iPhone vendrá con modelos de lenguaje basados en IA

Apple presentó OpeELM, una nueva familia de grandes modelos de lenguaje (LLM, por sus siglas en inglés) basados en inteligencia artificial (IA) que pueden ejecutarse con eficiencia en dispositivos personales como computadoras portátiles o smartphones.

OpeELM está disponibles en cuatro tamaños. Sus versiones son capaces de comprender y gestionar entre 270 millones y 3,000 millones de parámetros o instrucciones complejas. Cada algoritmo ofrece una versión de entrenada y una ajustada con instrucciones. La primera es útil para producir texto coherente mediante un proceso predictivo, con base en los datos de preentrenamiento.

La segunda permite añadir información personalizada para que el sistema responda con resultados más relevantes a solicitudes específicas del usuario.

OpenELM no requiere de una conexión constante a servidores en la nube, a diferencia de otros LLM. Puede ser ejecutado por completo en un dispositivo portátil, lo que sugiere tiempos de respuesta más rápidos y mayores garantías de privacidad.

La nueva familia de modelos de IA de Apple utiliza una estrategia de escalamiento por capas. Esto significa que el número de parámetros en cada capa del modelo transformer se ajusta de forma independiente, en lugar de utilizar un enfoque uniforme para todo el modelo.

La compañía asegura que con esta técnica es posible obtener resultados más precisos y con un menor consumo de recursos computacionales. “En su versión de 1,000 millones de parámetros, OpenELM muestra una mejora de 2,36% en la precisión en comparación con OLMo y requiere dos veces menos tokens de preentrenamiento”. OLMo es el LLM más reciente diseñado por el Instituto Allen para la IA .

Los ingenieros de Apple probaron la eficiencia de OpenELM en una MacBook Pro con chip M2 Max y 64GB de RAM, en un ordenador con procesador Intel Core i9-13900KF con 64GB en RAM y una GPU NVIDIA RTX 4090. Los resultados fueron satisfactorios en términos de precisión y razonamiento.

La variante OpenELM de 3,000 millones de parámetros obtuvo una

precisión de respuesta de 42.2% en el indicador ARC-C, diseñado para calificar los conocimientos y habilidades de razonamiento de los modelos de IA. Alcanzó un 26.7% en el punto de referencia MMLU que evalúa la comprensión de lenguaje de los algoritmos, y 73.2% en HellaSwag, un benchmark que determina la capacidad de razonamiento de sentido común de los sistemas de IA.

OpenELM está disponible en HuggingFace bajo una licencia de código abierto, que permite usos comerciales y la modificación de los algoritmos.

Sin embargo, Apple advierte que su nueva familia de IA “se pone a disposición sin ninguna garantía de seguridad. En consecuencia, existe la posibilidad de que estos modelos produzcan resultados inexactos, dañinos, sesgados u objetables en respuesta a las indicaciones de los usuarios”.

Con información de UN