

Comparative evaluation of the AlphaDigit document OCR models on olmOCR-Bench

Théo Hubert, *AlphaEdge*

October 2026 (v1.0)

Abstract

This document reports a comparative evaluation of the AlphaDigit family of document OCR models (AlphaDigitMax v2, AlphaDigitMedium v2 and AlphaDigitFast v2) on olmOCR-Bench, a public benchmark of about 8 400 binary unit tests spread over eight categories of PDF pages. All systems are scored with the same judge, built on the official olmOCR-Bench test implementations, after the same format-only normalisation. AlphaDigitMax v2 reaches the best overall score (89.4 %), 2.5 points ahead of the strongest competitor (Chandra, 86.9 %). It ranks first on tables, old scans and old scans with mathematics, and ties for first on headers/footers and the baseline. The gap is largest on old scans (60.6 % vs. 51.2 %). AlphaDigitMedium v2 (85.4 %) ranks third overall, ahead of Mistral OCR (83.8 %). AlphaDigitFast v2 (83.0 %) ranks fifth, ahead of DeepSeek-OCR, DeepSeek-OCR-2, LightOnOCR-2, GLM-OCR and Infinity-Parser2-Pro. AlphaDigitMax v2 also keeps first place when the headers/footers category is excluded from the average (88.0 % vs. 85.5 %).

1 Introduction

This report presents a benchmark of the AlphaDigit models, developed by AlphaEdge, for converting document pages (PDF or scanned images) into structured text: Markdown for running text, HTML for tables and $\text{\LaTeX}$ for mathematics. The aim is to position these models against a set of recent open-weight and commercial OCR systems, using a public benchmark with automatically verifiable tests.

Three variants are evaluated:

- **AlphaDigitMax v2**, the most accurate variant;
- **AlphaDigitMedium v2**, an intermediate variant;
- **AlphaDigitFast v2**, the lightest variant, designed for high throughput.

This report focuses on empirical evaluation. It does not describe the internal architecture of the models or the training data. This first version covers olmOCR-Bench only. Evaluations on other benchmarks are in progress (see Section 4).

2 Experimental protocol

2.1 Benchmark

olmOCR-Bench (Allen Institute for AI) contains 1 402 single PDF pages. Each page comes with deterministic unit tests that pass or fail. The tests cover eight categories:

Category	Tests	What is measured
Headers & footers	738	Running headers, footers and page numbers must <i>not</i> appear in the output.
Tables	1 019	Relative position of cells (left, right, above, below, header).
arXiv math	2 927	L ^A T _E X formulas from recent arXiv papers, checked after rendering.
Multi-column	884	Natural reading order across columns.
Long tiny text	438	Presence of text set in very small fonts on dense pages.
Old scans	525	Text presence and reading order on degraded historical scans.
Old scans math	458	Formulas on scanned mathematical documents.
Baseline	1 393	Output sanity checks: no empty page, no degenerate repetition, no invalid characters.

Table 1: olmOCR-Bench categories.

2.2 Compared systems

The systems compared are: AlphaDigitMax v2, AlphaDigitMedium v2, AlphaDigitFast v2, Chandra in *accurate* mode (API), Mistral OCR Latest (API), DeepSeek-OCR, DeepSeek-OCR-2, GLM-OCR, Infinity-Parser2-Pro and LightOnOCR-2.

2.3 Metrics

For each category, the score is the percentage of passed tests. Following the olmOCR-Bench convention, the **overall score** is the unweighted mean of the eight category scores. Because the headers/footers category rewards *omitting* content and strongly affects some systems (Section 3.4), we also report the mean over the seven other categories.

2.4 Inference

Each page is transcribed once (a single repeat per page) from the original PDF. No information from the test files is available at inference time. Locally executed models were run on a single NVIDIA RTX PRO 6000 GPU. API-based systems were queried through their official endpoints.

2.5 Output normalisation

The olmOCR-Bench tests are sensitive to purely typographic conventions. To compare transcription quality rather than writing conventions, every system’s output goes through the same steps before scoring. Two levels of normalisation are applied.

Format normalisation (all outputs, all categories).

