MedDocBench: A Benchmark for Medical Document Visual Question Answering

J Snehan✉ and Kunal Singh

Fractal Analytics
snehan2k2@gmail.com

Abstract. Large vision-language models (LVLMs) have remarkably improved document understanding capabilities, enabling the handling of complex layouts, longer contexts and a wider range of tasks. However, existing document understanding benchmarks are largely confined to general domains, cover only shallow single-page extraction and rarely reflect real-world clinical use, where answering a question often requires reasoning across many pages of a report. To close this gap, we introduce MedDocBench, a benchmark for medical document understanding built around multi-page clinical reports. It is organized into four primary task types, single-page understanding, single-page reasoning, cross-page reasoning, and answerability, which are further divided into 20 subtypes based on the task and the answer evidence. Furthermore, we develop a semi-automated construction pipeline with four stages, namely grounded QA synthesis, automated quality verification, model-based difficulty filtering, and expert human validation. We construct 2,484 high-quality question-answer pairs across medical documents (discharge summaries, lab reports, and prescriptions). Subsequently, we conduct comprehensive evaluation experiments on both open-source and closed-source models across different configurations. We further present an error analysis that identifies failure modes specific to medical documents such as reference-range interpretation, dosing-shorthand decoding and abstention on unanswerable clinical questions.

Keywords: Medical document VQA · Vision-language models · Medical document understanding · Benchmark · Multi-page reasoning

1 Introduction

The rise of LLMs and LVLMs has rapidly expanded what automated systems can do with documents. Modern models parse complex document elements, reason over their content and handle much longer context lengths than before. This has moved the field beyond traditional OCR towards document understanding and question answering. Despite these advances in model capability, the evaluation of complex document tasks has not kept pace. The gap is especially wide in the medical field, where reports are dense and multi-page and answering a question often depends on reasoning over clinical values rather than just locating a piece of text.

Table 1. Comparison of MedDocBench with existing document VQA benchmarks along element diversity, document length, task type, and domain.

Benchmark	Mixed Elements	Multi-page	Reasoning	Medical
DocVQA [18]	✓	✗	✗	✗
InfographicVQA [17]	✓	✗	✓	✗
ChartQA [16]	✗	✗	✓	✗
MP-DocVQA [28]	✓	✓	✗	✗
DUDE [14]	✓	✓	✓	✗
MMDocBench [35]	✓	✗	✓	✗
LongDocURL [7]	✓	✓	✓	✗
MMLongBench-Doc [15]	✓	✓	✓	✗
MedRepBench [27]	✓	✗	✗	✓
MedDocBench (ours)	✓	✓	✓	✓

A few recent efforts target the medical domain. MedS-Bench and its instruction set MedS-Ins [31] evaluate and train medical LLMs on clinical tasks such as report summarization and diagnosis, but work from text rather than document images. PubMed-OCR [11] provides large-scale OCR annotations for scientific medical articles to support layout-aware and OCR-dependent modeling. Closest to our setting are MeDocVL [30] and MedRepBench [27], both on Chinese medical documents. MeDocVL is a vision-language model for parsing medical invoices, while MedRepBench is a benchmark for structured field extraction from single-page report images.

Existing document benchmarks mainly remain narrow in three ways. They often target a single element type, such as tables or charts alone, rather than the mixed content of a real report. They tend to use short, often single-page inputs and to focus on simple question answering that only tests direct extraction. MedDocBench is built to close these gaps. Its documents mix richer elements, including paragraphs, tables, figures, and varied layouts, and run to multiple pages. Its questions call for reasoning over the content, including across pages, rather than extraction alone. Table 1 summarizes how MedDocBench compares with existing document VQA benchmarks.

The documents in MedDocBench are deliberately diverse, to mirror real clinical practice and to stress-test document understanding. The 100 reports (280 pages) span three types, 35 prescriptions, 15 discharge summaries, and 50 diagnostic reports covering blood counts, blood sugar, lipid profiles, liver and thyroid panels, vitamin assays, and dengue or malaria tests. The emphasis is on diversity rather than scale. They are sourced from private clinics, government hospitals, private specialty hospitals, and rural community health centers across ten Indian states, and span pediatric, adult, and senior patients across specialties including general medicine, pediatrics, dermatology, cardiology, respiratory medicine, orthopedics, ENT, and gynecology, with caps on documents per doctor and per center to limit source bias. Importantly for a vision benchmark, the collection mixes scanned pages, phone photographs, and born-digital PDFs, and spans

single-column, multi-column, and table-heavy layouts. This breadth, especially in layout and capture quality, packs many real-world variations into a single benchmark and makes MedDocBench a demanding testbed. All documents are fully de-identified and use synthetic patient information where required.

