MedDocBench: Benchmarking Vision-Language Models for Reliable Structured Extraction from Real-World Clinical Forms

Qijia Tian [*1] **Yushi Feng** [*1] **Shuangyu Yang** [2] **Weifa Yang** [†2] **Lequan Yu** [†1]

Abstract

Digitizing clinical workflows requires reliable conversion of paper and scanned forms into analyzable electronic records. Structured health data often begins at this ingestion step before entering downstream analysis, yet it is rarely benchmarked with the same rigor as the modeling that follows. Clinical forms pose challenges such as noisy handwriting, domain-specific abbreviations, ambiguous checkbox marks, and scarce annotated resources. We introduce MedDocBench, a focused bilingual benchmark of de-identified, IRB-approved clinical forms from an academic dental hospital. It covers free-form transcription and structured extraction of handwritten text and yes/no checkboxes. A central contribution is a normalization-aware evaluator with an explicit ON/OFF postprocessing switch for bilingual key mapping, value normalization, and fuzzy matching. Across 12 VLMs, toggling this switch shifts weighted scores by 0.13–0.46 across models; transcription performance does not reliably predict structured extraction quality. These results position MedDocBench as a diagnostic baseline for reliable clinical data ingestion. The benchmark and code are available at `https://github.com/ys-feng/MedDocBench`.

1. Introduction

Machine learning for health depends on structured inputs—diagnosis indicators, medication flags, and questionnaire responses—that feed cohort selection, risk modeling, and clinical operations (AlSaad et al., 2024; Ke et al., 2025). In practice, many of these fields begin as handwritten or scanned intake forms and pass through OCR or vision-language extraction before reaching any downstream table (Wang et al., 2025; Garrido-Munoz et al., 2026). Errors at this ingestion step can silently corrupt the records that follow, yet the step itself is rarely benchmarked with the same rigor as the modeling that consumes it.

Clinical forms pose four compounding extraction challenges that existing benchmarks do not jointly address. **Noisy handwriting** varies widely in stroke style, slant, and legibility, and current VLMs still struggle with unconstrained handwritten input (Garrido-Munoz et al., 2026; Crosilla et al., 2025). **Domain-specific abbreviations**, shorthand such as *Hx* or *c/o*, are ambiguous without clinical context and vary across specialties and institutions (Galfre et al., 2025; Wang et al., 2026). **Ambiguous checkbox marks** include ticks, crosses, circles, underlines, and overwritten strokes that resist clean binary classification (Jaume et al., 2019; Galfre et al., 2025). **Scarce annotated resources** reflect de-identification requirements and institutional access barriers that keep labeled clinical form corpora small relative to general document benchmarks (Yu et al., 2026; Jin et al., 2024).

We introduce **MedDocBench**, a bilingual English–Chinese benchmark from an academic dental hospital, covering two complementary tasks. The **transcription task** asks models to produce free-form plain text from a form image, while the **structured extraction task** asks for a JSON record of handwritten field text and yes/no checkbox values. The latter targets two core clinical form-understanding skills: checkbox interpretation and field-level recovery of sparse handwriting within dense printed layouts. We also provide a normalization-aware evaluator with postprocessing ablation over bilingual key mapping, yes/no value normalization, and fuzzy key matching, making evaluator effects transparent rather than absorbing them into a single reported score (Kovatchev & Lease, 2024; Alzahrani et al., 2024; Reuel et al., 2024).

Across 12 vision-language models and 100 documents, we find that postprocessing changes structured weighted scores by 0.13–0.46 and that transcription quality does not

[*]Equal contribution [†]Corresponding authors. [1]School of Computing and Data Science, The University of Hong Kong, Hong Kong SAR, China [2]Division of Oral and Maxillofacial Surgery, Faculty of Dentistry, The University of Hong Kong, Hong Kong SAR, China. Correspondence to: Weifa Yang <teddyrun@hku.hk>, Lequan Yu <lqyu@hku.hk>.

Proceedings of the Workshop on Structured Data for Health at the 43rd International Conference on Machine Learning, Seoul, South Korea. Copyright 2026 by the author(s).

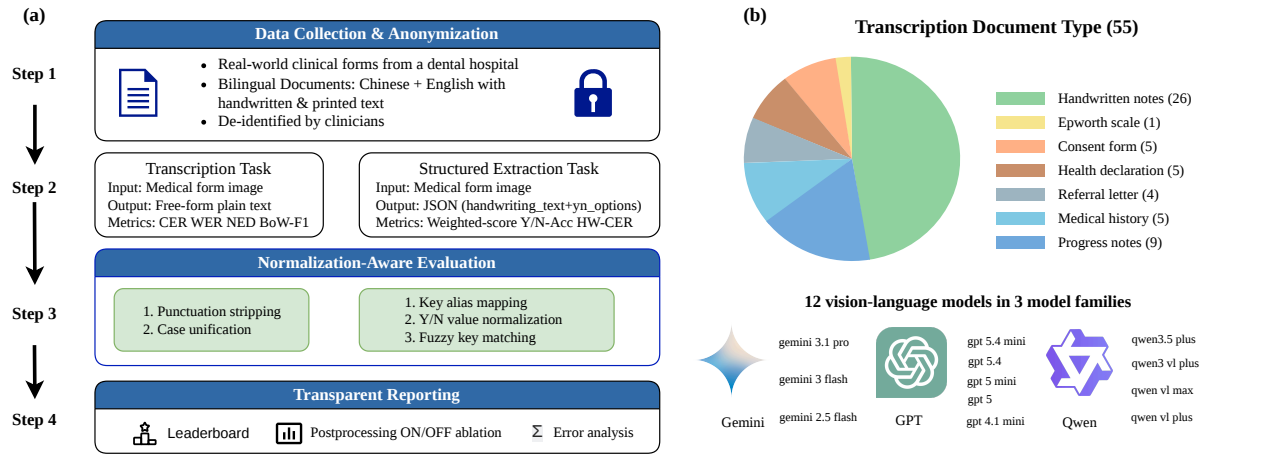

Figure 1. MedDocBench overview: data, two extraction tasks, and normalization-aware evaluation with postprocessing ON/OFF.

reliably predict structured extraction performance.

2. Related Work

OCR and document understanding. Recent OCR and document-understanding benchmarks evaluate increasingly broad capabilities, including text localization, visual-text reasoning, layout-aware extraction, and long-document question answering (Fu et al., 2025; Yang et al., 2025; Deng et al., 2025; Liu et al., 2026; Hu et al., 2024). Form-oriented datasets such as FUNSD and XFUND emphasize key-value structure in noisy or multilingual documents (Jaume et al., 2019; Xu et al., 2022), while receipt and layout benchmarks support post-OCR parsing and document-layout analysis (Park et al., 2019; Zhong et al., 2019; Li et al., 2020). These resources are valuable for general document intelligence, but they usually do not isolate the health-data ingestion setting in which sparse handwritten fields and checkbox marks must be converted into clinically meaningful variables.