- removal of reasoning tags (e.g. `<think>...</think>`) when a model emits them;
- Unicode NFC normalisation and folding of typographic punctuation (curly quotes and apostrophes, typographic dashes) to their ASCII equivalents;
- rejoining of words hyphenated at line breaks (e.g. “compe-\n tence” → “competence”), while keeping hyphens that do not fall at a line break;

- removal of lines that contain only page-number markup, and collapsing of runs of blank lines.

Math delimiters, table HTML and the textual content are left untouched.

Notation equivalences. A different character, digit or word is never corrected: for example, “< 6,000 km²” is never matched to “< 5,000 km²”. The equivalences are:

- *Table cells:* L^AT_EX commands vs. Unicode symbols ($\mu \rightarrow \mu$, $\pm \rightarrow \pm$, $\geq / \geq \rightarrow \geq$, $\leq / \leq \rightarrow \leq$, $\rightarrow \rightarrow$ the Unicode arrow), $\% \rightarrow \%$, $\text{km}^2 \rightarrow \text{km}^2$, ellipsis variants ($\dots, \ldots, \cdots \rightarrow \dots$), inline math delimiters and whitespace inside formulas, and empty checkboxes ($[] \rightarrow \square$).
- *Mathematics:* removal of $\tag$ and $\label$; $\leq / \leq$; $\dots / \ldots / \cdots$; $\frac / \dfrac / \tfrac$; $| / \text{vert}$; $:= / \coloneqq$; $\sim / \intercal$; $\rightarrow / \rightarrow$; removal of $\left / \right$ and spacing commands ($\backslash, \quad, \dots$). A formula split across several consecutive math delimiters is also compared in its joined form.
- *Multi-column reading order:* inline formulas are compared through their plain-text reading ($\mu \rightarrow \mu$, $\text{km}^2 \rightarrow \text{km}^2$). Spacing between words is not modified.

Headers and footers. The expected strings of the “absent” tests are *never* used to delete text from an output. Only one exception applies: a failed absence test counts as passed when the string occurs at least twice in the full transcription. In that case the string also legitimately appears in the body of the page (e.g. a running title repeated in the text).

2.6 Comparability notes

- **Denominators.** Test counts can differ by a few units between runs. For AlphaDigitFast v2 and GLM-OCR, the table category includes 1 021 tests instead of 1 019. The full counts are given in Appendix A.
- **Infinity-Parser2-Pro.** The model labels each output block with a layout class. Blocks that it labels itself as header, footer or page footnote are removed from its output before scoring, in the same way as page furniture is excluded by the AlphaDigit models.
- **LightOnOCR-2.** This is a full-page model without a document-layout stage. It transcribes page furniture whenever it is visible, so 71.4% of the headers/footers tests fail because the running text is faithfully reproduced. This reflects a design choice rather than a recognition error, and it weighs on the overall score in the same way as for the other full-page systems. Removing that text would require the benchmark annotations, which would make the measurement circular, so no such post-processing is applied.

2.7 Corrections applied to the test set

The expected values of olmOCR-Bench are partly built automatically, and an audit of the test set revealed a number of faulty tests, impossible to satisfy whatever the system output. They were reviewed in two passes: a systematic pass with Gemini 3.1 Pro, used only to flag suspicious tests, followed by a manual, visual check of every flagged test against the rendered page. The corrections were applied to the expected values, never to a system’s output, and the official judge was left unchanged. No correction favours a particular system: several of the corrected tests are tests that our own models get wrong. All systems, local or API-based, were scored against the identical corrected test set, so the comparison rests on a single common reference.

3 Results

3.1 Summary

AlphaDigitMax v2 obtains the best overall score at 89.4%, ahead of Chandra API (86.9%). **AlphaDigitMedium v2** (85.4%) ranks third, ahead of Mistral OCR (83.8%). **AlphaDigitFast v2** (83.0%) ranks fifth, ahead of DeepSeek-OCR-2 (79.9%), Infinity-Parser2-Pro (79.9%), DeepSeek-OCR (78.4%), LightOnOCR-2 (76.9%) and GLM-OCR (74.3%).

