NEURAL INFORMATION PROCESSING SYSTEMS

NEURIPS 2026 · EVALUATIONS & DATASETS TRACK · PARIS

GT-Free OCR Metrics

A Reference-Free Evaluation Framework for Document OCR Systems

Kshitij Singh

Contents

I · Motivation

4 Ground truth does not scale
5 The ranking task

II · Method

7 The page as ground truth
8 Mask, render, score
9 Five views of every page
10 Rendering differences

III · Benchmark

12 Derived from OmniDocBench
13 Building the benchmark
14 How success is measured
15 146 implementations

IV · Results

17 A clear ranking signal
18 Composite stacks lead
19 The winning stack
20 The DocSim similarity head

V · Use and outlook

22 Practical uses
23 Limitations

Resources

24 Code and data

SECTION I

Motivation

OCR metrics need ground truth, and ground truth does not scale

REFERENCE-BASED METRICS

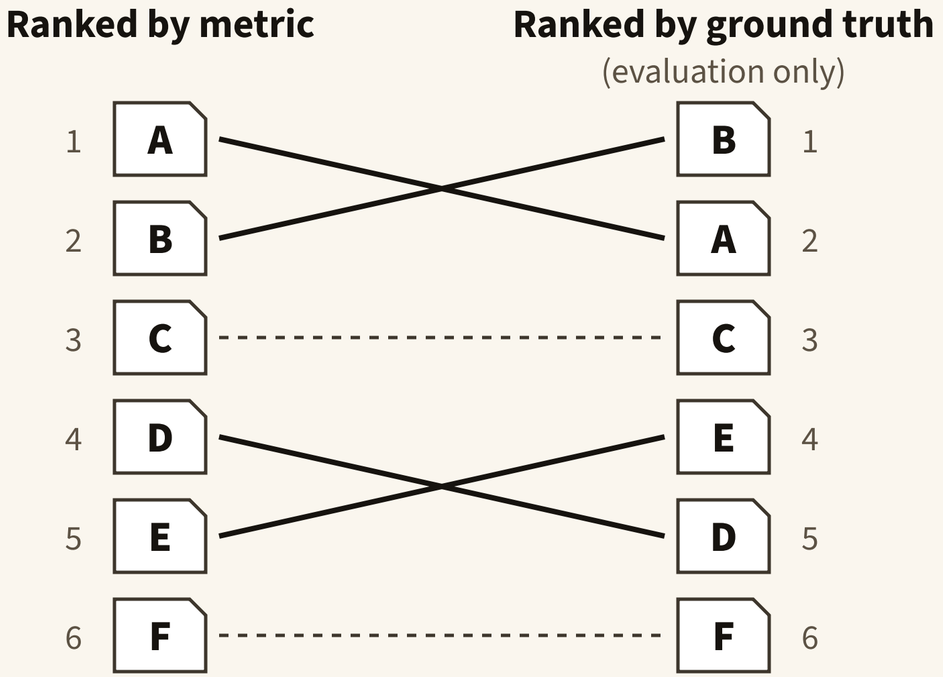
- **CER / WER** for text
- **CDM** for formulas
- **TEDS** for tables

Each needs ground-truth text aligned to every recognised element.

- Expert labour for every page
- Cost grows with the size of the corpus
- Annotators disagree on ambiguous content
- Goes stale as documents and OCR systems change
- Often impossible for sensitive or proprietary material

So production OCR is checked on **small audited samples** or proxy heuristics.

Can we rank pages by OCR quality with no ground truth?