Medical document benchmarks. Medical multimodal benchmarks increasingly study report understanding, clinical document QA, bilingual information extraction, and automated dental record generation (Jin et al., 2024; Shang et al., 2025; Do & Do, 2025; Jin et al., 2025; Yu et al., 2026; Wang et al., 2026). Related work on sick-leave certificates and handwritten text recognition also shows that real-world medical forms require more than page-level OCR (Galfre et al., 2025; Garrido-Munoz et al., 2026; Crosilla et al., 2025). MedDocBench complements these efforts by focusing on a narrower but operationally common problem: whether current VLMs can ingest real clinical forms into structured fields that downstream health systems can consume.

Evaluation robustness. Benchmark conclusions can be sensitive to scoring choices, missing outputs, and leaderboard construction (Kovatchev & Lease, 2024; Reuel et al., 2024; Himmi et al., 2024; Sälevä et al., 2025). This concern is amplified for bilingual clinical forms, where a semantically correct answer may use a different language, abbreviated key, or yes/no surface form. Our postprocessing ON/OFF protocol therefore treats normalization as an object of study rather than a hidden implementation detail.

3. Benchmark and Setup

Data. MedDocBench contains 100 clinical documents with bilingual English–Chinese labels, handwriting, printed text, and checkbox-heavy layouts. The benchmark covers two tasks: 55 documents for transcription and 50 for structured extraction, with 5 images shared across both. The transcription split spans multiple document categories, including medical history forms, handwritten notes, consent forms, and referral letters, whereas the structured extraction split consists entirely of medical history forms with a shared field schema, enabling consistent evaluation of handwritten-text extraction and yes/no checkbox recovery. Ground-truth annotations were initialized with model-generated drafts, manually corrected by technical staff, and verified by a certified physician. The structured extraction split contains 25 yes/no fields per document (1,250 total), with a 9% positive rate. The documents were collected from Prince Philip Dental Hospital, Faculty of Dentistry, The University of Hong Kong. All patient-identifying information was removed prior to analysis. The study was approved by the Institutional Review Board of The University of Hong Kong/Hospital Authority Hong Kong West Cluster (HKU/HA HKW IRB UW 25-420). Because the benchmark contains 100 documents

Table 1. Performance of multiple large models on the transcription and structured extraction tasks, averaged over three runs. All values are percentages. We bold the best results and underline the runners-up.

	Transcription Task				Structured Extraction Task			
Model	**CER↓**	**WER↓**	**NED↓**	**BoW F1↑**	**Weighted↑**	**Y/N Acc↑**	**HW CER↓**	**HW WER↓**
qwen-vl-plus	11.7	19.2	11.5	88.3	33.7	26.8	146.8	205.8
qwen3-vl-plus	9.0	16.6	8.7	90.7	69.6	64.0	33.9	77.8
qwen3.5-plus	8.6	12.5	8.3	92.9	74.8	**94.3**	44.7	58.8
qwen-vl-max	8.0	14.6	7.9	89.9	76.1	83.1	72.9	101.8
gpt-4.1-mini	11.0	17.0	10.9	88.2	70.0	74.1	38.1	58.4
gpt-5-mini	10.3	16.9	10.2	88.9	75.9	84.4	44.9	60.8
gpt-5.4-mini	12.1	20.3	12.1	87.3	66.0	76.3	45.6	64.9
gpt-5	55.2	74.9	43.4	56.2	50.8	63.0	71.5	110.5
gpt-5.4	8.4	14.2	8.3	91.9	64.9	78.3	48.4	64.9
gemini-2.5-flash	9.0	14.4	8.9	91.9	**84.4**	90.7	26.3	43.8
gemini-3-flash	**4.3**	**7.1**	**4.2**	**96.1**	84.2	93.1	26.1	**41.1**
gemini-3.1-pro	7.1	12.2	7.0	93.6	80.4	81.2	**23.0**	41.2

from one institution, we interpret it as a controlled diagnostic benchmark rather than a population-representative estimate of clinical-document performance. The fixed structured schema improves comparability across models, but it also limits claims about unseen form families. We therefore emphasize failure modes, evaluator sensitivity, and reproducible baselines over universal deployment claims.

Tasks. The **transcription task** asks the model to produce free-form plain text from a form image. The **structured extraction task** asks for a JSON record with two fields: `handwriting_text` and `yn_options`, where the latter maps checkbox labels to `Y/N`. We separate the tasks because preliminary runs showed that combining them in one prompt degraded both.

Metrics. For transcription we report character error rate (CER), normalized edit distance (NED), and bag-of-words F1 (BoW F1): CER and NED measure edit-distance error, while BoW F1 measures order-agnostic token overlap. For structured extraction we report checkbox accuracy, handwriting CER, and a weighted score computed per sample as

$$s_i = 0.5 \times \big(a_i + \max(0, 1 - \mathrm{CER}_i)\big), \qquad (1)$$

where a_i is sample-level checkbox accuracy and CER_i is handwriting CER; the handwriting term is clipped at zero for robustness to extreme outputs. We report the mean of s_i over samples and runs. Because checkbox positives are sparse, Appendix F additionally reports positive-class precision, recall, F1, and balanced accuracy. These auxiliary metrics are important because overall checkbox accuracy can be inflated by the 9% positive rate; we consequently avoid treating small differences in aggregate Y/N accuracy as definitive evidence of superior positive-case detection.

Normalization-aware evaluation. Both tasks support optional postprocessing before scoring. For transcription, postprocessing applies punctuation stripping and case unification symmetrically to prediction and ground truth. For structured extraction, additional steps handle (i) manually curated bilingual key alias mapping, (ii) yes/no value normalization to canonical `Y/N`, and (iii) fuzzy key matching using an ordered fallback of exact match, substring match, and character-bigram Jaccard similarity with a fixed threshold of 0.6. The `no-postprocess` flag disables these normalization steps. In the OFF setting, the evaluator still parses JSON output, but uses only strict key matching and literal `Y/N` value matching. We report both settings because their gap directly quantifies evaluator sensitivity (Kovatchev & Lease, 2024; Alzahrani et al., 2024). Importantly, the ON setting is deterministic and evaluator-side: it does not infer missing clinical content, repair invalid JSON beyond standard parsing, or use model-specific thresholds. Its role is to measure how much of the apparent error comes from surface-form mismatch rather than failed visual recovery. The OFF setting therefore remains useful as a strict interoperability test, while the ON setting approximates the normalization layer that would typically sit between OCR output and a downstream clinical table.