By category, AlphaDigitMax v2 is strictly first on **tables** (92.9%, +3.0 pt), **old scans** (60.6%, +9.4 pt) and **old scans math** (91.9%, +0.2 pt). It ties for first on **headers/footers** (99.5%) and the **baseline** (99.9%). It is close to the best system on long tiny text (95.4% vs. 95.9% for Chandra) and multi-column (86.2% vs. 86.8% for Mistral OCR). Its largest gap is on arXiv math (88.9% vs. 91.9% for LightOnOCR-2). The largest advantage is on degraded historical documents, the hardest category of the benchmark for every system.

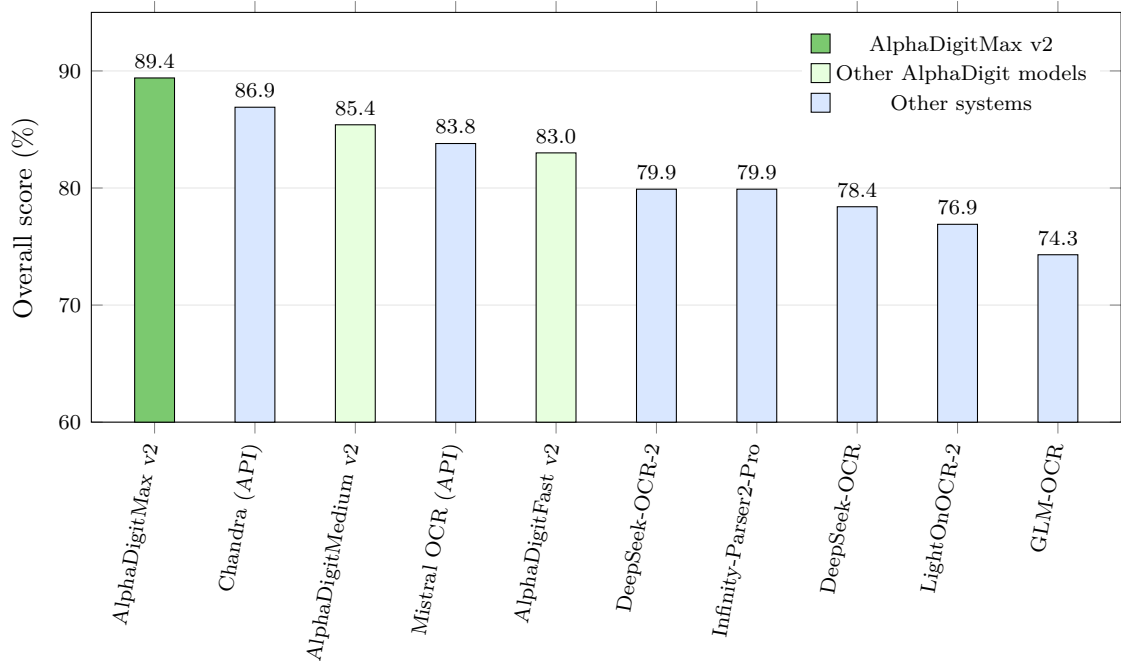
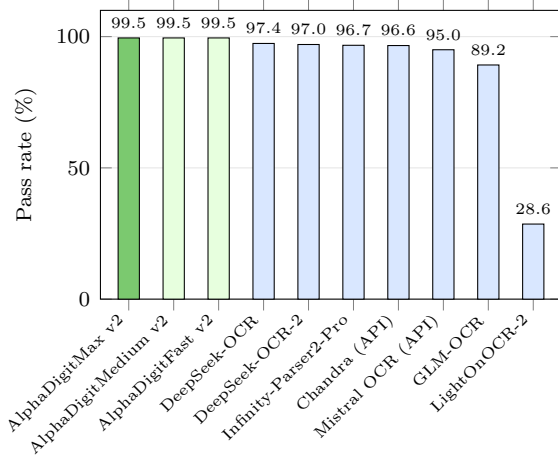
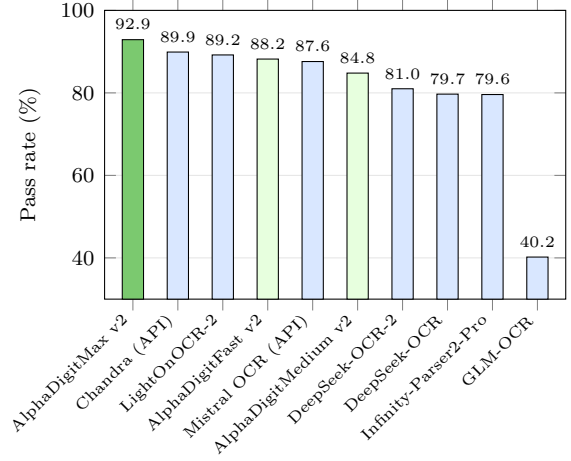


