Banca & Economía

2024

Edición 1455

Innovación financiera: el poder de los modelos LLM

16 de diciembre de 2024

Director:

Jonathan Malagón González

ASOBANCARIA:

Jonathan Malagón González
Presidente

Alejandro Vera Sandoval
Vicepresidente Técnico

Germán Montoya Moreno
Director Económico

- Los LLM son una de las mayores disrupciones tecnológicas en el mundo de la IA, habilitando la automatización del análisis complejo.
- Desde la detección proactiva de fraudes, pasando por la evaluación crediticia hasta la generación automatizada de reportes financieros y análisis regulatorio, los LLM pueden redefinir la eficiencia operativa de entidades financieras.
- Enfoques como el fine-tuning y RLHF permiten adaptar los LLM a necesidades específicas de las instituciones, optimizando desde contratos legales hasta interacciones éticas y alineadas con regulaciones.
- Tecnologías como RAG potencian a los LLM al combinar su capacidad generativa con acceso a datos cambiantes, garantizando decisiones precisas y actualizadas en entornos críticos.

Para suscribirse a nuestra publicación semanal Banca & Economía, por favor envíe un correo electrónico a bancayeconomia@asobancaria.com

Banca & Economía

2024

Edición 1455

Innovación financiera: el poder de los modelos LLM

Editor

Germán Montoya
Director Económico

Participaron en esta edición:

Adriana María Ovalle Herazo
Andrés Daniel Godoy Ortiz
Julián Rodríguez Vega

El sector financiero está ante un nuevo punto de inflexión en su transformación tecnológica con la incorporación de los Modelos de Lenguaje de Gran Escala (LLM, por sus siglas en inglés). Estas arquitecturas de Inteligencia Artificial Generativa, impulsadas por redes neuronales profundas y diseñadas para procesar lenguaje natural con precisión y escala, representan una oportunidad única para optimizar procesos, escalar operaciones y mantener su competitividad en un entorno de innovación constante.

Para líderes tecnológicos en el ámbito financiero, entender las capacidades y aplicaciones de los LLM no es solo una ventaja estratégica, sino una necesidad. Desde la implementación de modelos que generan reportes en tiempo real hasta el análisis regulatorio y la detección de fraudes, estas tecnologías no solo transforman la eficiencia operativa, sino también la calidad del servicio y la capacidad de anticiparse a las dinámicas del mercado. La clave radica en conocer cómo personalizarlos para casos de uso específicos y cómo integrarlos con fuentes de datos dinámicas y sistemas existentes.

Esta edición de Banca y Economía, organizada en tres secciones, profundiza en esta transformación. En primer lugar, explora las ideas fundamentales detrás de qué son los LLM, y cómo se construyen (o aprenden). A continuación, analiza casos de uso estratégicos que abarcan desde la detección de fraudes y el análisis regulatorio hasta la generación de reportes financieros y evaluación de crédito, destacando cómo estas soluciones pueden integrarse en entornos dinámicos. Finalmente, aborda las principales metodologías de personalización, como el fine-tuning clásico, el RLHF (Refuerzo por Retroalimentación Humana) y la generación aumentada por recuperación (RAG), proporcionando una guía práctica sobre cómo adaptar los LLM a las necesidades específicas de cada institución, con una mirada estratégica hacia el futuro. Con esta visión integral, esta edición ofrece una orientación para aprovechar al máximo el potencial de los LLM y consolidar su papel como catalizadores de la innovación en el sector financiero.

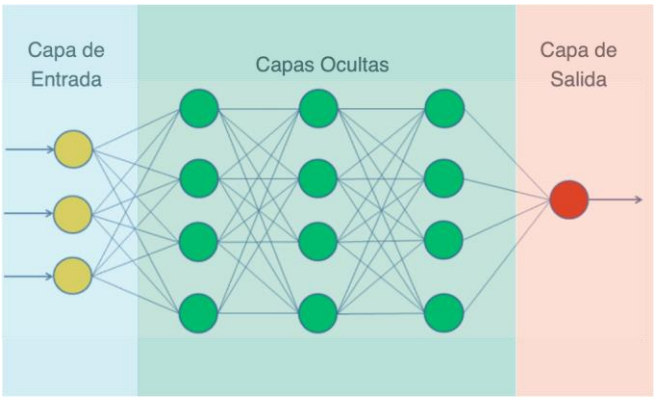
¿Qué es realmente un LLM?

En los últimos años, los Modelos de Lenguaje de gran escala o extensivos (LLM por sus siglas en inglés) han emergido como una de las tecnologías más transformadoras en el panorama de la inteligencia artificial (IA). Estos modelos, como GPT-4 de OpenAI, Claude de Anthropic o Gemini de Google, son sistemas de IA diseñados para procesar y generar texto de manera similar a cómo lo haría un humano, pero a una escala y velocidad inimaginables. Antes de profundizar en sus aplicaciones, es importante entender qué son, cómo funcionan y por qué están redefiniendo sectores completos, incluido el financiero.

En su núcleo, un LLM es una red neuronal profunda entrenada para manejar y generar texto. Las redes neuronales, inspiradas en el cerebro humano, son modelos matemáticos formados por nodos (neuronas artificiales) conectados entre sí mediante "pesos" o parámetros. Estos pesos son valores numéricos que determinan cómo una entrada (datos) particular se convierte en una salida

(predicción). Se puede pensar en una red neuronal como una vasta red de autopistas: los pesos son como señales de tráfico que dirigen el flujo de información, ajustándose dinámicamente para encontrar la mejor ruta hacia el destino deseado (Figura 1).

Figura 1. Red neuronal



Fuente: Recuperado de aprendeia.com/que-son-las-redes-neuronales-artificiales/

La capacidad de un LLM depende en gran medida de sus parámetros. Un modelo como GPT-4 de OpenAI, con más de 175 mil millones de parámetros, tiene la capacidad de almacenar y procesar información: conocimiento enciclopédico combinado con habilidades analíticas.