Models. We evaluate 12 API-accessible vision-language models from the Gemini, OpenAI GPT, and Qwen families, covering leading proprietary systems and open-source model families across different capability tiers. Each model receives the same task-specific prompt for each task and produces three independent outputs per sample; all reported metrics are averaged across the three runs to reduce response variance. No reranking or self-correction is applied, so score differences are more directly attributable to model capability and evaluator choice rather than prompt engineering or post-hoc parsing. Per-sample predictions and scoring artifacts are preserved for sample-level inspection.

4. Experimental Results

Table 1 reports the main results with postprocessing enabled. The benchmark reveals a clear mismatch between *free-form transcription* and *structured extraction*. Models

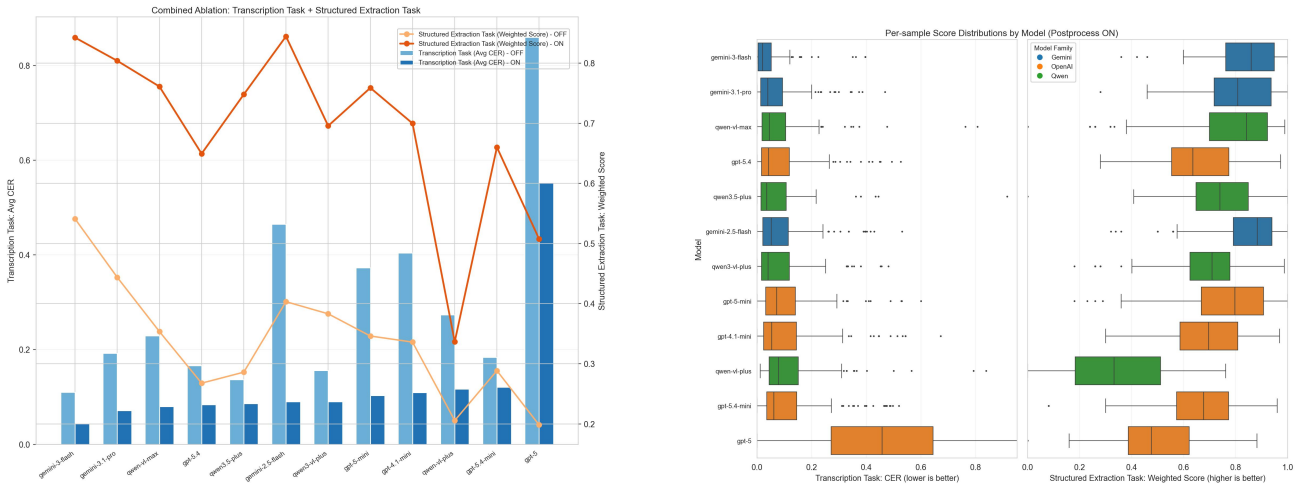

Figure 2. Evaluator sensitivity and model performance distributions. (a) Postprocessing ON/OFF ablation for transcription CER and structured extraction weighted score. (b) Per-sample score distributions under postprocessing ON.

that rank highly on transcription do not necessarily perform best on structured extraction, where success also depends on checkbox interpretation, handwritten-field recovery, and schema-consistent JSON generation.

On the transcription task, `gemini-3-flash` performs best across all three metrics, with a CER of 4.3%, an NED of 4.2%, and a BoW F1 of 96.1%. Most other models fall within a relatively narrow range, suggesting that OCR-style transcription is already reasonably strong for many current VLMs. The main exception is `gpt-5`, which performs substantially worse than the other models on this task.

The ranking changes on the structured extraction task. `gemini-2.5-flash` and `gemini-3-flash` form the top tier with weighted scores of 84.4% and 84.2%. Meanwhile, `gemini-3.1-pro` obtains the lowest handwriting CER at 23.0%. `qwen3.5-plus` reaches the highest Y/N accuracy at 94.3%, but its weaker handwriting CER reduces its overall weighted score relative to the top Gemini models. This pattern highlights the two-part nature of the structured extraction task, which requires both accurate checkbox interpretation and handwritten recognition. Because only 50 samples are available for structured extraction, close rank differences should be read cautiously. We therefore computed document-level bootstrap confidence intervals, resampling the 50 forms with all 25 checkbox fields kept together and averaging over the three runs. The top two weighted scores overlap substantially (`gemini-2.5-flash`: 0.844 [0.813, 0.873]; `gemini-3-flash`: 0.842 [0.816, 0.867]), while positive-class checkbox metrics show wider uncertainty under the 9.0% positive rate. Appendices D and F therefore report exploratory paired tests and positive-class diagnostics to support cautious interpretation.

These shifts across tasks are substantial. `gpt-5.4` performs strongly on transcription but reaches only 64.9% on structured extraction. In contrast, `gemini-2.5-flash` is not the top transcription model but achieves the strongest overall structured extraction result. At the low end, `qwen-vl-plus` performs poorly on both checkbox interpretation and handwritten-text recovery. Overall, the results show that strong OCR transcription alone does not imply reliable clinical form digitization.

5. Analysis

Figure 2 highlights two main findings. First, benchmark conclusions depend strongly on evaluator design. Second, the remaining errors are dominated by *structural* failures rather than OCR alone.

Figure 2(a) shows that postprocessing has a modest effect on the transcription task but a much larger effect on the structured extraction task. Turning postprocessing OFF increases CER for all models, yet the relative ordering on the transcription task is broadly preserved. By contrast, the weighted score on structured extraction drops sharply for nearly every model, indicating that raw matching penalizes not only true extraction failures but also representation mismatch, such as noncanonical keys, bilingual label variants, or yes/no surface forms. Evaluator sensitivity is therefore a central issue rather than an implementation detail: for clinical form extraction, a single reported score can conflate semantic recovery with output-format compliance.

Figure 2(b) helps explain why the transcription task and structured extraction task do not move together. Qualitative review of low-scoring samples suggests that the hard cases are not just difficult handwriting, but difficult *field binding*: models may mix nearby printed text into handwritten fields, drift across checkbox polarity in dense sections, or

assign recognized content to the wrong JSON key. These are structural errors, and they directly affect the machine-readable record even when much of the text content is visually recoverable. Taken together, the results suggest that the main bottleneck in clinical form extraction is not text recognition alone, but field-level structural reliability under imperfect output formats. The same pattern appears in checkbox extraction. Some models are conservative and miss positive marks, while others over-trigger on nearby strokes, printed option markers, or crossed-out annotations. These two failure modes can yield similar aggregate accuracy when the negative class dominates, but they imply different downstream risks: false negatives may exclude relevant patients from a cohort, whereas false positives may introduce unsupported risk factors into a record.

The per-sample distributions also show that average scores hide important reliability differences. Top models combine strong means with tighter spreads, while several mid-ranked models overlap substantially, suggesting that practical robustness is determined by a small number of challenging forms. For research, this argues for reporting both raw and normalized results, together with field-level diagnostics, rather than a single leaderboard number. For clinical use, the implication is that high-impact structured fields should be treated as verification targets: the near-term path is not fully automatic extraction, but risk-sensitive automation that prioritizes consistency checks and selective human review on difficult form layouts.