Figure 1: Overall olmOCR-Bench score (unweighted mean of the eight categories). The vertical axis starts at 60 % for readability.

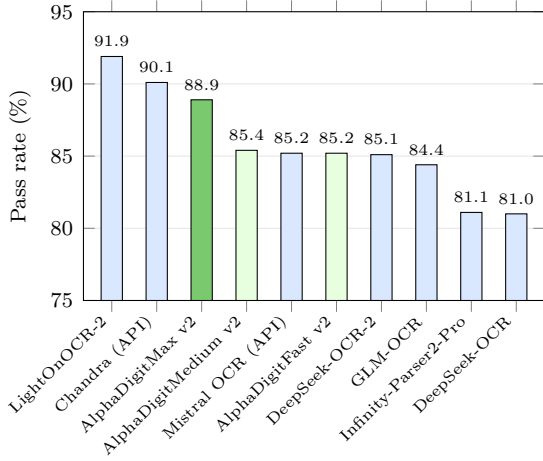
3.2 Results by category



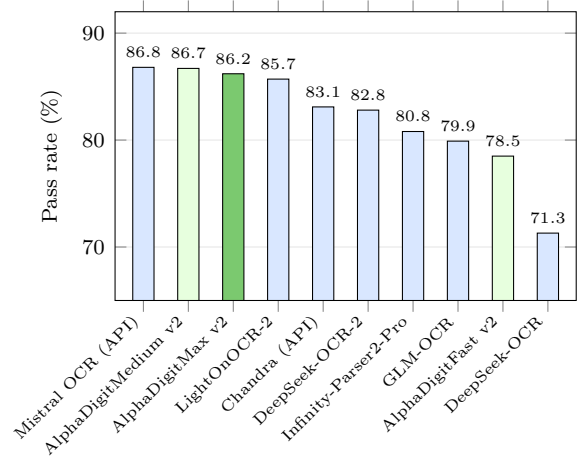
(a) Headers & footers



(b) Tables

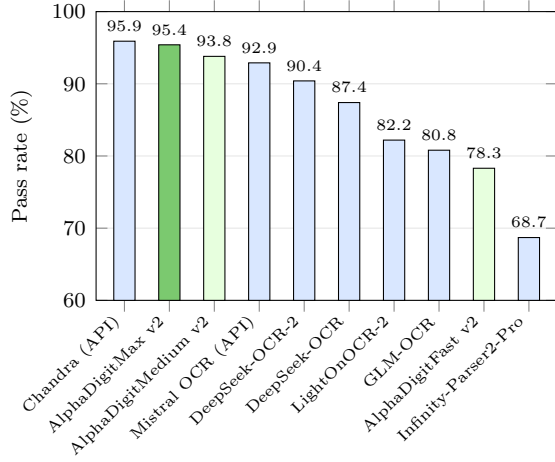


(c) arXiv math

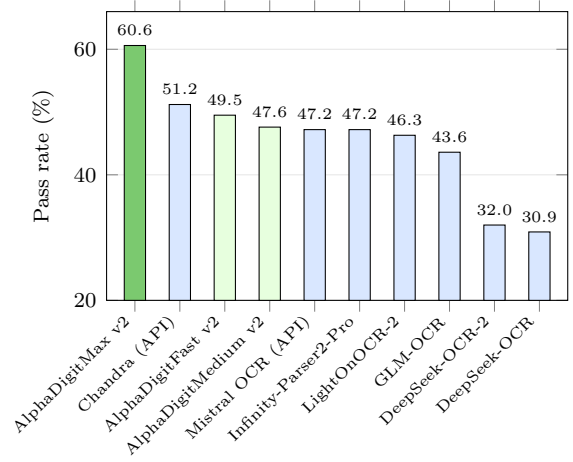


(d) Multi-column

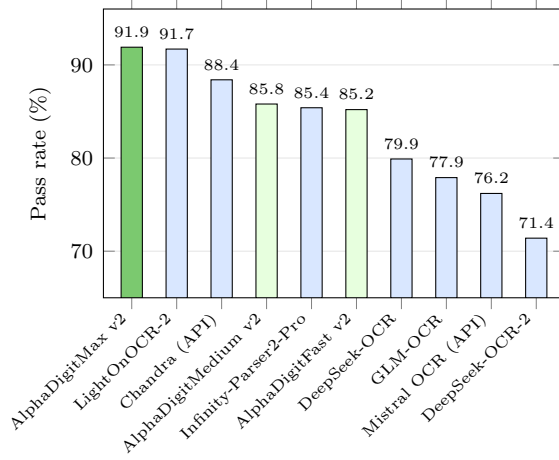
Figure 2: Pass rate by category (1/2): headers & footers, tables, arXiv math and multi-column. Dark green: AlphaDigitMax v2; light green: AlphaDigitMedium v2 and AlphaDigitFast v2; blue: other systems. Each panel has its own vertical axis range, chosen to make the differences readable. Systems are sorted by exact pass rate.



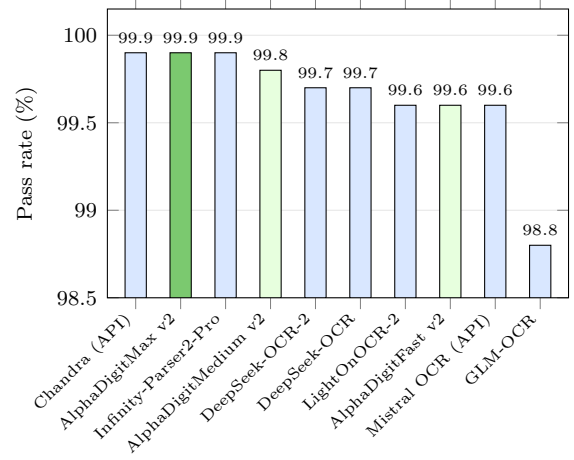
(a) Long tiny text



(b) Old scans



(c) Old scans math



(d) Baseline

Figure 3: Pass rate by category (2/2): long tiny text, old scans, old scans math and baseline. Same colour code as Figure 2.

