When Cases Get Rare: A Retrieval Benchmark for Off-Guideline Clinical Question Answering

Doeun Lee [1] **Muge Zhang** [1] **Yi Yu** [1] **Ashish Manne** [2] **Stephen Koesters** [2] **Frank Wen** [3] **Brady Buchanan** [2] **Lynda Villagomez** [2] **Oluwatoba Moninuola** [2] **James Lim** [2] **Kathryn Tobin** [2] **Andrew Srisuwananukorn** [2] **Ping Zhang** [1] **Sachin Kumar** [1]

Abstract

Across medical specialties, clinical practice is anchored in evidence-based guidelines that codify best studied diagnostic and treatment pathways. These pathways routinely fall short for the long tail of real-world care not covered by guidelines. Most medical large language models (LLMs), however, are trained to encode common, guideline-focused medical knowledge in their parameters. To address this gap, we introduce OGCAREBENCH, a free-form retrieval-focused benchmark aimed at evaluating LLMs at answering clinical questions that require going beyond typical guidelines. Extracted from published medical case reports and validated by medical experts, OGCAREBENCH contains long-form clinical questions requiring free-text answers, providing a systematic framework for assessing open-ended medical reasoning in rare, case-based scenarios. This work establishes a foundation for benchmarking and advancing both general-purpose and medical LLMs to produce reliable answers in challenging clinical contexts.

1. Introduction

Large language models are actively being explored in healthcare settings for many use cases with potential to transform clinical decision-making and ultimately enhance patient outcomes (Yan et al., 2024; Abrar et al., 2024; Shool et al., 2025). Realizing this potential requires evaluations that reflect the diversity and complexity of real clinical scenarios. Most current benchmarks, however, test models' recall of medical knowledge through exam-style questions (Ben Abacha & Demner-Fushman, 2019; Krithara et al., 2023), typically in multiple-choice settings (Jin et al., 2019; 2021; Pal et al., 2022; Hendrycks et al., 2021; Zuo et al., 2025). While free-form question-answering datasets exist, they are largely patient-oriented and not designed for clinician-facing decision support (Hosseini et al., 2024; Nguyen et al., 2023; Singhal et al., 2023a; Zhu et al., 2020). Furthermore, evidence grounding is especially crucial in medicine, where clinical guidance evolves rapidly, authoritative references are essential for trust, and patient care often involves rare conditions and atypical presentations. Memorization alone is thus insufficient; models must be able to integrate and synthesize knowledge from external sources to support real-world clinical decision-making.

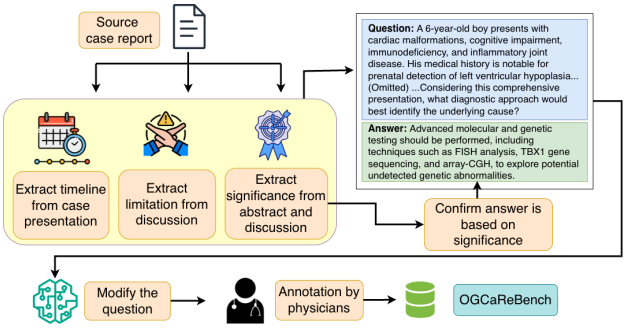

Figure 1. OGCAREBENCH creation pipeline.

We aim to benchmark LLMs in settings that reflect how physicians approach complex clinical problems. Such a benchmark must satisfy three key properties: (1) grounded in real patient cases, (2) free-form question-answering format, and (3) demanding expert-level domain knowledge. Guided by these principles, we focus on simulating scenarios in which physicians must consult external resources to determine appropriate clinical decisions for patients whose cases fall outside standard guidelines or involve rare, off-guideline presentations. For this purpose, we use published

[1]Department of Computer Science and Engineering, Ohio State University, Columbus, OH, USA [2]The Ohio State University Wexner Medical Center, Columbus, OH, USA [3]University of Chicago Medical Center, Chicago, IL, USA. Correspondence to: Doeun Lee <lee.11501@osu.edu>, Sachin Kumar <kumar.1145@osu.edu>.

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).

medical case reports. Case reports document novel, rare, or unprecedented clinical occurrences such as unusual case presentations, atypical diagnostic mechanisms or non-standard treatments. Physicians often consult them when typical guideline references, such as UpToDate (UpToDate, 2025) or standard specialty guidelines, are insufficient to manage complex or unusual cases. For each case report, we apply semi-automatic methods to extract a question and answer pair centered around the significant contribution of the report—which could be a novel diagnosis, novel treatment, or a test associated with a rare occurrence of a disease. We refer to this medical benchmark of **O**ff-**G**uideline **Ca**se **Re**ports as OGCAREBENCH. Our dataset contains cases across 10 medical specialties (see Table 2). All questions and answers are validated by experienced physicians to ensure accuracy and fidelity to real-world clinical reasoning.

Our evaluation of several state-of-the-art general-purpose and medical domain-specific models reveals that LLMs struggle to provide expected responses to rare cases. These results highlight the limitations of relying on parameteric memory of the models alone when handling rare cases, underscoring the necessity of retrieval augmentation in complex medical scenarios. Therefore, we expand our horizon to evaluating performances under retrieval, which is known to enhance the performance of medical question answering (Neha et al., 2025). We create a retrieval corpus of 53,617 case reports covering 12 medical specialties, drawn from publicly available reports. We find providing retrieved documents in the context of the question significantly increases model performance . While proprietary models generally outperform open-source models, they are not perfect. Even with oracle document provided, they continue to make significant number of errors, indicating limitations not only from retrieval but also in correctly integrating evidence into clinical questions.

In summary, our contributions are: (1) Introduce OGCAREBENCH, an expert-validated benchmark derived from published medical case reports, to evaluate language models on realistic rare clinical scenarios. (2) Empirically demonstrate the shortcomings of both medical and general-purpose models in open-ended rare-case reasoning, underscoring the limitations of their standalone use for supporting physicians in real clinical settings. (3) Show that retrieval augmentation enhances performance in expert-level tasks, emphasizing its necessity in building robust systems in the medical domain.

2. OGCAREBENCH: A benchmark of off-guideline medical cases

Medical case reports document novel or rare clinical occurrences. They are typically published to document and highlight unusual conditions, atypical disease courses, unexpected complications, new diagnosis mechanisms, or unique treatment strategies. To better understand how case reports are used in practice by physicians, we first conducted informal interviews with 10 physicians from different US based institutions with specialties ranging from emergency medicine, rheumatology, internal medicine, infectious diseases, oncology, and surgery. While not all domains rely on case reports—fields like infectious diseases or emergency medicine rarely need to consult them—specialties such as surgery, internal medicine, and oncology often turn to case reports when encountering cases that fall outside standard clinical guidelines[1]. This is supported by studies showing that only 55% to 57% of guideline-recommended treatments are implemented in routine practice (McGlynn et al., 2003; Runciman et al., 2012).

To construct a dataset that emphasizes such rare, patient-specific cases, we synthesize OGCAREBENCH from these reports. Starting from all open-access case reports available on PubMed Central (PubMedCentral, 2003), we filter for cases with novel content and persistent rarity. We then use LLMs to extract question-answer pairs detailing clinical procedures, especially highlighting the main topic of case reports. To simulate realistic clinical scenarios beyond the scope of the original reports, we apply controlled modifications to these questions, ensuring they are distinct from the documented cases. Finally, all modified questions undergo physician annotation to validate both accuracy and clinical relevance. We outline the benchmark construction in Figure 1 and detail it in Appendix F. An example of a case report along with the created question-answer pair is provided in Figure 5.

2.1. Evaluation Metric

To evaluate the performance of a model using OGCAREBENCH, we feed the question to the model with an instruction to generate a free-form natural language answer. To evaluate the alignment between the gold answers extracted from the case reports and the responses generated by the models, we use an LLM-as-a-judge to assess equivalence (specifically, we use GPT-5.2). To focus on the main clinical content, we prompt the judge with a few-shot example to output "equivalent" or "mismatch" (see Figure 19 for full prompt). Randomly selected 100 baseline results are validated by internal medicine physicians, yielding an agreement of 93%. See Appendix 3.3 for details about the decision boundary and expert validation.

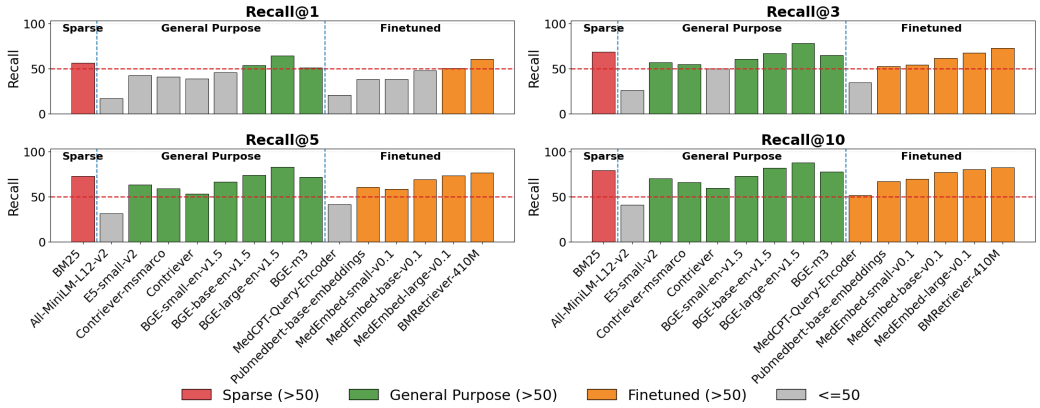

Figure 2. Retrieval results for all retrievers tested, with Recall@k measured as a percentage. Sparse (red), general-purpose (green), and medical finetuned (orange) models were tested. Models with accuracy below 50% are in gray. See Full retrieval result is in Appendix I.2.

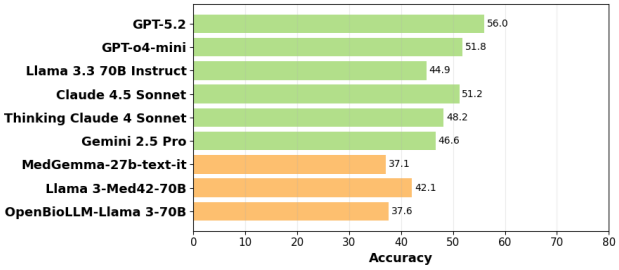

Figure 3. Overall baseline performance with general-purpose models in green and medical models in orange. Subfield-level results are provided in Appendix I.1.

3. Evaluation Setup

We consider two setups, (1) a baseline setup in which a model is expected to rely on its own parametric knowledge, and (2) a setup where we first perform retrieval on our datastore to find the most relevant case reports and provide the retrieved documents to the model's context to generate the answer. For the retrieval augmented generation (RAG) setup, we first validate the performance of different retrieval models and use the top-performing ones for final evaluation.

3.1. Baseline Evaluation

We benchmark six state-of-the-art general-purpose and three medical domain models as baselines. The full list is in Appendix H.1. We prompt the models to answer the question with one best answer and add length-related restrictions to medical models to avoid unusually lengthy output (see Figure 18).

[1]Clinical guidelines set by major societies like American College of Cardiology (ACC/AHA), the American College of Rheumatology (ACR) and the National Comprehensive Cancer Network (NCCN), and others codify large bodies of evidence and are regularly updated by broad expert panels.

3.2. Retrieval Augmented Evaluation

Evaluating Retrieval Methods To identify the most effective retrieval models for our downstream generation task, we evaluate a comprehensive set of 15 models encompassing sparse, general purpose, and biomedical models. To assess performance of retrieval, we report results using Recall@k, MRR, and nDCG with respect to the ground-truth case report (from which the question and answer are derived), which together capture different aspects of retrieval effectiveness. Instruction used for BMRetriever is in Figure 12. We explore various chunking methods and reranking methods to optimize retrieval performance, and ultimately use a maximum length of 512 tokens and a stride of 128. Full method exploration details are in Appendix I.4.

Retrieval Augmented Generation We select the best-performing retrievers (see Figure 2) from each of the three categories—sparse (BM25), general purpose (BGE), and biomedical (BMRetriever)—for retrieval augmented evaluation. Each of the nine LLMs used in baseline experiments is integrated into the pipeline. Deep research models are included with the retrieval corpus as research datastore. We evaluate the model performance using the top 1, 3, and 5 retrieved case reports as context, as well as an oracle setting in which the ground-truth source case report of the question is input. We test a maximum of 3 retrieved reports for OpenbioLLM and Llama 3-Med42 due to a small context window of 8K tokens (see Table 7a for detail).