6. Discussion

MedDocBench is designed as a focused benchmark for an under-studied step in the health-data pipeline: converting real-world clinical forms into structured variables. Its scope is intentionally controlled. The structured split uses one clinical form family with a shared schema, which enables consistent measurement of checkbox recovery, handwritten-field extraction, and evaluator sensitivity rather than mixing many document types into a broad OCR leaderboard. Although the data come from one clinical domain and one institutional document source, they provide a realistic diagnostic setting for field-level reliability under bilingual layouts, sparse handwriting, and ambiguous checkbox marks.

A key takeaway is that structured extraction is not a simple downstream formatting step after OCR. It requires the model to decide which visual marks correspond to clinical variables, bind sparse handwriting to the right field, and express the result in a schema that downstream systems can consume. This is precisely where page-level transcription metrics become insufficient: two outputs may look similarly readable to a human while producing very different structured records. This gap is central for health data, where a single field can affect cohort membership or risk-variable construction.

We view this benchmark as a clean baseline for current VLMs, not as the ceiling of a deployable clinical extraction system. We evaluate models in a single-turn, no-tool setting without self-correction, template-specific preprocessing, or human-in-the-loop repair. This controlled protocol makes errors easier to attribute and highlights a central lesson: clinical document understanding should be evaluated by the reliability of the structured variables produced, not only by page-level text recognition.

The benchmark also surfaces governance issues that are easy to miss in purely technical evaluations. Although all documents were de-identified and the study was IRB-approved, API-based evaluation still requires careful institutional oversight because medical-document images may retain residual privacy risks (Young, 2025). We therefore treat commercial model results as evidence about current extraction behavior, not as an endorsement of direct clinical deployment. Practical use would require local data-use agreements, audit trails, access controls, and human verification for high-impact fields.

Future versions of MedDocBench should broaden this diagnostic setting in three directions. First, multi-institutional data would test whether the observed failure modes persist across form templates, handwriting styles, and local abbreviation conventions. Second, additional structured schemas would separate schema-following ability from adaptation to a single medical-history form family. Third, prompt-engineering and multi-turn repair protocols could be evaluated as separate system settings rather than mixed into the baseline, with uncertainty estimates reported for sparse positive fields. We leave these extensions as natural follow-up work and keep the present benchmark focused on the clean single-turn comparison.

7. Conclusion

MedDocBench reframes clinical-form OCR as a structured health-data ingestion problem. Across 12 vision-language models, we show that free-form transcription performance does not reliably predict field-level structured extraction, where checkbox interpretation, handwritten-field binding, and schema alignment become critical failure points. We further find that evaluator design substantially affects measured performance: normalization has limited impact on transcription but can sharply change structured extraction scores. By separating visual text recovery, checkbox interpretation, field binding, schema alignment, and evaluator-side normalization, MedDocBench provides a compact diagnostic setting for auditing where current systems may fail before deployment.

Acknowledgements

This work was supported in part by the Research Grants Council of Hong Kong (17200125, C5055-24G, and T45-401/22-N), and the Hong Kong Innovation and Technology Fund (ITS/544/24FP and GHP/318/22GD).

References

AlSaad, R., Abd-alrazaq, A., Boughorbel, S., Ahmed, A., Renault, M.-A., Damseh, R., and Sheikh, J. Multimodal large language models in health care: Applications, challenges, and future outlook. *J Med Internet Res*, 26:e59505, 2024. doi: 10.2196/59505. URL https://doi.org/10.2196/59505.

Alzahrani, N., Alyahya, H., Alnumay, Y., et al. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024. URL https://aclanthology.org/2024.acl-long.744.

Crosilla, G., Klic, L., and Colavizza, G. Benchmarking large language models for handwritten text recognition. *Journal of Documentation*, 81(7):334–354, 2025. doi: https://doi.org/10.1108/JD-03-2025-0082. URL https://www.sciencedirect.com/science/article/pii/S0022041825000323.

Deng, C., Yuan, J., Bu, P., Wang, P., Li, Z.-Z., Xu, J., Li, X.-H., Gao, Y., Song, J., Zheng, B., and Liu, C.-L. LongDocURL: a comprehensive multimodal long document benchmark integrating understanding, reasoning, and locating. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T. (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1135–1159, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.57. URL https://aclanthology.org/2025.acl-long.57/.

Do, A.-D. and Do, T.-H. Adapting vision-language models for information extraction from bilingual medical invoices. In *2025 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 2424–2429, 10 2025. doi: 10.1109/APSIPAASC65261.2025.11249133.

Fu, L., Kuang, Z., Song, J., Huang, M., Yang, B., Li, Y., Zhu, L., Luo, Q., Wang, X., Lu, H., Li, Z., Tang, G., Shan, B., Lin, C., Liu, Q., Wu, B., Feng, H., Liu, H., Huang, C., Tang, J., Chen, W., Jin, L., Liu, Y., and Bai, X. OCRBench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. URL https://openreview.net/forum?id=Vb6i3Dp24N.

Galfre, M., Curcio, V., Dardouillet, P., and Le Van, O. Vision-language models for sick leave certificates: Beyond ocr in real-world form understanding. In *2025 3rd International Conference on Foundation and Large Language Models (FLLM)*, pp. 254–261, 2025. doi: 10.1109/FLLM67465.2025.11390982.

Garrido-Munoz, C., Rios-Vila, A., and Calvo-Zaragoza, J. Handwritten Text Recognition: A Survey . *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 48(04):4367–4387, 2026. doi: 10.1109/TPAMI.2025.3646002. URL https://doi.ieeecomputersociety.org/10.1109/TPAMI.2025.3646002.

Himmi, A., Irurozki, E., Noiry, N., Clemencon, S., and Colombo, P. Towards more robust nlp system evaluation: Handling missing scores in benchmarks. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 11759–11785, 2024.

Hu, A., Xu, H., Ye, J., Yan, M., Zhang, L., Zhang, B., Zhang, J., Jin, Q., Huang, F., and Zhou, J. mPLUG-DocOwl 1.5: Unified structure learning for OCR-free document understanding. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 3096–3120, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.175. URL https://aclanthology.org/2024.findings-emnlp.175/.

Jaume, G., Kemal Ekenel, H., and Thiran, J.-P. Funsd: A dataset for form understanding in noisy scanned documents. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 2, pp. 1–6, 2019. doi: 10.1109/ICDARW.2019.10029.

Jin, C., Zhang, M., Ma, W., Li, Y., Wang, Y., Jia, Y., Du, Y., Sun, T., Wang, H., Fan, C., Gu, J., Chi, C., Lv, X., Li, F., Xue, W., and Huang, Y. Rjuameddqa: A multimodal benchmark for medical document question answering and clinical reasoning. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 5218–5229, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704901. doi: 10.1145/3637528.3671644. URL https://doi.org/10.1145/3637528.3671644.

