GPT y agentes:

Conversando con tus apps

JOSÉ MANUEL PEÑA

Ingeniero industrial por la Universidad de Chile y Magíster en Applied Analytics por la Universidad de Columbia, Nueva York. Actualmente, es el manager del equipo de inversiones en Fintual, una fintech chilena con más de 700 millones de dólares en activos bajo gestión en Chile y México, donde también apoya al desarrollo de iniciativas de IA. Adicionalmente, es miembro del Comité de Inteligencia Artificial y Big Data del Instituto de Ingenieros de Chile.

manu@fintual.com

www.linkedin.com/in/jose-manuel-pena/

RESUMEN. Durante el último año, los modelos de lenguaje de gran tamaño como GPT han revolucionado la interacción humano-máquina, presentando capacidades avanzadas de procesamiento y generación de texto. Aunque estos modelos son poderosos, también tienen limitaciones como la propensión a "alucinaciones", la incapacidad de integración nativa a sistemas tradicionales y la carencia de manejar contenido propietario.

Ante estas limitaciones, nuevas soluciones basadas en el concepto de "agentes" intentan superarlas al integrar modelos de lenguaje con otros sistemas. Estos sistemas basados en agentes utilizan herramientas externas, como bases de datos o APIs, para enriquecer y verificar las respuestas del modelo.

Un ejemplo aplicado es Fintual Copiloto, un servicio de asesoría de inversiones que utiliza estos avances para interactuar con los usuarios y proporcionar información financiera.

GPTs, una revolución incompleta

A lo largo de este año, los colosos del procesamiento del lenguaje, comúnmente conocidos como LLMs (Large Language Models) y más específicamente GPTs (Generative Pretrained Transformers) han irrumpido dramáticamente en la sociedad. Su aparición ha desencadenado una auténtica "explosión cámbrica" de herramientas y aplicaciones, todas aprovechando su notable habilidad para procesar, comprender y generar texto lógico y realista.

Si bien algunos podrían argumentar que estos modelos de última generación no son estructuralmente muy distintos de los primeros modelos basados en mecanismos de atención, surgidos entre 2016 y 2017, su incremento en potencia los ha catapultado a un nuevo nivel.

El uso de menús y botones para navegar en servicios digitales pronto podría convertirse en cosa del pasado.

Aunado a esto, su elegante y simple interfaz —que se basa en mapear texto a texto en una modalidad similar a una conversación— ha sido un factor determinante. Estos dos elementos combinados han hecho que estos modelos entren al epicentro del debate global, tomando el podio como los emblemas de la inteligencia artificial contemporánea, por sobre el reconocimiento de imágenes y otras técnicas de *machine learning*.

Este enorme avance ha dado lugar a una nueva forma de interacción humano-máquina principalmente conversacional. El uso de menús y botones para navegar en servicios digitales pronto podría convertirse en cosa del pasado.

Sin embargo, a pesar de este creciente optimismo, los primeros intentos de implementar soluciones de preguntas y respuestas que dependen exclusivamente de los GPTs como una especie de "oráculo", han demostrado ser totalmente insuficientes para cualquier cosa más allá de la novedad inicial.

El dilema de las alucinaciones

La primera gran limitante del uso indiscriminado de GPTs en espacios productivos es el conocido problema de las "alucinaciones". En simples palabras, estos modelos pueden generar información que no está basada en datos reales. Esto ocurre porque estos modelos tienen como objetivo la imitación de texto realista replicando patrones de lenguaje, dejando la veracidad misma del mensaje en un segundo plano, generando respuestas plausibles sin una fuente de verdad objetiva.

Este problema hace que no podamos (y probablemente nunca) estimar a priori si la información vertida por un LLM es factualmente correcta, haciendo que todo dato generado internamente por un LLM sea considerado sospechoso o potencialmente falso. Esta limitación es un talón de aquiles sobre todo en los casos de uso donde la veracidad de la información es crítica (prácticamente en todas los casos, excepto para usos más creativos o artísticos).