3.3. Evaluation Metric

Model response is judged as equivalent if the primary clinical procedure recommended matches the procedure specified in the gold answer. Conversely, a response is considered a mismatch if the contents do not overlap or the gold answer appears in the output but not prioritized as the main procedure. Broad or vague recommendations are labeled as

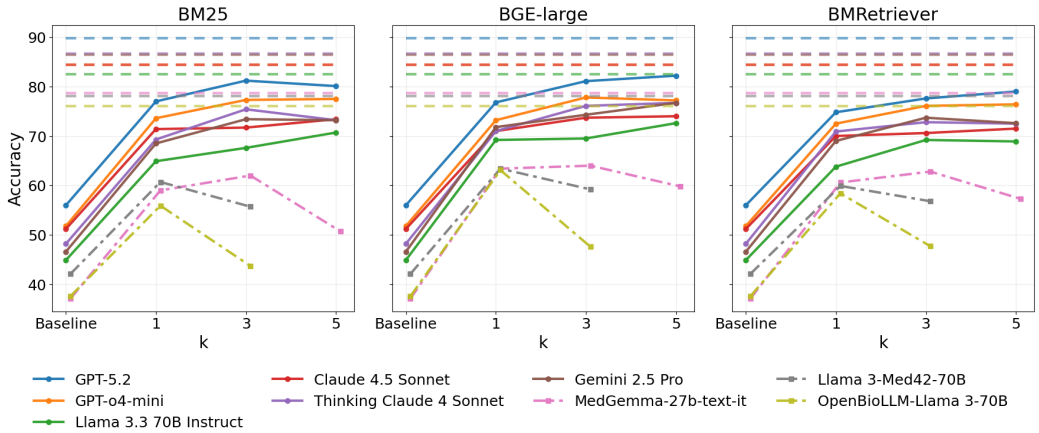

Figure 4. Performance of RAG in % accuracy with different retrieval methods and context lengths. Oracle performances are represented with dashed lines at the top of each graphs. Claude 4.5 Sonnet and GPT-o4-mini has similar oracle accuracy, as well as Thinking Claude 4 Sonnet and Gemini 2.5 Pro. Full RAG result is in Appendix I.3.

mismatches when the gold answer requires a specific clinical procedure, as our benchmark emphasizes detailed, case-based clinical reasoning. Our primary metric is a simple percentage of answers in the benchmark predicted correctly by the model. To validate this LLM-based evaluation, we randomly select 100 baseline results evenly spread across all models and setup to be validated by internal medicine physicians. We task them to label whether the GPT-5.2's evaluation of matching model-generated answers and gold answers reflects true clinical judgment, yielding an agreement of 93%.

4. Results and Analyses

Baselines without retrieval struggle. Figure 3 shows the baseline performance of the base models evaluated with OG-CAREBENCH without retrieval augmentation. Surprisingly, general-purpose models overall outperform medical specialized ones. These results show that both state-of-the-art general-purpose models and models for medical tasks struggle when presented with complex, rare medical questions. Subpar baseline performance suggests that memorization from pretraining alone is insufficient for such cases. The performance of GPT-5.2 and GPT-o4-mini suggests that reasoning offers some advantage in handling rare, case-based scenarios, while Thinking Claude 4 Sonnet doesn't follow this trend. We also speculate that OpenAI models' performance could be due to recent efforts in improving health-related information communication (OpenAI, 2025a;c) which may include training with domain-specific data.

Retrieving the right document is hard for complex medical queries. Figure 2 presents the retrieval performance of state-of-the-art retrievers. The results demonstrate that OG-CAREBENCH is a challenging retrieval benchmark, with most models achieving Recall@1 below 50%. For RAG systems, this indicates that the retrieved context is likely to miss crucial information more than half of the time, thereby reducing the quality and accuracy of generated answers. Although Recall@k approaches 100% at very high values of k (100–1,000), providing such a large number of documents as context to LLMs is impractical. Full retrieval result with other metrics are in Appendix I.2.

Retrieval augmentation improves results for large context models, but gap remains in others. GPT-5.2 and GPT-o4-mini are two models with the highest performance across all three context sizes with performance reaching up to ∼82%. With a context window of 5 reports (retrieved with BGE-large), GPT-5.2 even surpasses the oracle report performance of specialized medical models. Among medical models, MedGemma exhibits the most notable improvement; Llama 3-Med42 performs comparably. Deep research models exhibit substantially lower performance compared to conventional RAG with DR Tulu performing the worst at 19.9%, introducing a gap in general purpose agent's ability in domain specific tasks. Overall, we find three important aspects that affect model performance:

(1) **Retrieval performance is a critical factor influencing RAG performance.** Comparing the language models across retrievers, using BGE-large yields the highest accuracy in most models regardless of performance level, reflecting its highest retrieval quality. (2) **Context window size, as well as the number of documents in context affects the performance.** Models with limited context capacity (Llama 3-Med42, OpenBioLLM) of 8K tokens exhibit the lowest gains when using three case reports as context. In contrast, MedGemma shows significant improvements with RAG due to its large context window, even though it has the lowest baseline performance. (3) Finally, **model's reasoning ability effects RAG performance.** GPT-5.2 and GPT-o4-

mini achieve the best result across all retrievers and context lengths, followed by Thinking Claude 4 Sonnet and Gemini 2.5 Pro most of the time. This underscores the importance of reasoning in consulting external sources. Overall, RAG improves the performance of the models and thus proves essential for rare-case reasoning, transforming subpar baseline performance into clinically significant results. Full RAG and deep research results are in Appendix I.3.

Perfect retrieval does not close the oracle gap. The gains from RAG show that access to relevant case reports is important for rare case reasoning. However, the oracle results also show that retrieval is not the only bottleneck. Even when the oracle case is provided directly as a reference, the model still makes a significant number of errors. To better understand these particular failures, we define a number of failure modes as shown in Table 14 of Appendix K.1 and use an LLM judge (with manual spot checks) to classify each mismatch into one primary mode and one optional secondary mode.

We observe two major patterns as shown in Appendix K.1. First, across general purpose frontier models, the dominant failure mode is model performing a different task than what the question requires, although it may understand the oracle article (i.e., **objective misalignment**). This accounts for more than 40% of errors for Claude 4.5 Sonnet, Gemini 2.5 Pro, GPT-5.2, and Thinking Claude 4 Sonnet. Rather than producing the oracle-specified action, models bundle multiple interventions, append subsequent therapeutic or procedural steps beyond what the article designates, or add unprompted concurrent interventions. Second, models failing to extract, preserve, or utilize the article-supported answer (i.e., **document grounding**) and model omitting a required qualifier, condition, or component (i.e., **constraint/qualifier erosion**) are substantial across all models. Document grounding accounts for 35.3% of primary errors in MedGemma and 28.6% in Llama 3-Med42, while OpenBioLLM shows particularly high rates of constraint/qualifier erosion and granularity mismatch . The distribution also varies by model family. Frontier models show a higher tendency to generalize beyond the article's canonical answers, while specialized models show weaker extraction of article specific decisions and poorer preservation of nuanced answer constraints. These findings show that improving retrieval alone is unlikely to eliminate all errors without improving grounded decision fidelity.

5. Conclusion

Our work argues that reliable medical LLMs must move beyond memorization and towards benchmarks that reflect real-world clinical reasoning. OGCAREBENCH highlights rare, case-based scenarios where current models fall short. Retrieval augmentation fills this gap by curating the cases to focus on, exhibited by significantly enhanced performance. Together, OGCAREBENCH shows retrieval as a crucial component for building clinically reliable LLMs and establishes a new benchmark for supporting physicians when faced with uncommon clinical cases. Retrieval performance, context window, number of documents used as context, and the reasoning ability all play essential roles when it comes to RAG. We hope this benchmark expands the field of open-ended rare-case reasoning in the medical domain and thereby supports physicians.

References

Abrar, M., Sermet, Y., and Demir, I. An empirical evaluation of large language models on consumer health questions, 2024. URL https://arxiv.org/abs/2501.00208.

Ankit Pal, M. S. Openbiollms: Advancing open-source large language models for healthcare and life sciences. https://huggingface.co/aaditya/OpenBioLLM-Llama3-70B, 2024.

Anthropic. Introducing claude 4: Opus 4 and sonnet 4. https://www.anthropic.com/news/claude-4, May 2025a.

Anthropic. Introducing claude sonnet 4.5. https://www.anthropic.com/news/claude-sonnet-4-5, September 2025b.

Balachandran, A. Medembed: Medical-focused embedding models, 2024. URL https://github.com/abhinand5/MedEmbed.

Ben Abacha, A. and Demner-Fushman, D. A question-entailment approach to question answering. BMC Bioinformatics, 20(1), October 2019. ISSN 1471-2105. doi: 10.1186/s12859-019-3119-4. URL http://dx.doi.org/10.1186/s12859-019-3119-4.

Bolton, E., Venigalla, A., Yasunaga, M., Hall, D., Xiong, B., Lee, T., Daneshjou, R., Frankle, J., Liang, P., Carbin, M., and Manning, C. D. Biomedlm: A 2.7b parameter language model trained on biomedical text, 2024. URL https://arxiv.org/abs/2403.18421.

Chen, H., Zhao, Z., Zhou, S., and et al. Rarearena: a comprehensive benchmark dataset unveiling the potential of large language models in rare disease diagnosis. The Lancet Digital Health, 8, 2026.

Chen, S., Moreira, P., Xiao, Y., Schmidgall, S., Warner, J., Aerts, H., Hartvigsen, T., Gallifant, J., and Bitterman, D. S. Medbrowsecomp: Benchmarking medical deep research and computer use, 2025. URL https://arxiv.org/abs/2505.14963.

Chen, Z., Hernández-Cano, A., Romanou, A., Bonnet, A., Matoba, K., Salvi, F., Pagliardini, M., Fan, S., Köpf, A., Mohtashami, A., Sallinen, A., Sakhaeirad, A., Swamy, V., Krawczuk, I., Bayazit, D., Marmet, A., Montariol, S., Hartley, M.-A., Jaggi, M., and Bosselut, A. Meditron-70b: Scaling medical pretraining for large language models, 2023.

Christophe, C., Kanithi, P. K., Raha, T., Khan, S., and Pimentel, M. A. Med42-v2: A suite of clinical llms, 2024. URL https://arxiv.org/abs/2408.06142.

Dai, Z. and Callan, J. Deeper text understanding for ir with contextual neural language modeling. In Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '19, pp. 985–988. ACM, July 2019. doi: 10.1145/3331184.3331303. URL http://dx.doi.org/10.1145/3331184.3331303.

Dong, X., Zhu, W., Wang, H., Chen, X., Qiu, P., Yin, R., Su, Y., and Wang, Y. Talk before you retrieve: Agent-led discussions for better rag in medical qa. ArXiv, abs/2504.21252, 2025. URL https://api.semanticscholar.org/CorpusID:278208163.

Dorfner, F. J., Dada, A., Busch, F., Makowski, M. R., Han, T., Truhn, D., Kleesiek, J., Sushil, M., Lammert, J., Adams, L. C., and Bressem, K. K. Biomedical large languages models seem not to be superior to generalist models on unseen medical data, 2024. URL https://arxiv.org/abs/2408.13833.

Garc'ia-Ferrero, I., Agerri, R., Salazar, A. A., Cabrio, E., de la Iglesia, I., Lavelli, A., Magnini, B., Molinet, B., Ramirez-Romero, J., Rigau, G., Villa-Gonzalez, J. M., Villata, S., and Zaninello, A. Medmt5: An open-source multilingual text-to-text llm for the medical domain. In International Conference on Language Resources and Evaluation, 2024. URL https://api.semanticscholar.org/CorpusID:269042766.

Google DeepMind. Gemini 2.5: Our most intelligent ai model. https://blog.google/innovation-and-ai/models-and-research/google-deepmind/gemini-model-thinking-updates-march-2025/, March 2025.

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., and Poon, H. Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare, 3(1):1–23, October 2021. ISSN 2637-8051. doi: 10.1145/3458754. URL http://dx.doi.org/10.1145/3458754.

Han, T., Adams, L. C., Papaioannou, J.-M., Grundmann, P., Oberhauser, T., Figueroa, A., Löser, A., Truhn, D., and Bressem, K. K. Medalpaca – an open-source collection of medical conversational ai models and training data, 2025. URL https://arxiv.org/abs/2304.08247.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring massive multitask language understanding, 2021. URL https://arxiv.org/abs/2009.03300.