MedDocBench is organized around four task types, single-page understanding, single-page reasoning, cross-page reasoning and answerability, which are further split into 20 subtypes. We construct it with a four-stage semi-automated pipeline, grounded QA synthesis, automated quality verification, model-based difficulty filtering and expert human validation. In summary, our contributions are:

- We introduce MedDocBench, a medical document VQA benchmark built on diverse, multi-page clinical reports, spanning understanding, single- and cross-page reasoning, and answerability across 20 subtypes.
- We propose a four-stage semi-automated construction pipeline that uses complementary models for synthesis, verification, and difficulty filtering, followed by expert human validation, yielding 2,484 high-quality QA pairs.
- We evaluate a range of open and closed-source models and present an error analysis that identifies failure modes specific to medical documents, such as reference-range interpretation, dosing-shorthand decoding, and abstention on unanswerable clinical questions.

2 Methodology

MedDocBench is organized around four types of questions, each targeting a different capability needed to read real medical reports. **Single-page understanding** questions (858) test direct extraction, such as reading a labeled field or a value from a table, which reflects the everyday need to pull specific facts from a record. **Single-page reasoning** questions (1,261) go a step further and require inference over a page, such as comparing two values, counting, or checking a value against a printed reference range, since clinical use often depends on interpreting numbers rather than just locating them. **Cross-page reasoning** questions (146) require combining information from two or more pages, mirroring how findings, trends, and events in a real report are spread across the document. **Answerability** questions (219) ask about facts that are not present in the report, testing whether a model recognizes what it cannot answer instead of guessing, which matters when a wrong answer can be harmful in a clinical setting. In total, the benchmark contains 2,484 questions. Representative examples of each question type are provided in Section S1 of the Supplementary Material.

We build the benchmark with a four-stage semi-automated pipeline (Figure 1) that turns these reports into a human-validated set of questions, described in the following subsections. The full generation and verification prompts are listed in Section S6 of the Supplementary Material. For evaluation, a model is shown the page image (or all page images, for cross-page and answerability questions) and asked to give only the answer, which may be a single value or a list. Each answer is then compared to the gold answer using ANLS [5], a soft

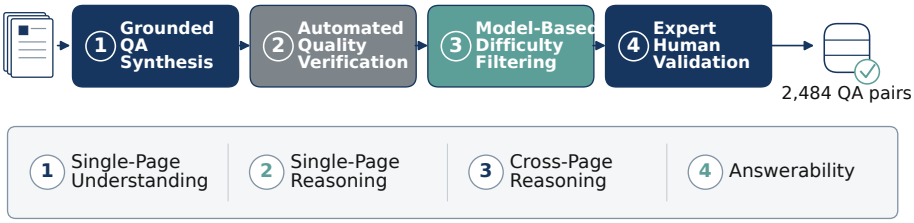

Fig. 1. Overview of the semi-automated data construction pipeline.

string-match score that gives partial credit for near-matches. The task-specific evaluation prompts are given in Section S5 of the Supplementary Material.

2.1 Grounded QA Synthesis

Each generator prompts a vision LLM with the page image, treated as the authoritative source, together with its Surya OCR [23] transcription as a reading aid. We use Surya for its strong OCR accuracy (83% on olmOCR-Bench [24]) and also because it is open-source. A system prompt defines the task, enumerates the allowed question sub-types and states the grounding rules, while the model returns a fixed Pydantic schema so every item carries a question, an answer, an answer type, a grounding field, and a difficulty label. Answers are constrained to be atomic and verifiable by ANLS, and each generator emits at most one question per applicable sub-type, preferring to skip a sub-type rather than fabricate content. Four task generators are used. Each sub-type mirrors a recurring way a clinician interprets or cross-references information when reading these reports.

Single-page understanding spans five sub-types: *key-value extraction* (e.g. an admission date), *table-cell extraction* and *table-row lookup* (a cell by row and column), *list extraction* (all items of an enumerable group), and *multi-field extraction* (two related values). Answers are copied verbatim, preserving casing, units, and OCR errors, and each item stores a verbatim evidence snippet that must contain every answer value.