Si bien existe la promesa de poder "dirigir" efectivamente las respuestas de un GPT a un contexto particular mediante el uso inteligente de *prompts* (el texto que se le entrega al modelo), este mecanismo sigue siendo estocástico por naturaleza y es difícil prever todos los posibles caminos que el modelo puede seguir: aunque la respuesta correcta puede existir dentro del corpus de datos sobre el cual el modelo se entrenó, no hay garantías de que el *prompt* entregado devuelva como resultado un texto basado en ese específico contenido. En otras palabras, estamos entregados a un universo de azar.

¿Y si lo entrenamos más? La verdad es que la naturaleza misma del problema hace ineficiente pensar que sólo con más entrenamiento (general o *fine-tuning*) garantice un funcionamiento 100% correcto. Más aún, el estimar cuánto entrenamiento extra (ya sea de todo el modelo o *fine-tuning*) necesitas para eliminar una "alucinación" es extremadamente difícil. Por lo que sólo nos queda tener una necesaria desconfianza de toda la información factual vertida por un GPT, siempre.

Un cerebro sin manos

En segundo lugar, parte del milagro de esta generación de modelos es también una gran limitante (a primera vista). Los GPTs, al enfocarse en mapear texto no estructurado a texto no estructurado, hacen que su integración en sistemas tradicionales sea problemático.

En la actualidad en el mundo digital, desde tu refrigerador hasta tus redes sociales, tú puedes interactuar fácilmente mediante algún tipo de lenguaje estructurado (APIs, interfaces, bases de datos, etc.). Esta limitación inicial hace que los GPTs, sin una mayor integración, sólo puedan crear conversaciones carentes de accionables, lo cual posiciona a estos modelos en el asiento del copiloto o de un consultor, incapaz de gatillar acciones útiles y sólo siendo capaz de sugerir y responder.

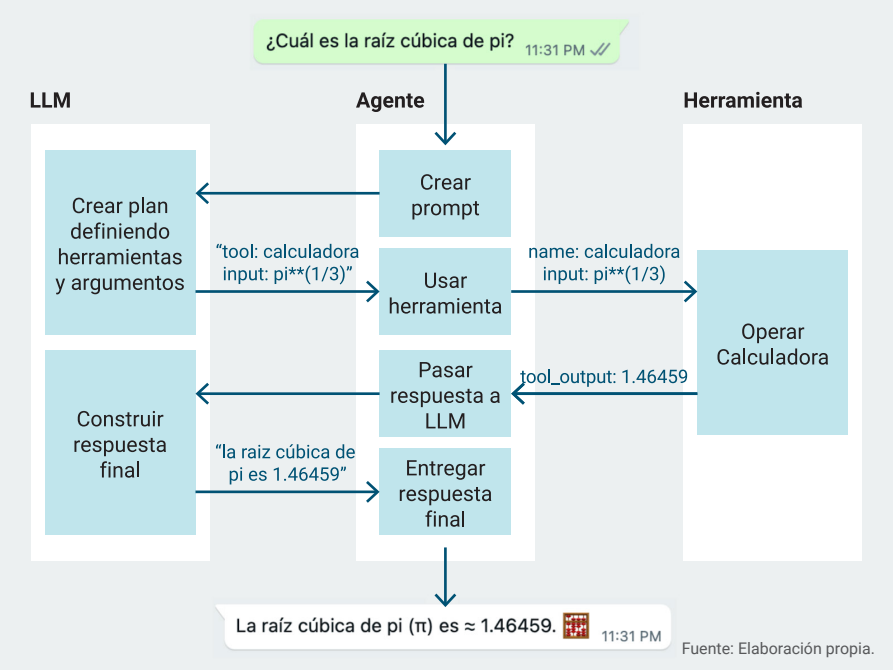


Figura 1. Diagrama estilizado de operación de un agente.

El contenido como diferenciador

Adicionalmente, desde un punto de vista productivo y comercial, las soluciones y aplicaciones basadas exclusivamente en GPTs de manera aislada no pueden sino ofrecer experiencias genéricas, carentes de detalle y expertise propias del contexto.

Alguien podría argumentar que la generación actual de LLMs ha sido entrenada con una cantidad tan vasta de información, que estos modelos pueden comportarse como "expertos" en casi cualquier área (desde pasar la prueba de admisión de abogacía en Estados Unidos, hasta completar con nota sobresaliente difíciles pruebas de conocimiento en medicina). Pero este conocimiento no abarca el universo de información privada que diferencia a una empresa: su *know-how*, datos propietarios, perfil de sus clientes/productos, etc.

