LLMs en la Industria

Oportunidades, Riesgos y Limitaciones

en entornos industriales de ingeniería

Rafael Larraz

Basado en: Prompt Engineering and the Transformation of Petroleum Refining

Springer, 2026

Agenda

01 **¿Qué son los LLMs y cómo funcionan?**
Fundamentos para ingenieros — 5 min

02 **Cómo comunicarnos con ellos**
Prompts, ventana de contexto, RAG, MCP — 7 min

03 **Problemas y limitaciones**
Alucinaciones, confidencialidad, ética — 8 min

04 **Aplicaciones en ingeniería**
Estado del arte con bibliografía verificada — 10 min

¿Qué es un LLM?

Definición, historia y tipos de arquitectura

Red neuronal entrenada con cantidades masivas de texto

Large (miles de millones de parámetros) • Language (lenguaje natural) • Model (aprende patrones)

Historia y Evolución

2013
Word2Vec
Embeddings

→

2017
Transformer
"Attention Is All You Need"

→

2020
GPT-3
175B params

→

2022-26
ChatGPT, Claude, Gemini
Multimodal

ENCODER-ONLY

BERT, RoBERTa

Bidireccional
Clasificación

DECODER-ONLY ★

GPT, Claude, Llama

Causal (unidireccional)
Arquitectura dominante

ENCODER-DECODER

T5, BART

Combina ambos
Traducción

¿Cómo Funciona un LLM?

Ejemplo: "El gato está en el ___"

1. INPUT

"El gato está en el"

→

2. TOKENIZACIÓN

"El" "gato" "está" "en" "el"

→

3. EMBEDDINGS

[0.23, -0.45, 0.89...]
[0.67, 0.34, -0.21...]

→

4. TRANSFORMER

Atención: "gato" 0.50
+ FFN (×12-96 capas)

→

5. PREDICCIÓN

"tejado" (45%)
"suelo" (30%)

→ BUCLE: El token generado se añade al input y el proceso se repite hasta completar la respuesta

Atención: ¿qué palabras importan?

Para predecir después de "el":

"El" atención: 0.05

"gato" atención ALTA: 0.50

"está" atención: 0.10

"en" atención alta: 0.35

Concepto clave

Los LLMs no "entienden" — predicen el siguiente token basándose en patrones estadísticos.

Pero a escala masiva, esta predicción captura relaciones semánticas profundas entre conceptos.

El proceso es AUTOREGRESIVO: genera un token, lo añade al input, y repite. Así construye respuestas completas palabra a palabra.

¿Cómo se Entrena un LLM?

Proceso completo desde datos hasta modelo desplegado

FASE 1: PRE-ENTRENAMIENTO

Aprender patrones del lenguaje a partir de datos masivos

1.1 Recolección

Libros, Internet, Wikipedia, Código, Artículos científicos

→

1.2 Limpieza

Eliminar duplicados, filtrar, normalizar, balance de datos

→

1.3 Tokenización

Dividir en tokens, crear vocabulario (50k-100k tokens)

→

1.4 Entrenamiento

Predecir siguiente token (Next Token Prediction)

FASE 2: FINE-TUNING

Supervised FT

Datos de alta calidad
Instrucciones + Respuestas

→

RLHF

Feedback humano
Alinear con preferencias

FASE 3: EVALUACION Y DESPLIEGUE

Evaluación

Benchmarks, seguridad
Tests de alucinación

→

Despliegue

Optimización, APIs
Monitoreo continuo

Recursos necesarios

Coste: €2M - €100M+ (modelos grandes) | Hardware: Miles de GPUs, semanas/meses | Tiempo: 6-18 meses desde inicio

02

Cómo Comunicarnos con Ellos

Prompts, ventana de contexto, RAG y MCP

Prompt Engineering

La interfaz con los LLMs — cómo comunicarse efectivamente

¿Qué es un Prompt? La entrada que le das al modelo. Determina completamente la salida. Cada palabra importa.

Anatomía de un Buen Prompt

1. CONTEXTO / ROL

"Eres ingeniero de procesos con 20 años..."

2. INSTRUCCIÓN CLARA

"Analiza causas de desactivación del catalizador..."

3. FORMATO DESEADO

"Responde en tabla con 3 columnas..."