La arquitectura que da vida a los LLM es el Transformer, introducido por Vaswani et al. en 2017[1]. Antes de su aparición, las redes neuronales tenían dificultades para comprender el contexto de una palabra más allá de unas pocas frases. Los Transformers resolvieron este problema al permitir que el modelo considere el contexto completo de una oración o documento a través de un mecanismo llamado atención. Este mecanismo evalúa la relevancia de cada palabra en relación con las demás.

El proceso de entrenamiento: afinando la guitarra

El entrenamiento de un LLM es como afinar una guitarra para que toque melodías perfectas. En lugar de cuerdas, el modelo ajusta sus "pesos". Este ajuste ocurre a lo largo de dos fases principales:

1. **Preentrenamiento**: Aquí, el modelo ajusta los "pesos" o parámetros. Aprende patrones generales del lenguaje utilizando cantidades masivas de datos textuales (libros, artículos, foros, etc.). Este proceso se basa en tareas como predecir la siguiente palabra en una frase incompleta. Por ejemplo, si el modelo ve: "En el mercado financiero, la diversificación es clave para minimizar el ___," probablemente prediga "riesgo", tal y como el cerebro humano lo hace. Este entrenamiento le permite desarrollar una comprensión general del lenguaje y las estructuras contextuales. Al final de este proceso, surge una red neuronal con miles de millones de parámetros que puede entender a partir de las palabras el mundo (o la fracción de este representada en los datos de entrenamiento).

2. **Ajuste fino (fine-tuning)**: Tras el preentrenamiento, el modelo se puede especializar en tareas o dominios específicos. Por ejemplo, un modelo ajustado para finanzas aprenderá a reconocer conceptos como "costo de capital", "derivados" o "estructura de deuda". Esto lo hace ideal para responder preguntas o realizar análisis complejos en ese sector. En este proceso los "pesos" se reajustan.

El entendimiento profundo y la capacidad de los LLM para generar texto los convierte en una herramienta indispensable para industrias como la financiera. Ya no solo hablamos de automatización, sino de proporcionar análisis y decisiones informadas en tiempo real, traduciendo la complejidad en claridad. No en vano, estos modelos trajeron consigo el concepto de "Agente", que es cada modelo particular derivado de esta tecnología, capaz de realizar tareas analíticas avanzadas que antes requerían la intervención humana. Desde la generación de informes de crédito hasta el análisis de riesgos regulatorios, los LLM prometen transformar cómo los bancos y otras instituciones financieras operan.

Casos de uso: analistas sí, chatbots no

El sector financiero se encuentra siempre en una encrucijada digital. Por un lado, enfrenta la presión de los clientes por servicios más rápidos y personalizados; por otro, debe cumplir con estándares regulatorios cada vez más estrictos. Aquí es donde los LLM entran en juego. Su capacidad para procesar lenguaje humano y conocimiento a gran escala pueden resolver múltiples desafíos clave:

Servicio al cliente automatizado

En el ámbito financiero, donde las interacciones con los clientes pueden ser tan simples como consultar un saldo o tan complejas como gestionar reclamaciones sobre operaciones internacionales, los LLM ofrecen una ventaja significativa sobre los sistemas tradicionales de atención al cliente. A diferencia de los *chatbots* convencionales basados en reglas rígidas, los LLM comprenden la intención detrás de una consulta, incluso cuando está redactada de forma ambigua o coloquial.

Un estudio reciente[2] muestra que los LLM no solo son capaces de manejar este tipo de interacciones, sino que también aprenden de cada una para mejorar su precisión y capacidad de respuesta. Además, estos modelos pueden generar traducciones precisas en tiempo real, lo que es crucial para instituciones financieras globales con clientes en diferentes regiones lingüísticas.

[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All You Need. Advances in Neural Information Processing Systems, 30, 5998-6008.
[2] Lin, X., Kundu, L., Dick, C., Galdon, M. A. C., Vamaraju, J., Dutta, S., & Raman, V. (2024). A Primer on Generative AI for Telecom: From Theory to Practice. *arXiv preprint arXiv:2408.09031*.

Si bien los *chatbots* son un caso de uso, reducir los LLM a este uso es similar a haber pensado en los computadores como simples calculadoras cuando estos se inventaron.

Detección del fraude

En un entorno donde los esquemas de fraude evolucionan constantemente, los LLM se han convertido en herramientas clave para la detección y prevención del fraude financiero. Su capacidad para entender el contexto y analizar datos complejos en tiempo real los hace únicos.

A diferencia de los modelos tradicionales, los LLM no solo identifican anomalías basadas en reglas predefinidas, sino que pueden acceder a registros dinámicos de técnicas de fraude recientes. Esto les permite combinar información actualizada de múltiples fuentes[3] —como bases de datos de IPs sospechosas, reportes de la policía o informes de fraude de otros bancos— con patrones históricos del cliente.

Además, ofrecen análisis adaptativos, explicando en lenguaje natural por qué una transacción parece sospechosa y cómo se relaciona con patrones emergentes, como el *phishing* avanzado o el *card testing.* Esta combinación de flexibilidad, dinamismo y capacidad generativa no solo ayuda a bloquear actividades ilícitas, sino que también mejora la experiencia de analistas y clientes al proporcionar contexto comprensible y decisiones bien informadas.

Análisis financiero y generación de reportes

En un entorno financiero complejo y cambiante, los LLM pueden automatizar la generación de reportes financieros, como análisis de riesgos, proyecciones económicas y balances de empresas en cuestión de segundos. Esto no solo ahorra tiempo, sino que también mejora la precisión al reducir errores humanos.

