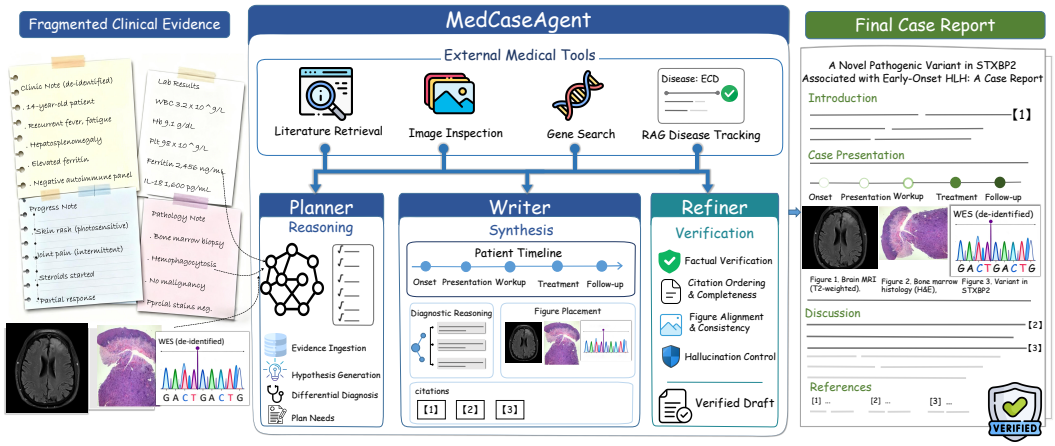
MEDCASEAGENT: FROM MULTIMODAL CLINICAL EVIDENCE TO RARE DISEASE CASE REPORTS

Anonymous authors
Paper under double-blind review



ABSTRACT

Although Large Language Models (LLMs) have demonstrated potential in clinical documentation, generating the rigorous, evidence-based reasoning required for unprecedented, atypical and rare disease case reports remains a significant challenge. We present MedCaseAgent, an end-to-end agentic reasoning paradigm that bridges this gap by synthesizing publication-quality reports from unstructured clinical notes. MedCaseAgent leverages a hierarchical structure, integrating patient-level data mining with iterative, tool-integrated reasoning. By dynamically querying medical literature and knowledge bases, the agent maintains high clinical fidelity and assesses case novelty. The framework is inherently modular, supporting various backbone LLMs in a plug-and-play fashion. To facilitate this research, we release MedCase-150K, a large-scale multimodal clinical case-report corpus, and MedCase-Bench, a robust temporally held-out evaluation suite. We also introduce an open-source model post-trained on Qwen3.5-9B optimized specifically for this task. Under both LLM-judged and rule-based evaluation, MedCaseAgent outperforms existing general-purpose and specialized medical LLMs and agents. Blinded expert reviews and ablations further support the system's clinical relevance and evidence grounding.

1 INTRODUCTION

A clinical case report, often referred to as a case study, is a comprehensive and highly structured account of a patient's clinical encounter (Cohen, 2006; Gagnier et al., 2014; Riley et al., 2017). The defining characteristic of a compelling case report is its professionalism, novelty, and structural consistency. The presentation must be sufficiently rare, unique, or clinically significant to impart new medical knowledge. By publicly detailing these findings, case reports enable global medical communities to cross-reference phenotypically similar patients, accelerating the diagnosis of rare diseases, uncovering underlying molecular mechanisms, and facilitating larger-scale statistical analyses. Consequently, they serve as a vital cornerstone of medical progress, with nearly all top-tier medical journals dedicating specific categories to them and receiving a massive volume of global submissions.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Despite their immense value, drafting these reports remains an arduous, entirely manual process for frontline physicians. Medical professionals must rigorously evaluate a case's novelty, meticulously synthesize patient histories and treatment plans from fragmented raw clinical notes, and conduct extensive literature reviews to identify homologous cases or genetic loci. Ultimately, they face the heavy cognitive burden of integrating specialty-specific reasoning with rigorous scientific citations to complete the narrative.

Recent years have witnessed significant advancements in large language models (LLMs), with systems such as GPT-5.4 and Gemini 3.1 Pro exhibiting remarkable reasoning capabilities. While their proficiency in structured problem-solving and logical inference has positioned them as the central cognitive engines for various clinical diagnostic agents (Chiu et al., 2025; Wang et al., 2025; Zhao et al., 2026), these systems primarily focus on sequential diagnostic deduction rather than the generative synthesis of academic medical knowledge. Consequently, generating clinical case reports for rare diseases remains an exceptionally demanding frontier. It requires rigorous, multi-hop reasoning across highly fragmented and multimodal data streams, ranging from unstructured clinical shorthand and complex patient histories to rare imaging modalities and genomic sequences. Crucially, a viable case report generation pipeline must adhere to stringent factual constraints. It must strictly circumvent clinical hallucinations, such as spurious gene-literature matching, fabricated citations, or inaccurate patient summaries, which would entirely compromise the clinical validity of the report.

To address this gap, we propose MedCaseAgent, an end-to-end agentic framework that synthesizes academic-grade case reports from unstructured clinical notes. Structured as an interactive planner-writer-editor workflow, the agent orchestrates specialized tools for literature retrieval, exact citation tracking, image analysis, and genetic search to guarantee clinical accuracy and reduce hallucinations. In parallel, we construct MedCase-150K from over 7 million PubMed articles published since 2000 and derive MedCase-Bench as the recent held-out evaluation split. By automatically stratifying cases by clinical rarity, these resources provide a real-world simulation that tests how effectively models interpret fragmented multimodal evidence to formulate complete reports.

Our contributions can be summarized as follows:

(1) We propose **MedCaseAgent**, a plug-and-play clinical case report agent that integrates multimodal foundation models with dynamic tool-use mechanisms to actively orchestrate and synthesize complex clinical narratives from clinical notes.

(2) We introduce **MedCase-150K**, a large-scale collection of rare and novel disease clinical case reports, and **MedCase-Bench**, a temporally held-out subset derived from MedCase-150K, both designed for multimodal medical case report generation and rigorously stratified by clinical rarity.

(3) We validate MedCaseAgent comprehensively on MedCase-Bench, where it surpasses mainstream MLLMs and medical agentic systems. Clinician evaluations and a widespread user study further highlight the reliability and usefulness of MedCaseAgent.

2 Related Work

Agentic Systems in Clinical Practice. Recent clinical agent research first shows strong progress in imaging. Across radiology and pathology, agentic systems support image-grounded reasoning (Du et al., 2026; Qiu et al., 2025; Chen et al., 2025a), evidence retrieval (Cao et al., 2026), report generation (Qi et al., 2026; Zhong et al., 2025b; Wang et al., 2026b), and quality control, but most remain designed for classification (Fard et al., 2025), segmentation (Liu et al., 2026), question answering (He et al., 2025; Li et al., 2024c; Xiao et al., 2025; Xu et al., 2025b; Jiang et al., 2025a), or planning (Zhang et al., 2025a). Text-centered agents then extend these ideas to longitudinal records (Li et al., 2025), simulated hospital workflows (Li et al., 2024b), clinical communication (Zhi et al., 2025), and safety oversight (Kim et al., 2025). Systems such as AgentEHR (Liao et al., 2026), MedAgentGym (Xu et al., 2025a), and Reflectool (Liao et al., 2025) study streaming EHRs (Croxford et al., 2025), retrospective summaries, and structured clinical tasks (Ren et al., 2025; Zhong et al., 2025a). A related direction studies clinical diagnosis and agent architecture search with agentic reasoning systems (Ghezloo et al., 2025; Zhao et al., 2026; Xia et al., 2025; Pushkar Mishra et al., 2025; Yang et al., 2026; Liu et al., 2025b; Wu et al., 2025b; Zhuang et al., 2025). These methods improve traceability through retrieval (Wang et al., 2026b; Deng et al., 2026; Li et al., 2026; Rose et al., 2025; Ahmed et al., 2026), verification (Chen et al., 2025a; Wang et al., 2026b; Li et al.,

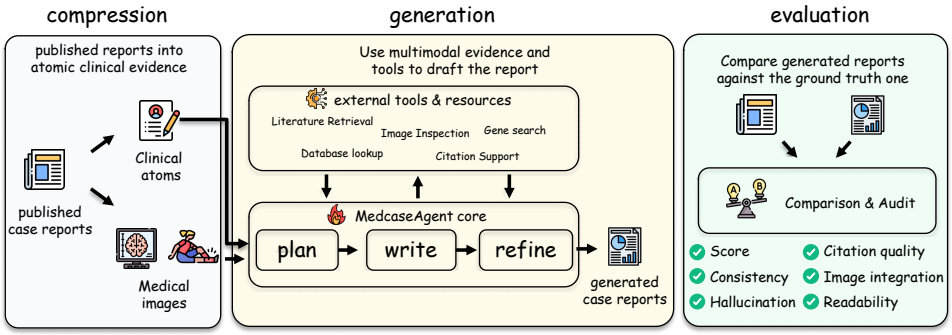


Figure 1: Overview of the MedCaseAgent report-generation workflow.

2026; Ahmed et al., 2026), and iterative refinement (Chen et al., 2025a; Wang et al., 2026b; Li et al., 2026; Rose et al., 2025), but usually stop at prediction (Xia et al., 2025; Pushkar Mishra et al., 2025; Deng et al., 2026; Yang et al., 2026; Liu et al., 2025b), recommendation (Wu et al., 2025b), or short-form clinical writing (Wang et al., 2026b). By contrast, MedCaseAgent targets end-to-end case report synthesis, linking multimodal evidence, timeline reconstruction, literature support, and manuscript-level writing.

Automated Scientific and Case-Report Writing. General-purpose research-writing agents such as the AI Scientist (Lu et al., 2024; Yamada et al., 2025), data-to-paper (Ifargan et al., 2024), CycleResearcher (Weng et al., 2024), and PaperOrchestra (Song et al., 2026) automate machine-learning or data-driven papers, but do not address patient-grounded clinical evidence, rarity assessment, or medical citation grounding. In medicine, prior studies prompted ChatGPT to draft individual case reports and concluded that the outputs required substantial expert revision and could contain fabricated references (Buholayka et al., 2023; Pinto et al., 2024); case-report corpora have otherwise been used for diagnostic reasoning (Wu et al., 2025a) or information extraction (Zhang et al., 2025b). MedCaseAgent instead treats case-report writing as a trainable, tool-grounded agentic task with a dedicated benchmark.

3 METHODOLOGY

Overview. We first construct MedCase-150K, a large-scale collection of peer-reviewed clinical case reports. From this collection, we define MedCase-Bench as a temporally held-out evaluation subset of 584 case reports published after January 2026. Both MedCase-150K and MedCase-Bench will be open-sourced under CC-BY-NC 4.0. We then propose MedCaseAgent (Figure 1), an agentic report-writing system that plans over the fragmented evidence, retrieves supporting medical literature, drafts the manuscript, and audits the final report.