Input: a page image and its OCR output

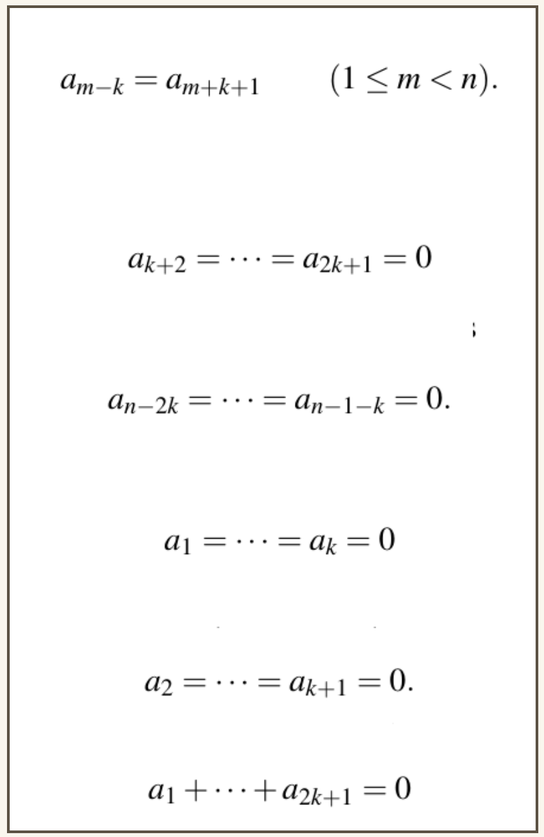
Output: one score per page

Goal: order pages the way a ground-truth metric would

SECTION II

Method

The page is its own ground truth



$$a_{m-k} = a_{m+k+1} \qquad (1 \leq m < n).$$

$$a_{k+2} = \cdots = a_{2k+1} = 0$$

;

