Adiós a escribir lento: la nueva IA que genera texto 3 veces más rápido

15/04/2026

I-DLM cambia la lógica de generación de texto al ofrecer paralelización sin sacrificar calidad. Este avance técnico promete reducir costos operativos en infraestructura de IA y acelerar la respuesta en productos con muchos usuarios concurrentes.

Qué hace y para que se puede utilizar la nueva IA que genera texto mucho más rápido que cualquier otro sistema

Hasta ahora los LLMs más usados operaban autoregresivamente generandotoken por token en secuencia. Ese enfoque **limita el throughput porque no permite calcular el token 50 hasta completar el 49**, aun con hardware potente. La

consecuencia: **despliegues caros y latencias altas para productos en producción**.

Los modelos de difusión prometían paralelizar la salida pero hasta ahora **les faltaba consistencia y calidad frente a contrapartes autoregresivas.** El equipo detrás de **I-DLM** detectó esa carencia y **diseñó un entrenamiento que añade verificación interna**, logrando que **la salida sea coherente sin renegar el rendimiento en benchmarks clave.**

Cómo funciona nueva IA que genera texto 3 veces más rápido que los otros chats

La decodificación Introspective Strided Decoding (ISD) es el núcleo del método. En cada pasada el modelo **propone tokens en posiciones enmascaradas y verifica otros 'limpios'** con una distribución ancla. Ese chequeo usa una **regla de aceptación probabilística** que asegura que **la distribución final sea equivalente a la de un modelo autoregresivo,** pero permitiendo generar varios tokens en paralelo.

En las pruebas públicas I-DLM mostró números contundentes, los números dieronuna tasa de aceptación **muy alta, overhead reducido** y una pendiente de infraestructura que supera por amplio margen a otros **DLMs**.

Estos resultados indican que, **en entornos limitados por memoria** el modelo logra **aproximadamente 3x de aceleración por token respecto a alternativas previas. I-DLM** permite además **convertir modelos autoregresivos preexistentes vía fine-tuning,** por lo que la adopción **puede ser práctica sin reconstruir todo el stack de IA desde cero.**

Fuente: La 100