Single-page reasoning spans seven sub-types: *multi-hop* (chaining several values, e.g. summing the components of a GCS score), *comparison*, *counting*, *out-of-range checks* against a printed reference interval, *conditional filtering*, *indirect reference*, and *negation*. The answer is derived rather than copied, so instead of an evidence snippet each item stores a step-by-step reasoning trace that cites the exact page values used.

Cross-page reasoning feeds the whole report to the model with each page delimited by a marker, and labels each question with one of eight cross-page link types (*temporal trend*, *aggregation*, *find-then-fetch*, *range transfer*, *ranking*, *event effect*, *spanning filter*, *summary-detail*). Every item records the source pages it draws on, and a post-hoc check discards any question that does not genuinely span at least two pages.

Answerability is built in two steps. A generator first writes plausible questions whose required fact is absent, each with the fixed gold answer `N/A`. A second verifier pass then re-reads the document and attempts each candidate, dropping any question it can actually answer so the remaining items are reliably unanswerable. The QA pairs reported in this benchmark were generated with GPT-5 [20].

2.2 Automated Quality Verification

The synthesized pairs are next checked by an automated verifier that keeps or drops each item on three criteria, namely grounding, answer shape, and task label. We describe the full procedure in Section S3 of the Supplementary Material.

2.3 Model-Based Difficulty Filtering

The questions that pass verification are next filtered to remove ones that are too easy to be useful. We answer each question with four small vision models: Qwen2.5-VL-3B [4], Phi-3.5-4B [1], Pixtral-12B [2], and InternVL2-26B [6], each given only the page image (or all page images for a cross-page question) and the question. Every answer is scored against the gold answer with ANLS, a soft string-match score in $[0, 1]$ where 1 means an exact match. A question is dropped only when all four models score an ANLS of 1. We use four models from different families and deliberately pick older, smaller ones, so that a question must be easy for a diverse set of weak models before we discard it.

2.4 Expert Human Validation

In the final stage, the items that survive difficulty filtering undergo expert human validation. Each annotator works from the source page images together with the question, its gold answer, and the automated verifier's reason, records a binary correctness judgment, and for every item judged incorrect fills two fields, a revised answer and the reasoning behind it, so that flawed items are corrected rather than discarded. Each item is reviewed independently by multiple annotators, and disagreements are resolved by majority vote, with a senior annotator adjudicating the remaining ties. Annotation is carried out by 10 postgraduate-level medical professionals.

3 Experiments

We evaluate four families of systems on MedDocBench, ordered by mean ANLS within each group: proprietary and open-source general-purpose VLMs, medical-specialty VLMs, and an OCR+LLM pipeline that answers from the Surya transcription rather than the page image. Table 2 reports ANLS on the three answerable task types (understanding, single-page reasoning, and cross-page reasoning) and answerability, averaged over two evaluation runs. Since the synthesis and verification models also appear in the evaluated panel, we examine the fairness of this overlap in Section S4 of the Supplementary Material.

3.1 Results

Frontier models excel at direct extraction, with the newer frontier models all scoring above 95 on single-page understanding and GPT-5 performing best at 96.7. They are similarly strong on reasoning, where Gemini 3.1 Pro leads with a 95.9 average, and the GPT-5 series improves clearly over the GPT-4 series, reflecting better visual reasoning. **Answerability is critical in the medical domain and remains an open problem**, and it is where these models are weakest, with Grok-4.3 attaining the highest score at 94.5. Even the strongest models retain a non-trivial error margin, which is consequential for clinical automation, where an incorrect answer can cause harm, and this gap indicates that the task is far from solved.

Among open-source models, Gemma4-31B achieves the strongest score at 94.4 on understanding and the highest abstention accuracy at 85.8. On reasoning, Qwen3.5-35B-A3B scores best with an average of 90.6, while most other open-source models fall away. The larger 30B-class open models reach a mean score of 82.8, well above the 67.4 of those under 10B. However, Qwen3.5-4B stays solid at an 83.5 mean score and surpasses the larger Gemma4-31B and Qwen2.5-VL-32B, while Gemma4-E2B collapses on reasoning. Overall, domain tuning offers no edge. The best medical model, Lingshu-7B, reaches a mean score of only 62.5, below the 70.7 of Qwen2.5-VL-7B, the model it is fine-tuned from, making it clear that **medical tuning does not directly help document understanding**. Document-specialist models are excluded from the comparison, as they are trained to parse and transcribe pages rather than answer questions. mPLUG-DocOwl2 [12] returns verbose, formatting-heavy outputs instead of atomic answers, and dedicated parsing models such as PaddleOCR-VL-1.6 [34] emit page transcriptions and score near zero under ANLS.

