EVOLUCIÓN DEL ML Y LLM EN RETAIL

TECH SUMMIT 2025

Historia del ML y Evolución hacia los LLM

1950s: Primeros algoritmos ML

Perceptrón (Rosenblatt)

1980s-1990s: Redes Neuronales

Backpropagation & CNN

2000s: Mejoras algorítmicas

SVM, Random Forests, GPU

2013: Word Embeddings

Word2Vec (Google)

2017: Arquitectura Transformer

Attention is All You Need

2018-2019: BERT & GPT-2

Preentrenamiento masivo

2020: GPT-3

175B parámetros

2022-2023: GPT-4, Claude

Multimodal & Razonamiento

2024-2025: LLM en Retail

Personalización & Decisiones

Factores Clave en la Evolución

- Poder computacional (GPUs/TPUs)
- Datos masivos para entrenamiento
- Arquitecturas de atención

Historia de Aplicaciones de ML en Retail

Análisis Predictivo Básico

1990s - 2000s

- Análisis de cestas de compra
- Reglas de asociación
- Pronóstico de inventario

Personalización

2000s - 2010s

- Sistemas de recomendación
- Segmentación de clientes
- Marketing personalizado

Computer Vision

2010s - 2015

- Reconocimiento de productos
- Mapeo de tiendas
- Análisis de comportamiento

NLP en Retail

2015 - 2020

- Asistentes virtuales
- Análisis de sentimientos
- Chatbots de atención

Era LLM

2021 - Presente

- Generación de contenido
- Hiperpersonalización
- Decisiones autónomas

Impacto en el Retail

Mayor ROI

+35% en campañas con ML vs. tradicionales

Reducción de stock

-20% en excedentes con predicción ML

Retención de clientes

+40% con personalización avanzada

Conceptos Clave: Tokens y LLM

¿Qué son los tokens?

Unidades básicas de información que los LLM procesan y generan.

Un token puede ser:

- `perro` Palabra completa
- `com-` Parte de palabra
- `,` Puntuación
- `[espacio]` Espacio en blanco

EJEMPLO DE TOKENIZACIÓN

"Los [ ] `clientes` [ ] `son` [ ] `import` `antes` `.` "

Importancia de los Tokens

Costo de Procesamiento
Los servicios de LLM cobran por tokens procesados (entrada + salida)

Límites de Contexto
Ventana limitada: GPT-4 (8K-128K), Claude (200K), Gemini (1M+)

Precisión en Texto
El modelo predice el siguiente token, no la siguiente palabra

Concepto Técnico Clave

Los tokens son el "lenguaje interno" de los LLM. Todo lo que el modelo "lee", "interpreta" y "responde" se procesa como secuencias de tokens, no como palabras completas.

Métricas de Evaluación Principales

Perplejidad (PPL) ↓
Incertidumbre del modelo al predecir la siguiente palabra

BLEU/ROUGE ↑
Coincidencia entre salida generada y referencia humana

BERTScore ↑
Similitud semántica entre texto generado y referencia

Eval. Humana ↑
Juicios sobre fluidez, coherencia y relevancia

Inteligencia Artificial Generativa (IAG)

Subdisciplina de IA que utiliza modelos de aprendizaje automático para **crear contenido nuevo y original** a partir de datos analizados. Genera texto, imágenes, audio, video o código, replicando patrones aprendidos durante el entrenamiento.

Características Principales

Creación Original

Genera contenido nuevo, no solo clasifica o predice información existente

Comprensión Contextual

Entiende y mantiene contexto en prompts e intercambios complejos

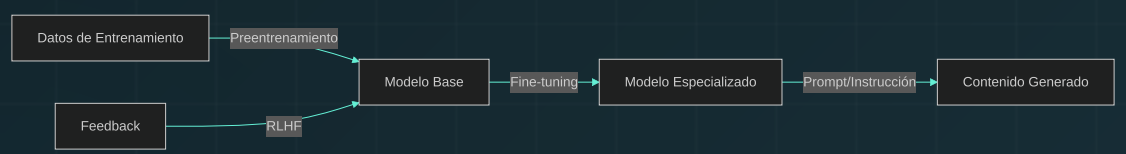
Adaptabilidad

Ajusta respuestas según el formato, estilo y propósito solicitados

Multi-modal

Procesa y genera diversos tipos de contenido (texto, imagen, audio)