3.1 TASK FORMULATION

Given unstructured clinical observations X and external medical resources $\mathcal{D}$, the goal is to generate a structured clinical case report Y_{report}. The inputs include textual clinical notes $X_{\text{notes}} = \{x_1, x_2, \ldots, x_n\}$ that describe patient demographics, symptoms, laboratory findings, diagnosis, treatment, and outcome, together with diagnostic images $X_{\text{img}} = \{i_1, i_2, \ldots, i_m\}$ such as MRI, CT, radiography, or histology scans. We represent the textual input as clinical atoms grouped into history, presentation, diagnostics, management, and outcome. This atom format is designed to approximate the first-hand information that clinicians receive before a manuscript is written: partial notes, test results, image findings, missed visits, and follow-up details, rather than a polished academic narrative (Abacha et al., 2023). Recent work also suggests that LLM-written summaries can form portable representations for irregular clinical time series (Ji et al., 2026); in our setting, atoms are used as controllable evidence for manuscript generation rather than as predictive embeddings. The atom format also gives a stable supervision unit for planner training and reward computation, since each generated plan can be checked for atom coverage. The target output is a complete report with source-driven manuscript sections such as Introduction, Case Description, Discussion, and References. During generation, the vision-language model acts as an agent. It issues tool calls and

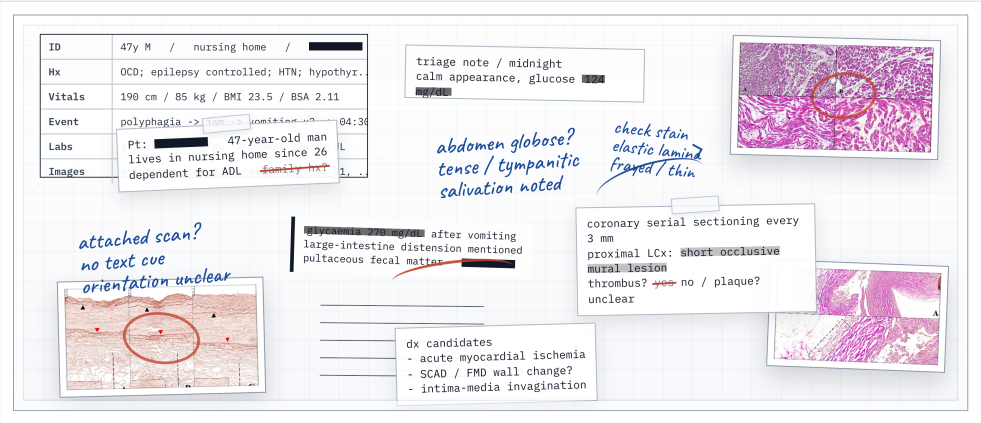


Figure 2: Illustration of fragmented first-hand clinical observations used as multimodal input to MedCaseAgent. The notes are patient-specific and incomplete, resembling the partial facts available before a formal academic case report is written.

search queries q_t to retrieve supporting knowledge $Z_{\text{retrieved}}$ from $\mathcal{D}$, and then links the raw clinical atoms with a complete manuscript. Figure 2 illustrates fragmented and partially structured clinical observations as the multimodal input for MedCaseAgent; Appendix B provides a fuller prompt-output atom-compression view.

3.2 MedCase-150K and MedCase-Bench Construction

While datasets like CaseReportBench (Zhang et al., 2025b) treat case reports as targets for information extraction, our setting requires a model to write a complete, well-cited report from fragmented multimodal inputs, which is a much harder problem. As there is no existing dataset for this domain, we construct MedCase-150K, a dataset including 154,872 case reports with unprecedented, atypical presentation, rare case reports and their variants (see Appendix C, Data Construction and Validation, for dataset overview and statistics). Each MedCase-150K entity contains a source published article, unpaired medical images, and the corresponding LLM-compressed clinical atoms. The construction pipeline of MedCase-150K is semi-automated. First, we manually identify case report PMC Open Access XML publications with full text and high-quality medical images. An LLM-based parser converts the final prose into unstructured clinical atoms X_{notes} by removing background literature, reference-driven claims, and final diagnostic exposition. The parser keeps patient-specific facts, including timeline details, tests, treatments, image findings, and outcomes.

This compression step is necessary because the original article already contains the target report. By turning the published manuscript back into noisy clinical atoms, we make the task closer to real clinical writing and create supervision that can be used for planner SFT and GRPO training. Source images accompany the atoms without captions or explicit image–text links. Human reviewers checked all 632 benchmark candidates for atom fidelity, then screened retained cases for rarity-label validity, retaining 584 cases (Appendix C.1). MedCase-Bench is derived from MedCase-150K as the evaluation subset: it uses a strict temporal holdout of peer-reviewed PMC Open Access case reports published after January 2026, reducing the risk that evaluated models have memorized the final reports.

3.3 MedCaseAgent: Agentic Medical Case Report Synthesis Pipeline

Given fragmented clinical atoms and paired medical images, MedCaseAgent divides report generation into three stages: Planner, Writer, and Editor. The same external tools are available throughout all three stages. In practice, tool use is intentionally concentrated in the Planner, where most literature search, image inspection, gene lookup, and prior-case retrieval calls are made before manuscript drafting begins.

Phase 1: The Planner (Multimodal Reasoning and External Tool Execution). The planner starts from the multimodal state $S_0 = (X_{\text{notes}}, X_{\text{img}})$. A multimodal vision-language model reads the

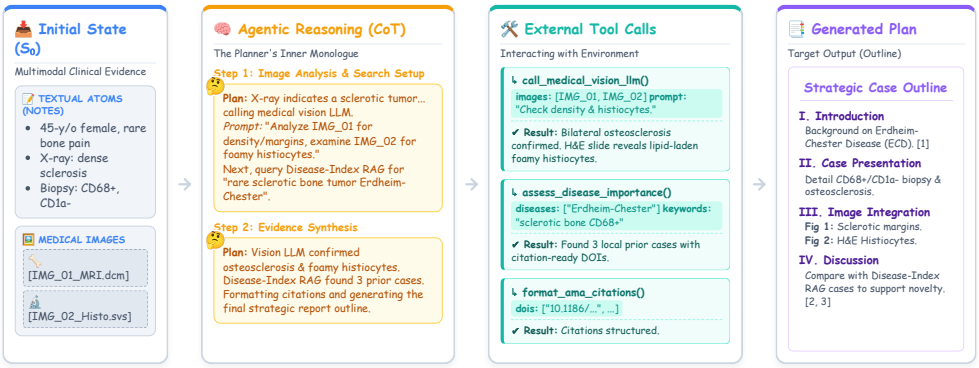


Figure 3: Planner phase from initial multimodal evidence to a generated case-report plan.

images and clinical atoms, then selects actions from a predefined action space (Yao et al., 2022; Akkiraju et al., 2024), following prior tool-mediated clinical agent systems (Yue et al., 2024). These tools include Literature Retrieval for PubMed search and citation grounding, Image Inspection for single-frame or multi-panel figure interpretation, Gene Search for ClinGen variant lookup, and Disease-Index RAG for retrieving similar MedCase-150K prior cases to support novelty planning. The Disease-Index RAG corpus excludes the evaluated cases' source articles and related literature to prevent information leakage. At step t, the planner generates a query q_t, retrieves context C_t, and updates working memory as $S_{t+1} = S_t \cup C_t$.

The planner returns three outputs. First, it maps each clinical atom to the source-driven manuscript section in which it should appear, so that patient history, diagnostic tests, procedures, outcomes, and unresolved details are not lost. Second, it builds an evidence plan by selecting literature search terms, verified citations, and Disease-Index RAG results that can support background and discussion claims. Third, it creates a figure plan by assigning each source image to a main-text callout and a short caption. This plan gives the writer a grounded intermediate representation between raw notes and manuscript structure. Figure 3 illustrates how the planner moves from the initial multimodal state through reasoning and tool calls to the generated section, citation, and figure plan.

Phase 2: The Writer (Draft Generation). The writer creates the first manuscript draft from the planner output and the accumulated evidence $Z_{\text{retrieved}} = \bigcup_{t=1}^{T} C_t$. It conditions on the clinical notes, images, and retrieved context:

$$P(Y_{\text{draft}} \mid X_{\text{notes}}, X_{\text{img}}, Z_{\text{retrieved}}) \tag{1}$$

The writer follows the planner section by section, which helps keep manuscript-length generation coherent in a setting where long-output generation remains challenging for many long-context LLMs (Bai et al., 2024). It writes patient-specific content only from the atoms and uses retrieved literature only for background, diagnostic criteria, treatment rationale, and comparison with prior work. This separation is important because patient facts should not be cited as if they came from the literature, and literature facts should not be copied into the case description as patient evidence.

Phase 3: The Editor (Refinement and Verification). The editor refines the draft and checks factual consistency. It creates required front matter such as the title and abstract, verifies that the final text Y_{report} is supported by the clinical atoms $x_i \in X_{\text{notes}}$, checks citation order and density, and confirms that every figure is described in the main text. The editor also checks that the draft follows the source section order, that broad case-report sections still contain CARE-style content, and that unsupported statements are removed or rewritten. In the implemented pipeline, later audit and citation-repair passes reuse the same evidence bank to correct missing references, citation order, and figure callouts. The final output is a verified manuscript.

3.4 Training Objective for the Planner

We train the phase 1 planner with a two-stage recipe implemented with SWIFT and LoRA adapters. Figure 4 summarizes the planner-training flow from multimodal inputs through SFT trajectories and GRPO reward feedback. This use of supervised trajectories followed by verifiable reward

feedback follows the GRPO-style reasoning-training line (Shao et al., 2024) and is aligned with recent medical reasoning work on verifiable problems, process supervision, and reinforcement learning (Chen et al., 2025b; Jiang et al., 2026; Liu et al., 2025a). First, we apply supervised fine-tuning (SFT) on teacher-generated planner trajectories. The data preparation converts the public phase 1 planner trajectory into agent messages, with tool calls, tool responses, and final plans used as supervision. Given an input case x and a teacher trajectory $\tau^* = (y_1^*, \ldots, y_T^*)$, the SFT loss is $\mathcal{L}_{\text{SFT}} = -\sum_{t=1}^{T} \log \pi_\theta(y_t^* \mid x, y_{<t}^*)$.

Second, we apply Group Relative Policy Optimization (GRPO) to improve tool-use decisions and plan quality. For each case, the current policy samples a group of planner trajectories $\{\tau_i\}_{i=1}^{G}$. Each trajectory is scored by a compact reward:

$$R(\tau) = w_{\text{schema}} R_{\text{schema}} + w_{\text{doi}} R_{\text{doi}} + w_{\text{evidence}} R_{\text{evidence}} + w_{\text{coverage}} R_{\text{coverage}} + w_{\text{image}} R_{\text{image}} + w_{\text{quality}} R_{\text{quality}}. \tag{2}$$

Here, R_{schema} checks allowed tool names and valid arguments; R_{doi} checks whether the generated search queries retrieve the scorer-only source DOI when available; R_{evidence} checks DOI validity, citation anchors, and overlap with non-source retrieved support; R_{coverage} checks clinical atom coverage and required section coverage; R_{image} checks structure and image placement; and R_{quality} uses a deterministic quality proxy or a stored judge score. R_{doi} provides a dense signal for query quality: recovering the one article that describes the same patient requires queries that capture the distinctive clinical features of the case, the same capability that retrieving analogous prior cases demands.

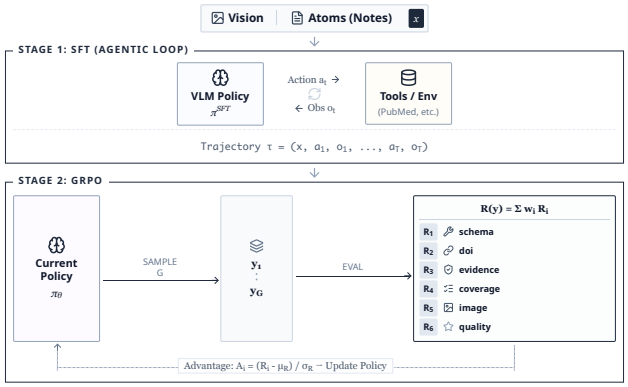


Figure 4: Training phase for the MedCaseAgent planner.

The group-relative advantage is computed as $A_i = (R(\tau_i) - \mu_R)/(\sigma_R + \epsilon)$, where μ_R and σ_R are the mean and standard deviation of rewards within the sampled group. The default weights prioritize clinical content and evidence: DOI retrieval, evidence support, coverage, and quality receive 0.20 each, while schema compliance and structure/image placement receive 0.10 each. Appendix D provides the rationale and reward-weight sensitivity analysis.

4 Experiments

4.1 Experimental Setup

Baselines. For benchmark evaluation, all systems are evaluated on MedCase-Bench. The input is the patient atom notes and unlabeled medical images. For a fair comparison, all baseline pipelines are equipped with the same tools. We reserve standalone model names for direct tool-equipped LLM/VLM baselines; in backbone-swap experiments, *MedCaseAgent (model)* denotes the full MedCaseAgent pipeline using that model as the centered LLM. MedCaseAgent-Base uses Qwen3.5-9B as its backbone. MedCaseAgent-Tuned uses the same Qwen3.5-9B backbone with the planner trained using the objective in Section 3.4; implementation and compute details are provided in Appendix E. MedCaseAgent-Max uses GPT-5.4.