Hicks, R. S., Trofimov, M., Lim, D., Arora, R. K., Tsimpourlas, F., Bowman, P., Sharman, M., Tong, C., Karthik, K., Dugar, A., Jagadeesh, A., Saab, K., Heidecke, J., Alexander, A., Gross, N., and Singhal, K. Healthbench professional: Evaluating large language models on real clinician chats, 2026. URL https://arxiv.org/abs/2604.27470.

Hosseini, P., Sin, J. M., Ren, B., Thomas, B. G., Nouri, E., Farahanchi, A., and Hassanpour, S. A benchmark for long-form medical question answering, 2024. URL https://arxiv.org/abs/2411.09834.

Hou, A. B., Weller, O., Qin, G., Yang, E., Lawrie, D., Holzenberger, N., Blair-Stanek, A., and Durme, B. V. Clerc: A dataset for legal case retrieval and retrieval-augmented analysis generation, 2024. URL https://arxiv.org/abs/2406.17186.

Izacard, G., Caron, M., Hosseini, L., Riedel, S., Bojanowski, P., Joulin, A., and Grave, E. Unsupervised dense information retrieval with contrastive learning, 2022. URL https://arxiv.org/abs/2112.09118.

Jeong, M., Sohn, J., Sung, M., and Kang, J. Improving medical reasoning through retrieval and self-reflection with retrieval-augmented large language models. Bioinformatics, 40:i119 – i129, 2024. URL https://api.semanticscholar.org/CorpusID:267312134.

Jin, D., Pan, E., Oufattole, N., Weng, W.-H., Fang, H., and Szolovits, P. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. Applied Sciences, 11(14), 2021. ISSN 2076-3417. doi: 10.3390/app11146421. URL https://www.mdpi.com/2076-3417/11/14/6421.

Jin, Q., Dhingra, B., Liu, Z., Cohen, W. W., and Lu, X. Pubmedqa: A dataset for biomedical research question answering, 2019. URL https://arxiv.org/abs/1909.06146.

Jin, Q., Kim, W., Chen, Q., Comeau, D. C., Yeganova, L., Wilbur, W. J., and Lu, Z. Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. Bioinformatics, 39(11), November 2023. ISSN 1367-4811. doi: 10.1093/bioinformatics/btad651. URL http://dx.doi.org/10.1093/bioinformatics/btad651.

Ke, Y. H., Jin, L., Elangovan, K., Abdullah, H. R., Liu, N., Sia, A. T. H., Soh, C. R., Tung, J. Y. M., Ong, J. C. L., Kuo, C.-F., Wu, S.-C., Kovacheva, V. P., and Ting, D. S. W. Retrieval augmented generation for 10 large language models and its generalizability in assessing medical fitness. npj Digital Medicine, 8(1):187, 2025. doi: 10.1038/s41746-025-01519-z. URL https://doi.org/10.1038/s41746-025-01519-z.

Krithara, A., Nentidis, A., Bougiatiotis, K., and Paliouras, G. Bioasq-qa: A manually curated corpus for biomedical question answering. Scientific Data, 10 (1):170, 2023. ISSN 2052-4463. doi: 10.1038/s41597-023-02068-4. URL https://doi.org/10.1038/s41597-023-02068-4.

Lee, J., Bang, J., Yang, S., Shim, K., and Chang, S. Chain-of-rank: Enhancing large language models for domain-specific rag in edge device. In North American Chapter of the Association for Computational Linguistics, 2025. URL https://api.semanticscholar.org/CorpusID:276558046.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., tau Yih, W., Rocktäschel, T., Riedel, S., and Kiela, D. Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021. URL https://arxiv.org/abs/2005.11401.

Li, Y., Li, Z., Zhang, K., Dan, R., Jiang, S., and Zhang, Y. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge, 2023. URL https://arxiv.org/abs/2303.14070.

McGlynn, E. A., Asch, S. M., Adams, J., Keesey, J., Hicks, J., DeCristofaro, A., and Kerr, E. A. The quality of health care delivered to adults in the united states. New England journal of medicine, 348(26):2635–2645, 2003.

Meta AI. Llama 3.3 model card and prompt formats. https://www.llama.com/docs/model-cards-and-prompt-formats/llama3_3/, 2024.

Neha, F., Bhati, D., and Shukla, D. K. Retrieval-augmented generation (rag) in healthcare: A comprehensive review. AI, 6(9), 2025. ISSN 2673-2688. doi: 10.3390/ai6090226. URL https://www.mdpi.com/2673-2688/6/9/226.

Nguyen, V., Karimi, S., Rybinski, M., and Xing, Z. MedRedQA for medical consumer question answering: Dataset, tasks, and neural baselines. In Park, J. C., Arase, Y., Hu, B., Lu, W., Wijaya, D., Purwarianti, A., and Krisnadhi, A. A. (eds.), Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 629–648, Nusa Dua, Bali, November 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.ijcnlp-main.42. URL https://aclanthology.org/2023.ijcnlp-main.42/.

Nori, H., Daswani, M., Kelly, C., Lundberg, S., Ribeiro, M. T., Wilson, M., Liu, X., Sounderajah, V., Carlson, J., Lungren, M. P., Gross, B., Hames, P., Suleyman, M., King, D., and Horvitz, E. Sequential diagnosis with language models, 2025. URL https://arxiv.org/abs/2506.22405.

OpenAI. Introducing gpt-5.2. https://openai.com/index/introducing-gpt-5-2/, 2025a.

OpenAI. Introducing openai o3 and o4-mini. https://openai.com/index/introducing-o3-and-o4-mini/, April 2025b. Published Apr 16, 2025.

OpenAI. Introducing healthbench. https://openai.com/index/healthbench/, 2025c.

OpenAI. o4-mini-deep-research. https://developers.openai.com/api/docs/models/o4-mini-deep-research, 2025d. OpenAI API documentation. Accessed: 2026-03-29.

OpenEvidence. Openevidence. https://www.openevidence.com/, 2024.

Pal, A., Umapathi, L. K., and Sankarasubbu, M. Medmcqa : A large-scale multi-subject multi-choice dataset for medical domain question answering, 2022. URL https://arxiv.org/abs/2203.14371.

PubMedCentral. Pmc open access subset. https://pmc.ncbi.nlm.nih.gov/tools/openftlist/, 2003. Bethesda (MD): National Library of Medicine. 2003– [cited YEAR MONTH DAY].

Qiu, P., Wu, C., Liu, S., Zhao, W., Zhang, Y., Wang, Y., and Xie, W. Quantifying the reasoning abilities of llms on real-world clinical cases. ArXiv, abs/2503.04691, 2025. URL https://api.semanticscholar.org/CorpusID:276812970.

Robertson, S. and Zaragoza, H. The probabilistic relevance framework: Bm25 and beyond. Found. Trends Inf. Retr., 3(4):333–389, April 2009. ISSN 1554-0669. doi: 10.1561/1500000019. URL https://doi.org/10.1561/1500000019.

Runciman, W. B., Hunt, T. D., Hannaford, N. A., Hibbert, P. D., Westbrook, J. I., Coiera, E. W., Day, R. O., Hindmarsh, D. M., McGlynn, E. A., and Braithwaite, J. Caretrack: assessing the appropriateness of health care delivery in australia. Medical Journal of Australia, 197 (2):100–105, 2012.

Sellergren, A., Kazemzadeh, S., Jaroensri, T., Kiraly, A., Traverse, M., Kohlberger, T., Xu, S., Jamil, F., Hughes, C., Lau, C., Chen, J., Mahvar, F., Yatziv, L., Chen, T., Sterling, B., Baby, S. A., Baby, S. M., Lai, J., Schmidgall, S., Yang, L., Chen, K., Bjornsson, P., Reddy, S., Brush, R., Philbrick, K., Asiedu, M., Mezerreg, I., Hu, H., Yang, H., Tiwari, R., Jansen, S., Singh, P., Liu, Y., Azizi, S., Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Riviere, M., Rouillard, L., Mesnard, T., Cideron, G., bastien Grill, J., Ramos, S., Yvinec, E., Casbon, M., Buchatskaya, E., Alayrac, J.-B., Lepikhin, D., Feinberg, V., Borgeaud, S., Andreev, A., Hardin, C., Dadashi, R., Hussenot, L., Joulin, A., Bachem, O., Matias, Y., Chou, K., Hassidim, A., Goel, K., Farabet, C., Barral, J., Warkentin, T., Shlens, J., Fleet, D., Cotruta, V., Sanseviero, O., Martins, G., Kirk, P., Rao, A., Shetty, S., Steiner, D. F., Kirmizibayrak, C., Pilgrim, R., Golden, D., and Yang, L. Medgemma technical report, 2025. URL https://arxiv.org/abs/2507.05201.

Shao, R., Asai, A., Shen, S. Z., Ivison, H., Kishore, V., Zhuo, J., Zhao, X., Park, M., Finlayson, S. G., Sontag, D., Murray, T., Min, S., Dasigi, P., Soldaini, L., Brahman, F., tau Yih, W., Wu, T., Zettlemoyer, L., Kim, Y., Hajishirzi, H., and Koh, P. W. Dr tulu: Reinforcement learning with evolving rubrics for deep research, 2025. URL https://arxiv.org/abs/2511.19399.

Shoham, O. B. and Rappoport, N. Medconceptsqa: Open source medical concepts qa benchmark, 2024. URL https://arxiv.org/abs/2405.07348.

Shool, S., Adimi, S., Saboori Amleshi, R., Bitaraf, E., Golpira, R., and Tara, M. A systematic review of large language model (llm) evaluations in clinical medicine. BMC Medical Informatics and Decision Making, 25(1): 117, 2025.

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Babiker, A., Schärli, N., Chowdhery, A., Mansfield, P., Demner-Fushman, D., Agüera y Arcas, B., Webster, D., Corrado, G. S., Matias, Y., Chou, K., Gottweis, J., Tomasev, N., Liu, Y., Rajkomar, A., Barral, J., Semturs, C., Karthikesalingam, A., and Natarajan, V. Large language models encode clinical knowledge. Nature, 620(7972):172–180, August 2023a. ISSN 1476-4687. doi: 10.1038/s41586-023-06291-2. URL https://doi.org/10.1038/s41586-023-06291-2.

Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Hou, L., Clark, K., Pfohl, S., Cole-Lewis, H., Neal, D., Schaekermann, M., Wang, A., Amin, M., Lachgar, S., Mansfield, P., Prakash, S., Green, B., Dominowska, E., y Arcas, B. A., Tomasev, N., Liu, Y., Wong, R., Semturs, C., Mahdavi, S. S., Barral, J., Webster, D., Corrado, G. S., Matias, Y., Azizi, S., Karthikesalingam, A., and Natarajan, V. Towards expert-level medical question answering with large language models, 2023b. URL https://arxiv.org/abs/2305.09617.

Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., and Gurevych, I. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models, 2021. URL https://arxiv.org/abs/2104.08663.

Toma, A., Lawler, P. R., Ba, J., Krishnan, R. G., Rubin, B. B., and Wang, B. Clinical camel: An open expert-level medical language model with dialogue-based knowledge encoding, 2023. URL https://arxiv.org/abs/2305.12031.

UpToDate. Uptodate: Trusted, evidence-based solutions for modern healthcare. https://www.wolterskluwer.com/en/solutions/uptodate, 2025.

Wang, L., Yang, N., Huang, X., Jiao, B., Yang, L., Jiang, D., Majumder, R., and Wei, F. Text embeddings by weakly-supervised contrastive pre-training, 2024. URL https://arxiv.org/abs/2212.03533.

Wu, C., Lin, W., Zhang, X., Zhang, Y., Wang, Y., and Xie, W. Pmc-llama: Towards building open-source language models for medicine, 2023. URL https://arxiv.org/abs/2304.14454.

Wu, K., Wu, E., Thapa, R., Wei, K., Zhang, A., Suresh, A., Tao, J. J., Sun, M. W., Lozano, A., and Zou, J. Medcasereasoning: Evaluating and learning diagnostic reasoning from clinical case reports, 2025. URL https://arxiv.org/abs/2505.11733.

Xiao, S., Liu, Z., Zhang, P., Muennighoff, N., Lian, D., and Nie, J.-Y. C-pack: Packed resources for general chinese embeddings, 2024. URL https://arxiv.org/abs/2309.07597.

Xiong, G., Jin, Q., Lu, Z., and Zhang, A. Benchmarking retrieval-augmented generation for medicine, 2024. URL https://arxiv.org/abs/2402.13178.

