Modelos MAI: La Siguiente Generación de IA Interna

Una plataforma de IA aplicada al servicio de la humanidad

Puerta al Conocimiento

Una interfaz confiable que habilita a individuos y organizaciones para lograr resultados extraordinarios.

Presencia de Apoyo

Diseñada para empoderar a cada persona del planeta, entendiendo necesidades únicas con personalidad y experiencia.

Productos Confiables

IA aplicada responsablemente, sirviendo como plataforma para definir nuevas categorías de productos profundamente confiables.

De la infraestructura fundacional a los modelos de propósito específico

Spinning the flywheel

2024: Construcción Base

Enfoque intensivo en el ensamblaje de un equipo de clase mundial y el despliegue de infraestructura de supercómputo global.

Hoy: Aplicación Directa

Transición hacia modelos creados con un propósito específico, necesarios para cumplir la visión de MAI en el ecosistema Copilot.

MAI-Voice-1 redefine la interfaz del futuro

Spec Card

Modelo	Naturaleza
MAI-Voice-1	Generación de voz natural de alta fidelidad.
Capacidad	
Audio expresivo para escenarios de un solo locutor y multilocutor.	

Visión MAI: La voz es la interfaz del futuro para los compañeros de IA.

La ecuación de eficiencia en generación de voz

1 sola GPU

< 1 Segundo

Tiempo de inferencia

60 Segundos

Audio de alta fidelidad generado

MAI-Voice-1 genera un minuto completo de audio de alta fidelidad en menos de un segundo de inferencia en una sola tarjeta gráfica, posicionándose como uno de los sistemas de voz más eficientes del mundo actual.

MAI-Voice-1 en producción activa y experimentación

Copilot Daily

Resúmenes diarios potenciados por una voz natural y conversacional.

Podcasts

Generación de contenido en formato largo, aprovechando la capacidad multilocutor.

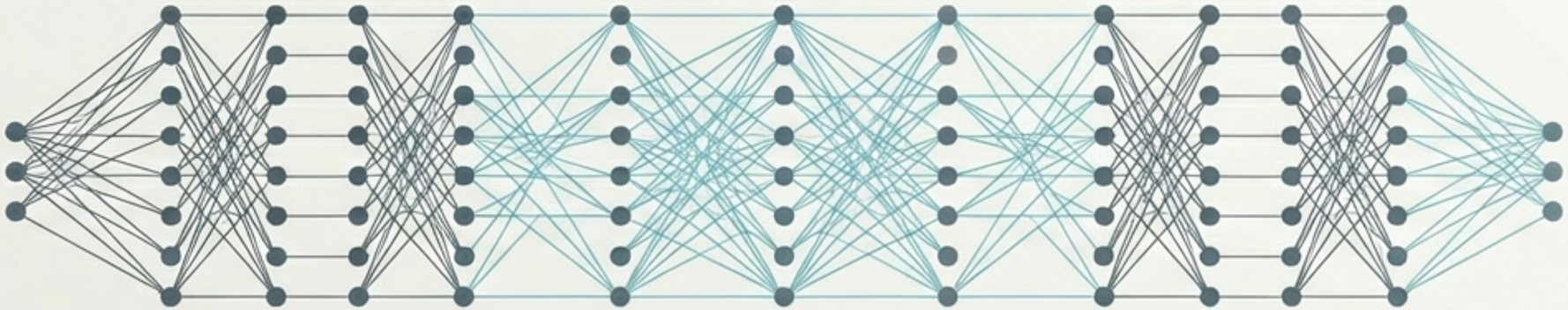
Copilot Labs

Demos interactivos para el usuario final.

Ejemplos activos:

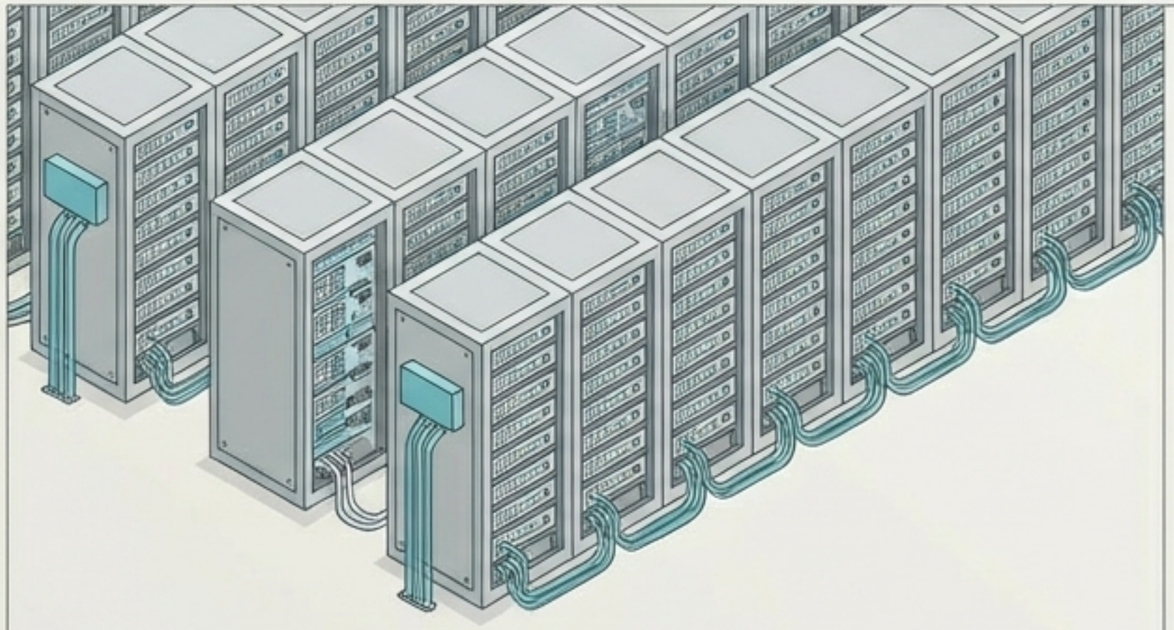
- Niño interactuando con un rudo capitán pirata.
- Demostración tecnológica con un vaquero escéptico.

MAI-1-preview es nuestro motor fundacional de texto



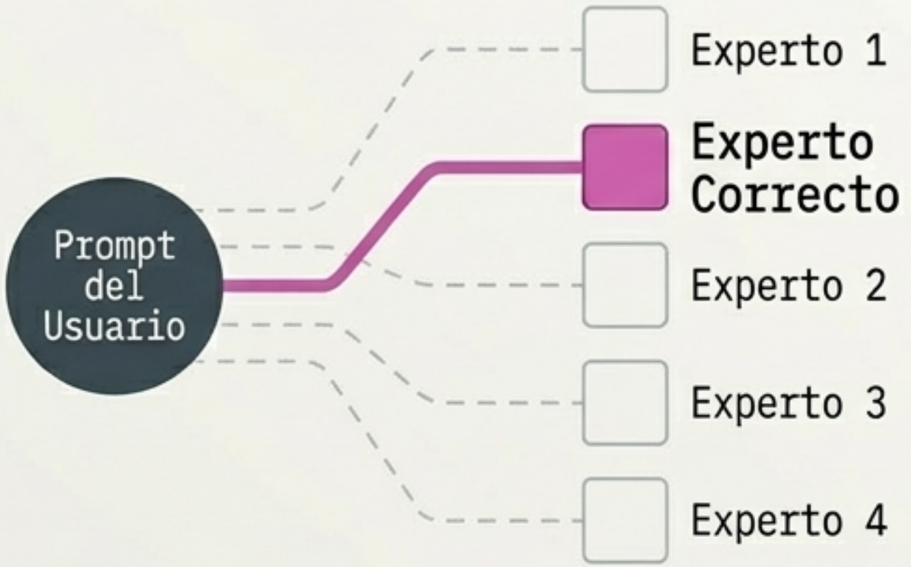
Modelo **MAI-1-preview**	Arquitectura Mixture-of-Experts (MoE) desarrollada internamente.
Entrenamiento Sistema entrenado End-to-end.	Propósito Especialización en seguimiento estricto de instrucciones y respuestas altamente útiles para consultas cotidianas complejas.

Escala masiva y enrutamiento inteligente (MoE)



~15,000 GPUs NVIDIA H100

Pre-entrenado y post-entrenado en uno de los clústeres de supercómputo más grandes del mundo.



La arquitectura MoE dirige inteligentemente cada tarea específica a la sub-red experta correspondiente, optimizando tanto el rendimiento de la respuesta como la eficiencia del clúster.

El ciclo de validación técnica en el mundo real

Pruebas Públicas

Despliegue en LMArena, la plataforma principal para evaluación de la comunidad global de IA.