Investigaciones recientes[4] han mostrado cómo los LLM pueden sintetizar informes extensos en resúmenes ejecutivos, identificar anomalías en balances financieros y sugerir estrategias de mitigación. Por ejemplo, al analizar los estados financieros de una empresa, un LLM puede identificar inconsistencias contables que podrían pasar desapercibidas en análisis manuales.

Así mismo, los LLM tienen la capacidad de actuar como asesores financieros digitales, ofreciendo recomendaciones personalizadas basadas en el historial financiero del cliente, sus objetivos y el entorno económico actual. Por ejemplo, un modelo puede sugerir estrategias de inversión adaptadas al perfil de riesgo de un consumidor o brindar recomendaciones sobre cómo mejorar su puntaje crediticio.

Esta capacidad, combinada con un aprendizaje continuo, permite a los bancos ofrecer servicios de asesoría asequibles y escalables para segmentos de mercado previamente desatendidos.

Análisis regulatorio y de contratos

En el sector financiero, los contratos son la columna vertebral de la operación: acuerdos de préstamo, términos y condiciones de servicios bancarios, cláusulas de derivados, convenios de confidencialidad y más. Históricamente, analizar y gestionar estos documentos ha sido una tarea intensiva en tiempo, costosa y propensa a errores humanos.

Los LLM, gracias a su capacidad de comprensión profunda del lenguaje, pueden procesar y analizar contratos legales y financieros. En particular pueden ser reentrenados (*fine-tuning*) en todo el marco normativo del sistema financiero o en todos los contratos de una entidad particular, para que actúen como agentes especializados. Es así como pueden detectar términos clave en un contrato, como tasas de interés, plazos, condiciones de incumplimiento, así como realizar análisis más avanzados para identificar cláusulas ambiguas, términos inconsistentes o riesgos potenciales. Bancos como JP Morgan ya están desarrollando modelos con estos propósitos[5].

En el estudio de Cao & Feinstein (2024)[6], puede observarse como en un entorno regulatorio cambiante, la flexibilidad de los LLM permitiría:

A. **Actualizar contratos existentes:** Cuando una nueva regulación entra en vigor, como la adopción de un nuevo índice de referencia de tasas de interés (por ejemplo, la transición de LIBOR a SOFR), el LLM puede analizar y sugerir enmiendas a todos los contratos afectados.

B. **Análisis previo a la firma:** Antes de que un banco firme un contrato con un cliente corporativo o inversor, el modelo puede realizar una revisión exhaustiva para garantizar que los términos sean equitativos y alineados con las políticas internas.

Evaluación alternativa de crédito

Uno de los grandes retos del sector financiero ha sido evaluar la capacidad crediticia de personas sin historial financiero tradicional. Esto es especialmente crítico en economías emergentes y sectores no bancarizados, donde millones de individuos carecen de acceso al crédito simplemente porque no tienen un historial crediticio formal. Los LLM pueden tener enfoque disruptivo en este ámbito por su facilidad para analizar datos no estructurados[7]: la evaluación crediticia basada en datos alternativos.

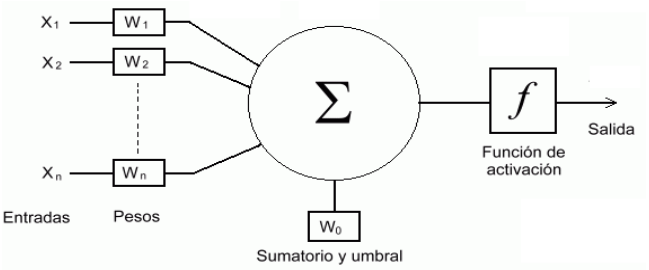
[3] https://blogs.nvidia.com/blog/ai-fraud-detection-rapids-triton-tensorrt-nemo/
[4] Godbole, A., George, J. G., & Shandilya, S. (2024). Leveraging Long-Context Large Language Models for Multi-Document Understanding and Summarization in Enterprise Applications. *arXiv preprint arXiv:2409.18454*.
[5] https://www.forbes.com/sites/janakirammsv/2024/07/30/jpmorgan-chase-leads-ai-revolution-in-finance-with-launch-of-llm-suite/
[6] Cao, Z., & Feinstein, Z. (2024). Large Language Model in Financial Regulatory Interpretation. arXiv preprint arXiv:2405.06808.
[7] Yin, Y., Yang, Y., Yang, J., & Liu, Q. (2023). FinPT: Financial Risk Prediction with Profile Tuning on Pretrained Foundation Models. arXiv preprint arXiv:2308.00065.

Métodos de integración: ¿cómo enseñar datos nuevos a estos modelos?

Una persona o un LLM solo tienen conocimiento de los datos que ha estudiado o en los que se ha entrenado. Si se quiere que sepa de contenido propio de una entidad o de un área de conocimiento especializada, se debe entrenar en eso.

Desde un punto de vista técnico, los LLM utilizan una representación vectorial para procesar palabras. En el entrenamiento del modelo, cada palabra se convierte en un vector numérico (que es el conjunto de valores de entradas en la Figura 2), que captura su significado y relación con otras palabras. En el espacio vectorial donde se ubican las palabras, la representación numérica de "banco" estará más cerca de "finanzas" que de "playa". A medida que estos vectores pasan por la red neuronal, los "pesos" y las capas del modelo aplican operaciones matemáticas como multiplicaciones y sumas, transformándolos para obtener un resultado que represente la predicción final (la siguiente palabra).

Figura 2. Red neuronal



Fuente: Elaboración Asobancaria.

Como se mencionó en la primera sección, el proceso de aprender de nuevos datos implica ajustar esos "pesos" para predecir mejor los nuevos datos. Si se quiere entrenar un modelo LLM en un conocimiento especializado hay tres caminos: entrenar desde cero, reentrenar (*fine-tuning*) un modelo base y generación mejorada por recuperación (RAG por sus siglas en inglés).