Jin, W., Sun, Y., Ji, K., Jiang, X., Hu, Y., Wang, J., and Liu, J. Medscreendental: Automated structured dental record generation via multimodal language model integration. *Displays*, 90:103119, 2025. ISSN 0141-9382. doi: https://doi.org/10.1016/j.displa.2025.103119. URL https://www.sciencedirect.com/science/article/pii/S0141938225001568.

Ke, W., Zheng, Y., Li, Y., Xu, H., Nie, D., Wang, P., and He, Y. Large language models in document intelligence: A comprehensive survey, recent advances, challenges and future trends. *ACM Transactions on Information Systems*, 44, 09 2025. doi: 10.1145/3768156.

Kovatchev, V. and Lease, M. Benchmark transparency: Measuring the impact of data on evaluation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1536–1551, 2024.

Li, M., Xu, Y., Cui, L., Huang, S., Wei, F., Li, Z., and Zhou, M. DocBank: A benchmark dataset for document layout analysis. In Scott, D., Bel, N., and Zong, C. (eds.), *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 949–960, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.82. URL https://aclanthology.org/2020.coling-main.82/.

Liu, Y., Yang, B., Liu, Q., Li, Z., Ma, Z., Zhang, S., and Bai, X. Textmonkey: an ocr-free large multimodal model for understanding document. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–12, 2026. doi: 10.1109/TPAMI.2026.3653415.

Park, S., Shin, S., Lee, B., Lee, J., Surh, J., Seo, M., and Lee, H. {CORD}: A consolidated receipt dataset for post-{ocr} parsing. In *Workshop on Document Intelligence at NeurIPS 2019*, 2019. URL https://openreview.net/forum?id=SJl3z659UH.

Reuel, A., Hardy, A., Smith, C., Lamparth, M., Hardy, M., and Kochenderfer, M. J. Betterbench: Assessing ai benchmarks, uncovering issues, and establishing best practices. *Advances in Neural Information Processing Systems*, 37:21763–21813, 2024.

Sälevä, J., Ataman, D., and Lignos, C. Beyond statistical significance: Quantifying uncertainty and statistical variability in multilingual and multitask NLP evaluation. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 2304–2321, Mumbai, India, December 2025. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics. URL https://aclanthology.org/2025.ijcnlp-long.125/.

Shang, F., Xia, Y., Yang, D., Wang, Y., and Yang, B. Medrepbench: A comprehensive benchmark for medical report interpretation, 2025. URL https://arxiv.org/abs/2508.16674.

Wang, W., Hu, H., Zhang, Z., Li, Z., Shao, H., and Dahlmeier, D. Document intelligence in the era of large language models: A survey, 2025. URL https://arxiv.org/abs/2510.13366.

Wang, W., Wu, W., Liu, Y., Zhao, Y., Lv, X., Diao, L., Fan, Z., Xie, W., Lin, Z., Shi, D., Huang, L., Xu, K., and Li, H. Medocvl: A visual language model for medical document understanding and parsing, 2026. URL https://arxiv.org/abs/2602.06402.

Xu, Y., Lv, T., Cui, L., Wang, G., Lu, Y., Florencio, D., Zhang, C., and Wei, F. XFUND: A benchmark dataset for multilingual visually rich form understanding. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 3214–3224, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.253. URL https://aclanthology.org/2022.findings-acl.253/.

Yang, Z., Tang, J., Li, Z., Wang, P., Wan, J., Zhong, H., Liu, X., Yang, M., Wang, P., Bai, S., Jin, L., and Lin, J. Cc-ocr: A comprehensive and challenging ocr benchmark for evaluating large multimodal models in literacy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21744–21754, October 2025.

Young, R. J. Vision token masking alone cannot prevent phi leakage in medical document ocr: A systematic evaluation, 2025. URL https://arxiv.org/abs/2511.18272.

Yu, A., Weile, J., and Courtot, M. Benchmarking llm-based information extraction tools for medical documents. *medRxiv*, 2026. doi: 10.64898/2026.01.19.26344287. URL https://www.medrxiv.org/content/early/2026/01/22/2026.01.19.26344287.

Zhong, X., Tang, J., and Jimeno-Yepes, A. Publaynet: Largest dataset ever for document layout analysis. *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pp. 1015–1022, 2019. URL https://api.semanticscholar.org/CorpusID:201124789.

A. Prompts for OCR Tasks

For reproducibility, we report the centralized prompts for the two OCR settings in MedDocBench. The transcription task uses a general-purpose text transcription prompt, whereas the structured extraction task uses a structured prompt that separates handwritten text from binary Y/N option detection.

Transcription task: standard text-based OCR (general purpose).

```
Please perform OCR on this image. Return the recognized text exactly as it appears.
CRITICAL RULES FOR FAIR EVALUATION:
1. Do NOT add any numbering (like '1.', '2.') unless they are printed in the image.
2. Do NOT add any markdown formatting, code blocks (```), tables, or bold text.
3. Do NOT fix typos, grammar, punctuation, or 'improve' the text.
Output exactly what is seen.
4. Do NOT add any extra punctuation or spaces to align text.
Preserve the original layout roughly via line breaks.
5. Do NOT try to format the output as a list if it's not a clear list in the image.
6. For selection options (like Yes/No or checkboxes),
represent the marked choice as 'Y' or 'N'.
7. Do NOT include any introductory or concluding remarks.
Start the output directly with the recognized text.
```

Structured extraction task: simple handwriting + Y/N extraction.

```
Perform OCR with two goals: (1) extract all handwritten text, and
(2) detect Y/N selections for any options or checkboxes, regardless of form format.

Return ONLY a JSON object with this structure:
{
  "handwriting_text": "...",
  "yn_options": { "Label 1": "Y/N", "Label 2": "Y/N" }
}

RULES:
1. handwriting_text: include ONLY handwritten content. Preserve line breaks when possible.
If none, use an empty string.
2. yn_options: use the visible option/question label as the key (any language).
Mark 'Y' only if the Y/Yes option is explicitly ticked/circled/checked.
If N/No is marked, or neither is marked, use 'N'.
3. Do NOT add extra text outside the JSON. Do NOT use markdown/code blocks.
```

B. Transcription Task Full Results

Table 2 reports per-model transcription task metrics across all 55 samples (multi-run mean over three runs). PP denotes whether postprocessing was enabled (✓) or disabled (×). Bold entries indicate the best value per column.

Table 2. Transcription task full results ($n = 55$ per run, 3 runs averaged per model). PP = postprocessing (✓ on, × off). ↓ lower is better; ↑ higher is better.