$$a_{n-2k} = \cdots = a_{n-1-k} = 0.$$

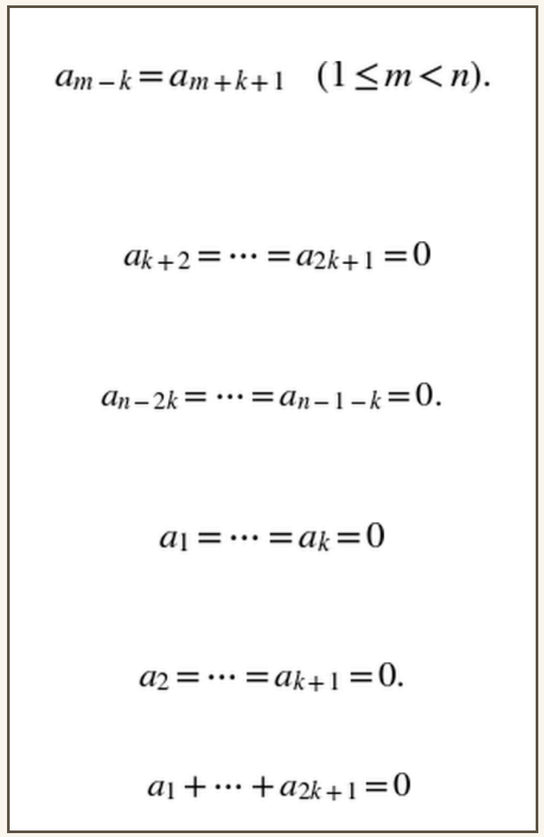
$$a_1 = \cdots = a_k = 0$$

$$a_2 = \cdots = a_{k+1} = 0.$$

$$a_1 + \cdots + a_{2k+1} = 0$$

Original page,
masked to formulas

≈



$$a_{m-k} = a_{m+k+1} \quad (1 \leq m < n).$$

$$a_{k+2} = \cdots = a_{2k+1} = 0$$

$$a_{n-2k} = \cdots = a_{n-1-k} = 0.$$

$$a_1 = \cdots = a_k = 0$$

$$a_2 = \cdots = a_{k+1} = 0.$$

$$a_1 + \cdots + a_{2k+1} = 0$$

OCR output,
rendered back

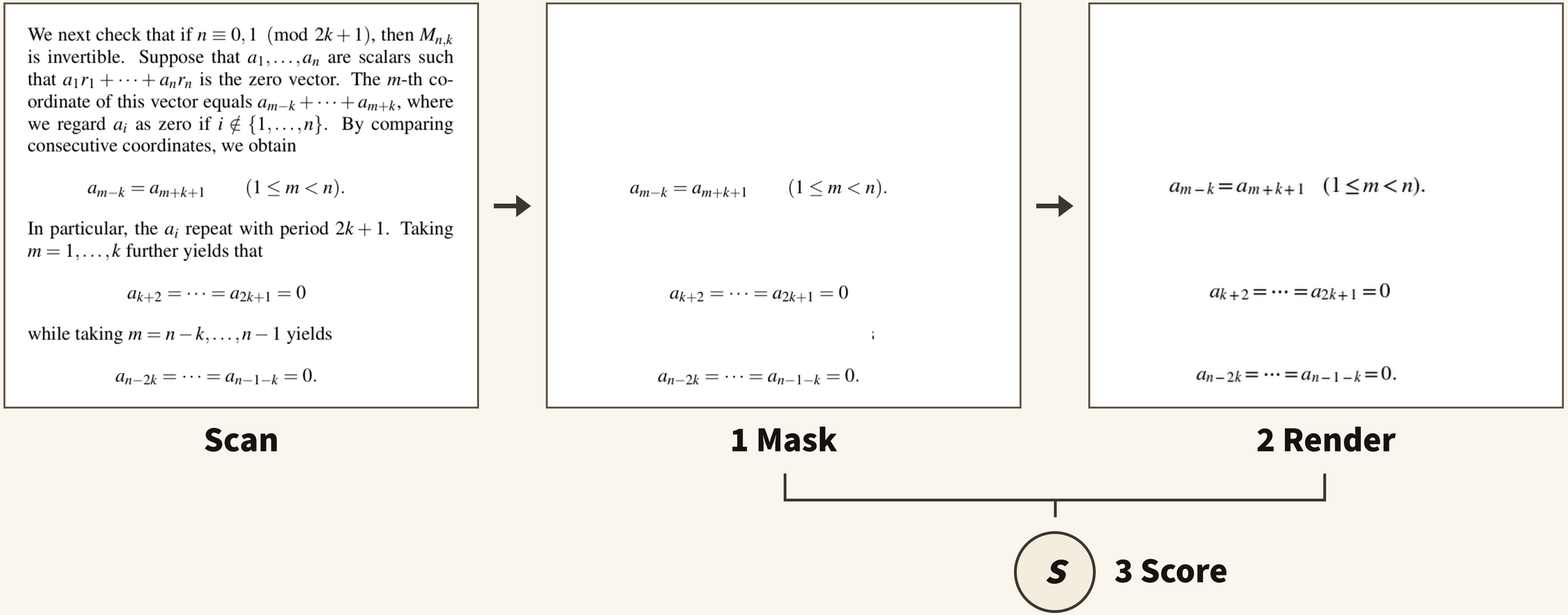
What the OCR returned for the first line:

```
$$a_{m-k} = a_{m+k+1} \quad (1
\leq m < n).$$
```

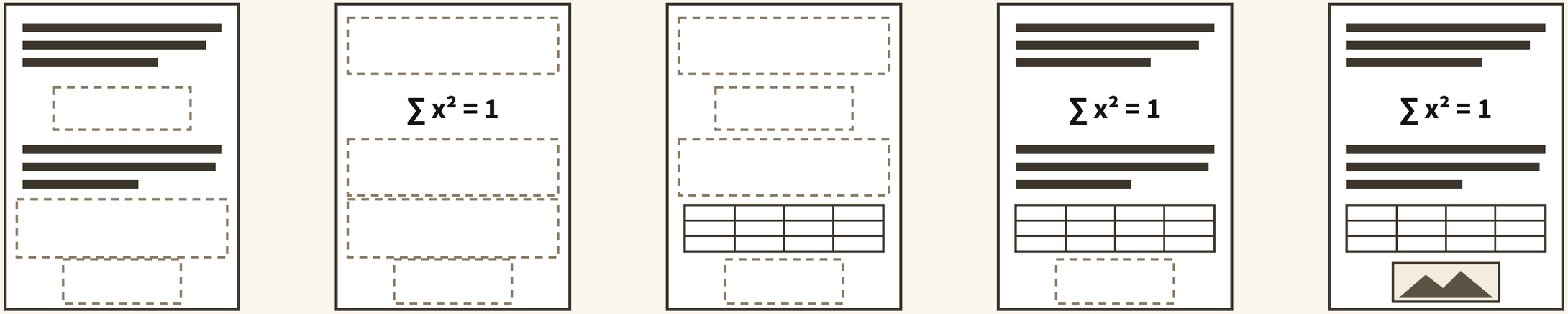
If the reading is right, the rendering looks like the page.

Page: OmniDocBench (Putnam 2008 solutions, p. 7).
OCR: Qwen3.5-122B-A10B.

Mask, render, score



Five views of every page



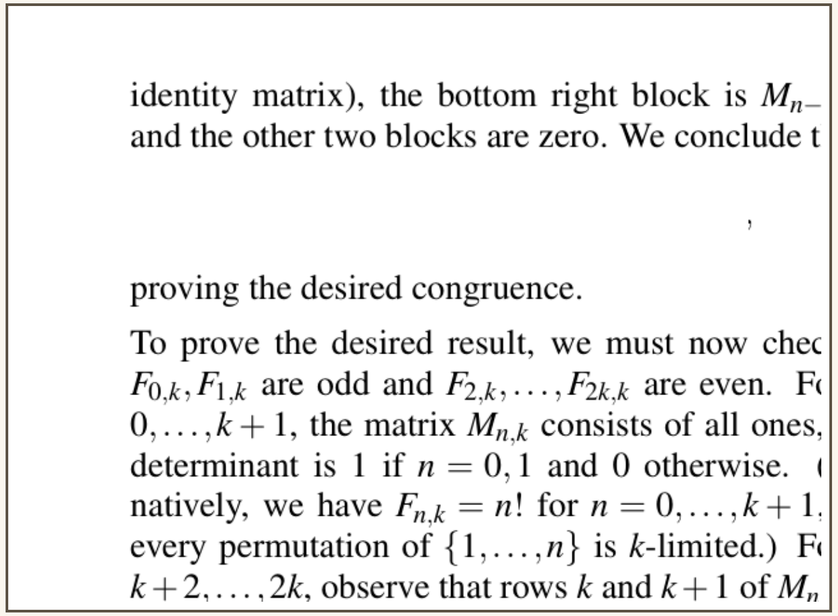
Text | **Formulas** | **Tables** | **All** | **All, no mask**