4. EJEMPLOS (opcional)

"Ejemplo: presión alta → verificar válvula"

Técnicas Principales

Zero-Shot
Sin ejemplos — instrucciones directas

Few-Shot Learning
2-5 ejemplos para guiar formato

Chain of Thought
"Piensa paso a paso" — razonamiento

Prompt Chaining
Divide en prompts secuenciales

Structured Output
JSON, tablas, formatos específicos

¿Cuándo Usar Cada Técnica?

Técnica	Complejidad	Mejor para...	Ejemplo industrial
Zero-Shot	Baja	Consultas rápidas	"¿Rango de temp. de un reformador?"
Few-Shot	Media	Formato, mayor precisión	Clasificar tipos de fallo con ejemplos
Chain of Thought	Alta	Problemas complejos	Diagnóstico de desactivación FCC
Prompt Chaining	Alta	Workflows multi-etapa	Riesgo → mitigación → informe
Structured Output	Baja	Integración, APIs	Extraer parámetros en JSON

Empieza simple (Zero-Shot), aumenta complejidad solo si necesario. Para tareas críticas: temp 0.1-0.3. Para creatividad: temp 0.7-0.9.

Context Window y Parámetros

La "memoria de trabajo" del LLM y cómo controlar su comportamiento

COMPUTADORA

RAM (8-32 GB)
Datos en uso AHORA

Disco Duro (1-2 TB)
Conocimiento permanente

CPU carga datos del disco a RAM
RAM limitada = no todo a la vez

LLM

Context Window (8k-200k+ tokens)
Tu prompt + conversación AHORA

Parámetros (7B-405B+)
Conocimiento aprendido (fijo)

Modelo aplica pesos al contexto
Contexto limitado = "olvida"

"El context window es como la RAM de una computadora" — Analogía de Andrej Karpathy

Temperature

0.0 = determinista
1.0+ = creativo

Top-K / Top-P

Limita palabras
candidatas

Max Tokens

Longitud máxima
de respuesta

Freq. Penalty

Penaliza
repeticiones

RAG y MCP: Ampliando Capacidades

RAG: Retrieval-Augmented Generation

Recuperar información relevante de tus documentos + dársela al LLM como contexto

1. Documentos
PDFs, manuales, base de datos

→

2. Chunking
Dividir en fragmentos

→

3. Embeddings
Convertir a vectores

→

4. Vector DB
Almacenar para búsqueda rápida

→

5. Retrieval
Buscar similares + inyectar en prompt

MCP: Model Context Protocol

Protocolo estándar (Anthropic, 2024) que permite a los LLMs conectarse con herramientas y fuentes de datos externas

CLIENTE
(Claude, GPT-4)

⟷

PROTOCOLO
MCP

⟷

SERVIDORES
Google Drive · Gmail · HYSYS
Databases · APIs

03

Problemas y Limitaciones

Alucinaciones, confidencialidad, ética y riesgos

Las Zonas de Confianza de un LLM

Basado en frecuencia de datos de entrenamiento

ZONA SEGURA

- Alta frecuencia en datos
- Alta confianza del modelo
- Bajo riesgo de alucinación
- *Ej: principios termodinámicos, reacciones conocidas*

ZONA DE INCERTIDUMBRE

- Frecuencia media en datos
- Mezcla hechos con invención
- Requiere verificación
- *Ej: parámetros de tecnologías emergentes*

ZONA DE ALUCINACIÓN

- Baja/nula frecuencia
- El modelo "confabula"
- Alto riesgo de error
- *Ej: datos económicos de tecnologías piloto*

← Alta frecuencia en datos de entrenamiento

Baja frecuencia →

Alucinaciones en Ingeniería

Información aparentemente fiable pero incorrecta — el riesgo más crítico

Cálculos numéricos

Errores en eficiencia energética, dimensionamiento de equipos y estimaciones económicas

Optimismo tecnológico

Tendencia sistemática a sobreestimar madurez y viabilidad de tecnologías emergentes

Datos y referencias fabricados

Generación de citas bibliográficas y métricas de rendimiento plausibles pero inexistentes

Salvaguardas: constraint-based prompting • primeros principios • incertidumbre explícita • verificación cruzada • límites de conocimiento

Limitaciones en el uso de LLMs

> *Igual que la reproducción mecánica de obras de arte transforma su significado, los LLMs reproducen patrones de conocimiento técnico despojándolos de su contexto experiencial original.*
>
> — Inspirado en John Berger, Ways of Seeing (1972)