Model	PP	CER↓	WER↓	NED↓	Prec.↑	Recall↑	BoW F1↑	Exact Match↑
gemini-3-flash	✓	**0.0434**	**0.0712**	**0.0424**	**0.9057**	**0.9066**	**0.9608**	**0.1455**
gemini-3-flash	×	0.1099	0.1050	0.1008	0.9037	0.8986	0.9333	0.0606
gemini-3.1-pro	✓	0.0712	0.1218	0.0703	0.8536	0.8533	0.9364	0.0849
gemini-3.1-pro	×	0.1921	0.1677	0.1566	0.8430	0.8416	0.8978	0.0182
qwen-vl-max	✓	0.0799	0.1464	0.0793	0.8039	0.7965	0.8990	0.0666
qwen-vl-max	×	0.2293	0.1790	0.1664	0.8076	0.8133	0.8728	0.0000
gpt-5.4	✓	0.0836	0.1422	0.0829	0.8474	0.8382	0.9194	0.1212
gpt-5.4	×	0.1663	0.1791	0.1581	0.8197	0.8079	0.8874	0.0000
qwen3.5-plus	✓	0.0860	0.1245	0.0827	0.8633	0.8652	0.9289	0.0666
qwen3.5-plus	×	0.1367	0.1548	0.1331	0.8694	0.8349	0.9030	0.0000
gemini-2.5-flash	✓	0.0899	0.1436	0.0887	0.8279	0.8274	0.9186	0.0485
gemini-2.5-flash	×	0.4647	0.1870	0.2532	0.7174	0.8045	0.8816	0.0121
qwen3-vl-plus	✓	0.0901	0.1660	0.0872	0.8104	0.8051	0.9070	0.0545
qwen3-vl-plus	×	0.1559	0.1959	0.1495	0.8295	0.8112	0.8837	0.0000
gpt-5-mini	✓	0.1034	0.1694	0.1020	0.7646	0.7485	0.8889	0.0727
gpt-5-mini	×	0.3731	0.2092	0.2674	0.7089	0.7874	0.8554	0.0000
gpt-4.1-mini	✓	0.1096	0.1700	0.1090	0.7648	0.7428	0.8820	0.0727
gpt-4.1-mini	×	0.4040	0.2127	0.2765	0.6990	0.7798	0.8469	0.0000
qwen-vl-plus	✓	0.1170	0.1918	0.1149	0.7313	0.7202	0.8832	0.0000
qwen-vl-plus	×	0.2737	0.2372	0.2114	0.7503	0.7490	0.8485	0.0000
gpt-5.4-mini	✓	0.1210	0.2033	0.1205	0.7851	0.7581	0.8725	0.0485
gpt-5.4-mini	×	0.1838	0.2415	0.1793	0.8036	0.7612	0.8404	0.0000
gpt-5	✓	0.5517	0.7489	0.4340	0.3282	0.3234	0.5619	0.0121
gpt-5	×	0.8596	0.7966	0.5419	0.3223	0.3701	0.5239	0.0000

C. Structured Extraction Task Full Results

Table 3 reports per-model structured extraction task metrics across all 50 samples (multi-run mean over three runs). HW CER, HW WER, and HW NED are effectively unchanged between PP-on and PP-off rows for each model, consistent with the postprocessing switch mainly affecting the Y/N extraction branch.

Table 3. Structured extraction task full results ($n = 50$ per run, 3 runs averaged per model). PP = postprocessing (✓ on, × off). ↓ lower is better; ↑ higher is better.

Model	PP	Weighted↑	Y/N Acc↑	HW CER↓	HW WER↓	HW NED↓
gemini-2.5-flash	✓	**0.8444**	0.9069	0.2627	0.4384	0.2043
gemini-2.5-flash	×	0.4033	0.0248	0.2627	0.4384	0.2043
gemini-3-flash	✓	0.8424	0.9309	0.2608	**0.4106**	0.2363
gemini-3-flash	×	0.5410	0.3283	0.2608	**0.4106**	0.2363
gemini-3.1-pro	✓	0.8038	0.8117	**0.2302**	0.4120	**0.1848**
gemini-3.1-pro	×	0.4438	0.0917	**0.2302**	0.4120	**0.1848**
qwen-vl-max	✓	0.7610	0.8312	0.7290	1.0184	0.2787
qwen-vl-max	×	0.3536	0.0165	0.7290	1.0184	0.2787
gpt-5-mini	✓	0.7588	0.8437	0.4488	0.6083	0.2674
gpt-5-mini	×	0.3462	0.0187	0.4488	0.6083	0.2674
qwen3.5-plus	✓	0.7479	**0.9432**	0.4474	0.5880	0.4474
qwen3.5-plus	×	0.2861	0.0197	0.4474	0.5880	0.4474
gpt-4.1-mini	✓	0.6997	0.7408	0.3813	0.5844	0.3311
gpt-4.1-mini	×	0.3361	0.0136	0.3813	0.5844	0.3311
qwen3-vl-plus	✓	0.6956	0.6400	0.3387	0.7777	0.2049
qwen3-vl-plus	×	0.3831	0.0149	0.3387	0.7777	0.2049
gpt-5.4-mini	✓	0.6601	0.7629	0.4555	0.6493	0.4411
gpt-5.4-mini	×	0.2886	0.0200	0.4555	0.6493	0.4411
gpt-5.4	✓	0.6494	0.7832	0.4843	0.6488	0.4843
gpt-5.4	×	0.2678	0.0200	0.4843	0.6488	0.4843
gpt-5	✓	0.5076	0.6299	0.7151	1.1054	0.5490
gpt-5	×	0.1984	0.0115	0.7151	1.1054	0.5490
qwen-vl-plus	✓	0.3367	0.2680	1.4682	2.0579	0.5538
qwen-vl-plus	×	0.2057	0.0059	1.4682	2.0579	0.5538

D. Full Pairwise Significance Tests

Table 4 reports paired t-tests ($n = 50$) for all model pairs on structured extraction task weighted score. Significance threshold: $\alpha = 0.05$. These pairwise tests are exploratory and are reported without multiple-comparison correction.

Table 4. Full pairwise paired t-test results on structured extraction task weighted score ($n = 50$).