The OCR+LLM pipeline behaves differently across tasks. On understanding, passing the page image directly outperforms the OCR input for every model, by about 9 points on average, because the transcription loses the visual cues and layout that the image preserves. On reasoning the effect is more interesting, OCR helps Gemma4-31B, while it hurts the stronger reasoners GPT-5.5 and Qwen3.5-35B-A3B. This shows that models already strong at visual reasoning lose useful information when the image is replaced by text, whereas Gemma4 benefits from the linearized layout that offloads its weaker layout parsing. On answerability, OCR input outperforms the image for every model, by about 8 points on average, because a clean transcription makes the absence of a fact explicit and removes the visual detail that tempts a model to fabricate a plausible value, so it abstains more reliably.

3.2 Error Analysis

On MedDocBench, some errors are the usual document reasoning errors, such as misreading a layout, losing a row, or a slip while counting. Apart from these, we find errors that are tied to clinical content. These errors are consequential, and they need medical knowledge or medical writing conventions to get right, so they

Table 2. Model performance on MedDocBench, reported as ANLS on the answerable tasks, while the answerability column reports abstention accuracy. Within each group, models are ordered by mean ANLS. The best result in each column is in **bold**.

Model	Understanding	Reasoning		Answerability
	Single-page	Single-page	Cross-page	
General-Purpose VLMs (Proprietary)				
Gemini 3.1 Pro [10]	95.1	**96.8**	95.0	88.4
GPT 5 [20]	**96.7**	95.4	**96.4**	84.7
Gemini 2.5 Pro [8]	96.1	95.1	87.4	91.5
GPT 5.5 [22]	95.6	96.0	96.0	82.0
Grok 4.3 [32]	94.8	94.0	91.9	**94.5**
Claude Opus 4.6 [3]	95.0	95.9	91.4	82.2
GPT 4.1 [21]	92.3	83.6	78.7	75.1
GPT 4o [19]	93.4	77.7	75.3	78.5
General-Purpose VLMs (Open-Source)				
Qwen3.5 35B A3B [25]	91.2	92.5	88.6	71.9
Qwen3.5 4B	87.8	84.0	80.0	66.4
Gemma4 31B [9]	94.4	73.4	73.2	85.8
Qwen2.5 VL 32B [4]	85.9	74.4	67.7	60.1
Qwen2.5 VL 7B	94.2	59.2	53.0	56.6
InternVL3.5 8B [29]	84.9	62.4	56.5	36.3
Gemma4 E2B	72.9	35.4	31.8	34.5
mPLUG-DocOwl2 9B [12]	39.4	31.2	27.7	28.3
Medical VLMs				
Lingshu 7B [33]	74.1	58.6	52.8	45.7
HuluMed 4B [13]	62.2	48.6	41.7	29.7
MedGemma 4B [26]	52.7	37.9	33.8	21.7
MedGemma 27B	52.5	31.7	29.4	6.2
OCR + LLMs				
OCR+GPT 5.5	90.4	90.9	91.0	87.4
OCR+Gemma4 31B	90.4	82.2	84.0	88.1
OCR+Qwen3.5 35B A3B	73.3	74.4	68.7	88.4

would not appear in a general document. A model can therefore score well on general document VQA and still make these mistakes. This is the main reason a medical-specific benchmark is needed. We describe five failure modes that are crucial for reading medical documents, detailing three here and the remaining two in Section S2 of the Supplementary Material.

Reference Range Interpretation. A lab result is normal or abnormal only when it is read against its reference interval. In real reports, the interval and the measured value sit right next to each other, sometimes without any label, and the report does not mark which values are out of range. The model has to find both numbers, line them up, and compare them, which general document

tasks rarely test. Models under-detect abnormal values. They report fewer out-of-range results than are actually present, and they miss values that sit just past a limit. Smaller models struggle even more, because reasoning over the interval itself is hard for them, and models also show directional confusion.