3.3 Full results

Model	Head.	Tables	arXiv	Multi-col.	Tiny	Old sc.	Old sc. math	Base.	Overall	Excl. head.
AlphaDigitMax v2	99.5	92.9	88.9	86.2	95.4	60.6	91.9	99.9	89.4	88.0
Chandra (API)	96.6	89.9	90.1	83.1	95.9	51.2	88.4	99.9	86.9	85.5
AlphaDigitMedium v2	99.5	84.8	85.4	86.7	93.8	47.6	85.8	99.8	85.4	83.4
Mistral OCR (API)	95.0	87.6	85.2	86.8	92.9	47.2	76.2	99.6	83.8	82.2
AlphaDigitFast v2	99.5	88.2	85.2	78.5	78.3	49.5	85.2	99.6	83.0	80.7
DeepSeek-OCR-2	97.0	81.0	85.1	82.8	90.4	32.0	71.4	99.7	79.9	77.5
Infinity-Parser2-Pro	96.7	79.6	81.1	80.8	68.7	47.2	85.4	99.9	79.9	77.5
DeepSeek-OCR	97.4	79.7	81.0	71.3	87.4	30.9	79.9	99.7	78.4	75.7
LightOnOCR-2	28.6	89.2	91.9	85.7	82.2	46.3	91.7	99.6	76.9	83.8
GLM-OCR	89.2	40.2	84.4	79.9	80.8	43.6	77.9	98.8	74.3	72.2

Table 2: olmOCR-Bench pass rates (%) by category. “Overall” is the unweighted mean of the eight categories. “Excl. head.” is the mean of the seven categories other than headers & footers. Best value per column in bold.