Por ahora, la primera opción es en general inviable financieramente, entrenar Chatgpt-4 costó cerca de 100 millones de dólares[8]. La segunda opción, el reentrenamiento de un modelo base como los de OpenAI, Google, Meta, Anthropic, etc., es el camino más viable cuando se quiere que el modelo haga un "curso" o "vaya a la universidad" para almacenar un conocimiento (o una capacidad de análisis) en sus parámetros que no se actualiza frecuentemente. Finalmente, RAG (Retrieval-Augmented Generation), es en los casos en los que requerimos que el modelo pueda acceder y analizar información que se actualiza constantemente, como noticias, datos financieros o económicos.

Reentrenar requiere hardware especializado, particularmente procesadores de alta gama y especializados para estas tareas, llamados GPU. Toda vez que el acceso a estos procesadores es costoso, la solución más asequible para las empresas es hacer uso de proveedores de hardware a distancia o servicios de "nube", tales como AWS, Azure o GCP. Descartando entrenar un modelo desde cero, a continuación profundizaremos en 3 tipos de fine-tunning o reentrenamiento - *Fine-tuning* clásico, *Instruction Fine-tuning* y Refuerzo por Retroalimentación Humana- y en RAG (Figura 3).

Fine-tuning clásico

Es uno de los enfoques más robustos para personalizar un Modelo LLM. Consiste en ajustar los parámetros internos del modelo base —los mismos que determinan cómo procesa y genera texto— para adaptarlo a un dominio específico, como el derecho o la medicina. En esencia, es como tomar a un experto generalista y entrenarlo exhaustivamente para que domine una disciplina en particular.

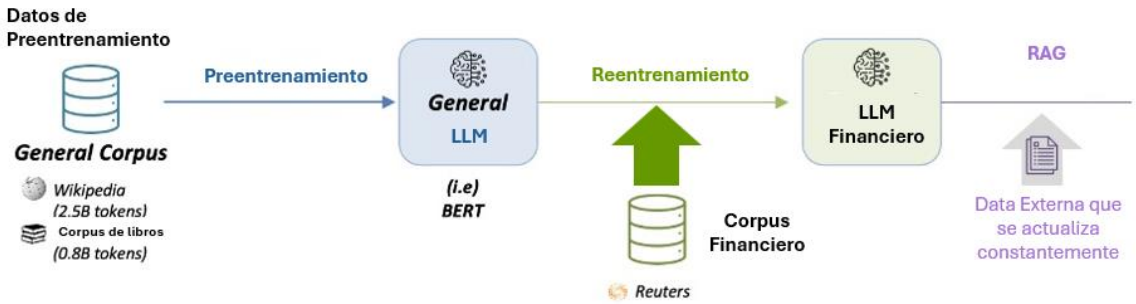
En términos simples, el modelo preentrenado pasa por una segunda etapa de aprendizaje utilizando datos específicos del dominio. Durante este proceso, los "pesos" del modelo, que son las conexiones matemáticas entre las neuronas artificiales, se ajustan para reflejar los patrones presentes en los nuevos datos. Esto se hace utilizando algoritmos de optimización avanzados que minimizan la diferencia entre las predicciones del modelo y los resultados esperados. En esta etapa, todas las capas del modelo se reconfiguran, lo que lo hace altamente efectivo, pero también computacionalmente costoso.

Por ejemplo, si un banco quiere que el modelo redacte contratos legales, se entrena el LLM con miles de ejemplos de contratos reales. Al finalizar el proceso, el modelo no solo comprende términos como "plazo de amortización" o "tasa fija", sino que también sabe contextualizarlos en el marco normativo y financiero relevante.

Antes de embarcarse en un proyecto de *fine-tuning*, es esencial contar con ciertos elementos clave. En primer lugar, se necesita un volumen considerable de datos de alta calidad. Para tareas específicas en el sector financiero, esto implica un volumen de datos considerable curados, como documentos regulatorios o contratos especializados. Estos datos deben estar limpios, organizados y representativos del problema que se quiere resolver.

[8] https://aws.amazon.com/blogs/machine-learning/efficient-continual-pre-training-llms-for-financial-domains/

Figura 3. Fases de entrenamiento de un modelo



Fuente: Lee, J., Stevens, N., Han, S. C., & Song, M. (2024). A survey of large language models in finance (finllms). arXiv preprint arXiv:2402.02315. Traducción Asobancaria.

Instruction Fine-Tuning

En muchos casos, es crucial enseñarles a estos modelos a interpretar instrucciones humanas con precisión y a responder de manera segura y alineada con valores predefinidos. Este es el propósito del *Instruction Fine-Tuning*, una técnica que entrena al modelo para seguir indicaciones explícitas.

En este caso, el modelo se reentrena con ejemplos diseñados específicamente para cubrir un rango amplio de tareas e instrucciones. Los datos de entrenamiento consisten en pares de instrucción-respuesta cuidadosamente curados, donde el modelo aprende a interpretar una solicitud y generar una salida coherente. En lugar de usar datos "crudos" del dominio, los ejemplos se presentan como preguntas, solicitudes o instrucciones, junto con sus respuestas ideales. Esto enseña al modelo a trabajar con lenguaje dirigido por el usuario.

Aprendizaje por refuerzo con retroalimentación humana (RLHF, por sus siglas en inglés)

Cuando hablamos de personalizar un modelo de lenguaje para aplicaciones críticas en el sector financiero, este enfoque destaca por su capacidad de combinar la potencia computacional de los LLM con el juicio humano. Este método no solo ajusta cómo responde el modelo, sino que asegura que lo haga de una manera alineada con los valores, normativas y objetivos del usuario o la entidad, utilizando la retroalimentación humana como guía.