COMPARED WITH THE GROUND TRUTH BY

text edit distance	formula edit distance	TEDS	mean of the three, where present

kept and compared | whited out in the scan

Even a faithful reading renders differently

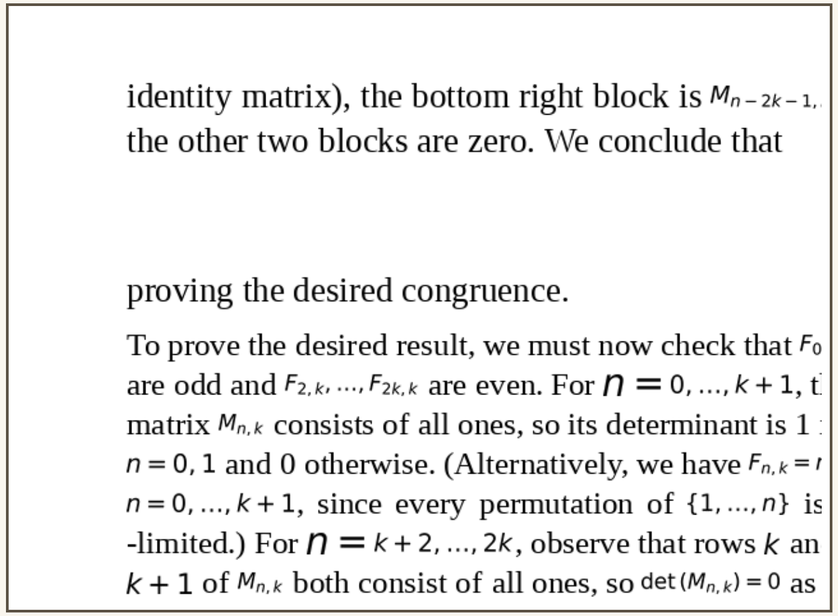


identity matrix), the bottom right block is M_{n-}
and the other two blocks are zero. We conclude t

proving the desired congruence.

To prove the desired result, we must now chec
$F_{0,k}, F_{1,k}$ are odd and $F_{2,k}, \ldots, F_{2k,k}$ are even. F
$0, \ldots, k+1$, the matrix $M_{n,k}$ consists of all ones,
determinant is 1 if $n = 0, 1$ and 0 otherwise.
natively, we have $F_{n,k} = n!$ for $n = 0, \ldots, k+1$,
every permutation of $\{1, \ldots, n\}$ is k-limited.) F
$k+2, \ldots, 2k$, observe that rows k and $k+1$ of M_n

Scan, text view (the formula is masked)



identity matrix), the bottom right block is $M_{n-2k-1,}$
the other two blocks are zero. We conclude that

proving the desired congruence.

To prove the desired result, we must now check that F_0
are odd and $F_{2,k}, \ldots, F_{2k,k}$ are even. For $n = 0, \ldots, k+1$, t
matrix $M_{n,k}$ consists of all ones, so its determinant is 1
$n = 0, 1$ and 0 otherwise. (Alternatively, we have $F_{n,k} = n$
$n = 0, \ldots, k+1$, since every permutation of $\{1, \ldots, n\}$ is
-limited.) For $n = k+2, \ldots, 2k$, observe that rows k an
$k+1$ of $M_{n,k}$ both consist of all ones, so $\det(M_{n,k}) = 0$ as

