

Harpia

Versión actual: harpia-1

Visión General

Amazônia IA es el principal modelo de visión-lenguaje (VLM) de la familia Amazônia IA. Integra un potente Modelo de Lenguaje Grande (LLM) con un sofisticado sistema de visión, ofreciendo un rendimiento de vanguardia al comprender y razonar sobre información visual entrelazada con texto.

- **Arquitectura del Modelo:** Combina un LLM base con un Vision Transformer (ViT). El ViT procesa imágenes en sus resoluciones nativas, permitiendo una comprensión visual detallada.
- **Características Principales y Capacidades Fundamentales:**
 - **Comprensión Multimodal:** Procesa y comprende nativamente texto, imágenes y videos entrelazados. Soporta análisis de videos de larga duración (más de una hora) con identificación de eventos (localización temporal).
 - **Análisis Visual y Documental Avanzado:** Realiza análisis detallado de imágenes, reconociendo elementos refinados, texto (OCR), gráficos, iconos y layouts complejos. Localiza objetos con precisión (grounding) vía cajas delimitadoras/puntos, proporcionando coordenadas JSON estables para razonamiento espacial. Destaca en el análisis de documentos complejos con layouts diversos (usando formatos como HTML), texto en varios idiomas, escritura a mano, tablas, gráficos y fórmulas.
 - **Interacción Agéntica:** Funciona como un agente visual, aprovechando la localización y el razonamiento para interactuar con interfaces de usuario en computadoras y dispositivos móviles.
 - **Razonamiento y Salida Mejorados:** Demuestra fuerte razonamiento matemático (visual/textual) y proporciona respuestas detalladas y bien formateadas alineadas con las preferencias humanas.

Especificaciones Técnicas

- **Modalidades de Entrada:** Texto, Imagen, Video
- **Modalidades de Salida:** Texto
- **Ventana de Contexto:** Soporta hasta 128K tokens. Máximo recomendado de 32K tokens para tareas de visión-lenguaje debido a patrones típicos de uso y costo

computacional. Contacte a WideLabs si ventanas de contexto mayores son esenciales.

- **Idiomas:** Multilingüe (decenas de idiomas). El rendimiento en idiomas específicos puede ser mejorado a través de fine-tuning.

Rendimiento

Alcanza resultados altamente competitivos en una amplia gama de benchmarks, demostrando rendimiento excepcional, especialmente en comprensión de documentos, razonamiento complejo y tareas agénticas.

Benchmark Category	Benchmark	Score	Notes
Text (Alignment)	IFEval	85.8	Evaluación de Seguimiento de Instrucciones
Text (Code)	HumanEval	85.4	Generación de Código
Text (General)	MMLU-Pro	70.6	Comprensión de Lenguaje Multitarea Masiva
Text (Math)	GSM8K	91.6	Problemas Matemáticos de Nivel Escolar
Text (Math)	Math	80.7	Resolución de Problemas Matemáticos
Text (Science)	GPQA	47.4	Preguntas y Respuestas de Nivel de Postgrado
Vision (Doc/Chart)	InfoVQA	84.9	Preguntas y Respuestas Visuales de Información
Vision (Doc/Chart)	DocVQA	94.0	Preguntas y Respuestas Visuales de Documentos
Vision (General)	MMMU	70.0	Problemas Multimodales de Nivel Universitario
Vision (General)	MMMU-Pro	50.2	Problemas Multimodales de Nivel Universitario más Difíciles
Vision (Math)	MathVision	37.3	Razonamiento Matemático

			Visual
Vision (Math)	MathVista	72.2	Razonamiento Matemático Visual
Vision (OCR & Reasoning)	OCR-BenchV2	60.0	OCR Integral y Razonamiento
Vision (OCR)	CC-OCR	76.7	Análisis relacionado con OCR

Implementación

- **Framework:** vLLM
- **Hardware:** 4xH100 80GB
- **Cuantización:** Versiones cuantizadas (ej: funcionando en 4x L40S o 4x H100) pueden reducir requisitos de hardware con degradación aceptable de rendimiento. Cuantización adicional es posible, pero puede impactar significativamente el rendimiento. Contacte a WideLabs para detalles sobre opciones de cuantización.
- **Integración de API:** API compatible con OpenAI
- **Formateo de Entrada:** Lista de diccionarios (`role`, `content`). El contenido del usuario es una lista de diccionarios (`type`: "text"/"image"/"video", `data`: string/URL/ruta/base64).
- **Ingeniería de Prompt:** Siga principios estándar de ingeniería de prompt. Aproveche la capacidad del modelo de procesar instrucciones detalladas y contexto visual simultáneamente. Utilice sus capacidades mejoradas de formateo en prompts cuando sea aplicable.
- **Fine-tuning:** Soportado y recomendado para optimizar el rendimiento en tareas o dominios específicos. Contacte a WideLabs para servicios de fine-tuning.