El RLHF comienza donde otros métodos de personalización terminan. Después de un proceso inicial de ajuste fino (*fine-tuning*), el modelo genera respuestas para tareas específicas. Estas respuestas son evaluadas por un equipo de expertos humanos, quienes las califican en función de criterios clave como relevancia, claridad y alineación con los objetivos deseados. Con estas evaluaciones, se entrena un modelo de recompensa, una segunda red neuronal que aprende a predecir qué tan "buena" es una respuesta basada en las calificaciones humanas.Una vez que este modelo de recompensa está entrenado, se utiliza para guiar al LLM principal mediante un proceso de optimización por refuerzo. En este paso, el modelo ajusta sus pesos internos para maximizar la puntuación del modelo de recompensa, lo que significa que prioriza generar respuestas que sean consistentes con las expectativas humanas.

En términos más sencillos, el RLHF no solo ajusta lo que el modelo "sabe", sino también cómo aplica ese conocimiento. En lugar de simplemente responder con precisión, aprende a responder con propósito, adecuándose al contexto en el que será utilizado.

RAG (*Retrieval-Augmented Generation*)

Uno de los grandes desafíos de los métodos de reentrenamiento, como el *fine-tuning* clásico o el *instruction fine-tuning*, es su naturaleza estática. Una vez entrenado, el modelo almacena el conocimiento en sus parámetros, lo que significa que cualquier cambio en los datos o en el contexto requiere volver a ajustar o reentrenar el modelo, un proceso costoso y lento. Aquí es donde entra RAG (*Retrieval-Augmented Generation*), un enfoque revolucionario que combina la generación de texto con la recuperación de información externa en tiempo real. En lugar de depender únicamente del conocimiento almacenado en el modelo, RAG amplía su alcance al integrar bases de datos, sistemas internos y otras fuentes dinámicas.

Con RAG, no necesitas enseñarle al modelo todo lo que necesita saber, así como un analista económico no necesita saberse todos los datos de la serie del PIB de los últimos 20 años. En su lugar, el modelo se convierte en un intérprete altamente capacitado que consulta la información más actualizada en el momento de generar una respuesta. Esto hace que sea ideal para sectores como el financiero, donde los datos cambian constantemente y la precisión es crítica. RAG es una arquitectura que combina dos componentes principales:

1. ***Retriever* (recuperador):** Este módulo busca información relevante en bases de datos externas, sistemas internos o fuentes en línea. Utiliza técnicas como *embeddings* vectoriales para identificar los documentos o datos más relacionados con la consulta del usuario.

2. ***Generator* (generador):** Una vez que se recupera la información, el modelo de lenguaje (como GPT-4) utiliza estos datos como contexto para generar una respuesta precisa y bien fundamentada.

Su éxito depende de configurar correctamente los sistemas de recuperación y establecer una integración fluida con el modelo generativo. Esto implica:

1. **Fuentes de datos bien estructuradas**: Se necesitan repositorios accesibles y actualizados que contengan la información necesaria. En el sector financiero, esto podría incluir bases de datos de clientes, sistemas de transacciones, políticas regulatorias y reportes financieros.

2. **Motor de búsqueda semántica**: El sistema debe ser capaz de recuperar información relevante basándose en la semántica de la consulta, no solo en coincidencias exactas de palabras clave. Tecnologías como FAISS (*Facebook AI Similarity Search*) o *Elasticsearch* son comunes para este propósito.

3. **Modelo generativo robusto**: Un LLM preentrenado que pueda procesar los fragmentos recuperados y generar respuestas contextuales, como GPT-4 o modelos especializados en finanzas.
4. **Infraestructura de integración**: Se requiere un sistema que conecte el motor de recuperación con el LLM, asegurando un flujo eficiente entre la consulta, la recuperación y la generación.

Cuadro 1. Comparación de todos los casos

Criterio	Fine-tuning Clásico	Instruction Fine-tuning	RLHF	RAG (Retrieval-Augmented Generation)
Naturaleza del Modelo	Estático: el conocimiento se almacena en los parámetros del modelo.	Estático, pero especializado para seguir instrucciones específicas.	Estático, pero con ajustes basados en retroalimentación humana.	Dinámico: el modelo accede a datos externos en tiempo real.
Objetivo Principal	Especialización profunda en un dominio específico.	Mejorar la capacidad de seguir instrucciones y responder tareas complejas.	Refinar las respuestas del modelo para alinearlas con criterios humanos.	Combinar generación de texto con información actualizada externa.
Uso Ideal	Tareas altamente específicas y especializadas (análisis de riesgo, contratos legales).	Interacciones conversacionales avanzadas y personalizadas (educación financiera, soporte normativo).	Interacciones donde la calidad, seguridad y alineación son críticas (asesoría personalizada, auditorías).	Consultas que requieren datos dinámicos o en tiempo real (reportes financieros, detección de fraude).
Datos Requeridos	Conjunto grande y especializado (50k-100k ejemplos).	Pares de instrucción-respuesta bien diseñados (10k-50k ejemplos).	Respuestas generadas por el modelo y calificadas por humanos (10k-20k ejemplos de retroalimentación).	Base de datos organizada e indexada, no necesita datos preetiquetados.
Costo Computacional	Alto: ajusta todos los parámetros del modelo base.	Alto, pero optimizado para instrucciones específicas.	Alto: incluye fine-tuning e integración del modelo de recompensa.	Bajo: no requiere reentrenamiento del modelo base.
Flexibilidad	Baja: requiere ajustes cada vez que cambian los datos o el contexto.	Moderada: funciona bien para tareas dirigidas por instrucciones.	Moderada: el modelo se adapta mejor a los valores y expectativas humanas, pero requiere curación constante.	Alta: se adapta fácilmente a diferentes tareas y datos dinámicos.
Ejemplo en Finanzas	Entrenar un modelo para redactar contratos.	Diseñar un asistente que explique procedimientos complejos.	Refinar un modelo para garantizar respuestas éticas y personalizadas.	Generar reportes financieros basados en datos actualizados del mercado.

