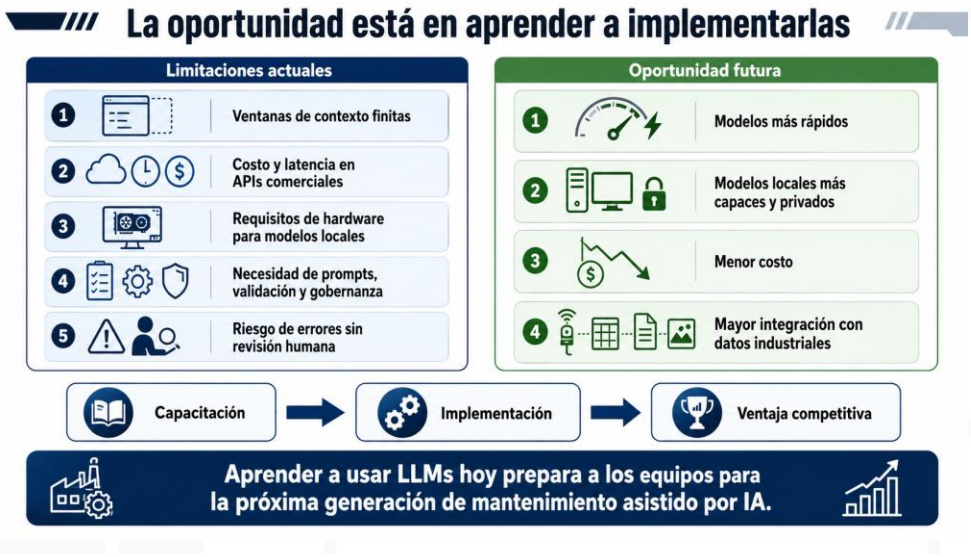
60

Qué Demostramos

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

61

Por Qué Aprender Esto Ahora ?



62


SESIÓN 13
ESCANEA EL CÓDIGO QR
RESPONDE UNA BREVE ENCUESTA
CONGRESO DE MANTENIMIENTO & CONFIABILIDAD COLOMBIA
3ª EDICIÓN
¡GRACIAS!
Enrique López Droguett
eald@ucla.edu

63