Proceso de Generación



Aplicaciones en Retail

Generación de Contenido

Descripciones de productos, emails promocionales, posts para redes sociales

Servicio al Cliente

Chatbots inteligentes, respuestas personalizadas, asistentes virtuales 24/7

Hiperpersonalización

Recomendaciones específicas según historial, preferencias y contexto

Análisis Predictivo Avanzado

Tendencias de mercado, patrones de comportamiento, optimización de inventario

Impacto en Negocio

+40%	-25%	+65%
Eficiencia operativa	Costo de adquisición	Personalización

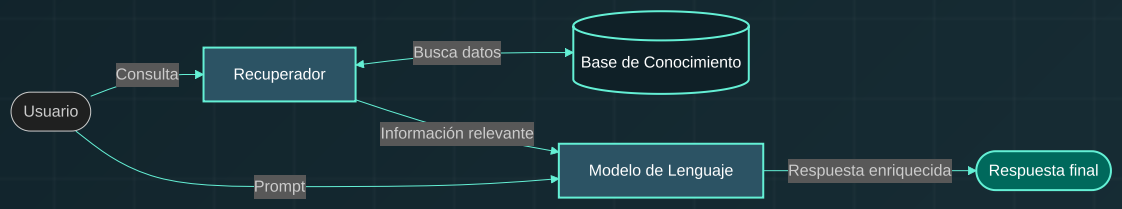
Retrieval-Augmented Generation (RAG)

¿Qué es RAG?

RAG combina **modelos de lenguaje** con **sistemas de recuperación de información** para proporcionar respuestas basadas en datos externos y actualizados.

En lugar de depender únicamente del conocimiento preentrenado, RAG busca información relevante en bases de datos o documentos y la integra en las respuestas generadas.

Funcionamiento de RAG



Ventajas Clave

Precisión Mejorada
Información actualizada y precisa

Reducción de Alucinaciones
Basado en fuentes verificables

Información Actualizada
No limitado a datos de entrenamiento

Adaptabilidad
Fácil de actualizar conocimiento

RAG vs. LLM Estándar

Característica	LLM Estándar	RAG
Fuente de información	Conocimiento entrenado	Conocimiento + fuentes externas
Actualización	Requiere reentrenamiento	Actualización de base de datos
Precisión en domain-specific	Limitada	Alta (con fuentes adecuadas)
Capacidad de citación	Pobre/Inexistente	Puede citar fuentes específicas

Aplicaciones en Retail

Asistentes de Compra Inteligentes
Combinan catálogos de productos actualizados con interacción natural para ofrecer recomendaciones precisas basadas en inventario real.

Soporte Técnico Especializado
Acceso a manuales de productos, historiales de casos y soluciones documentadas para problemas específicos.

Análisis de Mercado en Tiempo Real
Incorpora datos de tendencias, competencia y comportamiento de consumidor para generar insights estratégicos actualizados.

Impacto en ROI

-35%	+60%	+45%
Reducción de errores en respuestas	Satisfacción del cliente	Conversión en ecommerce

Cache-Augmented Generation (CAG)

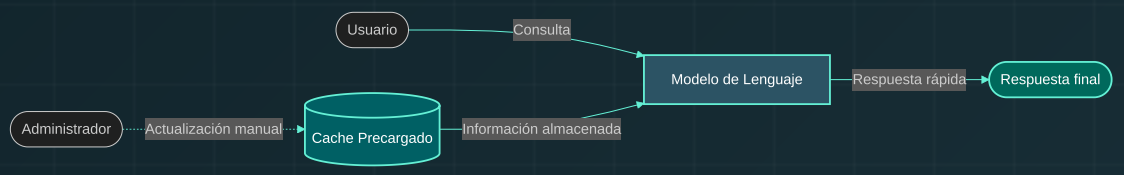
¿Qué es CAG?