Xu, R., Shi, W., Yu, Y., Zhuang, Y., Zhu, Y., Wang, M. D., Ho, J. C., Zhang, C., and Yang, C. Bmretriever: Tuning large language models as better biomedical text retrievers, 2024. URL https://arxiv.org/abs/2404.18443.

Xu, S., Huang, X., Wei, Z., Pang, L., Shen, H., and Cheng, X. Reverse physician-ai relationship: Full-process clinical diagnosis driven by a large language model, 2025. URL https://arxiv.org/abs/2508.10492.

Yan, L. K. Q., Niu, Q., Li, M., Zhang, Y., Yin, C. H., Fei, C., Peng, B., Bi, Z., Feng, P., Chen, K., Wang, T., Wang, Y., Chen, S., Liu, M., and Liu, J. Large language model benchmarks in medical tasks, 2024. URL https://arxiv.org/abs/2410.21348.

Zhao, W., Wu, C., Fan, Y., and et al. An agentic system for rare disease diagnosis with traceable reasoning. Nature, 651:775–784, 2026. doi: 10.1038/s41586-025-10097-9.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., and Stoica, I. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023. URL https://arxiv.org/abs/2306.05685.

Zheng, L., Guha, N., Arifov, J., Zhang, S., Skreta, M., Manning, C. D., Henderson, P., and Ho, D. E. A reasoning-focused legal retrieval benchmark. In Proceedings of the Symposium on Computer Science and Law on ZZZ, CSLAW '25, pp. 169–193. ACM, March 2025. doi: 10.1145/3709025.3712219. URL http://dx.doi.org/10.1145/3709025.3712219.

Zheng, Q., Sun, Y., Wu, C., Zhao, W., Qiu, P., Yu, Y., Sun, K., Zhang, J., Wang, Y., Zhang, Y., and Xie, W. End-to-end agentic rag system training for traceable diagnostic reasoning, 2026. URL https://arxiv.org/abs/2508.15746.

Zhu, M., Ahuja, A., Juan, D.-C., Wei, W., and Reddy, C. K. Question answering with long multiple-span answers. In Cohn, T., He, Y., and Liu, Y. (eds.), Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 3840–3849, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.342. URL https://aclanthology.org/2020.findings-emnlp.342/.

Zuo, Y., Qu, S., Li, Y., Chen, Z., Zhu, X., Hua, E., Zhang, K., Ding, N., and Zhou, B. Medxpertqa: Benchmarking expert-level medical reasoning and understanding, 2025. URL https://arxiv.org/abs/2501.18362.

Appendix

A. LLM Usage Disclosure

We used LLM during the creation of OGCAREBENCH and during evaluation. OGCAREBENCH was created by extracting information such as case presentation, significant point, and limitations from the source case report using GPT-5.2. Modifying the query was done by Claude Opus 4. The model output of OGCAREBENCH was evaluated using LLM-as-a-judge, with GPT-5.2.

B. Reproducibility Statement

The prompts for dataset construction process in Appendix F are in Appendix L, including significance and timeline extraction, question-answer pair creation, controlled modification, model prompts for evaluation, and evaluating answer and model response equivalence. The full dataset and code will be publicly released upon publication.

C. Limitations

OGCAREBENCH has a few limitations. First, the case reports are published and updated regularly, and it is likely that our filtered subset of 28,219 reports, as well as the cases included in OGCAREBENCH, will no longer satisfy the filter criteria as medical knowledge advances. This might make the rare cases part of the guideline reducing the utility of the benchmark in a few years. Therefore, future efforts may require updating the datasets or developing a dynamic version of OGCAREBENCH to mitigate data depreciation. This issue of benchmark saturation is however not unique to medical domain and has been explored extensively in the literature. It has additional challenges in medicine as expert annotation is required. Second, case reports vary in quality; some recommend the use of a specific product on behalf of a company. While we exclude such instances in OGCAREBENCH, they remain in the retrieval corpus. This may lead to the recommendation of certain products during RAG. Third, certain questions may have more than one clinically appropriate next step, and it is feasible that those are not captured by our answer. Although physician validation reduces this risk, multiple answers may still exist in some instances, given the complex nature of clinical practice. Finally, there are other venues to handle rare cases, such as platforms only accessible by physicians. They are the edge cases that are not covered by using case reports as our sources, and in practice RAG will not be able to solve them, as it is not included in any publicly available retrieval corpus.

D. Disclaimer

OpenBioLLM is included as one of our baselines as it has been included in multiple prior studies (Shoham & Rappoport, 2024; Dorfner et al., 2024). However, the model is released without an accompanying paper, data description, or detailed methodological description. Therefore, its performance should be interpreted with consideration and we include it for completeness and comparability with existing benchmarks.

E. Related Work

Models and Datasets Focused on Medicine Medical question-answering (QA) models have significantly evolved (Shool et al., 2025; Yan et al., 2024). A large portion focuses on and is mostly tested on multiple-choice question answering (Han et al., 2025; Wu et al., 2023; Singhal et al., 2023b; Bolton et al., 2024), often using exam-style benchmarks for evaluation (Shool et al., 2025; Krithara et al., 2023; Jin et al., 2019; 2021; Pal et al., 2022; Hendrycks et al., 2021; Zuo et al., 2025). Models and datasets with free-form answers are often patient-oriented (Li et al., 2023; Hosseini et al., 2024; Nguyen et al., 2023; Singhal et al., 2023a; Zhu et al., 2020) or based on general clinical knowledge (Garc'ia-Ferrero et al., 2024; Bolton et al., 2024; Krithara et al., 2023) rather than case-conscious reasoning. Nevertheless, there is increasing interest in more complex free form QA in medical domain recently (OpenAI, 2025c). Especially, there have been studies focusing on case-based models and datasets (Chen et al., 2026), many of them focusing on reasoning (Xu et al., 2025; Nori et al., 2025). Zhao et al. (2026) presents an agent system with reasoning traces for rare disease diagnosis. Most related to our work are Qiu et al. (2025) and Wu et al. (2025), which also focus on using case reports to construct a benchmark for final diagnosis, clinical reasoning, and treatments. However, majority of these evaluations are done with very limited expert validation which limits their trustworthiness. Our work, while similarly constructed, introduces modifications (see item F) to the case scenario to ensure the cases presented are unprecedented. The full dataset is annotated to confirm the modifications. We

Table 1. Comparison table of popular benchmark datasets in the medical domain that involve expert annotation.

Benchmark	Size (Annotated)	Format	Case Based?	Level
PubMedQA-Test Set (Jin et al., 2019)	500 (500)	Binary	✗	Expert
HealthSearchQA (Singhal et al., 2023a)	3173 (100)	Question Only	✗	Consumer
MedCaseReasoning (Wu et al., 2025)	14489 (100)	Stepwise Diagnosis	✓	Expert
SDBench (Nori et al., 2025)	304 (56)	Stepwise Diagnosis	✓	Expert
RareArena (Chen et al., 2026)	∼50000 (100)	Diagnosis	✓	Expert
HealthBench (OpenAI, 2025c)	5000 (5000)	Long Form	✗	Both
HealthBenchPro (Hicks et al., 2026)	525 (525)	Free Form	✗	Expert
MedBrowseComp (Chen et al., 2025)	∼1100 (∼1100)	Short Answer	✗	Expert
OGCAREBENCH (Ours)	639 (639)	Free Form	✓	Expert

provide a comparison table with other benchmarks involving expert annotation on Table 1. Additionally, we broaden the focus and convey the novelty of the case report, whether it may be diagnosis, treatment, or clinical examinations that are presented in a novel way.

Retrieval augmentation in expert domains Retrieval augmentation generation (RAG) is known to enhance the performance in knowledge-intensive tasks (Lewis et al., 2021), providing a promising foundation for domain-specific reasoning (Lee et al., 2025). Using RAG in areas requiring domain expertise mitigates the limitation of memorization by integrating curated professional context as shown by examples from legal domains (Zheng et al., 2025; Hou et al., 2024). With medicine, prior studies have shown that incorporating RAG enhances the performance in various medical QA, ranging from multiple choice to case-based reasoning (Xiong et al., 2024; Dong et al., 2025; Ke et al., 2025; Chen et al., 2025; Jeong et al., 2024). There is also an increasing interest in agentic models that use retrieval internally such as OpenEvidence (OpenEvidence, 2024) and Deep-DxSearch (Zheng et al., 2026). However, use of RAG targeting various rare-case scenarios and case-based retrieval corpus still remains a gap, and we address this by evaluating rare-case questions using RAG.

F. Dataset Creation

Step 1: Collecting and filtering case reports We start by collecting all open-access medical articles through PubMed Central (PMC) (PubMedCentral, 2003). We then leverage the open-access journal list provided by PMC and filter for case reports based on the venues in which these journals are published, focusing on those known to regularly feature case reports (see Table 15). This results in a total of 53,617 reports. We use this corpus as the data store for retrieval-augmented evaluation. To construct a dataset centered on rare cases, we filter this collection to remove reports which meet any of the following criteria: (1) the case report was published in or before 2022. (2) more than three articles cite the case report. (3) the case report is cited by more than one non-case report article. The rationale for these filters is as follows: recently published case reports retain novelty for a few years; case reports cited fewer than a conservative three times are assumed to represent persisting rarity; and case reports cited multiple times by non-case report articles are excluded, as these follow-up studies are indicators that the case has been further explored and potentially resolved into a standard guideline. This setup may introduce false negatives, as low citation counts may reflect low visibility rather than true persisting rarity. Such cases are later filtered during expert annotation. Based on these criteria, we obtain 28,219 reports.

Step 2: Extracting raw question-answer pairs Among the filtered case reports, we randomly select a subset of 1,500 to constitute the dataset. Using GPT-5.2 (`2025-12-11`), we extract three key elements from each report: (1) the case presentation as a timeline providing a sequence of procedures in patient care, (2) limitations that hinder the application of standard medical procedures, motivating the use of off-guideline approaches, and (3) the most significant contribution of the report, defined as the rationale for its publication. Reports whose significance reflects solely the application of standard procedures to a rare disease, rather than the introduction of a novel intervention, are discarded as the resulting questions and answers will be trivial. The question is then formulated by presenting all procedures preceding the decision point that reflect the significance, asking for the next appropriate step given the limitations. For example, the most significant contribution of a report might be in developing a new diagnostic test for a condition where all standard diagnostics are inconclusive. In this case, the question will include the patient's history up until the point where the new test was performed,

along with unavailable diagnostics, and ask what the next step should be. This procedure simulates a scenario where a physician encounters a similar case and has run out of standard guideline-recommended options. The corresponding answer for this question would be the subsequent significant step which in this example is the new diagnostic test. Finally, we use the LLM to verify that the answer was directly connected to the identified significance, thereby confirming the integrity of the question-answer generation process.

Step 3: Adding distractors to generated questions Questions generated in Step 2 are further modified to increase their realism. As our goal is to simulate a situation where physicians consult case reports as guidance for treating their own patients, it is essential that the questions represent unforeseen scenarios and are presented differently from the original reports from which they were derived. To achieve this, we introduce controlled modifications—referred to as distractions—using Claude 4 Opus. These modifications include altering patient demographics (e.g., age and ethnicity), substituting medical terminology with semantically equivalent expressions, integrating comorbidities that do not affect the original condition, and other related adjustments (see Figure 17 for the full prompt). Importantly, while the questions are modified, the answers are preserved; distractions are applied only to the extent that the clinical plausibility of the case remains intact and the correct answer remains unaffected. To validate this process, three internal medicine physicians are presented with subsets of the original question, modified question, and corresponding answer from an early version of the dataset. Their evaluation confirms the medical coherence of the modifications. Such distrations mirror the challenges physicians face in real-world settings where unrelated comorbidities may exist.

Step 4: Dataset verification by experts We assess the medical validity of the question-answer pairs using annotations from eight physicians in internal medicine or pediatrics. The experts are presented with the modified questions and asked to evaluate them based on the following criteria: (1) the question and answer should be medically aligned, and (2) the question should require domain-specific medical expertise rather than general medical knowledge held by the public. We ask them to rate the pairs on a likert scale of 1 to 5—1 indicating that question-answer pair is not realistic under any circumstances and 5 indicating that the question is realistic and the answer correctly addresses the question. Detailed instructions we provide the physicians are in Figure 20. Only question-answer pairs rated 4 or 5 are retained, yielding 639 instances in the final dataset. We do a bootstrap analysis to confirm that our benchmark shows stable results (see Appendix I.6).