3.4 Discussion

Headers and footers. This category rewards systems that leave page furniture out of the transcription. LightOnOCR-2 transcribes running headers and footers faithfully, so it scores

below 30 % here. This is a design choice rather than a recognition error, and it weighs heavily on its overall score. When this category is excluded, AlphaDigitMax v2 remains first (88.0 %), ahead of Chandra API (85.5 %), LightOnOCR-2 (83.8 %), AlphaDigitMedium v2 (83.4 %) and Mistral OCR (82.2 %). The AlphaDigit models exclude page furniture by design, which suits downstream uses such as retrieval and document understanding.

Degraded documents. Old scans is the hardest category for every system: no competitor exceeds 51.2 %. AlphaDigitMax v2 reaches 60.6 %, a 9.4-point lead, and also ranks first on old scans with mathematics (91.9 %). These results show robustness to noise, faded ink and historical typography, which are common in archival and back-office documents.

Remaining gaps. On long tiny text, AlphaDigitMax v2 (95.4 %) is within 0.5 point of the best system (Chandra, 95.9 %). On multi-column pages, the top systems are within about one point of each other (86.2 % vs. 86.8 % for Mistral OCR). The clearest remaining gap is on arXiv math, where LightOnOCR-2 leads (91.9 % vs. 88.9 %).

Lighter variants. AlphaDigitMedium v2 scores 85.4 % overall and ranks third, behind Chandra only, with strong results on multi-column (86.7 %) and long tiny text (93.8 %). AlphaDigitFast v2 scores 83.0 %, 0.8 point behind Mistral OCR, and ahead of the DeepSeek models, LightOnOCR-2, GLM-OCR and Infinity-Parser2-Pro. It keeps a high level on tables (88.2 %) and mathematics (85.2 % on arXiv, 85.2 % on old scans math). Its main weaknesses are long tiny text (78.3 %) and multi-column reading order (78.5 %).

4 Conclusion

Under the experimental protocol described here, AlphaDigitMax v2 obtains the best overall score on olmOCR-Bench (89.4 %) among the ten systems compared. It leads clearly on tables and on degraded historical scans, is within 0.5 point of the best system on long tiny text, and stays first when the headers/footers category is excluded. AlphaDigitMedium v2 ranks third, behind Chandra (API) only, and AlphaDigitFast v2 ranks fifth. This places the whole AlphaDigit family among the most accurate document OCR systems on this benchmark.

This report covers olmOCR-Bench only. Evaluations on other public document-understanding benchmarks, including **OmniDocBench** and **MDPBench**, are in progress and will be reported in a future version of this document.

A Raw test counts

Model	Head.	Tables	arXiv	Multi-col.	Tiny	Old sc.	Old sc. math	Base.
AlphaDigitMax v2	734/738	947/1019	2603/2927	762/884	418/438	318/525	421/458	1391/1393
AlphaDigitMedium v2	734/738	864/1019	2501/2927	766/884	411/438	250/525	393/458	1390/1393
AlphaDigitFast v2	734/738	901/1021	2494/2927	694/884	343/438	260/525	390/458	1388/1393
Chandra (API)	713/738	916/1019	2638/2927	735/884	420/438	269/525	405/458	1392/1393
Mistral OCR (API)	701/738	893/1019	2494/2927	767/884	407/438	248/525	349/458	1387/1393
DeepSeek-OCR-2	716/738	825/1019	2492/2927	732/884	396/438	168/525	327/458	1389/1393
DeepSeek-OCR	719/738	812/1019	2371/2927	630/884	383/438	162/525	366/458	1389/1393
LightOnOCR-2	211/738	909/1019	2689/2927	758/884	360/438	243/525	420/458	1388/1393
GLM-OCR	658/738	410/1021	2471/2927	706/884	354/438	229/525	357/458	1376/1393
Infinity-Parser2-Pro	714/738	811/1019	2373/2927	714/884	301/438	248/525	391/458	1391/1393

Table 3: Passed tests / total tests per category.