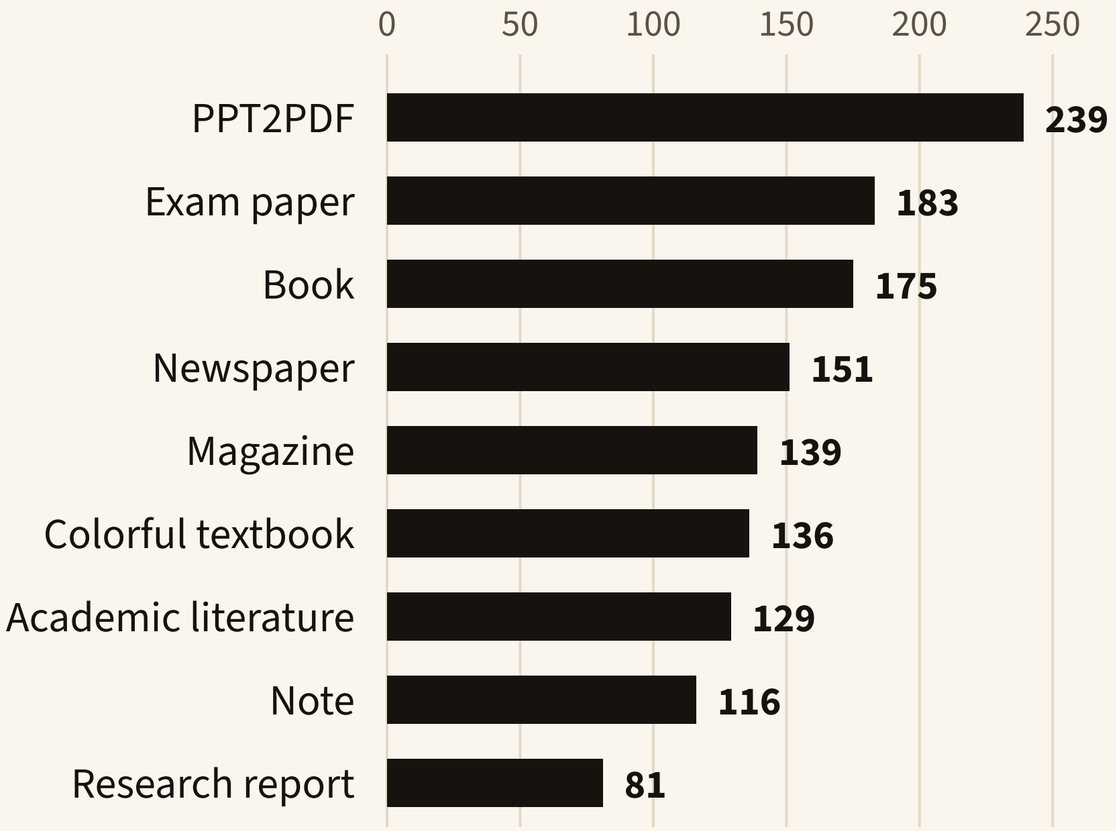
OCR text, rendered back

Typeface, line breaks and inline maths all change.
The comparison must ignore them, yet **notice a wrong or missing character**.

SECTION III

Benchmark

We derived our experimental dataset from OmniDocBench



BENCHMARK

For parsing diverse PDF pages

Ouyang et al., CVPR 2025

PAGES

English and Chinese

GROUND TRUTH

Layout boxes, text,
LaTeX formulas, HTML tables

How we built the benchmark from OmniDocBench

OCR

1,355

pages

- read by Qwen3.5-122B-A10B
- type, box and content of each element
- token log-probabilities

→

MASK AND RENDER

6,775

image pairs

- 5 variants of every page
- masked scan and rendering

→

REFERENCE

3

ground-truth measures

- text edit distance
- formula edit distance
- TEDS for tables

RELEASED ON HUGGING FACE

omnidocbench-render-compare
the 6,775 image pairs, with 20,280 DocSim triplets

omnidocbench-qwen-ocr-logprobs
the token log-probabilities of all 1,355 pages

Success means agreeing with the reference ranking

$$\bar{\rho} = \text{mean over 5 variants of Spearman } \rho(\text{metric, reference})$$

$\rho = 1$: same order as the reference · $\rho = 0$: no relation

PAGES SCORED PER VARIANT

Text	Formulas	Tables	All	All, no mask
1,349	**200**	**351**	**1,355**	**1,355**
with text	with formulas	with tables	all pages	all pages

WHY RANK CORRELATION

1. The question is ordering, not linear prediction
2. Robust to outlier pages in nine diverse document types
3. Metric scores live on different scales