G. Data Statistics

We summarize the dataset statistics split across medical specialties in Table 2. We use the original corpus for all 53617 case reports extracted in Step 1 as our retrieval store. The case reports collected are represented by 12 disciplines. Both the full set and OGCAREBENCH are heavily inclined toward Internal Medicine and Surgical Studies. This imbalance is expected and reflects the specialties where this benchmark is most applicable. For internal medicine, this is due to its overlap with other specialties and it also encompassing a variety of sub-disciplines such as hepato-biliary-pancreatic and vascular medicine. For surgical studies, each case is unique and hence more case reports are written about them. Each question in OGCAREBENCH consists of 1-2 paragraphs, and answers are often 1-2 sentences (length distribution is reported in Table 7a).

Table 2. Distribution of case reports across specialties

Specialty	**All Reports**	**OGCaRe-Bench**
Basic Sciences	4571	25
Dentistry	1678	14
External Health	3394	27
Intensive Care	2699	34
Internal Medicine	15606	284
Neurology	1872	17
OBGYN	1654	11
Oncology	1078	12
Orthopedics	2646	32
Pediatrics	1642	14
Surgical Studies	13216	169
General Medicine	1396	-
Others	2165	-
Total	53617	639

H. Models

H.1. Full List of Models Used

For general-purpose models, we evaluate: GPT-5.2 (OpenAI, 2025a), GPT-o4-mini (OpenAI, 2025b), Llama 3.3 (70B Instruct) (Meta AI, 2024), Claude 4.5 Sonnet (Anthropic, 2025b), Thinking Claude 4 Sonnet (Anthropic, 2025a), and Gemini 2.5 Pro (Google DeepMind, 2025). For models specializing in medical question answering, we evaluate: OpenbioLLM-Llama 3 (70B) (Ankit Pal, 2024), MedGemma (27B-text) (Sellergren et al., 2025), and Llama 3-Med42 (70B) (Christophe et al., 2024). These six general-purpose models and three medical models are used for both baseline and RAG. In addition to this, we evaluate deep research models using retrieval corpus as datastore. DR Tulu(Shao et al., 2025) and GPT-o4-mini-deep-research (OpenAI, 2025d) were tested, with web search enabled for GPT-o4-mini-deep-research. We include these models to attest to the importance of our retrieval store in solving this benchmark. .

H.2. Retrieval Models

For the sparse baseline, we employ BM25 (Robertson & Zaragoza, 2009), a model known for its strong performance across various benchmarks, including BEIR (Thakur et al., 2021). Our general-purpose models include All-MiniLM-L12-v2[2], E5-small-v2 (Wang et al., 2024), Contriever and Contriever-MSMARCO (Izacard et al., 2022), and the BGE family (Xiao et al., 2024), which integrates dense, sparse, and multi-vector strategies. For the biomedical domain, we assess MedCPT (Jin et al., 2023), PubMedBERT (Gu et al., 2021), MedEmbed series (Balachandran, 2024), and BMRetriever(Xu et al., 2024), a medically pre-trained and fine-tuned instruction-following model.

H.3. Additional Models Explored

We considered Meditron (70B) (Chen et al., 2023) and Clinical Camel (70B) (Toma et al., 2023), but we decided unsuitable since their primary focus is multiple-choice question answering and therefore generated responses by always outputting answer choices when given questions from OGCAREBENCH (even though no choices were provided in the question).

I. Ablation Study

I.1. Baseline in Subfields

Table 3. Baseline performance across specialties.

Model	Basic Sci.	Dent.	Ext. Health	Intensive	Int. Med.	Neuro.	OBGYN	Onc.	Ortho.	Peds.	Surgical
GPT-5.2	72.0	57.1	63.0	61.8	56.7	23.5	72.7	58.3	50.0	57.1	53.3
GPT-o4-mini	64.0	57.1	66.7	64.7	53.5	23.5	36.4	33.3	31.3	50.0	50.9
Llama 3.3 70B Instruct	56.0	57.1	37.0	58.8	47.5	17.7	18.2	33.3	40.6	35.7	43.2
Claude 4.5 Sonnet	52.0	28.6	63.0	52.9	54.2	35.3	36.4	33.3	46.9	64.3	49.1
Thinking Claude 4 Sonnet	48.0	35.7	48.2	64.7	52.1	23.5	36.4	33.3	28.1	57.1	46.8
Gemini 2.5 Pro	68.0	42.9	40.7	55.9	50.4	17.7	36.4	41.7	34.4	42.9	43.2
MedGemma-27b-text-it	36.0	35.7	48.2	50.0	38.0	11.8	27.3	25.0	28.1	28.6	37.9
Llama 3-Med42-70B	56.0	42.9	37.0	52.9	44.0	17.7	18.2	41.7	34.4	42.9	40.8
OpenBioLLM-Llama 3-70B	48.0	28.6	25.9	47.1	40.5	17.7	27.3	33.3	25.0	28.6	37.9

Table 3 shows baseline results for all specialties. Most models show better performance than the baseline for basic science and Intensive Care, but struggle to perform well in Neurology and Orthopedics. Between the two representative disciplines, Internal Medicine and Surgical Studies, all models exhibit better accuracy for internal medicine.

[2] https://huggingface.co/sentence-transformers/all-MiniLM-L12-v2.

I.2. Full Retrieval Result

Table 4. Full retrieval results across state of the art retrievers. Values are in percentage. R@k denotes recall at k.

Type	Model	Params	R@1	R@3	R@5	R@10	R@100	R@1000	MRR@5	nDCG@10
Sparse	BM25	N/A	56.0	68.7	72.6	79.3	91.1	95.9	62.6	67.3
General	All-MiniLM-L12-v2	33.4M	16.7	25.7	31.5	40.8	70.6	88.4	21.9	27.3
	E5-small-v2	33.4M	42.6	56.7	63.2	70.3	87.6	96.6	50.2	55.8
	Contriever-msmarco	110M	40.7	54.5	58.8	66.0	85.6	96.2	47.8	52.9
	Contriever	110M	38.5	49.8	53.2	59.2	82.5	95.6	44.4	48.5
	BGE-small-en-v1.5	33.4M	45.4	60.7	66.2	72.8	90.8	98.9	53.6	58.9
	BGE-base-en-v1.5	109M	53.8	67.1	73.9	82.0	92.8	99.4	61.3	67.1
	BGE-large-en-v1.5	335M	64.3	78.1	82.9	87.5	96.4	99.4	71.6	75.9
	BGE-m3	560M	51.0	64.8	71.7	77.6	92.3	98.0	58.6	63.8
Finetuned	MedCPT-Query-Encoder	109M	20.8	34.7	41.2	51.3	83.1	97.3	28.3	34.8
	Pubmedbert-base-embeddings	109M	38.2	52.6	60.3	66.8	87.2	97.5	46.4	51.9
	MedEmbed-small-v0.1	33.4M	37.9	54.0	58.4	69.5	89.4	98.1	46.1	52.8
	MedEmbed-base-v0.1	109M	47.9	61.5	69.2	77.0	94.1	99.5	55.6	61.5
	MedEmbed-large-v0.1	335M	50.4	67.3	73.1	80.4	94.4	99.4	59.2	65.1
	BMRetriever-410M	410M	60.6	72.9	76.5	82.2	95.0	99.1	66.7	70.9

I.3. Full RAG Result

Table 5. Performance of RAG with different retrieval methods and context length. Δ stands for the percentage improvement from baseline. For retrieval with BGE, BGE-large-en-v1.5 was used.

# Reports	Model	BM25	Δ BM25	BGE	Δ BGE	BMRet.	Δ BMRet.	Oracle
1	GPT-5.2	77.0	21.0	76.8	20.8	74.8	18.8	89.8
	GPT-o4-mini	73.6	21.8	73.2	21.4	72.5	20.7	84.4
	Llama 3.3 70B Instruct	64.9	20.0	69.2	24.3	63.8	18.9	82.5
	Claude 4.5 Sonnet	71.4	20.2	71.0	19.9	70.0	18.8	84.4
	Thinking Claude 4 Sonnet	69.3	21.1	71.0	22.8	70.9	22.7	86.7
	Gemini 2.5 Pro	68.5	21.9	71.8	25.2	69.0	22.4	86.5
	MedGemma-27b-text-it	59.0	21.9	63.4	26.3	60.6	23.5	78.7
	Llama 3-Med42-70B	60.7	18.6	63.4	21.3	59.9	17.8	78.1
	OpenBioLLM-Llama 3-70B	55.9	18.3	63.1	25.5	58.4	20.8	76.1
3	GPT-5.2	81.2	25.2	81.1	25.0	77.6	21.6	-
	GPT-o4-mini	77.3	25.5	77.8	26.0	76.1	24.3	-
	Llama 3.3 70B Instruct	67.6	22.7	69.5	24.6	69.2	24.3	-
	Claude 4.5 Sonnet	71.7	20.5	73.7	22.5	70.6	19.4	-
	Thinking Claude 4 Sonnet	75.4	27.2	76.1	27.9	72.8	24.6	-
	Gemini 2.5 Pro	73.4	26.8	74.3	27.7	73.7	27.1	-
	MedGemma-27b-text-it	62.0	24.9	64.0	26.9	62.8	25.7	-
	Llama 3-Med42-70B	55.7	13.6	59.2	17.1	56.8	14.7	-
	OpenBioLLM-Llama 3-70B	43.7	6.1	47.6	10.0	47.7	10.2	-
5	GPT-5.2	80.1	24.1	82.2	26.1	79.0	23.0	-
	GPT-o4-mini	77.5	25.7	77.2	25.4	76.4	24.6	-
	Llama 3.3 70B Instruct	70.7	25.8	72.6	27.7	68.9	23.9	-
	Claude 4.5 Sonnet	73.4	22.2	74.0	22.8	71.5	20.3	-
	Thinking Claude 4 Sonnet	73.2	25.0	76.7	28.5	72.5	24.3	-
	Gemini 2.5 Pro	73.2	26.6	76.7	30.0	72.6	26.0	-
	MedGemma-27b-text-it	50.7	13.6	59.8	22.7	57.3	20.2	-

Table 6. Deep research models' performance.

Model	Accuracy
DR Tulu	19.9
GPT-o4-mini-deep-research	**53.5**

I.4. Retrieval Method Exploration

Given the long lengths of our corpus and queries, along with the context-window limitations summarized in Table 7a and Table 7b, we employ a text processing strategy to optimize document representation. Documents are chunked with a maximum length of 512 tokens and a stride of 128. We then aggregate passage-level scores using a two-level Maximum Passage (MaxP) strategy (Dai & Callan, 2019).

Table 7. Overview of token length (left) and model context limits (right) guide the retrieval processing strategy.

(a) Token statistics for the retrieval corpus and queries and answers from OGCAREBENCH (Computed using Contriever's tokenizer (Izacard et al., 2022)).

Category	Avg. tokens	Max tokens
Corpus	2730.21	24271
Question	480.64	771
Answer	29.74	95

(b) Maximum context lengths of a subset of models we evaluate with RAG.

Model	Context Length
GPT-5.2	400K
GPT-o4-mini	200K
Llama 3.3 70B Instruct	128K
Claude 4.5 Sonnet	1M
Thinking Claude 4 Sonnet	1M
Gemini 2.5 Pro	1M
MedGemma-27b-text-it	128K
Llama 3-Med42-70B	8K
OpenBioLLM-Llama 3-70B	8K

I.4.1. COMPARISON OF CHUNKING METHODS

Table 8. Retrieval results using different chunking methods.

Model	Corpus	Query	Recall@1	Recall@3	Recall@5	Recall@10	Recall@100	Recall@1000	MRR@5	nDCG@10
BGE-large-en-v1.5	truncation	truncation	35.2	50.2	57.9	66.0	87.6	96.9	43.5	49.7
	chunk	truncation	62.9	76.4	81.5	86.7	96.1	**99.4**	70.2	74.7
	chunk	chunk	**64.3**	**78.1**	**82.9**	**87.5**	**96.4**	**99.4**	**71.6**	**75.9**
BMRetriever-410M	truncation	truncation	32.2	48.4	52.9	59.9	82.5	95.1	40.4	45.8
	chunk	truncation	**60.6**	72.8	**76.8**	**82.3**	**95.0**	**99.1**	**66.7**	**71.0**
	chunk	chunk	**60.6**	**72.9**	76.5	82.2	**95.0**	**99.1**	**66.7**	70.9

Table 8 presents the results of our chunking experiments on two top-performing retrievers, BGE-large and BMRetriever. We compared different combinations of chunking and truncation. Our chunking strategy used a maximum length of 512 tokens with a stride of 128, while truncation was a simple cut-off at 512 tokens. The results demonstrate that applying chunking to both the corpus and the query is essential for achieving high performance in our use case.