Direct baselines include tool-equipped general VLMs such as Gemma-4-31B (Google DeepMind, 2026) and specialist medical VLMs such as the MedGemma family (Sellergren et al., 2025; 2026). We compare with representative medical agentic systems, including MedAgents (Tang et al., 2024), MDAgents (Kim et al., 2024), MedAgent-Pro (Wang et al., 2025), and MMedAgent (Li et al., 2024a). These methods are strong reference points because they use agents to support clinical reasoning, diagnosis, and medical decision-making. We adapt them to the same report-generation setting and provide the same tool access, so the comparison focuses on system design rather than tool availability.

Table 1: Disease-subset evaluation on MedCase-150K, comparing tool-equipped general MLLMs and medical agentic systems across selected disease categories. Each disease column samples 100 cases, except Erdheim-Chester disease (ECD), where all 53 available MedCase-150K cases are used. Each score averages citation depth, patient-history nuance, clinical coherence, image integration, and readability.

Method	Centered LLM	COVID-19	Small-Cell Lung Cancer	Erdheim-Chester Disease	Rare Genetic Disorder	Ultra-Rare Disease	Average
Human Publication	-	82.96	76.84	73.75	73.28	71.07	75.58
Qwen3.5-9B (Yang et al., 2025)	-	62.46	55.84	54.35	53.44	53.79	55.98
Qwen3.5-397B-A17B	-	78.28	74.64	66.21	66.15	64.11	69.88
Gemma-4-31B	-	72.34	69.26	64.25	63.03	62.45	66.27
Claude Opus 4.6	-	82.05	75.18	74.62	73.58	67.75	74.64
GPT-5.4	-	84.41	79.41	74.34	70.65	73.36	76.43
Gemini 3.1 Pro	-	82.42	75.92	72.30	75.61	72.09	75.67
LLaVA-Med-7B (Li et al., 2023)	Mistral-7B	52.85	49.27	46.68	47.45	43.90	48.03
MedGemma-1.5-4B (Sellergren et al., 2026)	Gemma 3-4B	51.61	49.16	46.26	47.57	40.41	47.00
MedVLM-R1-2B (Pan et al., 2025)	Qwen2-VL-2B	46.88	40.84	39.50	38.08	34.31	39.92
MedAgents (Tang et al., 2024)	Qwen3.5-9B	66.37	60.03	58.00	57.37	55.64	59.48
MDAgents (Kim et al., 2024)	Qwen3.5-9B	66.22	61.80	59.18	58.40	56.93	60.51
MedAgent-Pro (Wang et al., 2025)	Qwen3.5-9B	70.58	64.52	62.61	61.68	58.16	63.51
MMedAgent (Li et al., 2024a)	Qwen3.5-9B	69.52	63.93	62.34	61.70	58.71	63.24
MedCaseAgent-Base	Qwen3.5-9B	72.88	66.40	64.23	63.43	58.95	65.18
MedCaseAgent-Max	GPT-5.4	87.98	82.97	82.80	83.71	78.37	83.17

Judge ensemble and report scoring. For report scoring, we use a four-model VLM judge ensemble (Qwen3.5-397B-A17B, Gemini 3.1 Pro, Claude Opus 4.6, and GPT-5.4), following clinical LLM-as-judge and automated medical-reasoning evaluation practice (Croxford et al., 2025; Zhou et al., 2025). The ensemble spans four model families, so no single family determines the score. Appendix F.2 reports agreement with clinical experts, and Appendix G.1 reports per-judge scores.

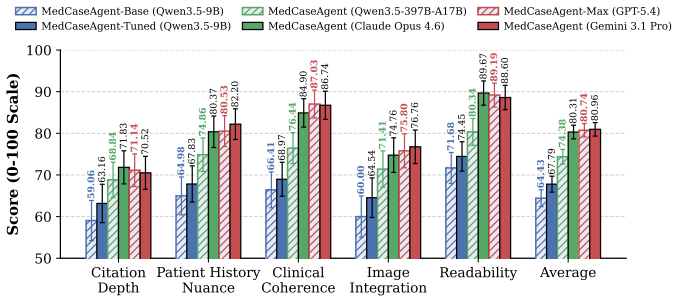


Figure 5: MedCaseAgent configuration comparison on MedCase-Bench across five rubric metrics and their average. Each legend entry is a MedCaseAgent variant, with the centered LLM shown in parentheses. Bars show means; error bars show standard deviations.

Evaluation metrics. We evaluate generated reports from complementary views. Rubric-level judge scores assess citation depth, patient-history nuance, clinical coherence, image integration, and readability on a 0–10 scale; we linearly scale them to 0–100 for display and report average quality as the mean of these five dimensions. Disease-subset scores test robustness across common, rare, and ultra-rare diagnoses; rule-based metrics audit retrieval, novelty recognition, figure ordering, and citation coverage; and clinician-reviewer A/B validation measures practical preference and authorship indistinguishability.

4.2 Report Quality and Disease-Stratified Evaluation

Figure 5 compares MedCaseAgent configurations on MedCase-Bench when the same agent pipeline is equipped with different backbone LLMs. After training, MedCaseAgent-Tuned improves over MedCaseAgent-Base in average score ($64.43 \rightarrow 67.79$), with clear gains in Citation Depth ($59.06 \rightarrow 63.16$) and Image Integration ($60.00 \rightarrow 64.54$). Stronger backbones further raise overall report quality within the MedCaseAgent workflow, with MedCaseAgent-Max (GPT-5.4), MedCaseAgent (Gemini 3.1 Pro), and MedCaseAgent (Claude Opus 4.6) forming the top group.

Table 1 stratifies performance across disease-specific subsets sampled from MedCase-150K and tests whether models remain stable as disease priors weaken. Although scores tend to decline from better-represented diseases such as COVID-19 and small-cell lung cancer to rarer settings such as Erdheim-Chester disease, rare genetic disorders, and ultra-rare diseases, MedCaseAgent-Max remains

Table 2: Comparison of rule-based and structural evaluation metrics across different models. All entries are percentages. Abbreviations: Exact DOI Retrieval (Exact DOI), Novelty Recognition (Novel Rec.), Figure Sequence Accuracy (Fig. Seq.), Authorship Indistinguishability (Auth. Indist.), and citation recall for must-cite (P0) and good-to-cite (P1) references.

Method	Exact DOI	Novel Rec.	Fig. Seq.	Auth. Indist.	P0 Recall	P1 Recall
Qwen3.5-9B	22.00	38.12	54.75	31.95	62.63	31.54
Qwen3.5-397B-A17B	36.27	57.39	66.93	45.53	66.70	35.32
Gemma-4-31B	31.11	52.88	63.36	38.92	65.87	34.87
Claude Opus 4.6	43.22	66.75	73.45	53.79	68.60	35.62
GPT-5.4	46.90	69.01	75.92	50.89	69.15	36.13
Gemini 3.1 Pro	44.50	68.01	76.56	52.76	68.83	35.59
MedCaseAgent-Base	30.60	52.40	61.10	42.84	64.62	34.16
MedCaseAgent-Max	84.59	88.35	87.34	69.18	74.14	44.82

the best-performing system across all disease subsets. Frontier closed-source LLMs score close to the human-publication reference on average, and MedCaseAgent-Max scores above it on every disease subset. In the blinded study, 26 of 32 experienced clinicians correctly identified the generated report in at most three of five pairs, and reviewers more often preferred the MedCaseAgent-Max report (Appendix F.3). MedCaseAgent-Base also stays above prior medical-agent baselines, indicating that the planner–writer–editor workflow improves disease-level robustness even before the strongest full configuration is used. Appendix G reports the corresponding LLM-judged blind A/B comparison across baselines, and Appendix H provides three complete case reports generated by MedCaseAgent for qualitative inspection.

4.3 Rule-Based Structural Metrics and Ablation

Table 2 reports rule-based and structural metrics for source retrieval, novelty recognition, figure ordering, authorship indistinguishability, and citation coverage. Exact DOI measures whether a keyword-based literature search retrieves the ground-truth PubMed source paper from which the atoms were derived. Note that even if the ground-truth paper is retrieved, it is masked during report generation to prevent information leakage. Exact DOI thus measures retrieval precision: whether the agent's queries isolate the single article describing the same patient among millions of PubMed records. Novelty Recognition is the percentage of cases where the model recognizes rare or first-case novelty even though this cue is omitted from the atoms. Figure Sequence Accuracy is binary per case: 1 only when the generated report uses the same figure order as the published reference, and 0 otherwise. Authorship Indistinguishability follows an LLM-based A/B test and measures how often generated reports are difficult to distinguish from the published ones. Inspired by PaperOrchestra (Song et al., 2026), we assess citation coverage using two reference tiers. P0 denotes must-cite analogous or claim-critical case-report sources. P1 denotes useful background, mechanism, treatment, or comparison sources. After canonical PubMed/DOI/title matching, tier k is scored as $\mathrm{Recall}_k = |C \cap G_k|/|G_k|$, where C is the generated citation set and G_k is the reference set for that tier. MedCaseAgent-Max has the strongest structural grounding (84.59% Exact DOI vs. 46.90% for the strongest direct baseline), the best novelty recognition, the closest figure sequencing, and the highest P0/P1 Recall.

Table 3 uses the same five report-quality dimensions as Figure 5 to study planner training and tool removal, and it adds Novelty Recognition, the same percentage metric abbreviated as Novel Rec. in Table 2, as the structural metric most directly tied to rarity and prior-case lookup. The first three rows isolate the effect of replacing the untuned Qwen3.5-9B planner with the trained planner and then with the strongest full system. Among tool ablations, removing literature retrieval causes the largest drop in average score and citation depth. Image inspection and disease-index retrieval contribute most to image integration and novelty recognition, respectively.

Reward-weight sensitivity. We compare four GRPO reward-weight profiles retaining the source-DOI reward from the same SFT checkpoint under fixed training and evaluation settings. The default achieves the highest four-metric average (71.68), with a 1.03-point spread across these

Table 3: Two ablation studies for planner training (upper part) and external tools (bottom part) using report-quality metrics and novelty recognition. Novelty Recognition is the same percentage metric as Novel Rec. in Table 2. The bottom part is performed on MedCaseAgent-Max.

Configuration	Citation Depth	Patient-History Nuance	Clinical Coherence	Novelty Recognition	Image Integration	Readability	Average
MedCaseAgent-Base	59.06	64.98	66.41	52.40	60.00	71.68	62.42
MedCaseAgent-Tuned	63.16	67.83	68.97	59.80	64.54	74.45	66.46
MedCaseAgent-Max	**71.14**	**80.53**	**87.03**	**88.35**	**75.80**	**89.19**	**82.01**
w/o Literature Retrieval	48.72	78.64	79.18	72.48	75.12	87.03	73.53
w/o Image Inspection	70.28	78.91	84.10	85.71	58.95	86.85	77.47
w/o Gene Search	67.92	79.64	84.86	81.69	75.14	88.10	79.56
w/o Disease-Index RAG	62.84	78.73	81.48	68.90	74.92	87.30	75.70

profiles (Appendix D, Table 6). Removing the source-DOI reward ($w_{\text{doi}} = 0$, remaining weights renormalized) lowers the average to 69.83, a decrease of 1.85 points, with novelty recognition declining from 59.08% to 55.78% and report quality from 67.23 to 65.37.

4.4 Clinical Evaluation and User Study

To assess whether MedCaseAgent can support case-report writing from real clinical data, we conducted a hospital-based study using 30 de-identified multimodal cases from 10 hospitals. All cases were selected as suitable for rare-disease case-report preparation, including cases with unpublished manuscripts intended for submission. Fourteen clinical experts assessed reports generated from raw records and report-derived atoms alongside clinician-written reports; each case was independently reviewed by three experts against the source records. Table 4 shows comparable overall quality for raw-record and atom-based generation (7.4 versus 7.6), with higher omission-control scores for raw input (7.8 versus 7.6). Appendix F.1 details the input differences. Because these cases have no published source article, they evaluate MedCaseAgent in its intended deployment setting. In a paired calibration, LLM judgments agreed with clinicians in 23–26 of 30 assessments across scored dimensions (quadratic-weighted κ = 0.69–0.76; ICC = 0.69–0.77) and in all 30 hallucination checks (Appendix F.2).