Fuente: Elaboración Asobancaria.

Conclusiones y consideraciones finales

Los modelos de lenguaje de gran escala (LLM) han demostrado ser herramientas transformadoras en la inteligencia artificial, con capacidades que van mucho más allá de la generación de texto. Su arquitectura basada en *Transformers* y su capacidad para entender el contexto del lenguaje los posicionan como catalizadores de innovación en sectores complejos como el financiero. Desde la automatización de interacciones con clientes hasta la detección de fraudes, los LLM están redefiniendo la manera en que las instituciones procesan información, mejoran la toma de decisiones y optimizan sus operaciones, destacándose no solo por su eficiencia sino también por su adaptabilidad.

Los métodos de personalización y uso de estos modelos permiten adaptarlos a necesidades específicas, desde entrenamientos especializados hasta sistemas dinámicos como RAG (*Retrieval-Augmented Generation*). Cada enfoque tiene sus propias fortalezas y limitaciones, lo que permite a las organizaciones elegir la estrategia más adecuada según sus objetivos y recursos. Por ejemplo, mientras el *fine-tuning* clásico es ideal para aplicaciones de nicho como la redacción de contratos, RAG destaca en entornos donde se necesita información actualizada en tiempo real. Estas capacidades subrayan el valor de los LLM como herramientas versátiles que pueden integrarse fácilmente en los flujos de trabajo existentes.

Finalmente, aunque los LLM presentan un potencial disruptivo, su implementación también implica desafíos, como la necesidad de usar masivamente servicios en la nube, entrenamientos cautelosos y consideraciones éticas en su uso. A medida que estas tecnologías evolucionan, será crucial fomentar la colaboración interdisciplinaria para maximizar sus beneficios y mitigar riesgos. En este contexto, los LLM no solo representan una revolución tecnológica, sino una oportunidad para transformar industrias enteras, haciendo la información y la inteligencia más accesibles y útiles en escenarios críticos.

Principales indicadores macroeconómicos

Colombia

	2021	2022					2023				2024*
	Total	**Total**	T1	T2	T3	T4	**Total**	T1	T2	T3	**Total**
Producto Interno Bruto											
PIB Nominal (COP Billones)	**1192,6**	**1462,5**	385,3	379,9	398,0	409,3	**1572,5**	398,9	406,5	432,4	**1708,4**
PIB Nominal (USD Billions)	**318,5**	**344,6**	81,0	85,8	98,4	99,5	**382,3**	103,8	103,7	105,5	**428,5**
PIB Real (COP Billones)	**907,4**	**907,4**	236,1	239,1	245,7	257,2	**978,2**	237,2	244,9	250,6	**995,8**
PIB Real (% Var. interanual)	**11**	**7,3**	2,9	0,1	-0,6	0,3	**0,9**	0,7	2,1	2	**1,8**
Precios											
Inflación (IPC, % Var. interanual)	**5,6**	**13,1**	13,3	12,1	11,0	9,2	**9,3**	7,4	7,2	5,8	**5,6**
Inflación sin alimentos (% Var. interanual)	**3,4**	**10**	11,4	11,6	11,5	5,0	**10,3**	8,8	7,65	6,5	**5,7**
Tipo de cambio (COP/USD fin de periodo)	**3981**	**4810**	4627	4191	4054	3822	**3822**	3842	3918	4164,2	**4014**
Tipo de cambio (Var. % interanual)	**16**	**20,8**	23,5	1,5	-10,6	-19,3	**-19,3**	-17,0	-11,43	3,6	**5**
Sector Externo											
Cuenta corriente (USD millones)	**-17951**	**-21333**	-2996	-2266	-1758	-2133	**-9715**	-1924	-1630	-1669	**-11140**
Déficit en cuenta corriente (% del PIB)	**-5,7**	**-6,2**	-3,7	-2,6	-1,8	-2,1	**-2,7**	-1,9	-1,6	-1,6	**-2,6**
Balanza comercial (% del PIB)	**-6,4**	**-4,8**	-2,7	-2,5	-1,5	-2,2	**-2,3**	-1,9	-2,2	-2,4	**-3**
Exportaciones F.O.B. (% del PIB)	**13,6**	**21,3**	21,1	19,2	17,6	17,3	**18,6**	15,8	16,6	16,5	**-17,2**
Importaciones F.O.B. (% del PIB)	**18**	**26,1**	23,8	21,7	19,0	19,5	**20,9**	17,7	18,7	18,9	**14,2**
Renta de los factores (% del PIB)	**-2,8**	**-4,9**	-4,8	-3,7	-3,8	-3,3	**-4,0**	-3,3	-3,1	-3,1	**-3,3**
Transferencias corrientes (% del PIB)	**3,4**	**3,6**	3,8	3,5	3,4	3,4	**3,6**	3,3	3,7	3,9	**3,4**
Inversión extranjera directa (pasivo) (% del PIB)	**3**	**4,9**	5,1	6,2	4,0	3,8	**4,8**	3,6	2,7	3,1	…
Sector Público (acumulado, % del PIB)											
Bal. primario del Gobierno Central	**-3,6**	**-1**	0,3	1,2	0,2	…	**-0,3**	…	…	…	**-0,9**
Bal. del Gobierno Nacional Central	**-7**	**-5,3**	-0,9	0,0	-0,7	-2,7	**-4,3**	-1,2	-2,1	…	**-5,6**
Bal. primario del SPNF	**-3,5**	**-1,4**	…	…	…	…	**1,5**	…	…	…	**-0,2**
Bal. del SPNF	**-7,1**	**-6**	…	…	…	…	**-2,7**	…	…	…	**-4,9**
Indicadores de Deuda (% del PIB)											
Deuda externa bruta	**53,9**	**53,4**	55,2	56,1	…	…	**53,6**	…	…	…	…
Pública	**32,2**	**30,4**	31,4	31,8	…	…	**30,9**	…	…	…	…
Privada	**21,7**	**23**	23,8	24,2	…	…	**22,8**	…	…	…	…
Deuda neta del Gobierno Central	**60**	**57,7**	54,1	52,2	52,2	53,8	**53,8**	51,5	55,4	…	**55,3**

