Benchmarking Table Extraction from Heterogeneous Scientific PDF Documents

Marijan Soric[1] Cécile Gracianne[2] Ioana Manolescu[3] Pierre Senellart[1]

[1]DI ENS, ENS, PSL University, CNRS, Inria [2]BRGM, Orléans [3]Inria, Institut Polytechnique de Paris

Motivation and Gap

Table Extraction (TE) turns tables in PDFs into structured data for querying, data lakes and analytics. It has two stages:

- **Table Detection (TD)**: locate each table with a bounding box.
- **Table Structure Recognition (TSR)**: recover rows, columns, merged cells and cell content as HTML.

Prior benchmarks[4,6,7,8,9,10] evaluate TD and TSR **separately**, on **gold-cropped** tables, using **only pages that contain a table**.

Real pipelines don't work that way. TD feeds TSR, so detection errors propagate downstream. And a model never shown a table-free page is never penalised for inventing one. Published TD and TSR scores therefore say very little about end-to-end quality.

Contributions

1. **An end-to-end evaluation framework** for TE, with formally justified metrics that score TD and TSR jointly rather than in isolation.
2. **Two heterogeneous datasets** with high-quality ground truth: **Table-arXiv** (37k pages, LaTeX) and **Table-BRGM** (2k pages, Word, FR/EN, hand-annotated). Both include table-free pages.
3. The first **broad comparison** of 15 methods, from classical libraries to modern Vision Language Models, on 4 datasets.
4. An analysis of **model calibration**, and of the **performance vs. computational cost** trade-off.

Benchmark datasets

Four datasets of *text-based scientific PDFs* with related TD & TSR ground truth, covering LaTeX, Word and professional publishing layouts.

Dataset	Source	# Pages	# Tables
PubTables[1]	Publishing (biomed)	46 942	55 990
Table-arXiv ⋆	LaTeX preprints	36 869	6 308
Table-BRGM ⋆	Word (geology, FR/EN)	2 003	361
ICDAR-2013[2]	Gov. reports	238	163

⋆ New in this work. **Table-arXiv** is generated automatically from 20 years of arXiv LaTeX sources across all domains. **Table-BRGM** is hand-annotated French/English geological survey reports with atypical layouts. Both contain **pages with no table**, so false positives are measurable.

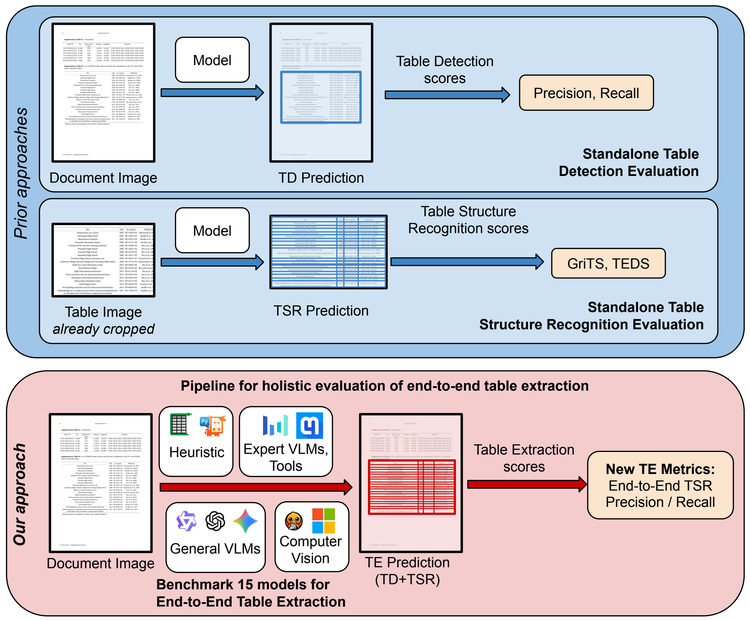
Paper Code Dataset Website Demo

marijan.soric@inria.fr KDD '26, Jeju Island, Republic of Korea

End-to-end evaluation pipeline

Blue: prior benchmarks score TD and TSR independently, each on its own gold-cropped input.

Red: our pipeline runs all 15 methods on the same raw pages, feeds the TD output into TSR, and scores the composite result with dedicated end-to-end metrics.



New end-to-end TE metrics

We replace binary TP/FP with **continuous, TSR-weighted** scores. A detection contributes its structure-recognition quality instead of a 1 or a 0, and TSR quality is measured on **real, possibly imperfect** detections rather than gold crops.