I.4.2. EFFECTS OF CONTEXT LENGTH

Table 9. Retrieval results using different context lengths.

Model	Max Len	Recall@1	Recall@3	Recall@5	Recall@10	Recall@100	Recall@1000	MRR@5	nDCG@10
BMRetriever-410M	512	**60.6**	**72.9**	**76.5**	**82.2**	**95.0**	**99.1**	**66.7**	**70.9**
	1024	48.2	62.1	67.3	72.0	88.9	97.8	55.6	60.1

To evaluate the impact of context length, we conducted an experiment with BMRetriever, which supports a maximum context length of 1,024 tokens and was tested with a fixed stride of 128. The results, presented in Table 9, indicate that merely increasing the context window does not necessarily yield improved performance—particularly for long, domain-specific medical queries such as those in our dataset.

I.4.3. Effect of stride values

Table 10. Retrieval results using different stride values.

Model	Stride	Recall@1	Recall@3	Recall@5	Recall@10	Recall@100	Recall@1000	MRR@5	nDCG@10
BGE-large-en-v1.5	128	**64.3**	**78.1**	**82.9**	**87.5**	**96.4**	**99.4**	**71.6**	**75.9**
	256	57.6	71.8	76.8	84.2	95.0	**99.4**	64.9	70.3
	384	56.7	69.8	75.3	81.1	95.1	**99.4**	63.7	68.5
	512	50.5	64.3	70.1	77.6	94.1	**99.4**	58.1	63.5
BMRetriever-410M	128	**60.6**	**72.9**	**76.5**	**82.2**	**95.0**	**99.1**	**66.7**	**70.9**
	256	51.8	65.6	70.0	75.9	92.6	98.9	59.0	63.7
	384	48.7	62.0	67.1	74.0	91.7	98.6	55.7	60.8
	512	43.0	56.7	63.5	69.8	89.4	98.0	50.5	55.8

To see the effect of different stride values, we conducted experiments on the two top-performing retrievers, BGE-large and BMRetriever models, with a fixed maximum context length of 512. The results, as detailed in Table 10, revealed that a stride of 128 consistently outperformed other configurations. Consequently, this stride value was selected for all subsequent experiments.

I.4.4. Retrieval Result Under Simple Truncation

Table 11. Retrieval results using simple truncation. Values are in percentage. R@k denotes recall at k.

Type	Model	Params	R@1	R@3	R@5	R@10	R@100	R@1000	MRR@5	nDCG@10
General	All-MiniLM-L12-v2	33.4M	5.8	10.3	12.4	17.4	37.6	70.0	8.2	10.8
	E5-small-v2	33.4M	15.6	26.4	31.6	38.3	66.0	87.9	21.3	26.0
	Contriever-msmarco	110M	15.2	23.5	28.6	34.9	63.7	87.0	20.0	24.2
	Contriever	110M	16.9	23.0	26.4	30.4	53.1	78.6	20.5	23.2
	BGE-small-en-v1.5	33.4M	30.2	45.1	49.5	56.7	82.8	96.1	37.6	42.9
	BGE-base-en-v1.5	109M	30.2	42.6	49.3	58.4	83.1	96.4	37.1	43.1
	BGE-large-en-v1.5	335M	35.2	50.2	57.9	66.0	87.6	96.9	43.5	49.7
	BGE-m3	560M	22.2	32.7	37.6	43.7	68.2	88.3	28.0	32.3
Finetuned	MedCPT-Query-Encoder	109M	18.0	28.2	33.0	42.7	71.8	92.3	23.4	28.9
	Pubmedbert-base-embeddings	109M	24.7	37.6	45.4	53.2	81.5	95.9	32.2	38.0
	MedEmbed-small-v0.1	33.4M	27.4	38.7	44.9	54.0	81.5	96.1	33.7	39.5
	MedEmbed-base-v0.1	109M	29.6	43.0	47.7	56.3	82.0	96.9	36.4	42.0
	MedEmbed-large-v0.1	335M	29.1	44.3	50.4	58.1	84.8	96.7	37.2	42.9
	BMRetriever-410M	410M	32.2	48.4	52.9	59.9	82.5	95.1	40.4	45.8

Table 11 reports retrieval performance under a simple truncation strategy with a maximum context length of 512 tokens for both corpus and queries. As expected, performance is consistently lower than with the chunking strategy, with no model achieving Recall@1 above 50%. This highlights the importance of chunking and reveal substantial room for improvement in modern retrievers, particularly for rare-case retrieval.

I.4.5. Retrieval Results Using Reranker

Table 12. Percentage improvement in retrieval using MedCPT-Cross-Encoder and BGE-reranker-large.

model	MedCPT-Cross-Encoder		BGE-reranker-large	
	Recall@10↑	NDCG@10↑	Recall@10↑	NDCG@10↑
All-MiniLM-L12-v2	61.3	115.4	-8.4	3.5
E5-small-v2	13.8	23.1	-19.2	-25.9
Contriever-msmarco	18.2	26.9	-12.6	-17.9
Contriever	27.8	34.2	-0.8	-6.0
BGE-small-en-v1.5	10.1	14.2	-33.3	-39.7
BGE-base-en-v1.5	2.1	6.7	-35.5	-42.1
BGE-large-en-v1.5	0.2	-3.2	-38.6	-48.8
BGE-m3	11.3	16.5	-30.0	-35.9
MedCPT-Query-Encoder	26.8	48.7	-10.1	-6.5
Pubmedbert-base-embeddings	14.5	25.4	-29.0	-31.4
MedEmbed-small-v0.1	15.5	30.5	-29.1	-32.9
MedEmbed-base-v0.1	6.5	13.5	-40.4	-47.4
MedEmbed-large-v0.1	5.1	9.6	-42.8	-50.3
BMRetriever-410M	7.4	5.2	-36.8	-48.0

As shown in Table 12, incorporating a strong biomedical reranker, MedCPT-Cross-Encoder, consistently improves the performance of most retrieval models. However, this improvement varies greatly according to retriever models, and several larger retriever models do not have a notable difference, suggesting that simply adding a specialized reranker is not always a guaranteed solution. BGE-reranker-large Meanwhile, (Xiao et al., 2024) exhibits a notable decline in performance across all retriever models. This result highlights that even state-of-the-art rerankers struggle to perform effective reranking within the context of OGCareBench.

I.5. LLM-as-a-Judge

Using the Wilson confidence interval, the 95% confidence interval for expert agreement of our LLM-as-a-judge evaluation is 85.6% to 96.9%. This surpasses the agreement levels in MT-Bench(Zheng et al., 2023), an extremely popular benchmark, which has human-human agreement of approximately 81% and LLM-human agreement of above 80%.

I.6. Bootstrap Analysis

Table 13. Bootstrap analysis of two representative setups.

(a) Baseline

Model	Accuracy	Mean	95% CI
GPT-5.2	56.0	55.97	52.27–59.94
GPT-o4-mini	51.8	51.81	47.73–55.71
MedGemma-27b-text-it	37.1	37.09	33.18–41.00
Llama 3-Med42-70B	42.1	42.05	38.34–45.70

(b) RAG (BGE-large, top-3)

Model	Accuracy	Mean	95% CI
GPT-5.2	81.1	81.04	78.09–84.19
GPT-o4-mini	77.8	77.80	74.80–80.91
MedGemma-27b-text-it	64.0	64.01	60.25–67.76
Llama 3-Med42-70B	59.2	59.14	55.24–62.76

We used bootstrap analysis, which is a resampling method of repeatedly drawing samples with replacement to approximate the sampling distribution, to confirm that our benchmark shows stable results. 1,000 bootstrap sampling of both the baseline and BGE retriever with 3 reports as context in Table 13 confirms that our benchmark has a stable performance estimate. The bootstrap mean is nearly identical to the reported accuracy. This shows that our clinically rich benchmark of 639 questions is sufficient for model comparison.

J. Example Data Creation

Bleeding anorectal varices treated by a direct puncture approach through the greater sciatic foramen: The utility of a steerable microcatheter for reverse catheterization
PMCID: PMC8829532 Internal Medicine

I. INTRODUCTION
Anorectal varices represent a developed portosystemic collateral pathway where portal venous blood flows from the superior rectal veins of the inferior mesenteric system to the middle and inferior rectal veins of the iliac system. Anorectal varices most commonly result from portal hypertension . . . (Omitted for brevity).

II. CASE PRESENTATION
The patient was a woman in her 70s with Child-Pugh class B primary biliary cholangitis. She was admitted to our hospital for treatment of intractable anorectal varices. Her gastroesophageal varices were repeatedly treated at another hospital, 5 and 2 years ago, respectively . . . (Omitted for brevity).

III. DISUCSSION
. . . (Omitted for brevity) It can be managed by using a variety of methods, including endoscopic therapy (EIS and EBL), IR approach, and surgical treatment, but no standard treatment has been established yet. In our case, large, dilated varices rendered both endoscopic options unsuitable. EIS was considered dangerous because, given the high blood flow . . . (Omitted for brevity) . . . TIPS and PTO cannot be conducted in cases, like ours, where ascites and portal vein thrombosis are present. Direct puncture access to the superior rectal vein approached through the greater sciatic foramen was pioneered by Kariya et al., who first used this pelvic approach for successful endovascular treatment for rectal varices and reported its efficacy and feasibility. . . . (Omitted for brevity).

QUESTION
A female patient in her eighth decade with moderate hepatic dysfunction secondary to primary biliary cirrhosis presents with life-threatening hemorrhoidal variceal bleeding. . . . (Omitted for brevity) . . . feeding vessel to the perianal varices, along with moderate volume ascitic fluid. Traditional interventional approaches including balloon-assisted venous occlusion (complicated by numerous outflow vessels), transjugular portosystemic shunting and percutaneous liver-directed obliteration (both contraindicated by portal thrombosis and peritoneal fluid) are all excluded. . . . (Omitted for brevity) . . . given the limitations precluding standard endoscopic or conventional radiologic therapies, *what emergent procedural intervention would provide optimal hemorrhage control?*

ANSWER
Proceed with variceal obliteration via direct puncture access to the dilated right superior rectal vein, approaching it through the left greater sciatic foramen.

Figure 5. Example of case report and corresponding final question-answer pair. Timeline is **bolded**, significance is in blue, and limitation is in red. Portions of original contents have been selectively omitted for brevity and irrelevance. The question asks the direct next step given the patient details (*italicized*), and the answer is related to the significant point of the case report. Link for full text: https://pmc.ncbi.nlm.nih.gov/articles/PMC8829532/.

K. Failure Mode

K.1. Taxonomy and Results

Table 14. Taxonomy of failure modes used to classify model errors on OGCareBench (oracle setting). Each case is assigned one primary and at most one secondary failure mode by a judge LLM (GPT-5.1).

Failure mode	Diagnostic question	Failure axis
Document grounding	Did the model fail to extract, preserve, or utilize the article-supported answer?	Source grounding
Objective misalignment	Did the model optimize for a different task than the question requires, although it may understand the oracle article?	Task objective
Granularity mismatch	Did the model answer the right question, but at the wrong level of specificity?	Answer specificity
Context/stage misbinding	Did the model select an action from the wrong stage of the workflow?	Workflow stage
Constraint/qualifier erosion	Did the model omit a required qualifier, condition, or component?	Answer completeness

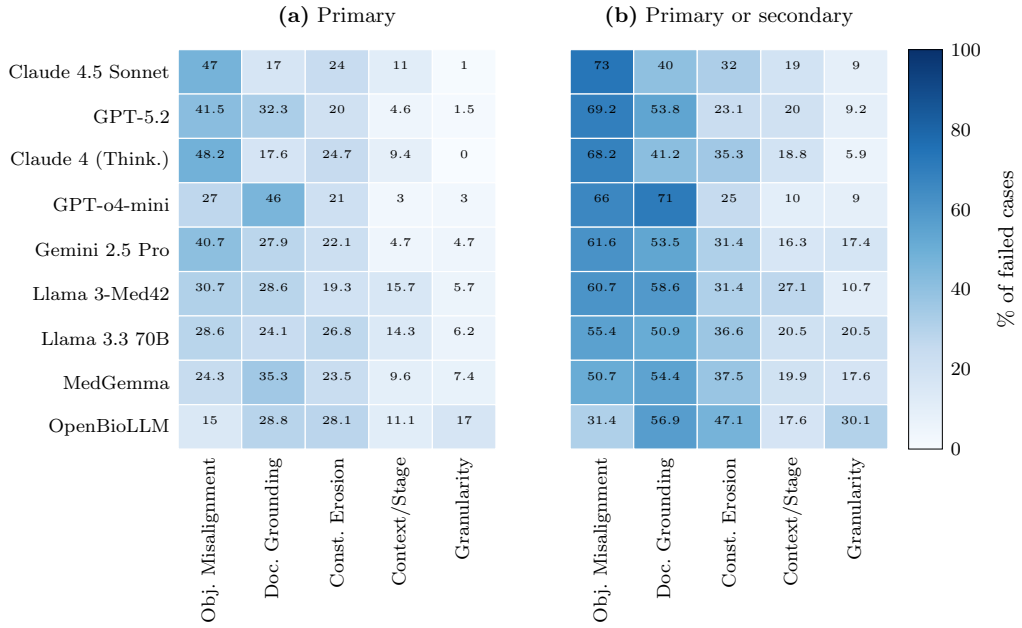

Figure 6. **Failure mode distribution on OGCareBench.** Cells show the percentage of failed cases for which each mode is primary (left) or primary/secondary (right).