Acceso Privado

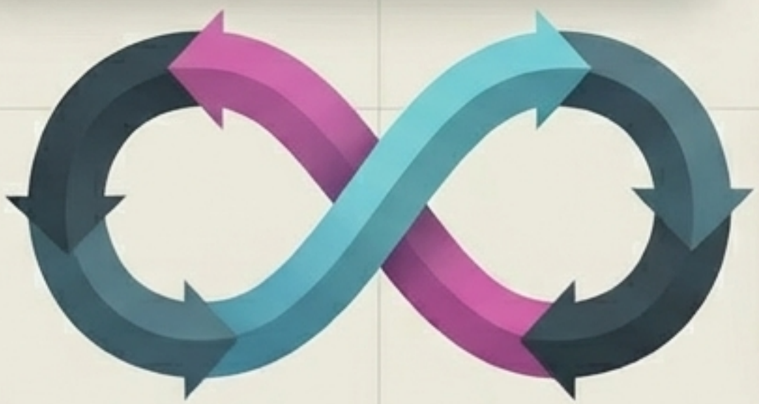
Despliegue seguro a través de API para Trusted Testers.

Feedback Activo

Recolección de datos sobre rendimiento y casos de uso cotidianos.

Despliegue en Producto

Próxima integración gradual en casos de uso de texto dentro de Copilot basada en los aprendizajes.