Reading Drug Charts and Dosing Shorthand. Discharge charts and prescriptions store drugs in a grid. The drug name is in one column, the dose for each time slot in the others, and the number of days and the instruction alongside, with many cells left blank. The dose and frequency are written in shorthand. Indian and Latin abbreviations are common: OD (once daily), BD (twice a day), TDS (three times a day), HS (at bedtime), and SOS (as needed). Answering a dose or frequency question needs two things at once. The model must find the right cell in the grid, and it must decode the shorthand. This is a joint spatial and medical decoding step, and models fail at it.

Clinical Negation. Clinical text often states a finding as a negative. "Anti-HCV – Non Reactive" means the test was run and came back negative, and "No growth after 48 hours" means the culture is normal. The negative word carries the result, so the model must treat it as a real finding and not as missing information. Models often fail here, and do not count a result such as "Non Reactive" as a negative finding. These negative-phrased results are typical of clinical writing and are rare elsewhere. Getting them wrong inverts or empties the finding.

Two further failure modes, detailed in Section S2 of the Supplementary Material, are tied to specific layouts and question types. On **dense, wide, multi-column panels** models skip rows and undercount. Under **fabrication on false clinical premises**, the most dangerous mode, models answer unanswerable questions with a plausible value instead of abstaining.

4 Conclusion

MedDocBench highlights several points for medical document understanding. Recognizing unanswerable questions is the weakest and most dangerous area, where models fabricate plausible clinical values instead of abstaining, and this remains an open problem. Several failure modes are specific to clinical reading, such as interpreting reference ranges, decoding dosing shorthand, and handling clinical negation, and strong general-document performance does not transfer to them. Above all, even the strongest models still make very costly errors, and in any automation such errors can cause direct harm, so these models require careful human oversight before clinical deployment.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Abdin, M., et al.: Phi-3 technical report: A highly capable language model locally on your phone. arXiv preprint arXiv:2404.14219 (2024)
2. Agrawal, P., et al.: Pixtral 12b. arXiv preprint arXiv:2410.07073 (2024)
3. Anthropic: Claude Opus 4.6. https://www.anthropic.com/claude (2026)
4. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
5. Biten, A.F., Tito, R., Mafla, A., Gomez, L., nol, M.R., Valveny, E., Jawahar, C.V., Karatzas, D.: Scene text visual question answering (2019). https://doi.org/10.1109/ICCV.2019.00439
6. Chen, Z., Wang, W., Tian, H., Ye, S., Gao, Z., Cui, E., Tong, W., Hu, K., Luo, J., Ma, Z., et al.: How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. arXiv preprint arXiv:2404.16821 (2024). https://doi.org/10.1007/s11432-024-4231-5
7. Deng, C., Yuan, J., Bu, P., Wang, P., Li, Z.Z., Xu, J., Li, X.H., Gao, Y., Song, J., Zheng, B., Liu, C.L.: Longdocurl: a comprehensive multimodal long document benchmark integrating understanding, reasoning, and locating (2024). https://doi.org/10.18653/v1/2025.acl-long.57
8. Gemini Team, Google: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
9. Gemma Team, Google DeepMind: Gemma 4. https://ai.google.dev/gemma (2026)
10. Google DeepMind: Gemini 3.1 Pro. https://deepmind.google/models/gemini/ (2026)
11. Heidenreich, H., Getachew, Y., Dinica, O., Elliott, B.: Pubmed-ocr: Pmc open access ocr annotations (2025)
12. Hu, A., Xu, H., Zhang, L., Ye, J., Yan, M., Zhang, J., Jin, Q., Huang, F., Zhou, J.: mplug-docowl2: High-resolution compressing for ocr-free multi-page document understanding. arXiv preprint arXiv:2409.03420 (2024). https://doi.org/10.18653/v1/2025.acl-long.291
13. Jiang, S., Wang, Y., Song, S., Hu, T., Zhou, C., Pu, B., Zhang, Y., Yang, Z., Feng, Y., Zhou, J.T., Hao, J., Chen, Z., Wu, R., Tang, T., Lv, J., Xu, H., Wang, H., Xiao, J., Feng, B., Zhu, F., Li, K., Xie, W., Sun, J., Wu, J., Liu, Z.: Hulu-med: A transparent generalist model towards holistic medical vision-language understanding (2025)
14. Landeghem, J.V., Tito, R., Borchmann, Ł., Pietruszka, M., Józiak, P., Powalski, R., Jurkiewicz, D., Coustaty, M., Ackaert, B., Valveny, E., Blaschko, M., Moens, S., Stanisławek, T.: Document understanding dataset and evaluation (dude) (2023). https://doi.org/10.1109/ICCV51070.2023.01789
15. Ma, Y., Zang, Y., Chen, L., Chen, M., Jiao, Y., Li, X., Lu, X., Liu, Z., Ma, Y., Dong, X., Zhang, P., Pan, L., Jiang, Y.G., Wang, J., Cao, Y., Sun, A.: Mmlongbench-doc: Benchmarking long-context document understanding with visualizations (2024). https://doi.org/10.52202/079017-3041
16. Masry, A., Long, D.X., Tan, J.Q., Joty, S., Hoque, E.: Chartqa: A benchmark for question answering about charts with visual and logical reasoning (2022). https://doi.org/10.18653/v1/2022.findings-acl.177