Model 1	Model 2	M1	M2	p-value	Winner
gemini-2.5-flash	gemini-3-flash	0.844	0.842	0.885	—
gemini-2.5-flash	gemini-3.1-pro	0.844	0.804	0.018	gemini-2.5-flash
gemini-2.5-flash	qwen-vl-max	0.844	0.761	0.010	gemini-2.5-flash
gemini-2.5-flash	gpt-5-mini	0.844	0.759	<0.001	gemini-2.5-flash
gemini-2.5-flash	qwen3.5-plus	0.844	0.748	<0.001	gemini-2.5-flash
gemini-2.5-flash	gpt-4.1-mini	0.844	0.700	<0.001	gemini-2.5-flash
gemini-2.5-flash	qwen3-vl-plus	0.844	0.696	<0.001	gemini-2.5-flash
gemini-2.5-flash	gpt-5.4-mini	0.844	0.660	<0.001	gemini-2.5-flash
gemini-2.5-flash	gpt-5.4	0.844	0.649	<0.001	gemini-2.5-flash
gemini-2.5-flash	gpt-5	0.844	0.508	<0.001	gemini-2.5-flash
gemini-2.5-flash	qwen-vl-plus	0.844	0.337	<0.001	gemini-2.5-flash
gemini-3-flash	gemini-3.1-pro	0.842	0.804	0.020	gemini-3-flash
gemini-3-flash	qwen-vl-max	0.842	0.761	0.011	gemini-3-flash
gemini-3-flash	gpt-5-mini	0.842	0.759	<0.001	gemini-3-flash
gemini-3-flash	qwen3.5-plus	0.842	0.748	<0.001	gemini-3-flash
gemini-3-flash	gpt-4.1-mini	0.842	0.700	<0.001	gemini-3-flash
gemini-3-flash	qwen3-vl-plus	0.842	0.696	<0.001	gemini-3-flash
gemini-3-flash	gpt-5.4-mini	0.842	0.660	<0.001	gemini-3-flash
gemini-3-flash	gpt-5.4	0.842	0.649	<0.001	gemini-3-flash
gemini-3-flash	gpt-5	0.842	0.508	<0.001	gemini-3-flash
gemini-3-flash	qwen-vl-plus	0.842	0.337	<0.001	gemini-3-flash
gemini-3.1-pro	qwen-vl-max	0.804	0.761	0.190	—
gemini-3.1-pro	gpt-5-mini	0.804	0.759	0.035	gemini-3.1-pro
gemini-3.1-pro	qwen3.5-plus	0.804	0.748	0.018	gemini-3.1-pro
gemini-3.1-pro	gpt-4.1-mini	0.804	0.700	<0.001	gemini-3.1-pro
gemini-3.1-pro	qwen3-vl-plus	0.804	0.696	<0.001	gemini-3.1-pro
gemini-3.1-pro	gpt-5.4-mini	0.804	0.660	<0.001	gemini-3.1-pro
gemini-3.1-pro	gpt-5.4	0.804	0.649	<0.001	gemini-3.1-pro
gemini-3.1-pro	gpt-5	0.804	0.508	<0.001	gemini-3.1-pro
gemini-3.1-pro	qwen-vl-plus	0.804	0.337	<0.001	gemini-3.1-pro
qwen-vl-max	gpt-5-mini	0.761	0.759	0.941	—
qwen-vl-max	qwen3.5-plus	0.761	0.748	0.706	—
qwen-vl-max	gpt-4.1-mini	0.761	0.700	0.035	qwen-vl-max
qwen-vl-max	qwen3-vl-plus	0.761	0.696	0.070	—
qwen-vl-max	gpt-5.4-mini	0.761	0.660	0.006	qwen-vl-max
qwen-vl-max	gpt-5.4	0.761	0.649	0.005	qwen-vl-max
qwen-vl-max	gpt-5	0.761	0.508	<0.001	qwen-vl-max
qwen-vl-max	qwen-vl-plus	0.761	0.337	<0.001	qwen-vl-max
gpt-5-mini	qwen3.5-plus	0.759	0.748	0.682	—
gpt-5-mini	gpt-4.1-mini	0.759	0.700	0.006	gpt-5-mini
gpt-5-mini	qwen3-vl-plus	0.759	0.696	0.006	gpt-5-mini
gpt-5-mini	gpt-5.4-mini	0.759	0.660	<0.001	gpt-5-mini
gpt-5-mini	gpt-5.4	0.759	0.649	0.002	gpt-5-mini
gpt-5-mini	gpt-5	0.759	0.508	<0.001	gpt-5-mini
gpt-5-mini	qwen-vl-plus	0.759	0.337	<0.001	gpt-5-mini
qwen3.5-plus	gpt-4.1-mini	0.748	0.700	0.055	—
qwen3.5-plus	qwen3-vl-plus	0.748	0.696	0.047	qwen3.5-plus
qwen3.5-plus	gpt-5.4-mini	0.748	0.660	<0.001	qwen3.5-plus
qwen3.5-plus	gpt-5.4	0.748	0.649	<0.001	qwen3.5-plus
qwen3.5-plus	gpt-5	0.748	0.508	<0.001	qwen3.5-plus
qwen3.5-plus	qwen-vl-plus	0.748	0.337	<0.001	qwen3.5-plus

Continued on next page

Table 4. Full pairwise paired t-test results on structured extraction task weighted score ($n = 50$) (continued).

Model 1	Model 2	M1	M2	p-value	Winner
gpt-4.1-mini	qwen3-vl-plus	0.700	0.696	0.846	—
gpt-4.1-mini	gpt-5.4-mini	0.700	0.660	0.148	—
gpt-4.1-mini	gpt-5.4	0.700	0.649	0.097	—
gpt-4.1-mini	gpt-5	0.700	0.508	<0.001	gpt-4.1-mini
gpt-4.1-mini	qwen-vl-plus	0.700	0.337	<0.001	gpt-4.1-mini
qwen3-vl-plus	gpt-5.4-mini	0.696	0.660	0.169	—
qwen3-vl-plus	gpt-5.4	0.696	0.649	0.135	—
qwen3-vl-plus	gpt-5	0.696	0.508	<0.001	qwen3-vl-plus
qwen3-vl-plus	qwen-vl-plus	0.696	0.337	<0.001	qwen3-vl-plus
gpt-5.4-mini	gpt-5.4	0.660	0.649	0.580	—
gpt-5.4-mini	gpt-5	0.660	0.508	<0.001	gpt-5.4-mini
gpt-5.4-mini	qwen-vl-plus	0.660	0.337	<0.001	gpt-5.4-mini
gpt-5.4	gpt-5	0.649	0.508	<0.001	gpt-5.4
gpt-5.4	qwen-vl-plus	0.649	0.337	<0.001	gpt-5.4
gpt-5	qwen-vl-plus	0.508	0.337	<0.001	gpt-5

E. Challenging Edge Cases

Our per-sample transcription analysis surfaces two failure modes: (i) systematically hard samples across model families, and (ii) model-specific catastrophic degradations on otherwise tractable inputs.

Systematically hard samples. Using cross-model mean CER from multirun outputs, we identify `hw14.png`, `hw13.png`, and `hw15.png` as the hardest samples. These files show CER ≥ 0.2 for 10–12 of 12 models, suggesting dataset-level difficulty rather than isolated decoding errors. Table 5 reports representative model CERs on these samples and each model's transcription-task mean CER.

Table 5. CER on systematically hard transcription samples (postprocessing on) and each model's transcription-task mean CER.