Black Box

No podemos trazar relaciones causa-efecto entre prompts y outputs de manera determinista

Capacidad computacional

Limitada para cálculos matemáticos complejos y simulaciones numéricas especializadas

Vigencia del conocimiento

Fecha de corte fija — crítico para tecnologías y regulación en evolución rápida

Profundidad técnica

Amplitud de conocimiento, pero falta profundidad de experiencia en áreas muy especializadas

Ética, Confidencialidad y Riesgos

Confidencialidad de Datos Industriales

- **Todo lo que envías a un LLM comercial viaja a servidores externos**
- Datos de proceso, parámetros operacionales, know-how propietario, información de patentes
- Algunos proveedores pueden usar tus datos para entrenar modelos futuros
- Soluciones: modelos on-premise, APIs con cláusulas de no retención, anonimización de datos sensibles

Consentimiento en datos

¿Dieron permiso los autores para entrenar con sus textos?

Atribución de autoría

¿Quién es el autor del texto generado por IA?

Sesgos del modelo

Reflejan sesgos de datos de entrenamiento, pueden amplificar estereotipos

Dependencia y autonomía

¿Delegamos demasiado nuestro pensamiento crítico al modelo?

"La ética no es un obstáculo para la innovación, sino el fundamento de una IA beneficiosa y sostenible"

04

Aplicaciones en Ingeniería

Estado del arte con bibliografía verificada (mayo 2026)

Mapa de Aplicaciones por Disciplina

Madurez basada en papers peer-reviewed con resultados cuantitativos verificados

Instrumentación y Control

Genera lógica PLC/DCS automáticamente desde especificaciones en texto natural. Spec2Control (ABB): 98.6% correcto

Alta

P&IDs

Permite consultar y analizar diagramas P&ID complejos en lenguaje natural usando grafos de conocimiento. ChatP&ID: 91% precisión

Alta

Simulación de Procesos

Convierte diagramas de proceso (sketches) en modelos de simulación HYSYS ejecutables. Conecta LLMs con AVEVA vía MCP

Media-Alta

Ing. Civil y Estructural

Diseña estructuras de hormigón armado con cumplimiento normativo automático. Genera planos en AutoCAD desde texto

Media

Sistemas de Potencia

Analiza redes eléctricas: flujo de potencia óptimo y contingencias N-1 mediante conversación natural (GridMind)

Media

Equipos Dinámicos

Mantenimiento prescriptivo: detecta anomalías en vibración y genera recomendaciones de acción. Predice vida útil remanente (RUL)

Media

Seguridad en Construcción

Evaluado en exámenes profesionales de seguridad (BCSP). GPT-4o: 84.6% precisión, supera umbral de certificación

Media

Gestión de Proyectos EPC

Potencial en análisis de viabilidad, estimación de costes y planificación. Brecha significativa entre potencial y adopción real

Baja

Resultados Destacados

Los números verificados más impactantes de la investigación actual

98.6%

conexiones correctas

Spec2Control

ABB Corporate Research

Genera lógica de control DCS desde especificaciones textuales. 98.6% de conexiones correctas, ahorro 94-96% mano de obra. Integrado en herramientas comerciales ABB.

Koziolek et al., arXiv:2510.04519

91%

precisión

ChatP&ID

TU Delft

Interfaz en lenguaje natural para consultar P&IDs complejos. Transforma DEXPI en grafos de conocimiento + RAG. +18% vs. imágenes, $0.004/tarea.

Alimin & Schweidtmann, arXiv:2603.22528

>0.96

consistencia corrientes

Sketch2Simulation

arXiv:2603.24629

Convierte diagramas de proceso en modelos HYSYS ejecutables vía API COM. Consistencia >0.96 en corrientes. Separa razonamiento LLM de ejecución determinista.

Bahamdan et al., arXiv:2603.24629

100%

soluciones correctas

GridMind

SC'25 Workshops

Sistema multi-agente para análisis de redes eléctricas: flujo de potencia óptimo + contingencias N-1 en lenguaje natural. 100% correcto en casos IEEE.

