ANDBANK
Private Bankers

Flash note
02/08/2024

Alex Fusté
@AlexfusteAlex
alex.fuste@andbank.com

Desarrollos Tecnológicos Relevantes en el Mercado II

Esta nota da continuidad a la nueva sección titulada "Desarrollos Tecnológicos Relevantes en el Mercado". Una serie de publicaciones con la que perseguimos compartir con ustedes nuestra valoración sobre los avances tecnológicos que consideramos poseen el potencial para continuar impulsando el mercado.

Desarrollos durante el mes de julio

1. **Nueva variante de GPT-4o**. Open AI ha dado un nuevo paso y ha lanzado el nuevo modelo de lenguaje llamado GPT-4o LONG OUTPUT. Se trata de una nueva variante con una capacidad de salida de tokens aún mayor. Diseñado para usuarios profesionales, investigadores, etc, que requieren respuestas y reflexiones más ricas y detalladas. Hasta ahora el modelo estaba limitado a 4000 tokens de salida (suficiente para muchas aplicaciones, pero insuficiente para necesidades más complejas. Con GPT-4o Long Output los usuarios más rigurosos (del ámbito de la ciencia) obtienen salidas de 64.000 tokens de profundización (equivalente a un libro de 200 páginas). El límite de contexto de 128.000 tokens se mantiene (es decir, la interacción que usa tokens de entrada y de salida no puede superar esa cifra). Aun así, esta actualización nos parece relevante y con implicaciones practicas importantes. Abre nuevas posibilidades para empresas y proyectos que precisan de modelos de IA cognitiva (procesos mentales involucrados en el conocimiento y la comprensión). ¿Permitirá esta actualización un salto en la cognición humana? No lo sé, pero doy por hecho que las funciones de adquirir, procesar, almacenar y utilizar conocimiento experimentarán un salto cualitativo. A partir de ahí, habrá que ver cómo afecta a la reflexión y la resolución de problemas. El precio del servicio será de 18$ por millón de tokens de salida. Un precio agresivo (bajo). Importante, pues hará que estas nuevas capacidades sean muy accesibles para un amplio espectro de usuarios e investigadores.

2. **Open AI empieza a desplegar el nuevo modo de voz de Chat GPT.** No es un lanzamiento masivo de la versión final, si no una versión para usuarios de GPT Plus. Lo hemos probado y solo puedo decir que hablar con la IA de forma totalmente natural es ya una realidad. Una de las mejoras relevantes frente al modo de voz original es que en esta versión puedes interrumpir la conversación y reconducirla de forma totalmente fluida. Otra de las mejoras es la posibilidad de mantener conversaciones emocionales (puedes pedirle al modelo que hable con el estilo de un comentarista de futbol argentino cuando Boca le mete un gol a River). Bromas aparte, para entender este progreso les diré que el modelo anterior convertía la voz en texto y luego GPT4 procesaba dicho texto para convertirlo en voz. Ahora, GPT4o, al ser un modelo multimodal lo procesa directamente

consiguiendo una latencia inapreciable. Muy natural. Este nuevo modo de voz no está limitado al inglés si no que está probado 45 idiomas. Con cuatro voces disponibles. A los fans de Scarlett Johanson, lamento decirles que la opción con la voz de la actriz no estará disponible (era la opción Sky). La actriz rechazó la oferta de Altman para poner su voz en el *chatbot* más famoso del mundo.

3. **Nvidia va a darle a China la solución que necesita para competir en IA:** En marzo, Nvidia presentó su Blackwell B200, un procesador bestial con 208.000 millones de transistores y que todas las grandes tecnológicas quieren. Debido a las restricciones impuestas por los EUA, China no tiene acceso a los últimos procesadores de Nvidia. Las firmas chinas están buscando alternativas que no son más que parches comparado con lo último de Nvidia. Pero ahora, Nvida ha alcanzado un acuerdo con el gobierno USA para poder vender a China una versión específica de su Blackwell B200. Esta versión se llama B20 y presenta un alto rendimiento (aunque menor) en centros de datos con refrigeración líquida (donde Super Micro Computer es líder). ¿El papel de China en la carrera de la IA? Desde la imposición de las restricciones con los Chips a China, Nvidia ha creado tres versiones que sí podían ser vendidas al mercado chino (L2, L20, H20). Ninguno era su procesador más avanzado, y se trataban de GPUs más débiles para que su venta a China fuera autorizados por el Departamento de Comercio de los EUA. Para que se hagan una idea, el modelo H20, que era el modelo más avanzado que Nvidia podía vender a China, es siete veces menos potente que el H100 que Nvidia comercializa en el resto del mundo. Es de esperar que la versión B20 guarde una relación similar respecto a la B200. Esto son buenas noticias para Nvidia y el ecosistema tecnológico (semiconductores), pues el mercado Chino representa todavía un 17% de los ingresos anuales de Nvidia (hace dos años era el 26%). Pero esto ya nos da una idea del lugar en donde está quedando China en esta carrera.

