Chandra OCR

Overview

Chandra is a unified end-to-end vision-language OCR model that converts PDFs/images into Markdown/HTML/JSON while preserving layout (tables, forms, math, figures, captions) and supports multiple languages. It ships as open-source code with model weights (modified OpenRAIL-M for weights; Apache-2.0 for code) and can run locally or behind a vLLM server. The published model card lists ~9B parameters (BF16) and tags it as `qwen3_vl` family, indicating a Qwen-3-VL style backbone with custom training for OCR tasks (Paruchuri and Nussbaum, 2025).

Inference & system design

- **Dual backends (strategy-switchable):** Chandra exposes an `InferenceManager` that lets you choose Hugging Face Transformers (local) or vLLM (server) under a unified API. Example snippets show batched `BatchInputItem(image, prompt_type="ocr_layout")` → `.generate(...)` → rich outputs (Markdown plus structured JSON/HTML) (Datalab, 2025).

- **Deployment modes:** CLI (`chandra`, `chandra_vllm`), a Streamlit demo (`chandra_app`), and a Dockerized vLLM server with environment switches for throughput and token limits (Paruchuri and Nussbaum, 2025).

- **Output artifacts:** per-file Markdown + HTML + a metadata JSON (page info, token counts, etc.) and extracted images (Paruchuri and Nussbaum, 2025).

Community write-ups also describe the internal design patterns—e.g., a Strategy abstraction over the two inference methods and typed data schemas for I/O—consistent with the official API surface (DeepWiki, 2025).

Architecture

- **Backbone & size:** Model card indicates Qwen-3-VL lineage with ~9B params (BF16), adapted for full-page OCR and layout reasoning, not block-by-block pipelines (Datalab, 2025).

- **Prompt types:** The public API centers on an `ocr_layout` prompt for layout-aware decoding (Datalab, 2025).

- **Licensing & usage constraints:** Code Apache-2.0, weights under modified OpenRAIL-M (free for research/personal/startups under a revenue cap; restrictions on competitive API use) (Paruchuri and Nussbaum, 2025).

Technical innovations

- **Full-page decoding for layout** (instead of Marker/Surya's segment-and-stitch pipelines), enabling accurate tables/forms with checkboxes, figure extraction with captions, and math fidelity (Datalab, 2025).

- **Layout-aware outputs** that reconstruct reading order and page structure directly as Markdown/HTML/JSON, including structured table data and extracted images (Paruchuri and Nussbaum, 2025).

- **Handwriting & old-scan robustness** emphasized in features and examples (Datalab, 2025).

Training details

Datalab's launch note describes real-world labeled data and synthetic augmentation, especially for math, to push accuracy on equations and handwritten math pages. It also mentions quantized 8B and 2B variants for high-throughput on-prem. Specific dataset names beyond examples aren't disclosed publicly (Datalab, 2025).

Performance

- **olmOCR-Bench (independent benchmark):** Chandra v0.1.0 reports 83.1 ± 0.9 overall, leading the table across diverse doc types (arXiv, old scans, math, tables, multi-column, long tiny text). The benchmark is widely used by OCR/VLM systems and consists of ~1,400 PDFs and ~7,000 unit tests for structure-faithful Markdown conversion (Allen Institute for AI (AI2), 2025a, Allen Institute for AI (AI2), 2025b).

- **Throughput:** Datalab reports up to ~4 pages/sec on an H100 (~345k pages/day) with quantized variants and "minimal accuracy degradation." (Vendor claim; no third-party latency audits yet.) (Datalab, 2025)

Practical deployment notes

- **Local vs. server trade-offs:** HF local mode is simpler; vLLM is recommended for batching, parallelism, and lower latency at scale. CLI and env vars expose max output tokens, page ranges, and concurrency controls (Paruchuri and Nussbaum, 2025).

- **Ecosystem interoperability:** Designed to slot into document ETL/RAG pipelines; outputs include bounding-box-traceable structure (Markdown/HTML/JSON), which pairs well with downstream retrieval or extraction. (General platform docs; specific Chandra SDK is in repo/card.) (Datalab, 2025)

References

1. Paruchuri, V. & Nussbaum, Z. (2025). *Chandra: Layout-Preserving OCR Model.* Datalab. Available at: https://github.com/datalab-to/chandra

2. Datalab. (2025). *datalab-to/chandra — Hugging Face Model Card.* Retrieved from https://huggingface.co/datalab-to/chandra

3. Paruchuri, V. (2025, October 30). *Introducing our newest model: Chandra.* Datalab Blog. Available at: https://www.datalab.to/blog/introducing-chandra

4. Allen Institute for AI (AI2). (2025a). *olmOCR-bench: A Benchmark for Layout-Preserving OCR to Markdown.*
   Dataset available at https://huggingface.co/datasets/allenai/olmOCR-bench

5. Allen Institute for AI (AI2). (2025b). *olmocr: Toolkit and Benchmark for Document OCR.* Source repository: https://github.com/allenai/olmocr

6. DeepWiki. (2025). *Chandra Architecture (overview).*
   Available at https://deepwiki.com/datalab-to/chandra/4-architecture