Table 4: Mean expert scores (1–10; ↑).

Report	Quality	Timeline	Omission control
Raw records	7.4	7.5	7.8
Atoms	7.6	7.6	7.6
Clinician	7.5	7.8	7.9

We additionally developed a web-based platform for user recruitment, user studies, and feedback collection, engaging more than 100 users. User feedback has been largely positive on the system's usefulness for case-report preparation. Detailed data-collection procedures, study protocols, blinded A/B results, and platform-based user feedback are provided in Appendix F.

5 Conclusion

We presented MedCaseAgent, a planner–writer–editor agent that turns fragmented clinical notes and medical images into publication-style rare-disease case reports, together with MedCase-150K and MedCase-Bench, a 584-case temporal holdout of reports published after the construction cutoff. Tool-grounded planning yields large gains in verifiable grounding: MedCaseAgent-Max retrieves the exact source article for 84.6% of cases against 46.9% for the strongest direct baseline, recognizes case novelty in 88.4% of cases, and attains the highest must-cite citation recall. It is the best-performing system on every disease subset, from COVID-19 to ultra-rare diseases, and scores above the human-publication reference under a four-judge ensemble spanning four model families. In blinded review, experienced clinicians more often preferred MedCaseAgent reports over the published originals, and most did not consistently identify which report was generated. The remaining frontier is diseases with no precedent in the retrieval sources: there, MedCaseAgent can organize the available evidence and its uncertainty, but the discovery is the clinician's.

AI USE STATEMENT

For manuscript preparation, generative AI tools were used for language polishing and improving readability. The authors reviewed the edits and take responsibility for the final manuscript. The use of LLMs for clinical-atom construction, planner-training trajectory generation, report generation, and automated evaluation is part of the research methodology described in Sections 3 and 4 and the appendices.

ETHICS STATEMENT

This research was conducted in accordance with the ICLR Code of Ethics. The hospital-based case study was conducted under existing institutional review board (IRB) approval, and all clinical materials were de-identified before analysis (Appendix F.1). MedCaseAgent is intended to assist clinicians in preparing case reports. Generated reports require clinical review of patient facts, image interpretations, and citation support before dissemination. Broader impacts and responsible-use considerations are discussed in Appendix I.

REPRODUCIBILITY STATEMENT

We describe the method and evaluation setup in Sections 3 and 4. Appendices B and C provide workflow and prompt displays, data-construction procedures, and validation details. Appendices D and E report reward-weight ablations, training hyperparameters, and computational resources. Appendices F and G describe the human-evaluation and LLM-judge protocols. We plan to release the MedCase-150K and MedCase-Bench datasets, together with the source code, data-construction scripts, evaluation prompts, metadata, and data splits, as described in Appendix J. The hospital cases used for expert evaluation are non-public.

REFERENCES

Show us the evidence for the value of medical ai. *Nature Medicine*, 32(4):1163–1163, Apr 2026. ISSN 1546-170X. doi: 10.1038/s41591-026-04389-4. URL https://doi.org/10.1038/s41591-026-04389-4.

Asma Ben Abacha, Wen-wai Yim, Yadan Fan, and Thomas Lin. An empirical study of clinical note generation from doctor-patient encounters. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 2291–2302, 2023.

Md Shamim Ahmed, Maja Dusanic, Moritz Nikolai Kirschner, Elisabeth Nyoungui, Jana Zschüntzsch, Lukas Galke Poech, and Richard Röttger. The provenance gap in clinical ai: Evidence-traceable temporal knowledge graphs for rare disease reasoning. *arXiv preprint arXiv:2604.17114*, 2026.

Rama Akkiraju, Anbang Xu, Deepak Bora, Tan Yu, Lu An, Vishal Seth, Aaditya Shukla, Pritam Gundecha, Hridhay Mehta, Ashwin Jha, et al. Facts about building retrieval augmented generation-based chatbots. *arXiv preprint arXiv:2407.07858*, 2024.

Rahul K Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, et al. Healthbench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*, 2025.

Yushi Bai, Jiajie Zhang, Xin Lv, Linzhi Zheng, Siqi Zhu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. Longwriter: Unleashing 10,000+ word generation from long context llms. *arXiv preprint arXiv:2408.07055*, 2024.

M. Buholayka, R. Zouabi, and A. Tadinada. The readiness of ChatGPT to write scientific case reports independently: A comparative evaluation between human and artificial intelligence. *Cureus*, 15(5): e39386, 2023. doi: 10.7759/cureus.39386.

Yan Cai, Linlin Wang, Ye Wang, Gerard De Melo, Ya Zhang, Yanfeng Wang, and Liang He. Medbench: A large-scale chinese benchmark for evaluating medical large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 17709–17717, 2024.

Ruike Cao, Shaojie Bai, Fugen Yao, Liang Dong, Jian Xu, and Li Xiao. Atpo: Adaptive tree policy optimization for multi-turn medical dialogue. *arXiv preprint arXiv:2603.02216*, 2026.

Haichao Chen, Zhengyun Zhao, Songchi Zhou, Shikai Hu, Jinyuan Wang, Ye Jin, Xianghong Jin, Yih Chung Tham, Xiaofei Wang, Weizhi Ma, et al. Rarearena: a comprehensive benchmark dataset unveiling the potential of large language models in rare disease diagnosis. *The Lancet Digital Health*, 8(2), 2026.

Jingyun Chen, Linghan Cai, Zhikang Wang, Yi Huang, Songhan Jiang, Shenjin Huang, Hongpeng Wang, and Yongbing Zhang. Pathagent: Toward interpretable analysis of whole-slide pathology images via large language model-based agentic reasoning. *arXiv preprint arXiv:2511.17052*, 2025a.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, and Benyou Wang. Towards medical complex reasoning with llms through medical verifiable problems. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 14552–14573, 2025b.

Xuanzhong Chen, Xiaohao Mao, Qihan Guo, Lun Wang, Shuyang Zhang, and Ting Chen. Rarebench: can llms serve as rare diseases specialists? In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 4850–4861, 2024.

Christopher Chiu, Silviu Pitis, and Mihaela van der Schaar. Simulating viva voce examinations to evaluate clinical reasoning in large language models. *arXiv preprint arXiv:2510.10278*, 2025.

Henry Cohen. How to write a patient case report. *American Journal of Health-System Pharmacy*, 63 (19):1888–1892, 2006.

Jean-Philippe Corbeil, Minseon Kim, Maxime Griot, Sheela Agarwal, Alessandro Sordoni, Francois Beaulieu, and Paul Vozila. Medriskeval: Medical risk evaluation benchmark of language models, on the importance of user perspectives in healthcare settings. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 5: Industry Track)*, pp. 513–524, 2026.

Emma Croxford, Yanjun Gao, Elliot First, Nicholas Pellegrino, Miranda Schnier, John Caskey, Madeline Oguss, Graham Wills, Guanhua Chen, Dmitriy Dligach, et al. Evaluating clinical ai summaries with large language models as judges. *npj Digital Medicine*, 8(1):640, 2025.

Yimin Deng, Zhenxi Lin, Yejing Wang, Guoshuai Zhao, Pengyue Jia, Zichuan Fu, Derong Xu, Yefeng Zheng, Xiangyu Zhao, Li Zhu, et al. Multidx: A multi-source knowledge integration framework towards diagnostic reasoning. *arXiv preprint arXiv:2604.24186*, 2026.

Yuexi Du, Jinglu Wang, Shujie Liu, Nicha C Dvornek, and Yan Lu. Care: Towards clinical accountability in multi-modal medical reasoning with an evidence-grounded agentic framework. *arXiv preprint arXiv:2603.01607*, 2026.

Reza Saadati Fard, Emmanuel Agu, Palawat Busaranuvong, Deepak Kumar, Shefalika Gautam, Bengisu Tulu, Diane Strong, and Lorraine Loretz. Ft-arm: Fine-tuned agentic reflection multimodal language model for pressure ulcer severity classification with reasoning. *arXiv preprint arXiv:2510.24980*, 2025.

Joel J Gagnier, Gunver Kienle, Douglas G Altman, David Moher, Harold Sox, David Riley, CARE group, et al. The care guidelines: consensus-based clinical case report guideline development. *Journal of clinical epidemiology*, 67(1):46–51, 2014.

Fatemeh Ghezloo, Mehmet Saygin Seyfioglu, Rustin Soraki, Wisdom O Ikezogwo, Beibin Li, Tejoram Vivekanandan, Joann G Elmore, Ranjay Krishna, and Linda Shapiro. Pathfinder: A multi-modal multi-agent system for medical diagnostic decision-making applied to histopathology. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23431–23441, 2025.

Google DeepMind. Gemma 4 model card. `https://ai.google.dev/gemma/docs/core/model_card_4`, 2026.

Shaowei Guan, Yu Zhai, Hin Chi Kwok, Jiawei Du, Xinyu Feng, Jing Li, Harry Qin, and Vivian Hui. Medpriv-bench: Benchmarking the privacy-utility trade-off of large language models in medical open-end question answering. *arXiv preprint arXiv:2603.14265*, 2026.

Yajie Vera He, Mohita Chowdhury, Jared Joselowitz, Aisling Higham, and Ernest Lim. Astrid-an automated and scalable triad for the evaluation of rag-based clinical question answering systems. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 16700–16716, 2025.

Tal Ifargan, Lukas Hafner, Maor Kern, Ori Alcalay, and Roy Kishony. Autonomous LLM-driven research from data to human-verifiable research papers. *arXiv preprint arXiv:2404.17605*, 2024.

Wisdom O Ikezogwo, Kevin Zhang, Mehmet Saygin Seyfioglu, Fatemeh Ghezloo, Linda Shapiro, and Ranjay Krishna. Medicalnarratives: Connecting medical vision and language with localized narratives. *arXiv preprint arXiv:2501.04184*, 2025.

Zongliang Ji, Yifei Sun, Andre Amaral, Anna Goldenberg, and Rahul G Krishnan. Can we generate portable representations for clinical time series data using llms? *arXiv preprint arXiv:2603.23987*, 2026.

Shuyang Jiang, Yusheng Liao, Zhe Chen, Ya Zhang, Yanfeng Wang, and Yu Wang. Meds^3: Towards medical slow thinking with self-evolved soft dual-sided process supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 31319–31327, 2026.

Songtao Jiang, Yuan Wang, Sibo Song, Tianxiang Hu, Chenyi Zhou, Bin Pu, Yan Zhang, Zhibo Yang, Yang Feng, Joey Tianyi Zhou, et al. Hulu-med: A transparent generalist model towards holistic medical vision-language understanding. *arXiv preprint arXiv:2510.08668*, 2025a.

Yixing Jiang, Kameron C Black, Gloria Geng, Danny Park, James Zou, Andrew Y Ng, and Jonathan H Chen. Medagentbench: a virtual ehr environment to benchmark medical llm agents. *Nejm Ai*, 2(9): AIdbp2500144, 2025b.

Roshan Kenia, Xiaoman Zhang, and Pranav Rajpurkar. Rex-mle: The autonomous agent benchmark for medical imaging challenges. *arXiv preprint arXiv:2512.17838*, 2025.

Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik S Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae W Park. Mdagents: An adaptive collaboration of llms for medical decision-making. *Advances in Neural Information Processing Systems*, 37: 79410–79452, 2024.

Yubin Kim, Hyewon Jeong, Chanwoo Park, Eugene Park, Haipeng Zhang, Xin Liu, Hyeonhoon Lee, Daniel McDuff, Marzyeh Ghassemi, Cynthia Breazeal, et al. Tiered agentic oversight: A hierarchical multi-agent system for healthcare safety. *arXiv preprint arXiv:2506.12482*, 2025.

Binxu Li, Tiankai Yan, Yuanting Pan, Jie Luo, Ruiyang Ji, Jiayuan Ding, Zhe Xu, Shilong Liu, Haoyu Dong, Zihao Lin, et al. Mmedagent: Learning to use medical tools with multi-modal agent. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 8745–8760, 2024a.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36: 28541–28564, 2023.

Junkai Li, Yunghwei Lai, Weitao Li, Jingyi Ren, Meng Zhang, Xinhui Kang, Siyu Wang, Peng Li, Ya-Qin Zhang, Weizhi Ma, et al. Agent hospital: A simulacrum of hospital with evolvable medical agents. *arXiv preprint arXiv:2405.02957*, 2024b.