4. **Se incorporan capacidades de Visión en un modelo IA.** Los modelos de visión son algo nuevo y diferente a los modelos de imagen (que permiten crear fotos y videos a partir de textos descriptivos). He visto alguna demo de este modelo de visión y presencié como Chat GPT ayudaba a unos niños con sus deberes observando sus cuadernos, o describiendo con detalles lo que hay en una sala, o cualquier espacio. Estas funciones están impulsadas por las capacidades de visión de GPT-4o y sospechamos pueden tener aplicación en muchas industrias (por ejemplo, conducción autónoma, cirugía, etc.). Este modelo se lanzará en una fecha aún por determinar.

5. **Friend. El primer wearable (colgante) que lo escucha todo durante 24 horas y está diseñado para convertirse en lo que su nombre indica. Tu mejor amigo.** Basado en el chatbot Claude 3.5 (versión más nueva y potente de Antrophic), dispone de varios micrófonos y una batería, funciona con el iphone y costará 99$ (sin suscripción asociada). Hemos visto algunos de los casos de uso y situaciones de interacción entre usuario y el dispositivo, y es capaz de inferir una respuesta adecuada y coherente en función del historial de escucha, por largo que sea. Algunos ejemplos: El dispositivo capta el contenido de TV que el usuario está viendo y comenta espontáneamente la escena o incluso el sentido completo de la película. Otro ejemplo interesante es cuando Friend regaña a una empleada que se está tomando un descanso más largo de lo permitido. O si estás jugando una partida y estás perdiendo, el dispositivo reacciona espontáneamente con una pequeña mofa. Pero el ejemplo que más me llamó la atención fue cuando un usuario derramó un poco de salsa sobre su dispositivo colgante

y éste respondió automáticamente con un "Yum, qué rico". Se supone que esta IA no dispone del sentido del gusto. Ello demuestra que el sentido de visión está extremadamente desarrollado. ¿Puede ser esto tu mejor amigo, como pretenden sus creadores? No lo sé pero ¿Se imaginan pedir consejo sobre cómo actuar a algo lógico que observa todo lo que ocurre y todo lo ocurrido a tu alrededor? A medida que asisto observante a estos desarrollos, no puedo evitar sentir una cierta inquietud sobre las implicaciones que esto pueda tener en las relaciones humanas. En fin. Nuevos tiempos.

6. **Meta presenta Llama 3.1. Un poderoso IA de código abierto para competir con GPT-4o.** Meta no quiere quedarse rezagado en la carrera de la IA y lanza el que será el primer modelo abierto (camino distinto al de Google i OpenAI, que optaron por desarrollar modelos cerrados como Gemini o GPT-4. Según Meta, su modelo Llama 3.1 es el modelo de lenguaje con más capacidad en el mundo, entrenada con 15 billones de tokens, priorizando no solo la cantidad, si no la calidad de respuestas. Meta afirma que la tardanza en sacar su producto se debía precisamente al fuerte proceso de entrenamiento. De hecho, rivaliza bien con GPT-4o o Claude 3.5 en matemáticas, traducción y conocimientos generales. Cosas extrañas del capitalismo. ¿No creen? No me mal interpreten. El capitalismo es ese sistema superior donde la libertad se mide por la cantidad de cosas que puedes comprar, aunque para conseguirlo tengas que hipotecar tu vida. No como el comunismo, donde la igualdad de oportunidades está asegurada... siempre y cuando te conformes con que todos tengan igual de poco.

