

Harpia

Current version: harpia-1

Overview

Amazônia IA is the premier vision-language model (VLM) in the Amazônia IA family. It integrates a powerful Large Language Model (LLM) with a sophisticated vision system, delivering state-of-the-art performance by understanding and reasoning about visual information interleaved with text.

- **Model Architecture:** Combines a base LLM with a Vision Transformer (ViT). The ViT processes images at their native resolutions, enabling detailed visual understanding.
- **Key Features & Core Capabilities:**
 - **Multimodal Understanding:** Natively processes and understands interleaved text, images, and videos. Supports long-duration video analysis (over an hour) with event pinpointing (temporal grounding).
 - **Advanced Visual & Document Analysis:** Performs detailed image analysis, recognizing fine-grained elements, text (OCR), charts, icons, and complex layouts. Accurately localizes objects (grounding) via bounding boxes/points, providing stable JSON coordinates for spatial reasoning. Excels at parsing complex documents with diverse layouts (using formats like HTML), multi-language text, handwriting, tables, charts, and formulas.
 - **Agentic Interaction:** Functions as a visual agent, leveraging grounding and reasoning to interact with user interfaces on computers and mobile devices.
 - **Enhanced Reasoning & Output:** Demonstrates strong mathematical reasoning (visual/textual) and delivers detailed, well-formatted responses aligned with human preferences.

Technical Specifications

- **Input Modalities:** Text, Image, Video
- **Output Modalities:** Text
- **Context Window:** Supports up to 128K tokens. Recommended maximum of 32K tokens for vision-language tasks due to typical use patterns and computational cost. Contact WideLabs if larger context windows are essential.

- Languages: Multilingual (dozens of languages). Performance in specific languages can be further enhanced via fine-tuning.

Performance

Achieves highly competitive results on a wide range of benchmarks, demonstrating exceptional performance, especially in document understanding, complex reasoning, and agentic tasks.

Benchmark Category	Benchmark	Score	Notes
Text (Alignment)	IFEval	85.8	Instruction Following Evaluation
Text (Code)	HumanEval	85.4	Code Generation
Text (General)	MMLU-Pro	70.6	Massive Multitask Lang. Understanding
Text (Math)	GSM8K	91.6	Grade School Math Word Problems
Text (Math)	Math	80.7	Mathematical Problem Solving
Text (Science)	GPQA	47.4	Graduate-Level Q&A
Vision (Doc/Chart)	InfoVQA	84.9	Information Visual Question Answering
Vision (Doc/Chart)	DocVQA	94.0	Document Visual Question Answering
Vision (General)	MMMU	70.0	College-level Multimodal Problems
Vision (General)	MMMU-Pro	50.2	Harder College-level Multimodal Problems
Vision (Math)	MathVision	37.3	Visual Math Reasoning
Vision (Math)	MathVista	72.2	Visual Math Reasoning
Vision (OCR & Reasoning)	OCR-BenchV2	60.0	Comprehensive OCR & Reasoning
Vision (OCR)	CC-OCR	76.7	OCR-related parsing

Deployment

- Framework: vLLM

- Hardware: 4xH100 80GB
- Quantization: Quantized versions (e.g., running on 4x L40S or 4x H100) can reduce hardware requirements with acceptable performance degradation. Further quantization is possible but may significantly impact performance. Contact WideLabs for details on quantization options.
- API Integration: OpenAI-compatible API
- Input Formatting: List of dictionaries (`role`, `content`). User content is a list of dicts (`type`: "text"/"image"/"video", `data`: string/URL/path/base64).
- Prompt Engineering: Follow standard prompt engineering principles. Leverage the model's ability to process detailed instructions and visual context simultaneously. Utilize its enhanced formatting capabilities in prompts where applicable.
- Fine-tuning: Supported and recommended for optimizing performance on specific downstream tasks or domains. Contact WideLabs for fine-tuning services.