Rumeng Li, Xun Wang, Dan Berlowitz, Jesse Mez, Honghuang Lin, and Hong Yu. Care-ad: a multi-agent large language model framework for alzheimer's disease prediction using longitudinal clinical notes. *npj Digital Medicine*, 8(1):541, 2025.

Shuyue S Li, Vidhisha Balachandran, Shangbin Feng, Jonathan S Ilgen, Emma Pierson, Pang W Koh, and Yulia Tsvetkov. Mediq: Question-asking llms and a benchmark for reliable interactive clinical reasoning. *Advances in Neural Information Processing Systems*, 37:28858–28888, 2024c.

Zheng Li, Jiayi Xu, Zhikai Hu, Hechang Chen, Lele Cong, Yunyun Wang, and Shuchao Pang. Medcorag: Interpretable hepatology diagnosis via hybrid evidence retrieval and multispecialty consensus. *arXiv preprint arXiv:2603.05129*, 2026.

Yusheng Liao, Shuyang Jiang, Yanfeng Wang, and Yu Wang. Reflectool: Towards reflection-aware tool-augmented clinical agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13507–13531, 2025.

Yusheng Liao, Chuan Xuan, Yutong Cai, Lina Yang, Zhe Chen, Yanfeng Wang, and Yu Wang. Agentehr: Advancing autonomous clinical decision-making via retrospective summarization. *arXiv preprint arXiv:2601.13918*, 2026.

Chi Liu, Derek Li, Yan Shu, Robin Chen, Derek Duan, Teng Fang, and Bryan Dai. Fleming-r1: Toward expert-level medical reasoning via reinforcement learning. *arXiv preprint arXiv:2509.15279*, 2025a.

Fenglin Liu, Zheng Li, Hongjian Zhou, Qingyu Yin, Jingfeng Yang, Xin Liu, Zhengyang Wang, Xianfeng Tang, Shiyang Li, Xiang He, et al. Dragent: Empowering large language models as medical agents for multi-hop medical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 15656–15668, 2025b.

Shengyuan Liu, Liuxin Bao, Qi Yang, Wanting Geng, Boyun Zheng, Chenxin Li, Wenting Chen, Houwen Peng, and Yixuan Yuan. Medsam-agent: Empowering interactive medical image segmentation with multi-turn agentic reinforcement learning. *arXiv preprint arXiv:2602.03320*, 2026.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 2511–2522, 2023.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024.

Junzhi Ning, Jiashi Lin, Yingying Fang, Wei Li, Jiyao Liu, Cheng Tang, Chenglong Ma, Wenhao Tang, Tianbin Li, Ziyan Huang, et al. Mmrarebench: A rare-disease multimodal and multi-image medical benchmark. *arXiv preprint arXiv:2604.10755*, 2026.

Jiazhen Pan, Che Liu, Junde Wu, Fenglin Liu, Jiayuan Zhu, Hongwei Bran Li, Chen Chen, Cheng Ouyang, and Daniel Rueckert. Medvlm-r1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 337–347. Springer, 2025.

D. S. Pinto, S. M. Noronha, G. Saigal, and R. M. Quencer. Comparison of an AI-generated case report with a human-written case report: Practical considerations for AI-assisted medical writing. *Cureus*, 16(5):e60461, 2024. doi: 10.7759/cureus.60461.

Pranav Pushkar Mishra, Mohammad Arvan, and Mohan Zalake. Teammedagents: Enhancing medical decision-making of llms through structured teamwork. *arXiv e-prints*, pp. arXiv–2508, 2025.

Tuoshi Qi, Shenshen Bu, Yingfei Xiang, and Zhiming Dai. Eviagent: Evidence-driven agent for radiology report generation. *arXiv preprint arXiv:2603.13956*, 2026.

Chuhan Qiao, Jianghua Huang, Daxing Zhao, Ziding Liu, Yanjun Shen, Bing Cheng, Wei Lin, and Kai Wu. Medconsultbench: A full-cycle, fine-grained, process-aware benchmark for medical consultation agents. *arXiv preprint arXiv:2601.12661*, 2026.

Pengcheng Qiu, Chaoyi Wu, Shuyu Liu, Yanjie Fan, Weike Zhao, Zhuoxia Chen, Hongfei Gu, Chuanjin Peng, Ya Zhang, Yanfeng Wang, et al. Quantifying the reasoning abilities of llms on clinical cases. *Nature Communications*, 16(1):9799, 2025.

Zhiyao Ren, Yibing Zhan, Baosheng Yu, Liang Ding, Pingbo Xu, and Dacheng Tao. Healthcare agent: eliciting the power of large language models for medical consultation. *npj Artificial Intelligence*, 1 (1):24, 2025.

David S Riley, Melissa S Barber, Gunver S Kienle, Jeffrey K Aronson, Tido von Schoen-Angerer, Peter Tugwell, Helmut Kiene, Mark Helfand, Douglas G Altman, Harold Sox, et al. Care guidelines for case reports: explanation and elaboration document. *Journal of clinical epidemiology*, 89: 218–235, 2017.

Daniel Philip Rose, Chia-Chien Hung, Marco Lepri, Israa Alqassem, Kiril Gashteovski, and Carolin Lawrence. Meddxagent: A unified modular agent framework for explainable automatic differential diagnosis. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13803–13826, 2025.

Samuel Schmidgall, Rojin Ziaei, Carl Harris, Eduardo Reis, Jeffrey Jopling, and Michael Moor. Agentclinic: a multimodal agent benchmark to evaluate ai in simulated clinical environments. *arXiv preprint arXiv:2405.07960*, 2024.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

Andrew Sellergren, Chufan Gao, Fereshteh Mahvar, Timo Kohlberger, Fayaz Jamil, Madeleine Traverse, Alberto Tono, Bashir Sadjad, Lin Yang, Charles Lau, et al. Medgemma 1.5 technical report. *arXiv preprint arXiv:2604.05081*, 2026.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Yiwen Song, Yale Song, Tomas Pfister, and Jinsung Yoon. Paperorchestra: A multi-agent framework for automated ai research paper writing. *arXiv preprint arXiv:2604.05018*, 2026.

Ivan Sviridov, Amina Miftakhova, Tereshchenko Artemiy Vladimirovich, Galina Zubkova, Pavel Blinov, and Andrey Savchenko. 3mdbench: Medical multimodal multi-agent dialogue benchmark. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 26625–26665, 2025.

Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. Medagents: Large language models as collaborators for zero-shot medical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 599–621, 2024.

Yuqi Tang, Jing Yu, Zichang Su, Kehua Feng, Zhihui Zhu, Libin Wang, Lei Liang, Qiang Zhang, Keyan Ding, and Huajun Chen. Clindef: A dynamic evaluation framework for large language models in clinical reasoning. *arXiv preprint arXiv:2512.23440*, 2025.

Kiril Vasilev, Alexandre Misrahi, Eeshaan Jain, Phil F Cheng, Petros Liakopoulos, Olivier Michielin, Michael Moor, and Charlotte Bunne. Mtbbench: A multimodal sequential clinical decision-making benchmark in oncology. *arXiv preprint arXiv:2511.20490*, 2025.

Ismael Villanueva Miranda, Zifan Gu, Donghan Yang, Kuroush Nezafati, Jingwei Huang, Peifeng Ruan, Xiaowei Zhan, Guanghua Xiao, and Yang Xie. Clinbench: A standardized multi-domain framework for evaluating large language models in clinical information extraction. *Advances in Neural Information Processing Systems*, 38, 2026.

Bichen Wang, Yixin Sun, Junzhe Wang, Hao Yang, Xing Fu, Yanyan Zhao, Si Wei, Shijin Wang, and Bing Qin. Care-bench: A benchmark of diverse client simulations guided by expert principles for evaluating llms in psychological counseling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 39378–39386, 2026a.

Rundong Wang, Wei Ba, Ying Zhou, Yingtai Li, Bowen Liu, Baizhi Wang, Yuhao Wang, Zhidong Yang, Kun Zhang, Rui Yan, et al. Qcagent: An agentic framework for quality-controllable pathology report generation from whole slide image. *arXiv preprint arXiv:2603.01647*, 2026b.

Ziyue Wang, Junde Wu, Linghan Cai, Chang Han Low, Xihong Yang, Qiaxuan Li, and Yueming Jin. Medagent-pro: Towards evidence-based multi-modal medical diagnosis via reasoning agentic workflow. *arXiv preprint arXiv:2503.18968*, 2025.

Yixuan Weng, Minjun Zhu, Guangsheng Bao, Hongbo Zhang, Jindong Wang, Yue Zhang, and Linyi Yang. Cycleresearcher: Improving automated research via automated review. *arXiv preprint arXiv:2411.00816*, 2024.

Kevin Wu, Eric Wu, Rahul Thapa, Kevin Wei, Angela Zhang, Arvind Suresh, et al. MedCaseReasoning: Evaluating and learning diagnostic reasoning from clinical case reports. *arXiv preprint arXiv:2505.11733*, 2025a.

Yicong Wu, Ting Chen, Irit Hochberg, Zhoujian Sun, Ruth Edry, Zhengxing Huang, and Mor Peleg. Lessons learned from evaluation of llm based multi-agents in safer therapy recommendation. *arXiv preprint arXiv:2507.10911*, 2025b.

Peng Xia, Jinglu Wang, Yibo Peng, Kaide Zeng, Zihan Dong, Xian Wu, Xiangru Tang, Hongtu Zhu, Yun Li, Linjun Zhang, et al. Mmedagent-rl: Optimizing multi-agent collaboration for multimodal medical reasoning. *arXiv preprint arXiv:2506.00555*, 2025.

Ziyan Xiao, Ruiyang Zhang, Yushi Feng, Lingting Zhu, Liang Peng, and Lequan Yu. A dynamic agent framework for large language model reasoning for medical and visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1143–1152, 2025.

Ran Xu, Yuchen Zhuang, Yishan Zhong, Yue Yu, Zifeng Wang, Xiangru Tang, Hang Wu, May D Wang, Peifeng Ruan, Donghan Yang, et al. Medagentgym: A scalable agentic training environment for code-centric reasoning in biomedical data science. *arXiv preprint arXiv:2506.04405*, 2025a.

Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, et al. Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. *arXiv preprint arXiv:2506.07044*, 2025b.

Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Shengran Hu, Chris Lu, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist-v2: Workshop-level automated scientific discovery via agentic tree search. *arXiv preprint arXiv:2504.08066*, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Cheng Yang, Hui Jin, Xinlei Yu, Zhipeng Wang, Yaoqun Liu, Fenglei Fan, Dajiang Lei, Gangyong Jia, Changmiao Wang, and Ruiquan Ge. Lungnoduleagent: A collaborative multi-agent system for precision diagnosis of lung nodules. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 29793–29801, 2026.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

Ling Yue, Sixue Xing, Jintai Chen, and Tianfan Fu. Clinicalagent: Clinical trial multi-agent system with large language model-based reasoning. In *Proceedings of the 15th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pp. 1–10, 2024.

Jie Zhang, Mengya Xu, Yiwei Wang, and Qi Dou. Csap-assist: Instrument-agent dialogue empowered vision-language models for collaborative surgical action planning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 139–148. Springer, 2025a.

Xiao Yu Cindy Zhang, Carlos R Ferreira, Francis Rossignol, Raymond T Ng, Wyeth Wasserman, and Jian Zhu. Casereportbench: An llm benchmark dataset for dense information extraction in clinical case reports. *Proceedings of machine learning research*, 287:527, 2025b.

Weike Zhao, Chaoyi Wu, Yanjie Fan, Pengcheng Qiu, Xiaoman Zhang, Yuze Sun, Xiao Zhou, Shuju Zhang, Yu Peng, Yanfeng Wang, et al. An agentic system for rare disease diagnosis with traceable reasoning. *Nature*, 651(8106):775–784, 2026.

Zhengyun Zhao, Qiao Jin, Fangyuan Chen, Tuorui Peng, and Sheng Yu. A large-scale dataset of patient summaries for retrieval-based clinical decision support systems. *Scientific data*, 10(1):909, 2023.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

Xiaoquan Zhi, Hongke Zhao, Likang Wu, Chuang Zhao, and Hengshu Zhu. Reinventing clinical dialogue: Agentic paradigms for llm enabled healthcare communication. *arXiv preprint arXiv:2512.01453*, 2025.