SIGNIFICANCE

$p < 0.001$ on every variant, from the best method down to the baseline

Two-sided Fisher *z*-test

146 implementations across 11 mechanism classes

CLASS	METHODS	WHAT IT DOES	EXAMPLE
Composite	**42**	Weighted mix of scores from different classes	DocSim + SSIM + OCR entropy
Deep feature	**36**	Cosine similarity of scan and rendering embeddings	CLIP, DINOv2 or DocSim cosine
Self-uncertainty	**33**	Only the OCR model’s own confidence, no images	Mean token log-probability
Per-element structure	**8**	Scores each formula or table cell separately	DocSim of the worst table cell
Hash	**1**	Several perceptual hashes combined into one score	pHash + wHash + dHash
Hash metric	**1**	A single perceptual hash of the whole page	256-bit pHash as a PDQ proxy
Perceptual	**1**	Shift-tolerant learned perceptual distance	ST-LPIPS
Perceptual latent	**1**	Perceptual error learned from human judgments	PieAPP
Classical image quality	**1**	Hand-crafted image statistics of scan vs rendering	Focus, contrast, edge density
Preprocessing	**1**	Erases pictures from the scan, then scores it	LaMa inpainting
Baseline	**1**	Naive full-page similarity, better of two metrics	CLIP cosine vs SSIM + MSE + LPIPS

Counts cover the 126-row catalogue: 125 scored implementations plus the baseline. The other 21 of the 146 were superseded or gave no signal.

SECTION IV

Results

A clear ranking signal, with no annotations at inference

SPEARMAN CORRELATION WITH THE REFERENCE RANKING

0.339

best full-page baseline

mean of the 5 variants

0.494

best composite (P2_210)