*Proyecciones de Asobancaria. Los datos fiscales corresponden a lo proyectado por el Gobierno Nacional en el MFMP 2024

Fuentes: DANE, Banco de la República, Ministerio de Hacienda y Crédito Publico

Estados financieros del sistema bancario Colombia

	dic-20	dic-21	dic-22	dic-23	oct-24 (a)	sep-24	oct-23 (b)	Var. real anual (b) - (a)
Activo	**729.841**	**817.571**	**924.121**	**959.797**	**988.141**	**971.243**	**949.731**	**-1,3%**
Disponible	53.794	63.663	58.321	64.582	52.334	50.299	64.136	-22,6%
Inversiones	158.735	171.490	180.818	189.027	215.041	207.566	182.542	11,8%
Cartera de crédito	498.838	550.204	642.473	655.074	669.857	665.969	657.564	-3,4%
Consumo	150.527	169.603	200.582	196.005	189.389	188.904	197.552	-9,1%
Comercial	263.018	283.804	330.686	338.202	352.416	350.102	341.244	-2,0%
Vivienda	72.565	82.915	95.158	102.972	108.801	107.945	100.988	2,2%
Microcrédito	12.727	13.883	16.047	17.896	19.251	19.018	17.780	2,7%
Provisiones	37.960	35.616	37.224	39.752	40.300	40.096	39.948	-4,3%
Consumo	13.729	12.251	15.970	18.644	18.311	18.423	18.721	-7,2%
Comercial	17.605	17.453	16.699	16.335	17.054	16.792	16.598	-2,5%
Vivienda	2.691	3.021	3.189	3.413	3.520	3.460	3.366	-0,8%
Microcrédito	1.133	913	858	1.181	1.343	1.341	1.092	16,7%
Pasivo	**640.363**	**713.074**	**818.745**	**856.579**	**877.274**	**860.737**	**846.365**	**-1,7%**
Depósitos y otros instrumentos	556.917	627.000	686.622	731.321	766.711	757.344	724.979	0,3%
Cuentas de ahorro	246.969	297.412	297.926	286.217	302.404	292.589	284.831	0,7%
CDT	154.188	139.626	207.859	272.465	291.810	290.727	270.780	2,2%
Cuentas Corrientes	75.002	84.846	80.608	75.483	74.209	72.120	74.689	-5,7%
Otros pasivos	9.089	9.898	11.133	10.841	10.967	10.284	11.388	-8,6%
Patrimonio	**89.479**	**104.497**	**105.376**	**103.218**	**110.867**	**110.506**	**103.366**	**1,8%**
Utilidades (año corrido)	**4.159**	**13.923**	**14.222**	**8.133**	**6.875**	**6.329**	**6.398**	**2,0%**
Ingresos financieros de cartera	45.481	42.422	63.977	91.480	72.338	65.439	76.006	-9,7%
Gastos por intereses	14.571	9.594	28.076	60.093	45.734	41.592	49.708	-12,7%
Margen neto de intereses	31.675	33.279	38.069	35.918	30.387	27.444	29.996	-3,9%
Indicadores (%)								
Calidad	**4,96**	**3,89**	**3,61**	**4,90**	**4,91**	**4,98**	**5,02**	**-0,10**
Consumo	6,29	4,37	5,44	8,10	7,25	7,45	8,13	-0,88
Comercial	4,55	3,71	2,73	3,42	3,86	3,85	3,71	0,15
Vivienda	3,30	3,11	2,47	3,03	3,51	3,49	2,89	0,62
Microcrédito	7,13	6,47	5,46	8,50	9,14	9,49	7,73	1,41
Cubrimiento	**153,5**	**166,2**	**160,6**	**123,8**	**122,4**	**121,0**	**121,1**	**-1,37**
Consumo	145,1	165,4	146,4	117,4	133,4	130,9	116,6	16,80
Comercial	147,1	165,6	184,7	141,2	125,3	124,4	131,2	-5,90
Vivienda	112,3	117,1	135,5	109,3	92,2	91,8	115,5	-23,34
Microcrédito	124,8	101,7	97,9	77,7	76,4	74,3	79,5	-3,10
ROA	0,6	1,7	1,5	0,8	0,8	0,9	0,8	0,03
ROE	4,6	13,3	13,5	7,9	7,5	7,7	7,5	0,01
Solvencia	16,3	20,5	17,1	16,5	17,1	17,3	16,0	1,07
IRL	213,1	204,4	183,7	194,0	188,5	190,1	195,7	-7,21
CFEN G1	0,0	113,5	109,6	115,5	114,5	114,0	114,6	-0,13
CFEN G2	0,0	134,4	127,3	134,4	130,8	131,4	132,2	-1,38

Fuente: Superintendencia Financiera de Colombia.
Nota: G1 corresponde a bancos con activos superiores al 2% del total y G2 a bancos diferentes a G1 que tengan cartera como activo significativo.