Yishan Zhong, Wenqi Shi, J Ben Tamo, Micky C Nnamdi, Yining Yuan, and May Dongmei Wang. Multi-agent llm reasoning for clinical procedure sequencing from high-granularity ehr data. In *Proceedings of the 16th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pp. 1–12, 2025a.

Zhusi Zhong, Yuli Wang, Jing Wu, Wen-Chi Hsu, Vin Somasundaram, Lulu Bi, Shreyas Kulkarni, Zhuoqi Ma, Scott Collins, Grayson Baird, et al. Vision-language model for report generation and outcome prediction in ct pulmonary angiogram. *NPJ digital medicine*, 8(1):432, 2025b.

Shuang Zhou, Wenya Xie, Jiaxi Li, Zaifu Zhan, Meijia Song, Han Yang, Cheyenna Espinoza, Lindsay Welton, Xinnie Mai, Yanwei Jin, et al. Automating expert-level medical reasoning evaluation of large language models. *NPJ Digital Medicine*, 9(1):34, 2025.

Yinghao Zhu, Ziyi He, Haoran Hu, Xiaochen Zheng, Xichen Zhang, Zixiang Wang, Junyi Gao, Liantao Ma, and Lequan Yu. Medagentboard: Benchmarking multi-agent collaboration with conventional methods for diverse medical tasks. *arXiv preprint arXiv:2505.12371*, 2025.

Yangyang Zhuang, Wenjia Jiang, Jia-Yu Zhang, Ze Yang, Joey Tianyi Zhou, and Chi Zhang. Learning to be a doctor: Searching for effective medical agent architectures. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 6996–7005, 2025.

APPENDIX

A Extended Related Work: Relevant Benchmarks **19**

B Prompt and Workflow Displays **20**

C Data Construction and Validation **23**

C.1 Human Verification of Clinical Atoms and Rarity Labels 23

C.2 Corpus Statistics and Data Construction Views 23

D GRPO Reward-Weight Sensitivity **27**

E Training and Compute Resources **28**

F Clinical Evaluation and User Study Details **30**

F.1 Hospital-Based Expert Evaluation 30

F.2 LLM–Clinician Agreement 31

F.3 Expert-Reviewer Blind A/B Test 32

F.4 Web-Based User Studies and Feedback 32

G LLM-Judged Blind A/B Comparison Across Baselines **34**

G.1 Judge Robustness and Self-Preference 35

H Example Case Reports Generated by MedCaseAgent **36**

I Broader Impacts **40**

J Release and LLM Usage Disclosure **41**

A Extended Related Work: Relevant Benchmarks

Recent benchmarks for clinical agents can be grouped by what they evaluate. One group studies clinical reasoning and decision making, including HealthBench (Arora et al., 2025), MedAgentBench (Jiang et al., 2025b), MedConsultBench (Qiao et al., 2026), ClinDEF (Tang et al., 2025), VivaBench (Chiu et al., 2025), RareBench (Chen et al., 2024), and MedAgentBoard (Zhu et al., 2025). A second group focuses on multimodal, imaging-centered, or sequential clinical decision-making evaluation, such as ReX-MLE (Kenia et al., 2025), MedicalNarratives (Ikezogwo et al., 2025), AgentClinic (Schmidgall et al., 2024), and MTBBench (Vasilev et al., 2025). A third group targets clinical workflow, dialogue, and system-level behavior, including MedBench (Cai et al., 2024), 3md-bench (Sviridov et al., 2025), Care-bench (Wang et al., 2026a) and MMRareBench (Ning et al., 2026). Other benchmarks examine safety, privacy, and structured extraction, such as MedPriv-Bench (Guan et al., 2026), MedRiskEval (Corbeil et al., 2026), and ClinBench (Villanueva Miranda et al., 2026). PMC-Patients (Zhao et al., 2023) provides patient summaries for patient and literature retrieval, while RareArena (Chen et al., 2026) benchmarks rare-disease diagnosis. Despite this growing benchmark ecosystem, most existing settings still focus on question answering, information extraction, summarization, diagnosis, report evaluation, or simulated interaction. Clinical case report generation brings these abilities together in a single task. A strong system must interpret fragmented multimodal evidence, reconstruct the patient course, link findings to prior literature, and write a rigorous scientific narrative.

B PROMPT AND WORKFLOW DISPLAYS

This appendix gathers the prompt-facing displays for the agentic report-generation pipeline: clinical-atom compression, planner organization, and the prompt stack used across the workflow. Figure 6 shows how polished source prose is reduced to patient-specific atoms before training and evaluation. Figure 7 expands the planner stage that connects atoms, images, retrieved context, citations, and figure placement. Figure 8 shows how instructions and evidence are separated across the agent workflow so that downstream drafting remains grounded in case-specific inputs.

Prompt and Output JSON for Clinical Atom Extraction

Prompt

```
Role: Senior Clinical Data Architect.
Task: Extract atomic clinical facts from the provided text into a strict structured JSON format.

JSON SCHEMA:
{
  "history": ["past medical history", "comorbidities", "demographics"],
  "presentation": ["symptoms", "timeline of current illness", "physical exam findings"],
  "diagnostics": ["lab results", "imaging findings", "pathology", "immunohistochemistry"],
  "diagnosis": ["definitive primary disease", "differential diagnoses"],
  "management": ["surgical interventions", "medications", "treatments"],
  "outcome": ["patient-specific results", "follow-up status"]
}

STRICT RULES:
1. ISOLATE THE DISEASE: Explicitly identify the primary disease/condition diagnosed and place it in the "diagnosis" array. If multiple conditions
or differential diagnoses are mentioned, list them as separate strings.
2. ATOMIC FACTS: Break down complex paragraphs into single, standalone concepts (e.g., "Histiocytic cells positive for CD68").
3. OUTCOME RESTRICTION: The "outcome" array must ONLY contain the patient's specific end-result or current status (e.g., "Patient is stable and
being followed-up").
4. NO FLUFF: DELETE any meta-commentary (e.g., "this report highlights", "this is the first case of") and do not include general epidemiological
facts about the disease.
5. COMPREHENSIVE EXTRACTION: Extract accurate facts from all provided text and metadata.
6. OUTPUT FORMAT: Return ONLY the raw JSON object. Do not include markdown code blocks (like ```json), conversational text, or explanations.

TEXT AND METADATA:
{raw_text}
```

Output JSON

```
{
  "history": [
    "59-year-old female patient",
    "Total hysterectomy 25 years prior due to fibroma",
    "Chronic seromucous maxillary sinusitis treated with antibiotics and endoscopic sinus
surgery",
    "Refractory arterial hypertension diagnosed at age 55",
    "Renal failure diagnosed at age 55",
    "Right renal artery stenosis",
    "Left renal artery occlusion",
    "Stenting of left renal artery",
    "Right aortorenal bypass surgery",
    "Chronic headaches for 5 years",
    "Chronic asthenia for 5 years",
    "Patient is a woman"
  ],
  "presentation": [
    "Sudden confusion",
    "Loss of vision",
    "Cortical blindness",
    "Anosognosia for blindness",
    "Amnesia",
    "Topographical disorientation suggestive of Dejerine-Botcazou syndrome",
    "No pyramidal signs",
    "No cerebellar ataxia",
    "No cranial nerve involvement",
    "Persistent chronic headaches",
    "Grade I hypertensive retinopathy"
  ],
  "diagnostics": [
    "Bilateral cerebral posterior infarction",
    "Hypodensity in territory of right postero-inferior cerebellar artery",
    "Severe narrowing of basilar artery (closed to 90%)",
    "Mild stenosis of M1 and M2 segments of left middle cerebral artery",
    "Severe stenosis of basilar artery confirmed by Transcranial Doppler",
    "Normal electrocardiogram",
    "Normal 24-hour electrocardiography",
    "Hypertrophic cardiopathy with severe left ventricular hypertrophy",
    "Circumferential dilatation of left atrium",
    "No valvulopathy",
    "Heterogeneous thickening of interatrial septum (1.5 cm)",
    "Hyperechogeneous thickening next to right atrium (1.9 x 1.3 cm)",
    "Right periatrial infiltration confirmed by cardiac CT angiography and MRI",
    "Thickening of wall of whole aortic arch with tardive enhancement of media",
    "Pericardial effusion",
    "Contrast enhancement on right and left atria on 18FDG PET",
    "Contrast enhancement on whole aorta, iliac arteries, and femoral arteries on 18FDG
PET",
    "Hyperleukocytosis (13,000/mm3)",
    "Neutrophilia (8890 neutrophils/mm3)",
    "Elevated C-reactive protein (24 mg/L)",
    "No monoclonal gammopathy",
    "Normal viral and bacterial serologies",
    "Normal immunological tests (antinuclear, anticardiolipin, anti-beta2-glycoprotein 1
antibodies)",
    "Normal metabolic tests (cholesterol, triglycerides, glycosylated hemoglobin,
homocysteine, TSH, transaminases)",
    "Normal cerebrospinal fluid analysis (1 white cell/mm3, protein 46 mg/dL, normal
glucose)",
    "Normal bone marrow biopsy",
    "Bilateral and symmetric cortical osteosclerosis of diaphyseal and metaphyseal regions
of lower limbs on plain radiographs",
    "Symmetric tracer uptake by long bones of lower extremities, maxillary, and frontal
bones on 99mTechnetium bone scintigraphy",
    "Xanthogranulomatous infiltrate composed of foamy histiocytes with fibrosis on
maxillary sinus biopsy",
    "Negative stains for CD1a and S-100 on maxillary sinus biopsy",
    "Positive immunostaining for CD68 and CD1a-negative on ECD diagnostic criteria"
  ],
  "diagnosis": [
    "Erdheim-Chester disease (ECD)",
    "Vertebrobasilar stroke secondary to infiltration and severe stenosis of the basilar
artery",
    "Hypertrophic cardiopathy",
    "Renovascular hypertension",
    "Diabetes insipidus"
  ],
  "management": [
    "Interferon-alpha therapy (3 million IU 3 times weekly, increased to 9 million IU 3
times weekly)",
    "Clopidogrel (75 mg/d)",
    "Desmopressin for diabetes insipidus",
    "Antibiotic therapy for sinusitis",
    "Endoscopic sinus surgery for sinusitis",
    "Stenting of left renal artery",
    "Right aortorenal bypass surgery"
  ],
  "outcome": [
    "Improvement of cognitive function, particularly visual function and memory",
    "Improvement of headaches",
    "Reduction of basilar artery stenosis from 90% to 50% after 3 months of interferon-
alpha",
    "Total recanalization of basilar artery 2 years after onset of treatment",
    "Neurological examination stable after 4 years",
    "Development of diabetes insipidus successfully treated with desmopressin"
  ]
}
```

Figure 6: Clinical atom compression example. A polished source case report is compressed into fragmented, patient-specific atoms that resemble real-world shorthand notes. The prompt-output view illustrates how background prose and final diagnostic exposition are removed while preserving clinical history, findings, tests, interventions, outcomes, and image-relevant details.