7. **Gemini (Google) se vuelve más rápido y poderoso con su última actualización (modelo Flash 1.5)**

8. **Grok, la IA de "X" (Twitter) podría tener ventaja sobre el resto.** La red social ha empezado a usar los datos de los usuarios para entrenar a su modelo de IA llamado Grok, que además de estar integrada en la red social (orientada a la monetización de contenidos), también está integrada en sus vehículos Tesla. Elon Musk empezó a usar los datos de los usuarios sin aviso previo, pero se puede revocar dicho permiso**.**

9. **IA y Corrupción.** No sé si habrán oído hablar del fraude del CEO. Yo mismo sufrí uno. Se trata de un caso en el que recibes un email de tu CEO indicándote que realices una transferencia urgente. Por su puesto en el mail te dice que en ese momento no puede hablar pero que más tarde te contactará para darte los detalles. Mi experiencia resultó algo cómica y por supuesto no se materializó. La cosa cambia ahora con la IA, ya que permite que recibas una llamada con la voz de tu CEO suplantada, o incluso una video llamada con tu mismísimo CEO en pantalla ordenando una transferencia urgente. ¡Claro! Puede darse el caso que no te atrevas a decirle que no a tu CEO, sin saber que en realidad es tu falso CEO. Yo mismo he creado un video sintético de más de 15 minutos de mí mismo. Los casos de *Deepfake* con IA de imagen ya están ocurriendo en todos los países. Algunos con éxito, por cierto. El más sonado ha ocurrido hace pocos meses, con la policía de Hong Kong reportando un caso de una multinacional financiera en la que un ejecutivo fue convocado a una reunión de videollamada por sus superiores (que en realidad no lo eran). La experiencia resultó en un robo de USD25 millones según las autoridades.

10. **Ritmo de penetración de la IA en la sociedad**. El banco JPMorgan ya ha desplegado una herramienta de IA (LLM Suite), una suerte de Chat GPT interno para ser usado por sus empleados y analistas y ayudarles en sus tareas diarias. Esta será una dinámica

constante, en la que el último en abrazar estas capacidades quedará relegado. Esta es una de las razones por las que considero que va a haber un salto importante en inversión y provisión de estos servicios. En mi opinión, un factor necesario y suficiente para que los drivers que han impulsado el mercado continúen.

11. **IA de imagen: Runway** lanza su nueva versión de videos a partir de *promtps* de texto y que aumenta mis capacidades creativas de imagen. Se llama Gen3 y la hemos probado. Con la versión económica puedo hacer una serie de clips (los que quiera) de 10 segundos cada uno y con el nivel de fantasía que uno desee, y después unirlos todos mediante otra aplicación (Cap cut), editarlo, reeditarlo, poner música, voz, etc… y acabar realizando una composición larga original. **VozoRewrite & Redub:** Aplicaciones de IA que me permiten transformar narrativas de videos existentes, como cambiar el tono, el idioma o incluso el mensaje completo. Esto abre la veda para narrativas rediseñadas, redoblar con voz clonada, editar voz con texto, sincronización labial avanzada, etc. Puede parecer divertido, pero no debe resultar tan divertido para el presidente de la nación española (por ejemplo) cuando ve en redes un video suyo diciendo boberías.

12. **Seguridad para las empresas que integran la IA en sus operaciones.** Dioptra es una herramienta desarrollada por el NIST (National Institute of Standards and Technology) de EEUU diseñada para asegurar el desarrollo confiable de la IA. Evita que los modelos de IA basados en datos erróneos puedan actuar de forma impredecible. Por ejemplo, una empresa que fabrica coches autónomos entrenados a partir de datos puede ser objeto de ataques a su base de datos de entrenamiento, que puede ser manipulada de forma maliciosa (desde inyecciones de datos erróneos hasta las manipulaciones más sutiles que puedan alterar el comportamiento del sistema). En tal caso, puede que sus coches no reconozcan correctamente una señal de tráfico. Dioptra permitirá a empresas, agencias y gobiernos responder a ataques malintencionados, centrándose en la seguridad de los datos para el buen funcionamiento de los modelos de IA. Dioptra será gratuito, permitiendo a pequeñas y medianas empresas proteger sus modelos de IA sin necesidad de grandes inversiones. Esto es muy relevante en un contexto donde muchas organizaciones empiezan a integrar la IA en sus operaciones diarias.

Cordiales saludos,