17. Mathew, M., Bagal, V., Tito, R.P., Karatzas, D., Valveny, E., Jawahar, C.V.: Infographicvqa (2021). https://doi.org/10.1109/WACV51458.2022.00264
18. Mathew, M., Karatzas, D., Jawahar, C.V.: Docvqa: A dataset for vqa on document images (2021). https://doi.org/10.1109/WACV48630.2021.00225
19. OpenAI: GPT-4o system card. arXiv preprint arXiv:2410.21276 (2024)
20. OpenAI: GPT-5. https://openai.com/index/introducing-gpt-5/ (2025)
21. OpenAI: Introducing GPT-4.1 in the API. https://openai.com/index/gpt-4-1/ (2025)
22. OpenAI: GPT-5.5. https://openai.com/ (2026)
23. Paruchuri, V., Team, D.: Surya: A lightweight document ocr and analysis toolkit. https://github.com/datalab-to/surya (2025)
24. Poznanski, J., Borchardt, J., Dunkelberger, J., Huff, R., Lin, D., Rangapur, A., Wilhelm, C., Lo, K., Soldaini, L.: olmOCR: Unlocking trillions of tokens in PDFs with vision language models. https://arxiv.org/abs/2502.18443 (2025)
25. Qwen Team: Qwen3.5. https://github.com/QwenLM/Qwen3 (2026)
26. Sellergren, A., Kazemzadeh, S., Jaroensri, T., Kiraly, A., Traverse, M., Kohlberger, T., Xu, S., Jamil, F., et al.: Medgemma technical report. arXiv preprint arXiv:2507.05201 (2025)
27. Shang, F., Xia, Y., Yang, D., Wang, Y., Yang, B.: Medrepbench: A comprehensive benchmark for medical report interpretation (2025)
28. Tito, R., Karatzas, D., Valveny, E.: Hierarchical multimodal transformers for multi-page docvqa (2023). https://doi.org/10.1016/j.patcog.2023.109834
29. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., et al.: Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265 (2025)
30. Wang, W., Wu, W., Liu, Y., Zhao, Y., Lv, X., Diao, L., Fan, Z., Xie, W., Lin, Z., Shi, D., Huang, L., Xu, K., Li, H.: Medocvl: A visual language model for medical document understanding and parsing (2026)
31. Wu, C., Qiu, P., Liu, J., Gu, H., Li, N., Zhang, Y., Wang, Y., Xie, W.: Towards evaluating and building versatile large language models for medicine (2024). https://doi.org/10.1038/s41746-024-01390-4
32. xAI: Grok 4.3. https://docs.x.ai/developers/models/grok-4.3 (2026)
33. Xu, W., Chan, H.P., Li, L., Aljunied, M., Yuan, R., Wang, J., Xiao, C., Chen, G., Liu, C., Li, Z., et al.: Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. arXiv preprint arXiv:2506.07044 (2025)
34. Zhang, Z., Liu, H., Liang, S., Zhang, Y., Xiang, Y., Liu, J., Sun, T., Lin, M., Zhang, Y., Zhou, C., Gao, T., Cui, C., Liu, Y., Yu, D., Ma, Y.: Paddleocr-vl-1.6: Expanding the frontier of document parsing with under-optimized region refinement and progressive post-training (2026)
35. Zhu, F., Liu, Z., Ng, X.Y., Wu, H., Wang, W., Feng, F., Wang, C., Luan, H., Chua, T.S.: Mmdocbench: Benchmarking large vision-language models for fine-grained visual document understanding (2024). https://doi.org/10.1007/978-981-95-6950-2_6