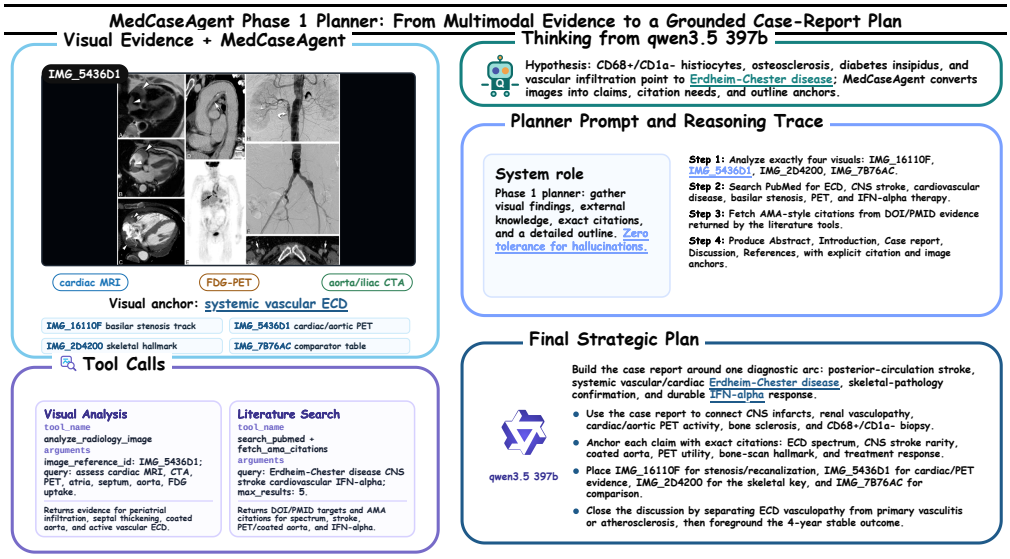


Figure 7: Detailed view of the MedCaseAgent planning phase. The planner organizes fragmented multimodal case evidence, retrieves relevant medical context, and produces a structured plan that guides downstream report writing and editing.

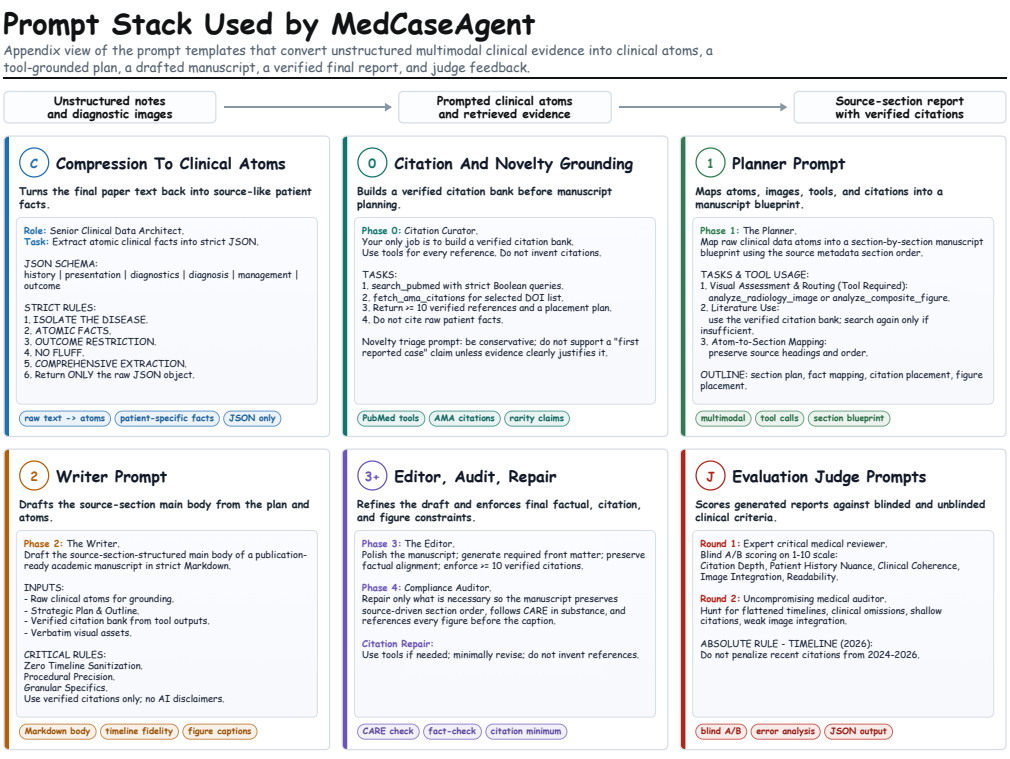


Figure 8: Prompt stack used by MedCaseAgent across the agentic report-generation workflow. The stack separates role instructions, case-specific evidence, retrieved context, output constraints, and audit requirements so that each phase receives grounded inputs for clinical case report synthesis.

C Data Construction and Validation

C.1 Human Verification of Clinical Atoms and Rarity Labels

During MedCase-Bench construction, 23 PhD researchers and domain experts with medical or biomedical backgrounds manually reviewed all 632 candidate cases. Each case was independently reviewed by at least two experts, with pre-adjudication judgments retained. The review proceeded in two stages: clinical-atom fidelity, followed by rarity/novelty-label validity for the cases retained after the first stage. Reviewers excluded cases that failed verification; they did not edit the compressed atoms or relabel the rarity categories.

For atom fidelity, reviewers checked each patient-specific fact against the source report for factual alterations, unsupported additions, temporal errors, and potentially misleading ambiguities. For rarity/novelty labels, they checked the source evidence against the predefined `rare_disease`, `unprecedented_presentation`, `atypical_presentation`, and `standard_case` categories.

Table 5: Sequential human verification of clinical atoms and rarity/novelty labels during MedCase-Bench construction. Cases failing either stage were excluded rather than edited or relabeled; exclusion and retention percentages use the number reviewed at that stage. Each case was reviewed by at least two reviewers. Agreement is measured before adjudication.

Review stage	**Reviewed (N)**	**Excluded n (%)**	**Retained n (%)**	**Pre-adjudication agreement**
Clinical-atom fidelity	632	30 (4.75%)	602 (95.25%)	94%
Rarity/novelty label	602	18 (2.99%)	584 (97.01%)	87%

Table 5 summarizes the sequential screening. Atom-fidelity review excluded 30 of 632 candidates, retaining 602 (95.25%); rarity-label review excluded 18 of these 602 cases, retaining 584 (97.01%). The final MedCase-Bench therefore contains 584 cases, with an overall retention rate of 92.41% (584/632). Pre-adjudication agreement was 94% for atom fidelity and 87% for rarity labels. This procedure verifies fidelity to the published source; comparison with original hospital records is described separately in Appendix F.1.

The disease-specific subsets in Table 1 are sampled separately from MedCase-150K: each category uses 100 cases, except Erdheim–Chester disease (ECD), which uses all 53 available cases in that corpus.

C.2 Corpus Statistics and Data Construction Views

This section visualizes the corpus-level quality checks behind MedCase-150K and MedCase-Bench. Figure 9 summarizes the rarity labels assigned during data construction. Figure 10 shows the temporal coverage used for the post-January-2026 evaluation split. Figure 11 samples the visual modalities that make the benchmark multimodal. Figure 12 checks whether the corpus is concentrated in a few journals or distributed across many source venues.

Rarity and novelty classification. Rare disease status is not always a stable binary property: the prevalence, clinical significance, and publication novelty of a condition can change across time, geography, and medical context. For publications identified as case reports, we assign one of four rarity labels. `rare_disease` means that the underlying disease or condition is epidemiologically rare. `unprecedented_presentation` means that the report describes a first-of-its-kind or completely novel manifestation. `atypical_presentation` means that a known disease appears in an uncommon form, location, demographic group, or clinical context. `standard_case` means that the report is a useful teaching case but does not contain a major rare-disease or novel-presentation element. For stratified corpus views, `rare_disease`, `unprecedented_presentation`, and `atypical_presentation` are grouped as rare or novelty-bearing case reports, while `standard_case` is retained as the non-rare/non-novel comparison group.

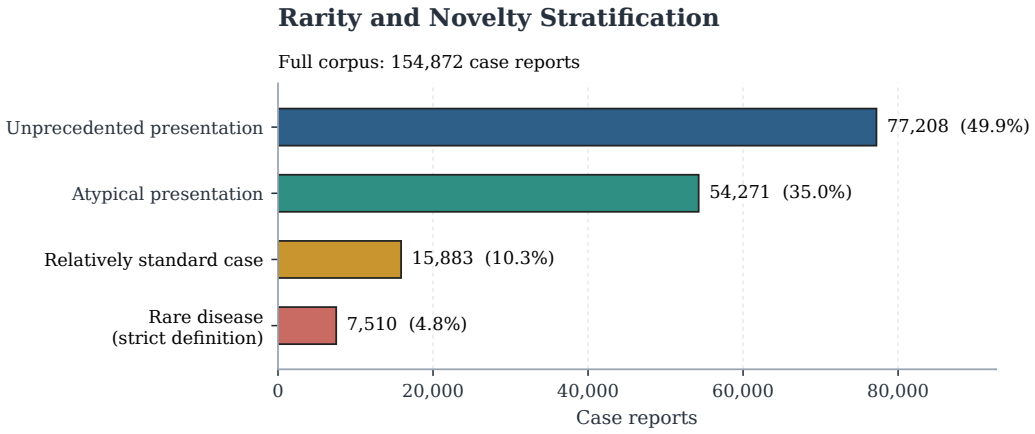


Figure 9: Rarity and novelty stratification for the curated case-report corpus. Bars summarize rarity labels assigned during data construction, separating rare diseases in the strict epidemiological sense, unprecedented presentations, atypical presentations, and relatively standard cases.

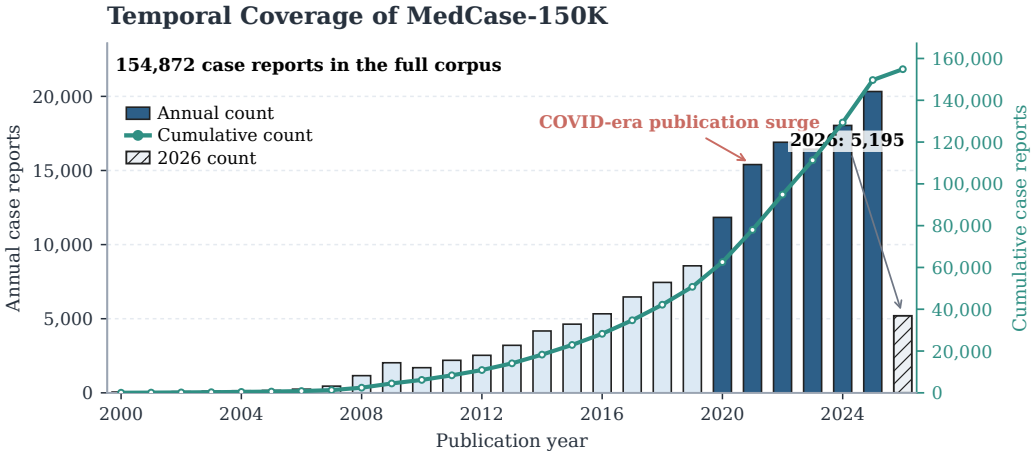


Figure 10: Temporal coverage of the MedCase-150K case-report corpus. Bars show annual case-report counts and the line shows cumulative coverage.



Figure 11: Representative visual complexity in MedCase-Bench. Rows group examples by imaging or figure modality, illustrating the breadth of radiology, clinical photography, ophthalmology, dentistry, pathology, and structured graphical evidence that the benchmark requires multimodal agents to interpret.

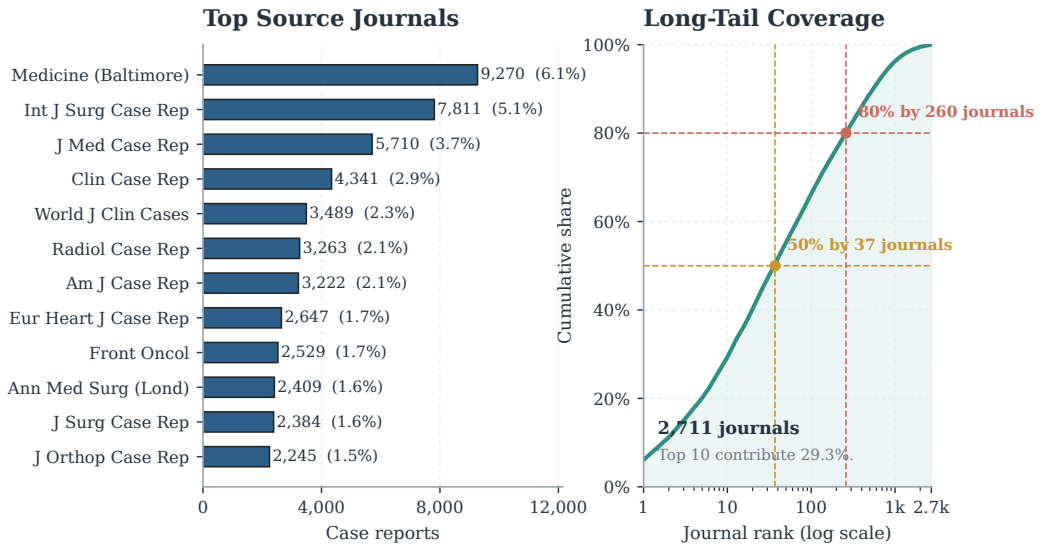


Figure 12: Source breadth across journals. The left panel reports the leading source venues, while the right panel shows that coverage remains broadly distributed across the long tail of journals.