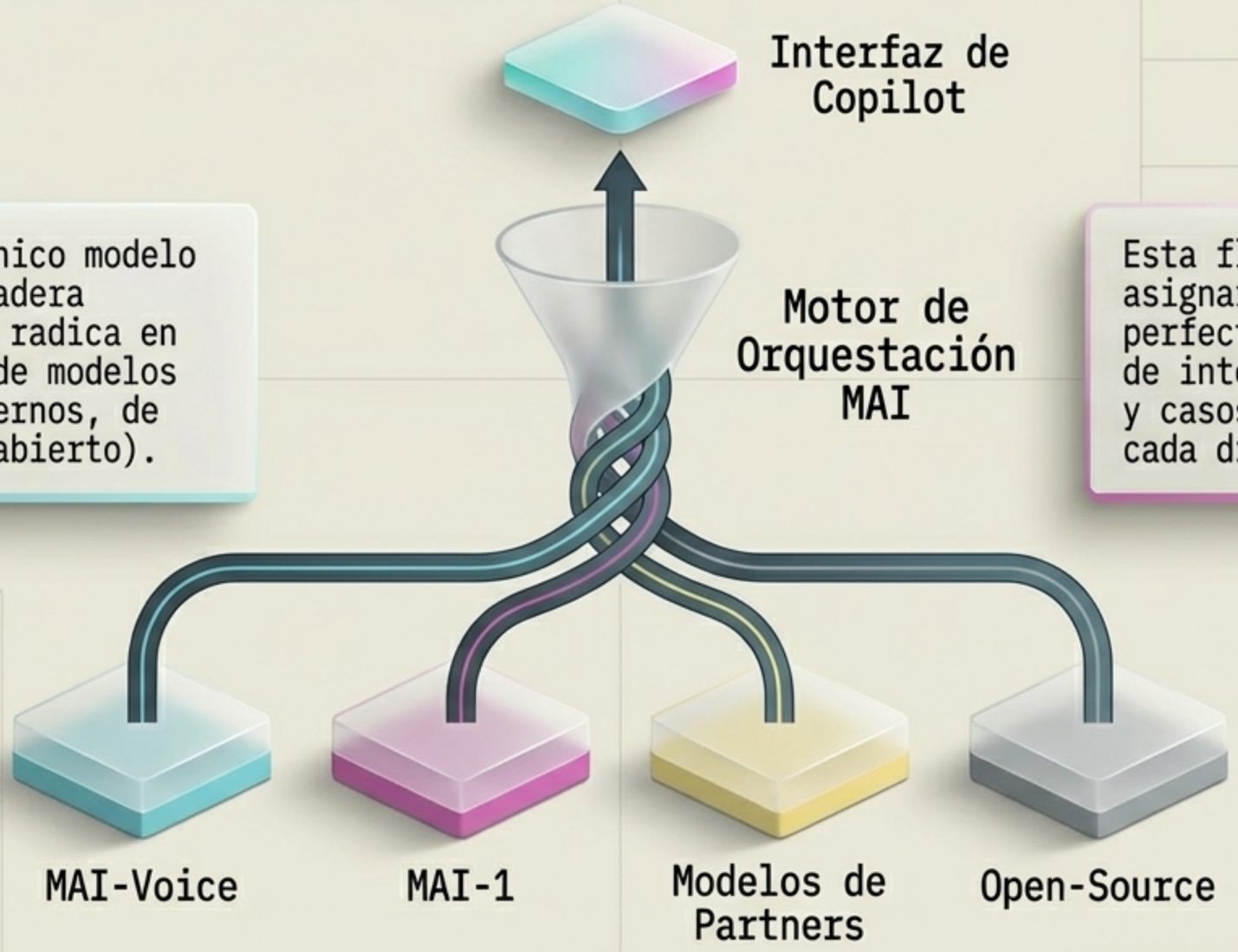
Anatomía del ecosistema bimodal actual

Diagnostic Table

	Naturaleza	Arquitectura	Hito de Cómputo	Estado de Despliegue
MAI-Voice-1	Voz de alta fidelidad	Generador de inferencia ultra-rápida	1 min de audio en < 1s en 1 GPU	Producción (Copilot Daily / Labs)
MAI-1-preview	Texto e instrucciones	Mezcla de Expertos (MoE) entrenada end-to-end	Entrenado en ~15,000 H100 GPUs	Pruebas Públicas (LMArena / API Testers)

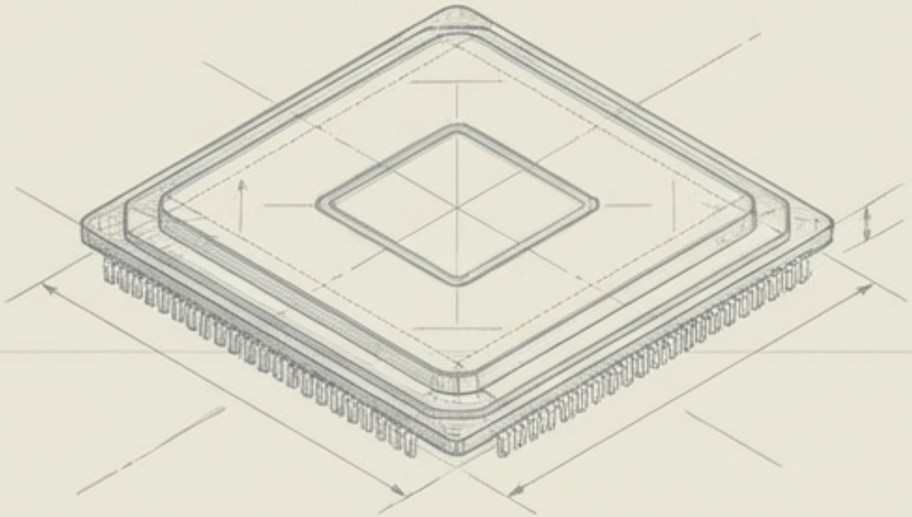
Orquestación estratégica sobre modelos monolíticos

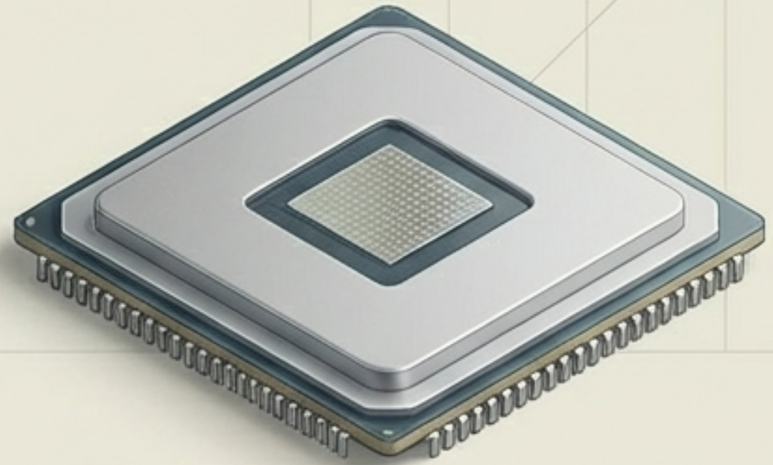
No perseguimos un único modelo monolítico. La verdadera generación de valor radica en orquestar una gama de modelos especializados (internos, de socios y de código abierto).



Esta flexibilidad permite asignar la herramienta perfecta para millones de intenciones de usuario y casos de uso únicos cada día.

El futuro de nuestra capacidad de cómputo ya está en marcha





Infraestructura GB200

Hoja de Ruta

Una hoja de ruta agresiva y ambiciosa.

Estado Actual

Nuestro clúster GB200 de próxima generación ya está completamente operativo.

El Objetivo

Impulsar el desarrollo de las siguientes fronteras de inteligencia artificial con un techo de cómputo drásticamente superior.

Construye el futuro con el laboratorio más ágil del mundo

Somos un laboratorio ágil y de rápido movimiento compuesto por algunas de las mentes más brillantes del mundo. Buscamos individuos brillantes, altamente ambiciosos y con bajo ego para impactar a miles de millones de usuarios.

Aplica como Trusted Tester (API Access)

Explora posiciones abiertas en Ingeniería y Producto