K.2. Failure Mode Illustrative Examples

Acute myocardial infarction following radiofrequency catheter ablation in a child
PMCID: PMC11049577 Internal Medicine

Case summary
A pre-adolescent female underwent catheter ablation for drug-refractory AVNRT. Post-procedurally, inferior ST-segment elevations developed; emergency coronary angiography confirmed right coronary occlusion at the AV-nodal artery. Repeated vasodilator administration failed; patency was restored by wire manipulation, and intravascular imaging was performed at that time. The team deferred stenting given the patient's age. With persistent uncertainty about the mechanism of obstruction, and intravascular imaging already completed, further characterization was sought.

Question
... With persistent uncertainty about the pathophysiology despite restored flow, *what diagnostic modality would best characterize the vascular abnormality?*

Oracle answer
Obtain coronary computed tomography angiography (CCTA) for further evaluation.

Model output
Intravascular ultrasound (IVUS) allows detailed visualization of the vessel wall ... clarifying whether the stenosis is due to direct thermal injury or extrinsic compression.

The article specifies CCTA as the next step after intravascular imaging has already been performed. The model recommends IVUS, the modality the case report states was already completed, failing to stay anchored to the document-specific investigative sequence.

Figure 7. **Document grounding failure.** The model substitutes a previously completed diagnostic modality (IVUS) for the article-specified next investigation (CCTA), indicating insufficient grounding to the oracle document's stated clinical course. Single underlines mark content directly informing the oracle answer; wavy underlines mark constraining clinical context the model disregarded.

Late-onset bleeding from an intercostal artery pseudoaneurysm after hemostasis using soft coagulation
PMCID: PMC10899919 Surgical Studies

Case summary
A 68-year-old male who previously underwent right lower lobe lobectomy and a subsequent wedge resection presented emergently with hemoptysis and severe hypoxemia (SpO_2 70%). Contrast-enhanced CT demonstrated an 11 cm right hemothorax with active extravasation from a pseudoaneurysm of the sixth intercostal artery. Three-dimensional vascular reconstruction confirmed the false aneurysm.

Question
... Given active vascular extravasation from an intercostal arterial false aneurysm causing hemodynamic compromise and severe hypoxemia, *what intervention should be pursued urgently?*

Oracle answer
Arrange immediate emergency angiography through the Radiology/Interventional Radiology service.

Model output
The patient should undergo urgent angiography **with embolization** to stop the bleeding from the intercostal arterial false aneurysm.

The oracle specifies the immediate next action: arranging angiography via IR. The model correctly identifies this step but appends embolization, a subsequent therapeutic decision that follows angiographic confirmation, thereby answering a broader task (the full interventional procedure) rather than the designated next action.

Figure 8. **Objective misalignment failure.** The model includes the oracle-specified step but expands its answer to encompass a downstream intervention not designated by the article, reflecting optimization toward clinical completeness rather than oracle fidelity. Single underlines mark content directly informing the oracle answer.

Conversion surgery for advanced jejunal adenocarcinoma with multiple peritoneal metastases
PMCID: PMC10435430 Surgical Studies

Case summary
A 64-year-old male with Stage IV jejunal adenocarcinoma and peritoneal carcinomatosis achieved a near-complete response after 4.5 years of systemic therapy, with laparoscopic confirmation of absent peritoneal disease. He now presents with chronic hemorrhage from the prosthesis site causing transfusion-dependent iron-deficiency anemia, and deteriorating performance status precluding further cytotoxic therapy. Exploratory laparoscopy confirmed no residual abdominal metastases.

Question
. . . Given uncontrolled prosthesis-related blood loss and chemotherapy intolerance following laparoscopic confirmation of absent metastases, *what intervention should be pursued?*

Oracle answer
Proceed with partial jejunectomy.

Model output
The intervention that should be pursued in this case is conversion surgery.

Both answers identify surgery as the required intervention class. The model uses the category-level label (conversion surgery) without naming the specific operative procedure. The oracle requires the exact operation (partial jejunectomy); the model's answer is correct in type but insufficiently specified to be equivalent.

Figure 9. **Granularity mismatch failure.** The model correctly identifies the intervention class (surgery) but answers at a higher level of abstraction than the oracle requires, naming the strategic category rather than the specific procedure. Single underlines mark content directly informing the oracle answer; wavy underlines mark constraining clinical context the model disregarded.

Use of the Stingray Re-Entry System in two complex cases of occluded superficial femoral arteries
PMCID: PMC3420382 Internal Medicine

Case summary
A 77-year-old female with Rutherford category 3 PAD and complete proximal right SFA occlusion underwent percutaneous revascularization. A CrossBoss device was advanced to the popliteal level via a subintimal route, then exchanged for the Stingray Re-Entry System over a Miracle guidewire. The Stingray catheter is now positioned subintimally at the popliteal artery following device exchange.

Question
. . . With the wire and catheter positioned subintimally after device exchange to the Stingray system, *what represents the most suitable immediate subsequent procedural maneuver?*

Oracle answer
Inflate the Stingray Re-Entry System balloon to stabilize the catheter within the subintimal space at the popliteal artery.

Model output
Advance the dedicated Stingray Re-Entry System guidewire through the catheter's distal port to penetrate the intima into the true lumen.

The article defines two sequential steps: (1) balloon inflation for subintimal stabilization, followed by (2) guidewire advancement for true-lumen re-entry. The question asks for step 1; the model provides step 2, selecting an action from the immediately subsequent procedural stage.

Figure 10. **Context/stage misbinding failure.** The model skips the current procedural step (balloon inflation for stabilization) and provides the immediately subsequent maneuver (guidewire advancement), selecting an action bound to the wrong point in the intervention sequence. Single underlines mark content directly informing the oracle answer.

Newly diagnosed AIDS patient with cerebellar JC virus
PMCID: PMC10461121 Neurology

Case summary
A 57-year-old male with AIDS-related progressive multifocal leukoencephalopathy (PML) achieved HIV suppression on antiretroviral therapy but continued to deteriorate neurologically, with repeat CSF showing JC virus burden rising from 2,000 to 14,000 copies/mL. Empiric glucocorticoids for suspected immune reconstitution syndrome provided no benefit. He subsequently developed multi-domain neurological deficits including new confusion, dysphagia, language impairment, and increasing left-sided weakness.

Question
... Given ongoing neurological decline with expanding deficits despite sustained HIV control and failure of glucocorticoids, *what immediate therapeutic intervention should be pursued?*

Oracle answer
Re-hospitalize him and initiate pembrolizumab therapy, with repeat neuroimaging to assess for progression.

Model output
The immediate therapeutic intervention that should be pursued for this patient is the initiation of pembrolizumab therapy.

The oracle specifies a three-component bundle: re-hospitalization, pembrolizumab initiation, and repeat neuroimaging. The model isolates the key therapeutic agent while dropping the organizational (re-hospitalization) and monitoring (neuroimaging) components, producing an answer that is related but not equivalent to the oracle.

Figure 11. **Constraint/qualifier erosion failure.** The model identifies the correct therapeutic agent but omits two required concurrent actions, producing an answer that is a strict subset of the oracle rather than equivalent to it. Single underlines mark content directly informing the oracle answer; wavy underlines mark constraining clinical context the model disregarded.

L. Prompts and Instructions

BMRetriever Instruction

Query template. Given a question, retrieve relevant documents that best answer the question. <query>
Passage template. Represent this passage\n passage: <passage>

Figure 12. Instruction used by BMRetriever.

GPT-5.2 prompt for significance extraction

Please carefully read the provided case report text or abstract. Identify whether the report describes any unique clinical actions from the following list:

- Novel treatment or drug introduced
- Existing treatment used in a new way or indication
- New surgical or procedural technique applied
- Innovative combination of treatments or devices
- Novel diagnostic test or imaging method used
- Advanced molecular/genetic testing guiding treatment
- Unique point-of-care or biomarker test employed
- Use of AI or machine learning for diagnosis or treatment planning
- Novel intervention to manage unexpected treatment complications
- Off-label drug use with unique dosing or delivery method
- Personalized or precision medicine approach in therapy
- New integrated multidisciplinary care strategy
- Innovative use of telemedicine or remote monitoring in clinical management
- New rehabilitation or follow-up protocol applied
- Novel preventive or screening intervention implemented
- Ethical or legal decision-making impacting treatment
- Use of newly developed medical devices or technologies for treatment

For each identified action, briefly summarize what it is and how it is unique or novel in this case report. If none apply, output exact string "no" (without quote marks) only. If there are multiple points, only describe the MOST SIGNIFICANT POINT, most likely mentioned in the abstract.

Figure 13. GPT-5.2 Prompt we used to extract significance from the case reports. The output is used to create the question-answer pairs.

GPT-5.2 prompt for timeline extraction

You will receive a clinical case presentation. Your task is to carefully parse and organize all clinical information strictly as a chronological timeline, listing events sequentially based on when they occurred or inferred from context. Include ALL significant events: arrival, initial presentation, interventions, medications administered, diagnostic tests and results, progression events, clinical decisions, and final diagnosis.

- Patient Arrival: Infer when and how the patient initially presented.
- Initial Clinical Presentation: Initial symptoms, vital signs, mental status, or general condition upon arrival.
- Early Medications and Interventions: Any initial treatments or interventions (medications, fluids, oxygen, etc.) performed upon or soon after presentation.
- Diagnostic Tests and Results: Tests ordered, procedures performed, and their outcomes. Infer order based on context if not explicitly stated.
- Clinical Progression: Any changes in the patient's condition over time, including improvements, deteriorations, or significant clinical events.
- Final Diagnosis or Impression: The final diagnosis, clinical assessment, or conclusion.

Avoid categorizing by types of events; instead, present them sequentially as they unfold in real clinical practice. Use exact sentences from the case report for building the timeline. DO NOT include any information that is not the a part of case presentation. Within the case presentation, include any information necessary for clinical decision making, such as the patient refusing or requireing certain diagnostic tools.

Output should be formatted in similiar way as this for parsing purpose:
- patient information
- initial symptoms and patient state
- following timeline in chronological order
...
- final timeline of the case report

Figure 14. GPT-5.2 Prompt we used to extract timeline from the case reports. The output is used to create the question-answer pairs.

GPT-5.2 prompt for limitation extraction

You are given (1) a medical case report, and (2) a paragraph describing the clinical significance or novelty of the case (i.e., why the case was written), referred to as ”significance”.

Task:
Your task is to extract and summarize explicitly case-specific clinical or contextual limitations that prevented the use of standard or traditional management and motivated the use of novel approach described in the significance.
A ”limitation” must be an explicitly stated barrier to eligibility, access, feasibility, or safety of standard medical procedure. Disease severity, clinical difficulty, rarity, or the fact that the case is ”hard” are NOT LIMITATIONS unless the case report explicitly states that they prevented use of standard management.
Prior treatment failure counts as a limitation only if explicitly stated as a reason standard management could not be used.
Limitation is stated in the Discussion section of the case report, and is different from the section header ”limitation”.

Some case reports do not have limitations specifically stated.