Evolución de RAG que optimiza la generación de respuestas **almacenando previamente información relevante** en la memoria del modelo, eliminando la necesidad de recuperar datos en tiempo real.

Pros y Contras

- Mayor velocidad de respuesta
- Arquitectura más simple
- No se actualiza automáticamente
- No ideal para datos cambiantes

Funcionamiento de CAG



Ideal Para:

Información Estática
Datos que no cambian frecuentemente

Respuestas Rápidas
Cuando la velocidad es crítica

Consultas Repetitivas
Preguntas frecuentes predecibles

Recursos Limitados
Entornos con restricciones

RAG vs. CAG

Característica	RAG	CAG
Actualización de datos	Dinámica y en tiempo real	Estática, actualización manual
Velocidad de respuesta	Moderada (depende de recuperación)	Alta (acceso directo)
Complejidad técnica	Alta (requiere sistemas de búsqueda)	Moderada (estructura más simple)
Casos de uso ideales	Información variable y actualizada	Información estable y frecuente
Costo operativo	Mayor (procesamiento en tiempo real)	Menor (procesamiento más eficiente)

Aplicaciones en Retail

FAQ sobre Productos Populares
Respuestas rápidas y consistentes para preguntas comunes sobre productos líderes en ventas.

Políticas de Tienda
Información sobre devoluciones, garantías y programas de fidelidad que cambian con poca frecuencia.

Asistencia en Punto de Venta
Apoyo rápido para cajeros sobre procedimientos estándar y consultas habituales.

Rendimiento vs RAG

5x
Mayor velocidad de respuesta

-40%
Costo computacional

+25%
Consistencia en respuestas

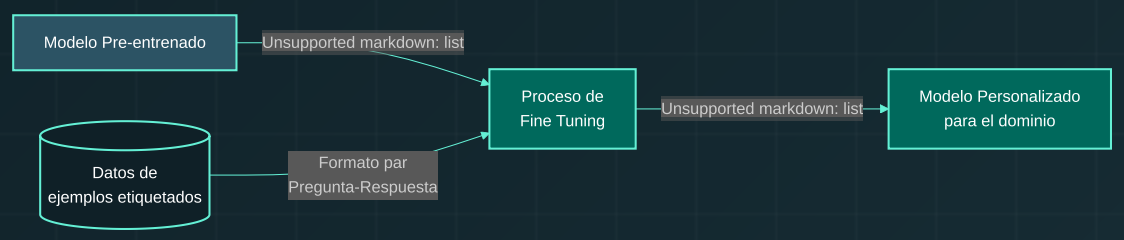
Fine Tuning (Ajuste Fino)

¿Qué es Fine Tuning?

Proceso de **reentrenar un modelo de lenguaje** con un conjunto de datos específico para **adaptar su comportamiento** a tareas o dominios particulares.

Permite al modelo aprender estilos de escritura, terminología especializada y formatos específicos para un caso de uso concreto.

Proceso de Fine Tuning



Pros y Contras

Ventajas

- Mayor precisión en dominio
- Adaptación a estilo y tono
- Menor necesidad de prompts

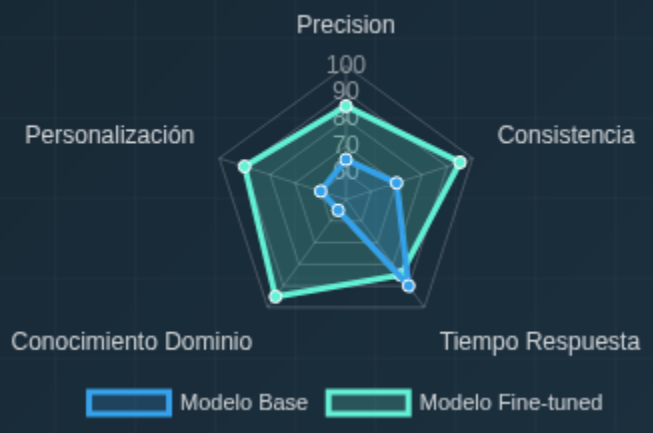
Desventajas

- Alto costo computacional
- Requiere datos etiquetados
- Potencial sobreajuste

Fine Tuning vs. Prompt Engineering

Aspecto	Fine Tuning	Prompt Engineering
Implementación	Reentrenamiento del modelo	Instrucciones avanzadas
Costo inicial	Alto (computación)	Bajo (diseño de prompts)
Costo por uso	Bajo (tokens reducidos)	Alto (tokens extensos)
Consistencia	Alta (comportamiento aprendido)	Variable (según prompt)

Mejora de Rendimiento



Aplicaciones en Retail

Atención al Cliente Especializada

Chatbots entrenados con historiales de soporte para responder según políticas y tono de la marca.