Principales indicadores de inclusión financiera

Colombia

	2021	2022					2023					2024		
	Total	T1	T2	T3	T4	**Total**	T1	T2	T3	T4	**Total**	T1	T2	T3
Profundización financiera - Cartera/PIB (%) EC + FNA	**50,9**	50	49,4	48,6	48,3	**48,3**	47,1	46,8	46,7	46,2	**46,2**	45,9	45,4	45,3
Efectivo/M2 (%)	**17**	16,2	15,9	15,6	16,3	**16,3**	14,7	14,3	13,9	15	**15**	14,2	14,1	14,5
Cobertura														
Municipios con al menos una oficina o un corresponsal bancario (%)	**100**	100	100	100	100	100	-	-	-	100	100			
Municipios con al menos una oficina (%)	**79,5**	79,1	77,8	77,8	78,7	78,7	-	-	-	78,7	78,7			
Municipios con al menos un corresponsal bancario (%)	**92,7**	98,6	98,7	99,6	100	100	-	-	-	100	100			
Acceso														
Productos personas														
Indicador de bancarización (%) SF*	**90,5**	91,2	91,8	92,1	92,3	**92,3**	-	-	-	94,6	**94,6**			
Adultos con: (en millones)														
Al menos un producto SF	**33,5**	33,8	34,2	34,4	34,7	**34,7**	-	-	-	36,1	**36,1**			
Cuentas de ahorro	**28,9**	29,2	29,5	29,6	29,9	**29,9**	-	-	-	30,8	**30,8**			
Cuenta corriente SF	**1,9**	1,9	1,9	1,8	1,8	**1,8**	-	-	-	-	-			
Cuentas CAES SF							-	-	-					
Cuentas CATS SF	**21**	21,7		23	23,5	**23,5**	-	-	-	27,5	**27,5**			
Depósitos electrónicos							-	-	-					
CDT	**-**	0,8	0,8	0,9	0,9	**0,9**	-	-	-	-	-			
Al menos un producto de crédito(en millones)	**12,6**	12,8	13,2	13,5	13,6	**13,6**	-	-	-	-	**13,5**			
Crédito de consumo SF	**6,9**	7,1	7,4	7,7	7,8	**7,8**	-	-	-	7,3	**7,3**			
Tarjeta de crédito SF	**7,9**	8	8,2	8,4	8,5	**8,5**	-	-	-	6,6	**6,6**			
Microcrédito SF	**2,3**	2,3	2,34	2,36	2,3	**2,3**	-	-	-	2,4	**2,4**			
Crédito de vivienda SF	**1,2**	1,23	1,25	1,27	1,3	**1,3**	-	-	-	1,2	**1,2**			
Crédito comercial SF	**0,2**	0,46	0,45	0,44	0,5	**0,5**	-	-	-	-	-			
Uso														
Productos personas														
Adultos con: (%)														
Algún producto activo SF	**74,8**	76,2	76,9	77,7	77,2	**77,2**	-	-	-	82,7	**82,7**			
Cuentas de ahorro activas SF	**65,7**	65,9	65,2	64,9	51,9	**51,9**	-	-	-	54,5	**54,5**			
Cuentas corrientes activas SF	**73,7**	76,9	76,5	76,3	74,5	**74,5**	-	-	-					
Cuentas CAES activas SF														
Cuentas CATS activas SF	**76,3**	77,8		80,2	78,6	**78,6**	-	-	-	-	**80,1**			
Depósitos electrónicos														
Productos de ahorro a término (CDTs)	**-**	77,5	79,3	80,1	73,2	73,2	-	-	-	-	-			

Fuentes: Banca de las Oportunidades,

Principales indicadores de inclusión financiera

Colombia

	2021					2022					2023				
	T1	T2	T3	T4	**Total**	T1	T2	T3	T4	**Total**	T1	T2	T3	T4	**Total**
Acceso															
Productos empresas															
Empresas con: (en miles)															
Al menos un producto SF	926,3	924,2	923,8	1028,6	**1028,6**	1029,0	1038,7	1065,7	1077,1	**1077,1**	-	-	-	1169,6	**1169,6**
*Productos de depósito SF	899,2	897,6	898,2	997,9	**998,9**	1004,0	1013,0	1039,8	1046,4	**1046,4**	-	-	-	1166,4	**1166,4**
*Productos de crédito SF	368,9	287,4	282,8	280,2	**280,2**	289,6	294,2	300,6	380,2	**380,2**	-	-	-	417,6	**417,6**
Uso															
Productos empresas															
Empresas con: (%)															
Algún producto activo SF	68,3	68,2	68,1	70,5	**70,5**	71,4	71,2	72,1	72,4	**72,4**	-	-	-	**-**	**-**
Operaciones (semestral)															
Total operaciones (millones)	-	4.939	-	6.222	**11.161**	-	6.668	-	7.769	**14.397**	-	7.500	-	7.808	**15.308**
No monetarias (Participación)	-	55,4	-	56,7	**56,1**	-	55,4	-	56,0	**55,8**	-	49,2	-	39,0	**44,1**
Monetarias (Participación)	-	44,6	-	43,3	**43,8**	-	44,6	-	44,0	**44,2**	-	50,8	-	61,0	**55,9**
No monetarias (Crecimiento anual)	-	-8,7	-	12,4	**2,3**	-	34,0	-	23,2	**27,9**	-	29,4	-	39,2	**34,7**
Monetarias (Crecimiento anual)	-	30,5	-	29,3	**29,1**	-	33,1	-	27,1	**29,8**	-	1,1	-	-29,9	**-15,7**
Tarjetas															
Crédito vigentes (millones)	14,9	14,6	15,0	15,6	**15,6**	15,9	16,0	16,1	16,0	**16,0**	15,8	15,5	15,4	15,0	**15,0**
Débito vigentes (millones)	39,2	38,4	39,7	40,8	**40,8**	41,1	42,6	43,7	45,8	**45,8**	46,2	46,4	47,1	47,2	**47,2**
Ticket promedio compra crédito ($miles)	197,6	208,2	201,4	219,9	**219,9**	215,3	225,2	209,5	225,6	**225,6**	211,1	211,8	200,0	212,6	**212,6**
Ticket promedio compra débito ($miles)	116,8	118,1	114,5	124,9	**124,9**	119,1	116,5	112,5	108,1	**108,1**	100,6	100,7	96,0	111,1	**111,1**

Fuentes: Banca de las Oportunidades, Superintendencia Financiera de Colombia.