Jin, Kim & Kwon, arXiv:2509.02494

El Patrón Arquitectónico Dominante

Confirmado en TODOS los papers revisados

LLM

Razonamiento semántico

- Interpreta diagramas
- Sintetiza código
- Planifica secuencias
- Genera explicaciones

+

Herramientas Deterministas

- HYSYS, OpenSeesPy
- Compiladores IEC 61131
- APIs COM, MATLAB
- Solvers numéricos

+

Supervisión Humana

- Valida outputs
- Aporta experiencia
- Decide y asume responsabilidad

El LLM NO hace los cálculos — orquesta y razona.
Las herramientas deterministas garantizan la corrección numérica.
La supervisión humana es reconocida como necesaria en el 100% de los trabajos revisados.

Caso Industrial: Emerson × Aramco

Modelos híbridos Aspen en planificación de refinería — el patrón en acción

¿Qué están haciendo?

Modelos híbridos que combinan:

- Modelos first-principles
- Simulación clásica Aspen
- Datos históricos de planta
- ML/IA para no linealidades

La capa IA NO reemplaza Aspen

- Corrige desviaciones
- Mejora predicción de yields
- Reduce gap planificación ↔ operación

Balances, constraints, convergencia y física del proceso → modelos deterministas

Unidades desplegadas

CCR, Platformers, expansión a Hydrocrackers
Unidades muy no lineales, sensibles a feed, con comportamiento catalítico complejo donde pequeños errores de predicción cuestan millones

98.5% prediction accuracy

Pero lo importante no es el número.
Es la optimización multi-site, multi-period, feedstock blending y margin forecasting global.

NO es un chatbot que sabe química. SÍ es un sistema híbrido donde IA orquesta modelos industriales existentes.
Esto es lo que viene: surrogate modeling + physics-informed ML + AI-assisted planning.

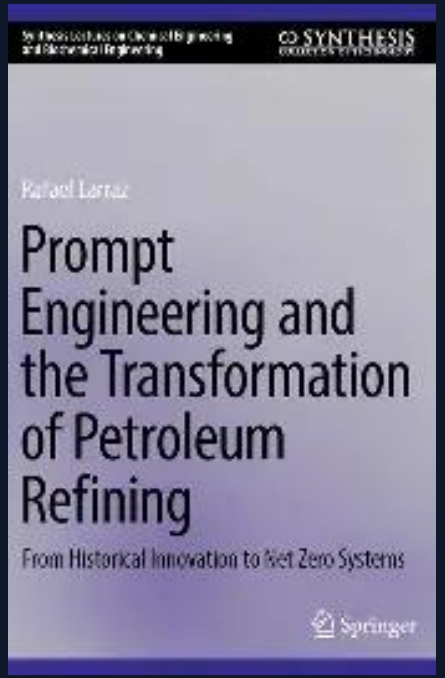
Referencias (1/2)

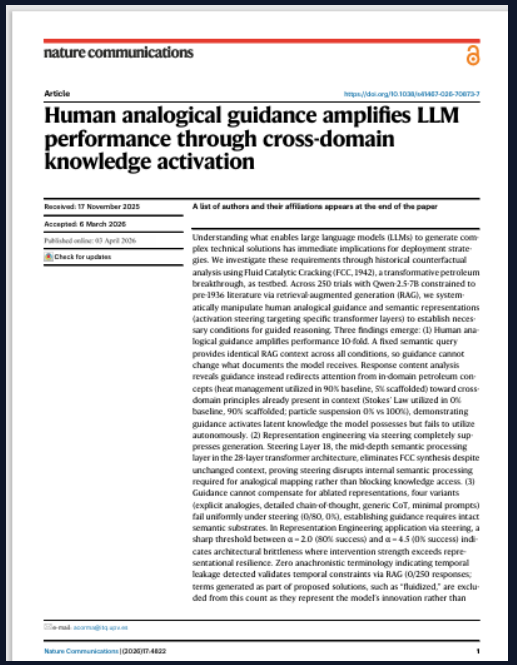
#	Paper	Autores	Fuente	ID
1	ChatP&ID	Alimin, Schweidtmann	arXiv, mar. 2026	arXiv:2603.22528
2	P&IDs como grafos	Alimin et al.	arXiv, feb. 2025	arXiv:2502.18928
3	Sketch2Simulation HYSYS	Bahamdan et al.	arXiv, mar. 2026	arXiv:2603.24629
4	LLM agent AVEVA (MCP)	Liang, Groll, Sin	arXiv, ene. 2026	arXiv:2601.11650
5	Challenges process sim.	Du, Yang	Front. Chem. Sci. Eng.	DOI:10.1007/s11705-025-2587-5
6	Spec2Control (PLC/DCS)	Koziolek et al. (ABB)	arXiv, oct. 2025	arXiv:2510.04519
7	Agents4PLC	Liu et al.	arXiv, oct. 2024	arXiv:2410.14209
8	Multi-agente hormigón	Chen, Bao	Autom. Constr. 177	DOI:10.1016/j.autcon.2025.106331