Generación de Descripciones de Producto

Modelos ajustados para crear textos de catálogo con el estilo y terminología específica del retailer.

Análisis de Sentimiento Personalizado

Detección de satisfacción/insatisfacción adaptada a expresiones propias del sector.

Prompt Engineering

¿Qué es Prompt Engineering?

Es el **diseño y optimización de instrucciones** específicas para **guiar el comportamiento** de modelos de lenguaje sin modificar sus parámetros internos.

Permite adaptar la salida de los LLM para aplicaciones específicas aprovechando al máximo sus capacidades existentes.

Técnicas Principales

Role Prompting

Asignar un rol específico al modelo para obtener respuestas desde perspectivas especializadas.

Chain of Thought

Guiar al modelo para que razone paso a paso antes de llegar a una conclusión final.

Few-Shot Learning

Proporcionar ejemplos de entradas y salidas deseadas para establecer patrones.

Structured Output

Solicitar formatos específicos como JSON, tablas o listas para facilitar procesamiento.

EXAMPLE

```
Rol: Actúa como un experto en servicio al cliente de retail.
Tarea: Responde a esta queja sobre un producto defectuoso.
Formato: 1) Empatía, 2) Solución, 3) Compensación, 4) Conclusión.
Tono: Profesional pero cálido.
```

Buenas Prácticas

- **Ser específico y claro**
  Instrucciones precisas y detalladas producen mejores resultados.
- **Estructurar jerárquicamente**
  Dividir instrucciones complejas en componentes manejables.
- **Iterar y experimentar**
  Refinar los prompts basándose en los resultados obtenidos.
- **Añadir constraints**
  Establecer límites claros para evitar respuestas no deseadas.

Aplicaciones en Retail

Descripciones de Producto

Generación de textos persuasivos con características y beneficios de productos.

Atención al Cliente

Respuestas personalizadas a preguntas frecuentes en tienda o online.

Análisis de Tendencias

Extracción de insights del mercado a partir de datos no estructurados.

Comunicación Interna

Transformación de informes técnicos a formatos comprensibles para equipos.

Ventajas vs Implementación Tradicional

10x	-90%	+70%
Velocidad de implementación	Costos de desarrollo	Adaptabilidad a cambios

Consejo Clave

La efectividad del prompt engineering depende del balance entre ser específico y dejar espacio para la creatividad del LLM. Demasiadas instrucciones pueden limitar el potencial de la IA.

Casos de Éxito en Retail

NotCo

IA para Desarrollo de Alimentos

IA "Giuseppe" que analiza alimentos a nivel molecular y encuentra combinaciones vegetales para replicar sabores, texturas y nutrientes de productos animales.

Resultados Empresariales:

- **10x** Aceleración en desarrollo de productos
- **1000+** Combinaciones evaluadas automáticamente
- **-60%** Costos de investigación

Tecnologías:

Deep Learning | Machine Learning | Análisis Molecular

Walmart

"Text to Shop" & IA Conversacional

Asistente conversacional que permite comprar mediante voz o texto, funcionando como un "personal shopper" virtual personalizado.

Resultados Empresariales:

- **+38%** Satisfacción del cliente (Chile)
- ↑↑ Aumento en ventas online
- **-85%** Tiempo de respuesta al cliente

Tecnologías:

NLP | Chatbots | Rec. de Voz | Personalización

Amazon

Recomendaciones Personalizadas

Sistema de LLM para analizar preferencias y generar recomendaciones altamente específicas y descripciones de producto personalizadas.