En conclusión, para desarrollar sistemas basados en GPTs o LLMs que entreguen información veraz, puedan proveer acciones más allá de simple texto, y puedan operar sobre contextos específicos de manera exitosa y robusta, debemos ser capaces de unir estos modelos con múltiples sistemas, tanto de obtención de información y datos, como de ejecución.

Agentes y herramientas: separando el conocimiento de la razón

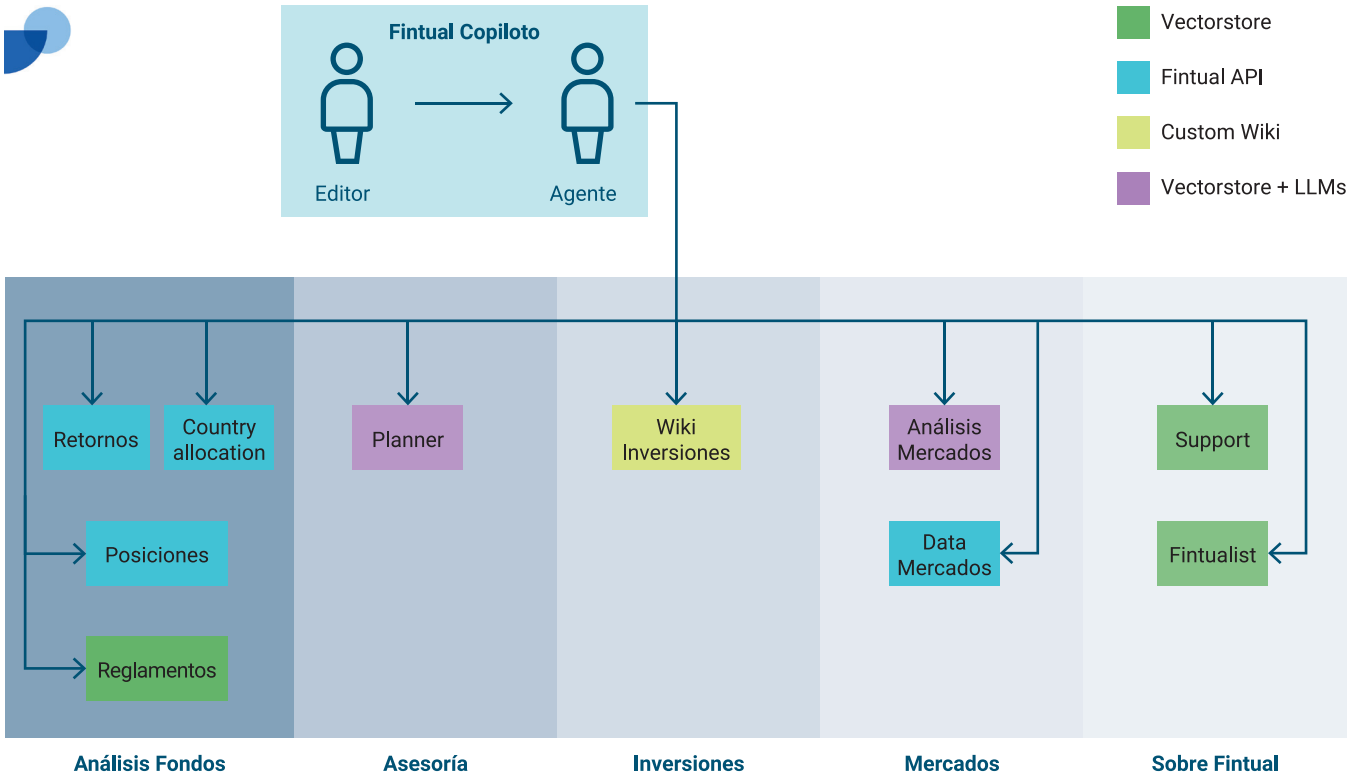
Ante este desafío, han aparecido múltiples proyectos, como Auto-GPT o LangChain entre otros, que buscan potenciar las capacidades de GPTs al integrarlas a múltiples sistemas, convirtiéndolas en ecosistemas capaces de planear y gatillar el uso de herramientas para cumplir los requerimientos de sus usuarios.