Referencias (2/2)

#	Paper	Autores	Fuente	ID
9	LLM análisis OpenSeesPy	—	arXiv, abr. 2025	arXiv:2504.09754
10	LLM planos AutoCAD	—	arXiv, jul. 2025	arXiv:2507.19771
11	Review LLMs AEC	Kampelopoulos et al.	Springer AI Review	DOI:10.1007/s10462-025-11241-7
12	IA seguridad construcción	Adil et al.	ASCE JCEM, oct. 2025	arXiv:2411.08320
13	GridMind (potencia)	Jin, Kim, Kwon	SC'25, pp. 560-568	arXiv:2509.02494
14	LLM electrónica potencia	—	Comput. Electr. Eng.	DOI:10.1016/j.compeleceng.2025.110248
15	PARAM mantenimiento	—	arXiv, ago. 2025	arXiv:2508.04714
16	LM4RUL vida útil	—	arXiv, ene. 2025	arXiv:2501.07191

Todas las fuentes verificadas directamente en arXiv, Springer, Elsevier o ASCE. Se han eliminado datos numéricos no confirmados.





nature communications

Article

https://doi.org/10.1038/s41467-026-70873-7

Human analogical guidance amplifies LLM performance through cross-domain knowledge activation

Received: 17 November 2025

Accepted: 6 March 2026

Published online: 03 April 2026

Check for updates

A list of authors and their affiliations appears at the end of the paper

Understanding what enables large language models (LLMs) to generate complex technical solutions has immediate implications for deployment strategies. We investigate these requirements through historical counterfactual analysis using Fluid Catalytic Cracking (FCC, 1942), a transformative petroleum breakthrough, as testbed. Across 250 trials with Qwen 2.5-7B constrained to pre-1936 literature via retrieval-augmented generation (RAG), we systematically manipulate human analogical guidance and semantic representations (activation steering targeting specific transformer layers) to establish necessary conditions for guided reasoning. Three findings emerge: (1) Human analogical guidance amplifies performance 10-fold. A fixed semantic query provides identical RAG context across all conditions, so guidance cannot change what documents the model receives. Response content analysis reveals guidance instead redirects attention from in-domain petroleum concepts (heat management utilized in 90% baseline, 5% scaffolded) toward cross-domain principles already present in context (Stokes' Law utilized in 0% baseline, 90% scaffolded; particle suspension 0% vs 100%), demonstrating guidance activates latent knowledge the model possesses but fails to utilize autonomously. (2) Representation engineering via steering completely suppresses generation. Steering Layer 18, the mid-depth semantic processing layer in the 28-layer transformer architecture, eliminates FCC synthesis despite unchanged context, proving steering disrupts internal semantic processing required for analogical mapping rather than blocking knowledge access. (3) Guidance cannot compensate for ablated representations, four variants (explicit analogies, detailed chain-of-thought, generic CoT, minimal prompts) fail uniformly under steering (0/80, 0%), establishing guidance requires intact semantic substrates. In Representation Engineering application via steering, a sharp threshold between $\alpha = 2.0$ (80% success) and $\alpha = 4.5$ (0% success) indicates architectural brittleness where intervention strength exceeds representational resilience. Zero anachronistic terminology indicating temporal leakage detected validates temporal constraints via RAG (0/250 responses; terms generated as part of proposed solutions, such as "fluidized," are excluded from this count as they represent the model's innovation rather than

e-mail: acorma@itq.upv.es

¿Preguntas?

Rafael Larraz

rafaellarraz@icloud.com

Prompt Engineering and the Transformation of Petroleum Refining (Springer, 2026)

R. Larraz, A.Corma,
Nature Communicatios (2026) 17:4822
https://www.nature.com/articles/s41467-026-70873-7