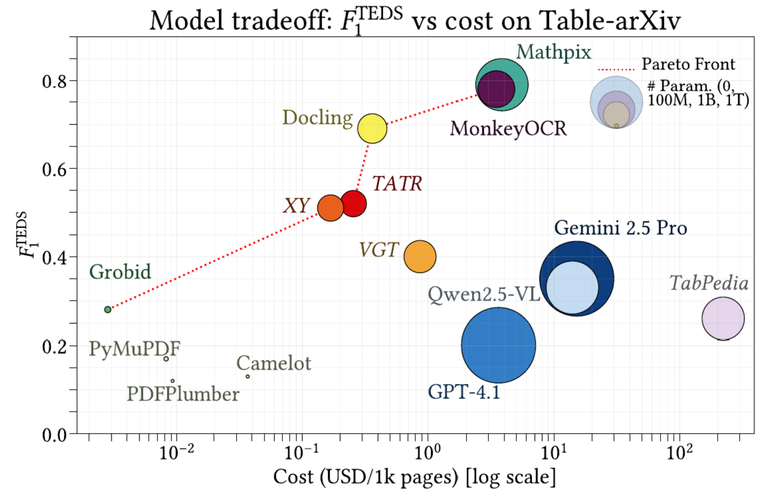
$$P^{\text{TSR}} = \frac{1}{|\mathcal{P}^+|} \sum_{i \in \mathcal{P}^+} s_i^{\text{TSR}} \mathbf{1}_{[\text{IoU}_i > \theta_J]} \qquad R^{\text{TSR}} = \frac{1}{|\mathcal{G}|} \sum_{i \in \mathcal{P}^+} s_i^{\text{TSR}} \mathbf{1}_{[\text{IoU}_i > \theta_J]}$$

$\mathcal{P}^+$: predicted tables; $\mathcal{G}$: ground-truth tables. $s_i^{\text{TSR}} \in [0,1]$ is structure quality: **GriTS**[3] (grid similarity) or **TEDS**[4] (tree edit distance on HTML). IoU_i is against the matched ground-truth table, $\theta_J = 0.5$. We also report **Average Precision** for probabilistic models and **D-ECE**[5] calibration error.

15 methods across 5 families

Family	Methods
Baseline libraries	Camelot, PyMuPDF, PDFPlumber, Grobid
Computer vision	TATR, XY+TATR, VGT+TATR, Docling
General VLMs (general-purpose)	Qwen2.5-VL, GPT-4.1, Gemini 2.5 Pro
Expert VLMs (table specialised)	TabPedia, GOT, MonkeyOCR
Commercial	Mathpix

Performance vs. cost trade-off



The dotted **Pareto front** for **Table-arXiv**, composed of Grobid, PyMuPDF, XY, TATR, Docling, MonkeyOCR and Mathpix, spans **four orders of magnitude** in cost.

Key findings

- **Cheap-to-mid cost wins:** expert VLM **MonkeyOCR**, computer-vision **Docling** and commercial **Mathpix** give near-perfect table detection across all datasets.
- **General VLMs disappoint:** despite their scale, GPT-4.1, Gemini 2.5 Pro and Qwen2.5-VL **hallucinate tables** on table-free pages, and cost 10–100× more for lower end-to-end quality.
- **Heuristics stay competitive:** On Word-typeset Table-BRGM, PyMuPDF and Camelot rival VLMs at a fraction of the cost.
- **Calibration is uneven:** VGT is best calibrated (D-ECE = 0.04); TATR is badly miscalibrated on full pages.
- **Table extraction is unsolved:** no method reaches $F_1^{\text{TEDS}} \geq 90\%$ on any dataset.

Conclusion

Table extraction **remains challenging** across heterogeneous data. Our framework enables the first *direct end-to-end* comparison of diverse methods, revealing that performance depends heavily on document style, that scale does not guarantee quality, and that robustness, calibration and content accuracy remain open problems.

References

[1] Smock, B. et al. (2022). PubTables-1M: Towards Comprehensive Table Extraction From Unstructured Documents. *CVPR*, 4634–4642.
[2] Göbel, M. et al. (2013). ICDAR 2013 Table Competition. *ICDAR*, 1449–1453.
[3] Smock, B. et al. (2023). GriTS: Grid Table Similarity Metric for Table Structure Recognition. *arXiv:2203.12278*.
[4] Zhong, X. et al. (2020). Image-based Table Recognition: Data, Model, and Evaluation. *ECCV*.
[5] Kuppers, F. et al. (2020). Multivariate Confidence Calibration for Object Detection. *CVPRW*.
[6] Li, M. et al. (2020). TableBank: Table Benchmark for Image-based Table Detection and Recognition. *LREC*, 1918–1925.
[7] Chi, Z. et al. (2019). Complicated Table Structure Recognition. *arXiv:1908.04729*.
[8] Long, R. et al. (2021). Parsing Table Structures in the Wild. *ICCV*.
[9] Gao, L. et al. (2019). ICDAR 2019 Competition on Table Detection and Recognition (cTDaR). *ICDAR*, 1510–1515.
[10] Zheng, X. et al. (2021). Global Table Extractor (GTE): A Framework for Joint Table Identification and Cell Structure Recognition Using Visual Context. *WACV*, 697–706.