mean of the 5 variants

0.605

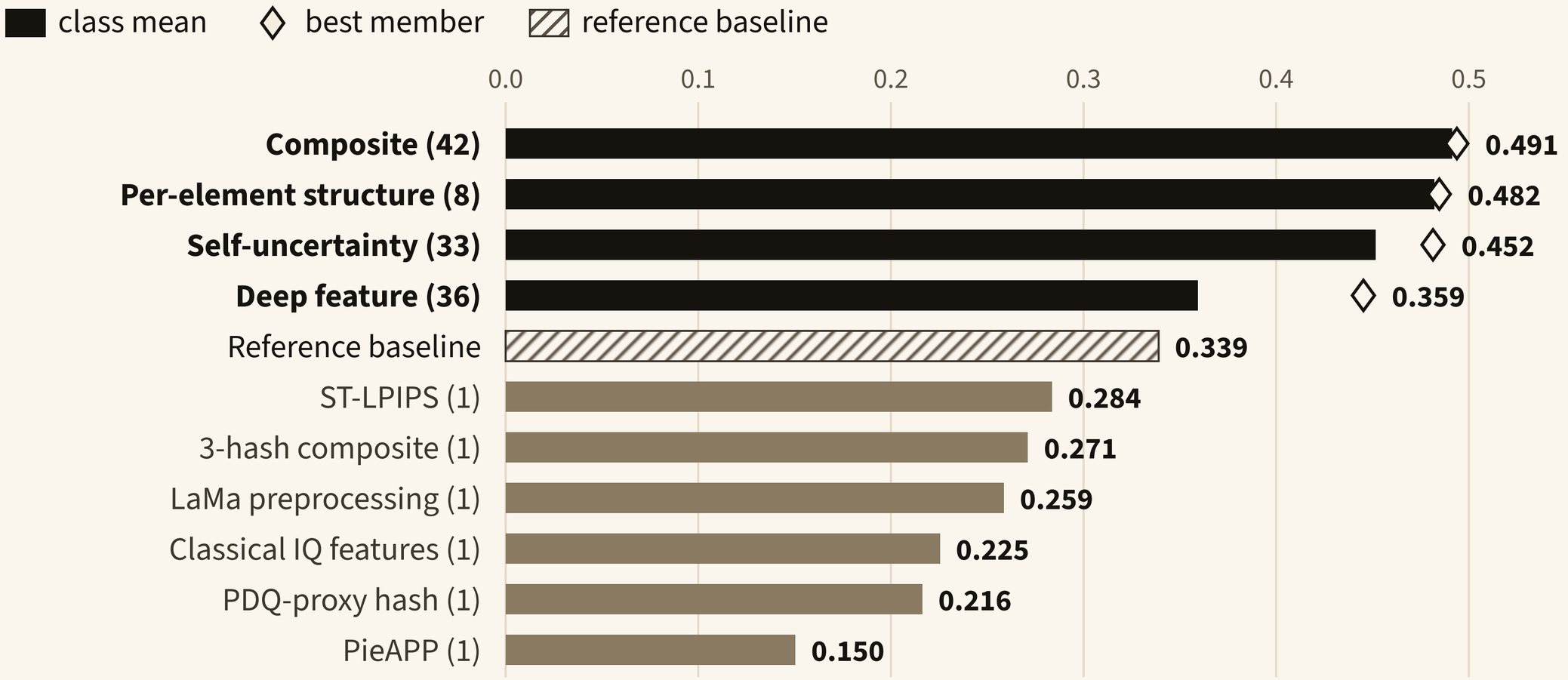
peak variant:
formulas only

p < 0.001 from the best method down to the baseline

1,355 pages

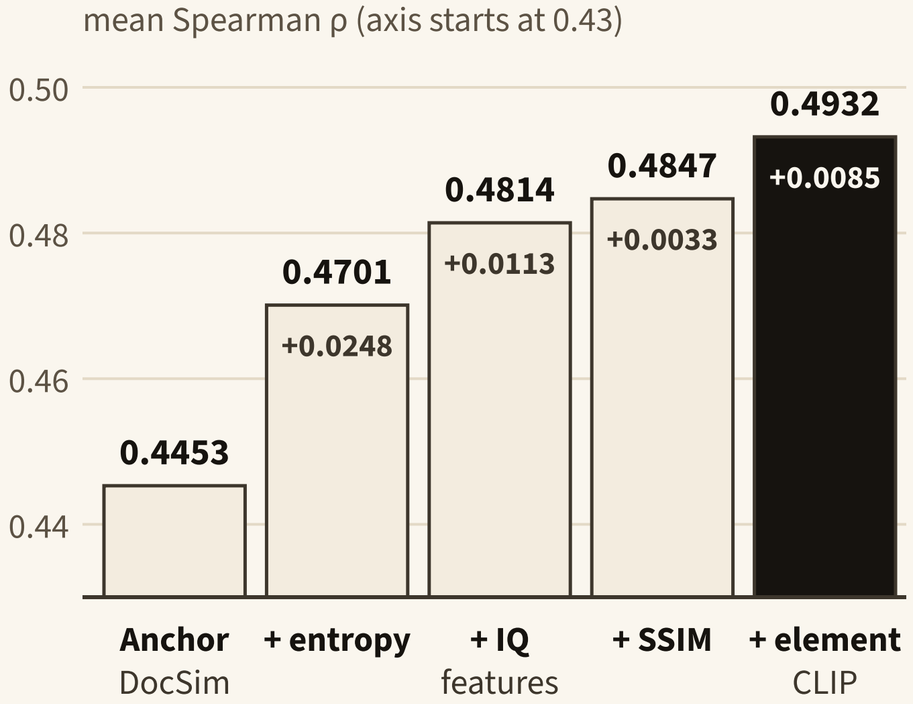
Baseline with CLIP cosine only: **0.269**

Composite stacks lead the field



The winner stacks five complementary signals, tuned per variant

1. **DocSim**: a small CLIP + DINOv2 head trained on page triplets to judge how alike scan and rendering are
2. **Entropy** of OCR token log-probabilities: picks out the stretch of tokens the model was least sure of
3. **Classical image-quality features**: sharpness, gradient and edge density of scan vs rendering
4. **SSIM**: pixel-level structural similarity of the greyscale images, with a weight tuned per variant
5. **Per-element CLIP cosine**: CLIP scores each element; the page takes the near-worst (5th percentile)



DocSim: a small learned similarity head for document pages

If we divide by Δx, we get

$$\frac{\Delta(uv)}{\Delta x} = u\frac{\Delta v}{\Delta x} + v\frac{\Delta u}{\Delta x} + \Delta u\frac{\Delta v}{\Delta x}$$

If we now let $\Delta x \to 0$, we get the derivative of uv:

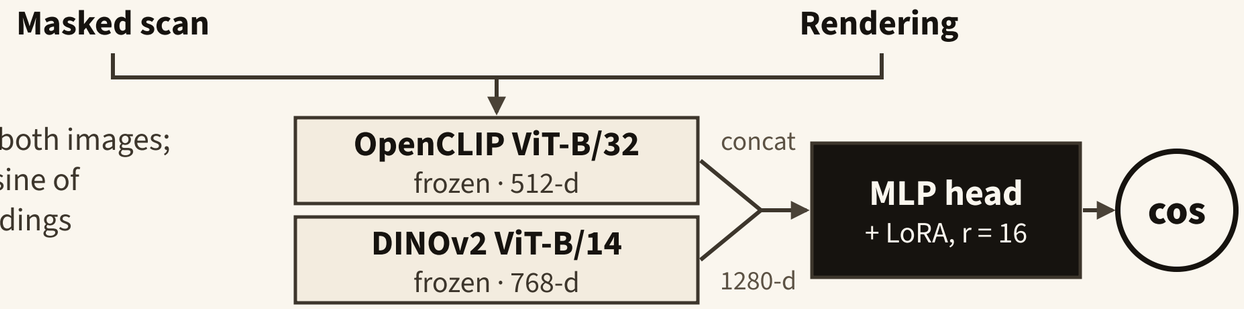
$$\begin{aligned}\frac{d}{dx}(uv) &= \lim_{\Delta x\to 0}\frac{\Delta(uv)}{\Delta x} = \lim_{\Delta x\to 0}\left(u\frac{\Delta v}{\Delta x} + v\frac{\Delta u}{\Delta x} + \Delta u\frac{\Delta v}{\Delta x}\right)\\ &= u\lim_{\Delta x\to 0}\frac{\Delta v}{\Delta x} + v\lim_{\Delta x\to 0}\frac{\Delta u}{\Delta x} + \left(\lim_{\Delta x\to 0}\Delta u\right)\left(\lim_{\Delta x\to 0}\frac{\Delta v}{\Delta x}\right)\\ &= u\frac{dv}{dx} + v\frac{du}{dx} + 0\cdot\frac{dv}{dx}\end{aligned}$$