Estas librerías no reemplazan los LLMs, sino que en base a llamar a uno o múltiples LLMs, se genera un esquema de trabajo que replica un agente "inteligente" capaz de comunicarse con datos y generar acciones mediante conectores usualmente llamados *plugins* o *herramientas*.

Para entender bien cómo funcionan estos sistemas basados en agentes, primero vamos a definir algunos de los conceptos básicos presentes en la mayoría de estas librerías. En general, un sistema basado en agentes se compone de al menos de 2 elementos: el agente y sus herramientas (ver Figura 1).

Herramientas

Las herramientas son conexiones a sistemas externos al LLM, como lo es una base de datos, una API o cualquier otro sistema que se pueda encapsular en una función que pueda entregar su resultado en forma de texto plano legible para el LLM. Este concepto es muy potente, ya



Fuente: Elaboración propia.

Figura 2. Fintual Copiloto en simple: posee 2 agentes, uno enfocado en edición y estilo, y otro enfocado en contenido. El agente enfocado en contenido posee acceso a una serie de herramientas enfocadas en análisis, asesoría y educación apalancando las APIs internas y públicas de Fintual, acceso a servicios externos y bases de datos documentales para búsqueda semántica.

que cualquier acción, mientras pueda ser expresada como una función, puede ser escrita como una herramienta que recibe y entrega texto. El listado de herramientas posibles crece cada día, desde conexiones con tus herramientas de gestión para manipular correos o presentaciones, hasta la conexión a otros tipos de modelos de inteligencia artificial (como es el caso de HuggingFace Agents) donde puedes invocar el análisis o creación de imágenes, audio, etc.

Agentes

Por otra parte, un agente no es más que un uso inteligente de *prompts* donde, en vez de pedir a un LLM responder directamente una pregunta, se le pide que escriba una estrategia para resolverla en base a una lista de recursos (*tools* o *plugins*), definiendo cuáles usaría y qué argumentos le entregaría. Estas instrucciones luego son ejecutadas y sus resultados se le devuelven al LLM para que genere una respuesta final.

Para ilustrar esto, pensemos en una solicitud simple donde necesito conocer la multiplicación de 2 números. Uno podría directamente pasar a un LLM la consulta "`Cuál es la raíz cúbica de pi?`", lo cual va a depender de la capacidad aritmética implícita del modelo y que puede ser inexacta.

Por el contrario, un agente con acceso a una calculadora daría un contexto donde se pasaría al LLM un *prompt* como éste:

```
Eres un asistente y tienes
acceso a una herramienta
"Calculadora" la cual se invoca
con el siguiente formato:
- tool: 'Calculadora'
- input: sintaxis de la
  operación

Usa esta herramienta para
responder la siguiente
pregunta:
- ¿Cuál es la raíz cúbica
  de pi?
```

Esta manera de formular el mensaje fuerza al LLM a no buscar la respuesta directamente, sino a generar un comando explícito con la herramienta a usar y el argumento necesario para generar la respuesta.

Fuente: Elaboración propia.

Figura 3. Ejemplos de algunas conversaciones en Fintual Copiloto (sic).

Este es un ejemplo excesivamente simple, pero refleja la potencia de este esquema, donde el LLM podría orquestar y combinar el uso de una multitud de herramientas, como bases de datos, documentos, APIs o herramientas de cálculo para generar respuestas factualmente correctas o acciones a pedido del usuario.

Esta manera de formular el mensaje fuerza al LLM a no buscar la respuesta directamente, sino a generar un comando explícito con la herramienta a usar.

Agentes en la asesoría de inversiones: Fintual Copiloto

¿Qué es Fintual?

Fintual es una plataforma digital de ahorro e inversión con presencia en Chile y México que, a través de la tecnología y automatización de procesos, busca simplificar y dar acceso al sector financiero reduciendo barreras como las altas comisiones y montos mínimos. Fintual nace el 2016 de la mano de sus 4 fundadores y hoy cuenta con más de 100 mil usuarios activos y administra más de 700 millones de dólares.