Resultados Empresariales:

- ↑↑ Relevancia en resultados de búsqueda
- **+35%** Tasa de conversión estimada
- **+25%** Compras recurrentes

Tecnologías:

LLMs | Análisis Predictivo | IA Generativa

Lección Clave: Enfoque Práctico de IA/ML en Retail

Las implementaciones exitosas resuelven problemas específicos de negocio, no solo aplican tecnología por novedad. El verdadero valor surge al integrar IA/ML directamente en la experiencia del cliente o la eficiencia operativa.

Soluciones IAG para Retail

Asistentes RAG para Atención al Cliente

Implementaciones rápidas de asistentes virtuales con arquitectura RAG (Retrieval-Augmented Generation), integrados con bases de conocimiento internas.

VALOR PARA RETAIL

- Reduce costos operativos de call centers
- Aumenta satisfacción con respuestas precisas
- No requiere entrenamientos extensos

PROPUESTA AL CLIENTE

"Transforma tu centro de atención con un asistente inteligente que responde con la información más reciente de tu negocio, sin entrenamientos extensos."

Generación de Contenido para E-commerce

Software e integraciones de LLM para generación automática de descripciones de productos, títulos SEO y contenidos para redes sociales.

VALOR PARA RETAIL

- Acelera la carga de nuevos productos
- Mejora posicionamiento SEO
- Personaliza experiencias a escala

PROPUESTA AL CLIENTE

"¿Tienes miles de productos sin describir? Te ayudamos a automatizar y escalar tu catálogo digital con textos que venden y posicionan."

Análisis Conversacional para BI

Soluciones que integran modelos generativos con plataformas de BI para permitir consultas en lenguaje natural.

VALOR PARA RETAIL

- Empodera personal no técnico
- Acelera toma de decisiones estratégicas
- Democratiza acceso a insights

PROPUESTA AL CLIENTE

"Deja que tus gerentes hablen con sus datos como si conversaran con un analista senior. La IA generativa convierte tus reportes en respuestas accionables."

Potencial de Mercado: IAG en Retail

$15.8B	42%	3.5x	68%
Mercado IAG Retail 2025	Reducción costos atención	ROI promedio soluciones IAG	Retailers adoptarán IAG en 2025

Ejemplo Práctico: Asistente RAG para Retail

Caso de Implementación

Desarrollo e implementación de un asistente virtual RAG para una cadena de tiendas de electrónica con +200 productos y políticas específicas.

Retos a Resolver:

- Alto volumen de consultas repetitivas (30% del total)
- Información técnica compleja y cambiante
- Necesidad de personalización por ubicación
- Integración con sistemas existentes de CRM

Etapas de Implementación

1. **Preparación de Datos**
   Catalogación y estructuración de manuales, FAQs y políticas
2. **Vectorización de Contenido**
   Conversión a embeddings utilizando modelos de encaje semántico
3. **Diseño de Prompts Maestros**
   Creación de plantillas de instrucciones para diferentes escenarios
4. **Integración y Testing**
   Implementación en canales (web, app, WhatsApp) y pruebas piloto
5. **Monitoreo y Mejora Continua**
   Análisis de interacciones y feedback para optimización

Arquitectura de la Solución

Stack Tecnológico

FRONT-END

- React
- TypeScript
- WebSockets

BACK-END

- Node.js
- FastAPI
- Redis Cache

MODELOS IA

- GPT-4 Turbo
- text-embedding-ada-002
- LangChain

BASES DE DATOS

- Pinecone
- PostgreSQL
- Prometheus

Factores Críticos de Éxito

- Calidad y estructura de datos
- Actualización automática de conocimiento
- Capacitación de personal
- Gobernanza y supervisión humana

Resultados Esperados

-45% Reducción del volumen de tickets	**+92%** Precisión en respuestas
24/7 Disponibilidad	**3.5x** ROI estimado (18 meses)

Evolución de Eficiencia

Timeline de Proyecto

- **Semanas 1-2** Preparación de datos y vectorización
- **Semanas 3-4** Desarrollo e integración
- **Semanas 5-6** Testing y ajuste fino
- **Semanas 7-8** Implementación y capacitación

Tiempo total de implementación: **8 semanas**

Cliente