Model	hw14	hw13	hw15	Mean CER
gemini-3-flash	0.2142	0.1991	0.1021	0.0434
gemini-3.1-pro	0.2816	0.2255	0.1999	0.0712
gpt-5-mini	0.5398	0.4227	0.2935	0.1034

Model-specific catastrophic failure. A distinct failure mode appears on `cf3.png`: `gemini-3-flash` achieves CER 0.0052, while `gpt-5` reaches CER 3.4120 (absolute gap = 3.4068, ratio = 652.8×). This large gap is consistent with occasional output instability in gpt-5 on this benchmark and motivates caution when interpreting its aggregate rank.

F. Checkbox Confusion Metrics

Because positive checkbox marks are sparse in the structured split (9.0% positive; 1,250 checkbox fields in total, with 113 positive and 1,137 negative fields per run), Table 6 reports field-level confusion statistics and positive-class metrics for postprocessed runs. We separate unmatched or invalid checkbox outputs into Missing+ (ground-truth Y) and Missing− (ground-truth N); under this definition, Y/N accuracy is exactly

$$\frac{TP + TN}{TP + TN + FP + FN + \text{Missing}^{+} + \text{Missing}^{-}}.$$

G. Ethics and Data Handling

The study was approved by the relevant institutional review board, and all documents were de-identified before analysis. We evaluated commercial API-accessible models only on de-identified document images and did not intentionally submit protected health information. This de-identification step reduces privacy risk but does not eliminate the governance

Table 6. Field-level checkbox confusion statistics and positive-class metrics on the structured extraction task (PP on, 3 runs averaged per model). Counts are run means, so they need not be integers.

Model	TP	TN	FP	FN	Miss+	Miss−	Y/N Acc↑	Y Recall↑	Y F1↑	Bal. Acc↑
gemini-2.5-flash	94.7	1039.0	85.0	18.0	0.3	13.0	0.9069	0.8378	0.6496	0.8758
gemini-3-flash	107.3	1056.3	57.7	5.7	0.0	23.0	0.9309	**0.9499**	0.7748	**0.9395**
gemini-3.1-pro	97.3	917.3	4.7	11.0	4.7	215.0	0.8117	0.8614	**0.9047**	0.8341
qwen-vl-max	87.3	951.7	59.7	21.3	4.3	125.7	0.8312	0.7729	0.6727	0.8049
gpt-5-mini	73.3	981.3	71.3	39.0	0.7	84.3	0.8437	0.6490	0.5702	0.7560
qwen3.5-plus	101.7	1077.3	12.7	11.0	0.3	47.0	**0.9432**	0.8997	0.8944	0.9236
gpt-4.1-mini	84.7	841.3	116.0	26.0	2.3	179.7	0.7408	0.7493	0.5402	0.7446
qwen3-vl-plus	70.0	730.0	322.3	42.3	0.7	84.7	0.6400	0.6195	0.2783	0.6308
gpt-5.4-mini	106.3	847.3	150.3	6.3	0.3	139.3	0.7629	0.9410	0.5764	0.8431
gpt-5.4	106.7	872.3	28.3	5.7	0.7	236.3	0.7832	0.9440	0.8603	0.8556
gpt-5	20.3	767.0	35.7	63.7	29.0	334.3	0.6299	0.1799	0.2413	0.4273
qwen-vl-plus	96.0	239.0	561.7	11.3	5.7	336.3	0.2680	0.8496	0.2491	0.5299

concerns associated with external model providers. Any deployment of clinical document extraction in real workflows would require institution-approved data-use agreements, auditability, access control, and safeguards appropriate for the local regulatory setting.

The benchmark is intended for evaluation of document-ingestion reliability, not for clinical decision making. Model outputs should be treated as candidate structured records that require validation, especially for high-impact fields used in cohort definition, risk modeling, or downstream clinical operations.

H. Evaluator Normalization Rationale

The postprocessing switch is intended to separate semantic extraction errors from format-only mismatches. In the structured JSON task, a single unstable key can cause many checkbox values to be scored as wrong, even when the model recovered the intended field. This is common for bilingual forms: a ground-truth key may contain both English and Chinese labels, while a model may output only the English label, only the Chinese label, or a shortened variant. The manually curated alias map and fixed fuzzy-matching threshold therefore reduce cascading key mismatches rather than repairing clinical content.

We report both postprocessing ON and OFF results to make this evaluator dependence explicit. The OFF setting reflects strict schema compliance, whereas the ON setting approximates practical normalization that would usually be applied before downstream tabular use. The gap between the two settings should therefore be interpreted as evaluator sensitivity, not as a single definitive leaderboard score.

I. Task Details

All models were evaluated between 19/4/2026 and 21/4/2026 using the provider API versions available at that time. We report the exact model IDs in the main tables. For each benchmark run, we additionally archive runtime metadata including UTC timestamp, model ID, SDK version, and API base URL.

Prompting protocol. Each sample is evaluated as a single-image, single-turn request with one task-specific prompt (Appendix A). We do not use multi-turn clarification, self-correction, reranking, or majority-vote decoding within a run, because these strategies introduce additional design choices that can obscure whether failures come from visual recognition, field binding, or output repair. The prompts are used mainly to specify the output schema and fair-evaluation constraints rather than to optimize model-specific performance. Each model is still queried three times per sample, and metrics are averaged across runs to reduce response variance.

Run averaging. Each model is queried three times per sample. Metrics are first computed for each run independently and then averaged across runs to obtain the reported model-level scores. If a transient API failure occurs (e.g., timeout, connection issue, or rate limit), that sample is not recorded as a completed prediction in that run and is retried in a later

resumed run; we therefore average only over completed runs stored in the released artifacts.

Image handling. Inference is run on one image at a time. We do not apply task-specific cropping, manual rotation, denoising, or template-dependent image editing. Images are passed to provider APIs in their original file format, except for the OpenAI pipeline, which applies automatic EXIF orientation correction and resizes the image so that its longest side is at most 1600 pixels before JPEG re-encoding for upload.

Decoding and API settings. We use the providers' default decoding behavior unless otherwise required by the API. In the OpenAI pipeline, we set a maximum output budget of 2048 tokens and, for GPT-5-family models, a low reasoning setting to stabilize OCR-style outputs. We do not fix a generation seed because the evaluated commercial APIs do not consistently expose one across providers.

JSON parsing and invalid outputs. For the structured extraction task, the evaluator expects a JSON object. If the model output is wrapped in a markdown `json` code block, we extract the fenced content before parsing. We then apply standard JSON parsing only; outputs are not repaired by heuristic post-hoc correction. If parsing fails, the prediction is treated as empty for scoring, which typically yields no credit on the structured fields.

Postprocessing OFF setting. The `no-postprocess` flag disables evaluator-side normalization during scoring, but it does not disable output parsing. In particular, JSON parsing is still applied in the structured extraction task; the ablation isolates normalization and matching decisions rather than raw transport formatting alone.