If we divide by Δx, we get

$$\frac{\Delta(uv)}{\Delta x} = u\frac{\Delta v}{\Delta x} + v\frac{\Delta u}{\Delta x} + \Delta u\frac{\Delta v}{\Delta x}$$

If we now let $\Delta x \to 0$, we get the derivative of uv:

$$\begin{aligned}\frac{d}{dx}(uv) &= \lim_{\Delta x\to 0}\frac{\Delta(uv)}{\Delta x} = \lim_{\Delta x\to 0}\left(u\frac{\Delta v}{\Delta x} + v\frac{\Delta u}{\Delta x} + \Delta u\frac{\Delta v}{\Delta x}\right)\\ &= u\lim_{\Delta x\to 0}\frac{\Delta v}{\Delta x} + v\lim_{\Delta x\to 0}\frac{\Delta u}{\Delta x} + \left(\lim_{\Delta x\to 0}\Delta u\right)\left(\lim_{\Delta x\to 0}\frac{\Delta v}{\Delta x}\right)\\ &= u\frac{dv}{dx} + v\frac{du}{dx} + 0\cdot\frac{dv}{dx}\end{aligned}$$



Same network for both images; the score is the cosine of their 256-d embeddings

TRAINING

- **20,280** cross-page triplets
- anchor: masked scan
- positive: low-error rendering
- negative: high-error rendering of another page
- triplet margin loss
- **40,960** trainable parameters

How it learns: the loss makes each scan more similar, by a margin, to its own page's low-error rendering than to another page's high-error one.

SECTION V

Use and outlook

Practical uses

Check the worst pages first

Sort pages by score so that reviewers start with the pages most likely to be badly read.

Re-run weak pages

Send low-scoring pages to a stronger engine to be read again.

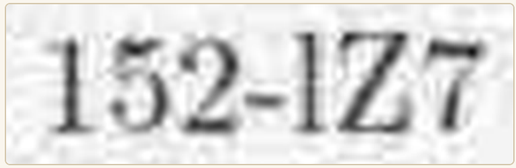
Hold back weak pages

Keep low-scoring pages out of a search index or training data.

CHARACTER LEVEL DETECTION EXAMPLE

Characters such as `1`, `l` and `I`, or `0` and `O`, differ by only a few pixels. Each one is drawn both ways and compared with the page; the closer drawing wins.

On the page



As the OCR read it (“`I`”)



match 0.83

With the look-alike (“`l`”)



match 0.93

“`l`” matches the page better, so the OCR’s “`I`” is flagged and “`l`” is suggested as the fix.

Limitations

WHAT THE SCORE CANNOT TELL

- A **ranking signal**: one page's score does not give its accuracy
- **No visual metric** can see an element the OCR labels with the wrong type, such as a formula read as text: the Case-2 ceiling
- Least sensitive to **single-symbol errors** inside otherwise correct elements
- Engine comparison is reliable when quality **differs clearly**; near-identical engines need finer metrics

SCOPE

- Developed on **one OCR engine** and on English and Chinese pages; other languages are untested
- Positive on every document type, but **weaker on handwritten notes** and design-heavy pages such as slides and magazines
- On a new benchmark, olmOCR-bench, the weights **must be re-tuned**; the frozen CLIP comparison still carries over
- The confidence term needs the OCR's **token probabilities**, which some engines do not provide
- Ground truth is still needed **once, to train DocSim**; scoring needs none

RESOURCES

Thank you!

Code

github.com/kshitij2605/GT-free-OCR-metrics

Datasets and DocSim

huggingface.co/spaces/gt-free-ocr-metrics/dataset-collection

For any questions feel free to email the author at: puku1304@gmail.com