El desafío de la asesoría financiera y el nacimiento de Copiloto

En la misión de Fintual de acercar el mundo de las inversiones y el ahorro a las personas, resulta crítico ser capaz de entregar ayuda oportuna y simple para que sus usuarios entiendan y confíen en los servicios que se entregan. En esta línea, tradicionalmente el mundo de las inversiones se ha basado en tener ejércitos de vendedores (más conocidos como asesores, captadores, etc.) los cuales, haciendo un trabajo 50% asesoria 50% venta, buscan responder dudas, planificar una estrategia de ahorro y recomendar productos para invertir. Esta dependencia en la asesoría 1:1 incrementa los costos de administración enormemente, reduciendo el retorno final que perciben las personas y alejando la inclusión financiera de grupos de bajo patrimonio.

Por esta razón, Fintual desde sus inicios nunca ha contado con fuerza de ventas. Más bien tiene un buen sistema de soporte (donde todos los empleados de

Es ante este desafío donde en Fintual nace la idea de apalancar todo su contenido financiero y herramientas internas de análisis e información mediante el uso de LLMs y agentes para desarrollar un "Copiloto de Inversiones".

Fintual dedicamos tiempo a resolver dudas) y constantemente genera contenido financiero simple para el público general (a través de la revista Fintualist). Desde estos dos flancos ha logrado enseñar e introducir al buen hábito de invertir a miles de personas, donde en más del 75% de los usuarios Fintual es su primera inversión.

Pese a los esfuerzos anteriores, existe un factor de individualización y cercanía que un post de inversiones nunca va a tener, ni una disponibilidad 24/7 inclusive si toda la empresa se dedicara al soporte. Es ante este desafío donde en Fintual nace la idea de apalancar todo su contenido financiero y herramientas internas de análisis e información mediante el uso de LLMs y agentes para desarrollar un "Copiloto de Inversiones", *aka* Fintual Copiloto.

Copiloto: un asesor 24/7

Fintual Copiloto (ver Figuras 2 y 3) se trata de un servicio de asesoría en el que con sólo enviar un mensaje al WhatsApp de Fintual podrás conversar con nuestro agente de inversiones (por ahora sólo disponible en México). Sus objetivos principales son ayudar a organizar las finanzas personales de los clientes, informar sobre los productos de Fintual y resolver dudas sobre el mercado, economía y las inversiones en general.

Visión a futuro: de vuelta al quiosco

En vista de esta evolución de modelos, desde sus orígenes intentando resolver simples tareas de procesamiento de texto, hasta convertirse en los potentes LLMs actuales capaces de obedecer órdenes y operar como agentes conversacionales, ¿cómo se ve el futuro relativo a nuestra relación con servicios digitales?

Una analogía útil para ver lo que se viene es la siguiente. El paradigma actual de diseño y desarrollo, y aplicaciones, se parece mucho a una tienda por departamentos o supermercado; los productos (información o acciones) que necesitan los usuarios son ubicados dentro de una aplicación la cual, mediante la navegación vía menús y botones, buscan encontrar lo que necesitan, similar a como tú navegas con tu carrito de compras entre pasillos, leyendo la señalética para encontrar lo que necesitas.

En el futuro, la capacidad de conversar hace posible el que estos mismos activos, al ser conectados a agentes, puedan ser encontrados y accionados simplemente pidiéndolos. Es más, el usuario ya no requiere saber a priori cuál es el activo específicamente que necesita, sino que en base a la conversación con su asesor, llega a definir qué es lo que más le conviene. Este comportamiento lo conocemos desde hace mucho tiempo, ya que es como opera cualquier tienda de la esquina o quiosco, donde quien conoce a cabalidad los productos es el administrador y es él quien recomienda y pone a disposición lo que necesita el usuario.

En conclusión, los LLMs son herramientas revolucionarias pero lejos de ser perfectas. Sin embargo, su integración en sistemas está permitiendo una nueva forma de interactuar de forma natural y desestructurada con datos y aplicaciones. Y si bien todavía estamos en sus primeras iteraciones, con un poco de optimismo es posible que pasemos de un presente de menús y botones, a un futuro como el de las películas, donde a la tecnología se le habla. ▪