Limitations may include (but not restricted to):
- patient-specific factors (e.g., age, frailty, comorbidities, contraindications)
- system-level constraints (e.g., limited resources, emergency, long wait times)
- disease-specific constraints blocking standard medical procedure(e.g., anatomical complexity preventing surgery)
- explicitly stated ineligibility for standard medical procedure

The following are NOT limitations on their own:
- disease severity
- treatment difficulty
- failed prior methods, unless stated as blocking standard care and reason why significance of the case report was used
- general background or motivation of the case
- novelty or importance of the case

Instructions:
- identify limitations only if they are explicitly stated in the report (from the Discussion section)
- DO NOT infer based on medical knowledge
- DO NOT restate significance
- summarize only the reasons or constraints, not the procedure itself
- if multiple limitations are mentioned, list all of them concisely
- if no explicit limitation is stated, output exact string ”no” (without quote marks) only

Figure 15. GPT-5.2 Prompt we used to extract limitations from the case reports. The output is used to create the question-answer pairs.

GPT-5.2 prompt for question-answer pair generation

You are given a timeline and significant point of a medical case report. The significance refers to the new technique that the case report introduces. It can be surgical method, treatment, etc. Significance mentions the "new" item and explains why it is new. Timeline is an ordered list of bullet-style events exactly as they appear in the case report.

Create one question-answer pair asking for the immediate next step at the moment just before the action happens. The question should be the steps before taking the "new" action introduced in the significance. It should include all the steps before the decision point of "new" action. If there are multiple significant points in the input, choose the most complicated and unique one.

The question should not be based on an outcome but asking what is the significance or next step in treatment, or whether certain treatment/test could be done for this specific case, what to keep in mind during the treatment, etc. It should not ask about the past.

Extract a concise answer, using only exact sentence(s) from the timeline. The answer must be verbatim, not paraphrased.

This is a text only communication, so neither question nor answer should be referring to images, tables, or any other non-text media.

Output:

Context: Clearly list EVERY step of the timeline BEFORE this decision point in chronological order without leaving out any bulletpoints. ANY POINTS AFTER THE DECISION POINT SHOULD NOT BE IN THE CONTEXT. patient arrival, symptoms, vital signs, prior history, treatments/interventions, medications administered, diagnostic tests and their results, and clinical progression events—strictly up to this decision point should be included. If two consecutive elements of timeline are "decided to do xxx" and "xxx was performed", state it only once.

Question: Given this information, what is the immediate next appropriate clinical step? (Note: Adapt this to fit the situation, e.g., "next appropriate test", "next appropriate treatment", etc.)

Answer: State the exact next clinical step taken, using precise procedural or diagnostic terminology. Only use information explicitly stated in the timeline. THIS SHOULD NOT BE INCLUDED IN THE END OF THE CONTEXT. THIS IS THE **NEXT** STEP AFTER THE QUESTION STATEMENT.

The response should be outputted in json format. [Format omitted for brevity]

Figure 16. GPT-5.2 Prompt we used to create question-answer pairs from the case reports. Output "Context" and "Question" are combined to form the question.

Claude 4 Opus prompt for controlled question modification

You are given three items: (1) a detailed clinical query, (2) its corresponding concise query, (3) the answer.

Your task is to rewrite the **detailed query** following the instructions below:

Instructions:
1. Identify all overlapping content between the detailed and concise queries.
2. You must **preserve** the meaning of all overlapping content exactly yet **modify** the words or expressions.
- Keep the core meaning **unchanged**, but vary the surface form:
- Use synonyms, abbreviations, or different phrasing.
- Do NOT alter the medical intent or expected answer.
- Example: ”Management of acute MI” → ”Initial treatment of a heart attack”.
3. Identify the non-overlapping parts of the detailed query:
- Use synonyms, abbreviations, or different phrasing.
- Adjust numerical values by adding or subtracting within medically reasonable ranges.
- Altering the logical flow, sentence structure, or clinical context.
4. Add **extra distracting medical content** that are medically plausible but irrelevant to the answer:
- Comorbidities
- Symptoms, tests, and treatments.
- Background information.
- Past but resolved medical history.
- Family history that does not affect the answer.
- Redundant or vague phrases.

IMPORTANT:
- The revised detailed query should look **substantially different** from the original, while remaining **medically plausible**.
- Dilute keywords and disperse critical information across the query.
- It must still be **answerable by the original answer**.

- Return **only the modified detailed query**.

Figure 17. Claude 4 Opus prompt we used to “distract” questions. The detailed question is the full question from Figure 16 output. The concise question is the part that is necessary to derive the answer, while a detailed question usually conveys details unnecessary to reach the answer. Concise question remains semantically unchanged while other details are notably modified.

(a) General-purpose model prompt

You are a helpful medical assistant answering expert-level medical questions. You will receive a detailed clinical question about a patient case. Answer the question with one best answer. Do not generate multiple answers.

(a) Prompt for general-purpose models.

(b) Medical domain model prompt

You are a helpful medical assistant answering expert-level medical questions. You will receive a detailed clinical question about a patient case. Answer the question with one best answer. Do not generate multiple answers. Do not include analysis, steps, or thoughts, and restrict response to less than 100 words.

(b) Prompt with content and length restriction, for medical domain models

Figure 18. Prompts for generating responses with the questions from OGCAREBENCH.

GPT-5.2 prompt for evaluating answer and model response equivalence

You are an expert clinical decision maker. You will be given two texts:
Gold: the gold standard next step in medical procedure for a specific patient and situation
Response: a model's proposed next step for the patient and situation

Your task: decide whether Gold and Response have matching medical intent.

Criteria are following:

Equivalence:
Return "Equivalence" if Response intends same clinical procedure as Gold, even if:
- worded differently
- different descriptions of the same surgical step or technique (e.g., "laparoscopic cholecystectomy" vs "surgical removal of gallbladder")
- minor details differ as long as the core action is the same

Mismatch:
Return "Mismatch" if Response and Gold differs in any of these high-impact ways:
- different action type (e.g., monitoring vs. intervention, surgery vs. medication)
- significantly different order of action. A minor deviance is allowed
- different major additional action that changes the plan (e.g., Gold: "order CT"; Response: "start lobectomy)

Follow these steps:

1. Compare two core medical actions:
Extract the core medical action(s) from Gold. Express them as concise medical actions (e.g., "start chemotherapy," "perform lobectomy," "order CT scan"). Ignore details such as dose, frequency, or surgical technique/words unless they fundamentally change the type of action.
Extract the core medical action(s) from Response. Express them as concise medical actions (e.g., "start chemotherapy," "perform lobectomy," "order CT scan"). Ignore details such as dose, frequency, or surgical technique/words unless they fundamentally change the type of action.
2. Compare action, target, and clinical intent.
3. If unsure, default to Mismatch.
4. The output format should be one word: "Equivalence" or "Mismatch".

Do not output any other texts.

Few-shot examples
[Omitted for brevity]

Figure 19. GPT-5.2 prompt we used to evaluate the answer and LLM equivalency. The few-shot examples are drawn from early version of the dataset.

Instruction for the annotators

There are 7 columns in the spreadsheet:
- **Title**: title of the case report where the query and answer were derived from.
- **Classification**: the topic of the case report.
- **Link**: link to the case report. If you have any confusion or want to review the case report for validation, please use this link. If the query is straightforward, you don't need to validate with the case report.
- **Query**: question presenting the case that is similar to the case report and asking the next step at a potentially confusing decision point, related to the significant point, or the reason why the case report was written. The query will not be a direct iteration of the case report, as we added distractions to make the query more interesting and difficult to answer.
- **Answer**: answer to the query, derived from the case report.
- **Rating**: your rating of the query-answer pair, in 1 to 5 scale.
- **Comments**: your comments about the query-answer pair. Feel free to leave it blank.

Rating and Comments columns are for you to fill with your opinions. We want to simulate a situation where a doctor is looking up a case report for reference, and whether the action will be taken or not is up to the doctor. Our goal is to test if the models are reliable for that purpose using our dataset. The gold answers on the spreadsheet reflect the answer presented by the case report, whether it is the standard or not. We want to confirm that our information extraction was successful and that our queries are not too easy or obvious and correct.

To serve these purposes, the rating should be based on these criteria:
- The answer should be answering the query.
- The query should not be asking too "easy" question. The query should require knowledge from medical professionals to be answered.
- The provided answer is one of the most plausible acceptable answer given clinical context, ideally, the only answer. We understand that there will be multiple possible medical procedures given the characteristics of the medical field, but we want the answer provided to be at least one of the most reasonable answers.
- There are some hallucinated data (e.g., answer is already stated on the question). We did our best to filter them using a model, but if you see malformed data, feel free to rate it as 1 without further consideration.

These are okay:
- The answer is not gold standard (The goal is to create a dataset reflecting the case report's action taken.)
- The answer is not detailed
- The answer is too specific to the patient (Some information, such as numbers, may be too specific to the patient, but the doctor potentially searching this case will also see the same information.)

Figure 20. Instruction given to three annotators to verify question-answer pairs.

M. Case Report Venues

Table 15. Journal names of the corpus of 53,617 case reports. Extracted from PMC commercially available file list provided by: https://ftp.ncbi.nlm.nih.gov/pub/pmc/oa_non_comm_use_pdf.csv

Case Reports Plast Surg Hand Surg	Case Rep Cardiol	SAGE Open Med Case Rep
Int J Med Pharm Case Reports	JACC Case Rep	Case Rep Genet
J Clin Cases Rep	Case Rep Clin Med	Case Rep Dent
J Clin Case Rep	Case Rep Surg	Int Med Case Rep J
Case Rep Pancreat Cancer	Trauma Case Rep	Case Rep Ophthalmol
Arch Clin Case Rep	Case Rep Pulmonol	Autops Case Rep
Respir Med Case Rep	Clin Med Insights Case Rep	J Med Case Reports
Case Rep Neurol Med	Clin Pract Cases Emerg Med	Gynecol Oncol Case Rep
Case Rep Transplant	Case Rep Endocrinol	Clin Med Rev Case Rep
J Vasc Surg Cases Innov Tech	Med Case Rep Short Rev	Psychiatry Res Case Rep
Clin Nephrol Case Stud	Clin Case Rep Rev	Gen Thorac Cardiovasc Surg Cases
Arch Med Case Rep	Case Rep Emerg Med	MOJ Clin Med Case Rep
Endocrinol Diabetes Metab Case Rep	Prof Case Manag	IDCases
J Cardiol Case Reports	Ann Clin Case Rep	J Cardiol Cases
Am J Med Case Rep	Case Reports Immunol	Spinal Cord Ser Cases
Med Mycol Case Rep	Int Clin Med Case Rep J	Oxf Med Case Reports
Case Rep Psychiatry	IJU Case Rep	Case Rep Otolaryngol
J Surg Tech Case Rep	JCEM Case Rep	Case Rep Ophthalmol Med
Clin Med Case Rep	Case Rep Dermatol	JAAD Case Rep
Ann Clin Med Case Rep	J Pediatr Surg Case Rep	ACG Case Rep J
Surg Case Rep	Am J Ophthalmol Case Rep	Case Rep Crit Care
Case Rep Orthop	Case Rep Vet Med	Clin Case Stud
Case Rep Perinat Med	GMS Ophthalmol Cases	CASE (Phila)
Case Rep Radiol	Case Rep Womens Health	Eur Heart J Case Rep
Open J Clin Med Case Rep	Case Rep Gastrointest Med	Case Rep Infect Dis
Case Stud Eng Fail Anal	Case Rep Oncol Med	Cases J
BJR Case Rep	J Surg Case Rep	Case Stud Chem Environ Eng
Case Rep Dermatol Med	Clin Case Rep	Indian J Ophthalmol Case Rep
Case Rep Pediatr	BMJ Case Rep	Urol Case Rep
CEN Case Rep	Case Rep Anesthesiol	J Med Case Rep
Case Rep Urol	Case Reports Hepatol	Int J Surg Case Rep
Case Rep Obstet Gynecol	Int J Case Rep Imag	J Investig Med High Impact Case Rep
Asploro J Biomed Clin Case Rep	Case Rep Nephrol	Case Rep Hematol
Respirol Case Rep	Case Rep Pathol	Int J Clin Case Rep Rev
HeartRhythm Case Rep	Neurocase	AACE Clin Case Rep
Retin Cases Brief Rep	Cold Spring Harb Mol Case Stud	Case Stud Transp Policy
JBJS Case Connect	J Endourol Case Rep	Case Rep Rheumatol
Case Rep Med	Oral Health Case Rep	Arch Clin Med Case Rep
JMM Case Rep	Radiol Case Rep	Epilepsy Behav Case Rep
Case Rep Vasc Med	APSP J Case Rep	