D GRPO REWARD-WEIGHT SENSITIVITY

Reward design and weight selection. Each of the six reward components is bounded to $[0, 1]$ before aggregation, and the default training configuration uses deterministic rewards without invoking a live LLM judge. The default weights form a two-tier heuristic based on the task objectives. Exact DOI retrieval, evidence support, clinical-atom/section coverage, and the overall quality proxy represent evidence retrieval, evidential validity, coverage of patient facts, and plan quality, respectively, and each receives 0.20. Tool-schema compliance and structural/image-placement constraints are necessary for valid outputs but tend to saturate early; each receives 0.10 so that formatting rewards do not outweigh clinical content and evidence.

Experimental setup. We evaluate sensitivity by repeating GRPO from the same SFT checkpoint under three alternative normalized weight profiles, holding the training data, optimization budget, sampling configuration, and evaluation procedure fixed. Uniform assigns equal weight to all six components; Retrieval-heavy increases exact DOI retrieval and evidence support; and Quality/image-heavy increases the quality proxy and image placement. A separate No-DOI ablation removes the source-DOI reward and renormalizes the remaining weights. In the component order $[\mathrm{schema}, \mathrm{doi}, \mathrm{evidence}, \mathrm{coverage}, \mathrm{image}, \mathrm{quality}]$, the five evaluated profiles are

$$\begin{aligned}
\mathbf{w}_{\text{Default}} &= [0.10,\ 0.20,\ 0.20,\ 0.20,\ 0.10,\ 0.20],\\
\mathbf{w}_{\text{Uniform}} &= [\tfrac{1}{6},\ \tfrac{1}{6},\ \tfrac{1}{6},\ \tfrac{1}{6},\ \tfrac{1}{6},\ \tfrac{1}{6}],\\
\mathbf{w}_{\text{Retrieval-heavy}} &= [0.10,\ 0.30,\ 0.30,\ 0.10,\ 0.10,\ 0.10],\\
\mathbf{w}_{\text{Quality/image-heavy}} &= [0.10,\ 0.10,\ 0.10,\ 0.10,\ 0.30,\ 0.30],\\
\mathbf{w}_{\text{No-DOI}} &= [0.125,\ 0,\ 0.25,\ 0.25,\ 0.125,\ 0.25].
\end{aligned}$$

Table 6: GRPO reward-weight sensitivity from the same SFT checkpoint, with training data, optimization budget, sampling configuration, and evaluation procedure held fixed. Bold indicates the best result in each column.

Weight setting	Atom coverage	Image placement	Novelty recognition	Report quality	Average
Default	**86.30**	74.12	59.08	**67.23**	**71.68**
Uniform	85.08	73.95	58.21	66.47	70.93
Retrieval-heavy	84.79	72.49	**60.02**	66.48	70.95
Quality/image-heavy	83.25	**74.63**	57.85	66.86	70.65
No DOI reward ($w_{\mathrm{doi}} = 0$)	84.26	73.91	55.78	65.37	69.83

Results. Table 6 reports the average of atom coverage, image placement, novelty recognition, and report quality. The default weights achieve the highest average (71.68), atom coverage, and report quality. Retrieval-heavy leads novelty recognition, while Quality/image-heavy leads image placement. Across the four profiles retaining the source-DOI reward, the averages range from 70.65 to 71.68, a spread of 1.03 points.

Removing the source-DOI reward. To assess the contribution of R_{doi}, the No-DOI profile sets $w_{\mathrm{doi}} = 0$ and renormalizes the default weights of the remaining components. Its four-metric average is 69.83, compared with 71.68 for Default. Novelty recognition decreases by 3.30 percentage points (59.08% to 55.78%), and report quality decreases by 1.86 points (67.23 to 65.37). These results support a contribution of the source-DOI reward to novelty recognition and report quality under the evaluated training configuration.

E TRAINING AND COMPUTE RESOURCES

The tuned MedCaseAgent planner reported in the experiments is initialized from Qwen3.5-9B and trained with a two-stage SFT-to-GRPO recipe using the SWIFT training stack. The planner training set contains 3,000 phase-1 planning trajectories derived from the MedCase-150K training split. These trajectories are generated by a Qwen3.5-397B-A17B teacher planner: for each case, the teacher reads the compressed clinical atoms and mapped images, issues image-analysis and literature-search actions, records public tool observations, builds an evidence table with DOI-linked citations, and writes the final section-level case-report plan. The data-preparation step converts each teacher trace into SWIFT agent messages and removes private thinking fields. For the GRPO and evaluation records, source DOI identifiers are kept only as scorer-side labels; if search hits the source article for a case, its title, abstract, and article text are masked from policy-visible observations, while non-source PubMed records remain available as retrieved support.

We train LoRA adapters rather than full model weights, targeting the attention and MLP projection modules (q, k, v, o, gate, up, and down projections) with rank 16, alpha 32, and dropout 0.05. Training uses 4-bit bitsandbytes quantization, bfloat16 compute, and gradient checkpointing.

Item	Setting
Base model	Qwen3.5-9B
Trajectory teacher	Qwen3.5-397B-A17B
Training data	3,000 teacher-generated planning trajectories from the training split
Hardware	8 NVIDIA H100 GPUs, 80GB memory each
Wall-clock time	Approximately one day for the final planner adaptation run
Estimated compute (planner adaptation only)	Approximately 192 H100 GPU-hours
SFT objective	Teacher-forced planner trajectory learning
SFT effective batch	128 examples (8 GPUs $\times$ batch 1 $\times$ accumulation 16)
SFT epochs	2
SFT learning rate	2.0×10^{-5} with cosine decay and 0.03 warmup ratio
GRPO rollout group	4 candidate planner trajectories per prompt
GRPO effective prompt batch	64 prompts per update (8 GPUs $\times$ batch 1 $\times$ accumulation 8)
GRPO maximum steps	500 optimizer updates
GRPO sampling	Temperature 0.8, top-p 0.95, KL coefficient $\beta = 0.02$
GRPO learning rate	5.0×10^{-6} with cosine decay and 0.03 warmup ratio
GRPO rewards	Weighted sum of schema, DOI retrieval, evidence support, atom coverage, image placement, and quality proxy rewards

The compute estimate covers the reported SFT+GRPO LoRA adaptation run and does not include teacher-model inference, preliminary prompt or tool development, or inference-only evaluation.

The GRPO stage follows the configured 500-step cap and scores each sampled trajectory with

$$R(\tau) = 0.10R_{\text{schema}} + 0.20R_{\text{doi}} + 0.20R_{\text{evidence}} + 0.20R_{\text{coverage}} + 0.10R_{\text{image}} + 0.20R_{\text{quality}}, \quad (3)$$

where all component scores are clipped to $[0, 1]$ before aggregation. The default training path uses deterministic rewards and does not call a live judge. The reward components and weights used by the planner adaptation are:

Component	Weight	Reward signal
Tool schema	0.10	Valid tool names, required arguments, valid image identifiers for image-analysis calls, DOI-bearing citation-fetch calls, and non-duplicate tool calls.
Exact DOI retrieval	0.20	Recovery of scorer-only source DOI labels when available, with source title, abstract, and article text withheld from policy-visible retrieval context.
Evidence support	0.20	DOI format validity, citation-anchor validity, and overlap between generated citations and non-source retrieved support DOIs.
Plan coverage	0.20	Coverage of clinical atom groups and required report sections, reasonable plan length, and plan-like formatting.
Structure and images	0.10	Parse success, required-section coverage, placement of provided image identifiers, and citation-anchor validity.
Quality proxy	0.20	Deterministic blend of structure, evidence, coverage, explicit gap awareness, and avoiding unsupported citation-from-memory patterns; a precomputed judge score can replace this proxy when supplied.

F Clinical Evaluation and User Study Details

Section 4.4 summarizes the human evaluation. This appendix describes hospital-case collection and expert assessment, differences between raw clinical records and report-derived atoms, paired LLM–clinician agreement, the blinded A/B study, and the web-based user studies.

F.1 Hospital-Based Expert Evaluation

Case collection. We retrospectively assembled 30 original, non-public, de-identified multimodal cases from 10 hospitals through an ongoing case-report effort coordinated by the authors' institution under existing IRB approval. Each case involved a rare or clinically complex presentation and included longitudinal clinical records, multimodal evidence, and a corresponding report prepared and reviewed by the clinical team. These reports included both published articles and complete unpublished manuscripts prepared for submission. Cases were selected for sufficiently complete records and a matched clinician-written report; consolidating records across hospitals and care episodes required clinical coordination and manual review. All materials were de-identified before analysis.

Inputs and source verification. For each patient, MedCaseAgent generated one report directly from the raw clinical records (E_{raw}) and another from report-derived clinical atoms (E_{atom}). The atoms were obtained by compressing the clinician-written manuscript, rather than by directly extracting facts from the raw clinical files; images were supplied separately without captions or explicit atom-level links. Experts compared both generated reports with the clinician-written report and used the original records to verify patient facts. Raw records contained redundancy, irregular timelines, uncertainty, and cross-source conflicts, requiring filtering, deduplication, source reconciliation, and timeline reconstruction. Report-derived atoms retained the core history, examinations, treatment, and follow-up in a more concise form, but compressed some source context.

Expert assessment. Fourteen clinical experts assessed the 30 cases, with every case independently reviewed by three experts. Experts assessed overall report quality, patient-timeline reconstruction, and omission control on a 1–10 scale, with higher scores indicating better performance, while checking the reports against the source records. Table 4 reports the mean scores for raw-record generation, atom-based generation, and clinician-written reports.

Raw-record versus atom-based generation. In the order of overall quality, patient timeline, and omission control, mean scores were 7.4/7.5/7.8 for raw-record generation, 7.6/7.6/7.6 for atom-based generation, and 7.5/7.8/7.9 for clinician-written reports. Atom-based generation received slightly higher quality and timeline scores than raw-record generation, while raw records provided better omission control. Both input forms reach quality on par with clinician-written reports, supporting atoms as a controlled benchmark input whose results carry over to real hospital records.

Same-patient input comparison. Table 7 illustrates the input differences using a de-identified pediatric cerebral visual impairment case. Its original source package comprised two DOCX files, three PDFs totaling 76 pages, and seven standalone PNG images. The corresponding clinician-written manuscript contained four embedded figures. The comparison shows how report-derived atoms retain central patient facts while compressing temporal context, source provenance, conflicting records, and acquisition differences. The model must still associate images, organize evidence, retrieve literature, and construct a report, while raw inputs additionally require document filtering and cross-source reconciliation.

Table 7: Same-patient example of atomization for visual dysfunction following encephalitis. The atoms were derived from a clinician-written manuscript, not directly from the original clinical files. The abbreviated comparisons show source context compressed by the corresponding report-derived atoms.

Raw documentation	Corresponding report-derived atom	Information compressed
Dated entries separately record fever, marrow findings, impaired consciousness, visual loss, and later motor milestones.	Visual dysfunction persisting for over four months after EBV-associated encephalitis.	The central presentation is retained, while earlier events, care transitions, and temporal context are compressed.
Visual findings occur in narrative text and photographed reports. Several BAEP, P-VEP, and OCT screenshots contain demographic details inconsistent with the index patient.	P-VEP showed bilateral NPN waveforms with prolonged latencies; fundoscopy showed no obvious structural abnormality.	Summarized findings are retained, while provenance and possible wrong-patient contamination may be lost.
A summary specifies occipital rTMS at 90–100% motor threshold and 5–6 Hz. The contemporaneous plan records 20% intensity, a 25% threshold, 5 Hz, and 750 stimuli.	Occipital rTMS at 90–100% resting motor threshold, 5–6 Hz, and 20 minutes per day.	A single-version atom may obscure conflicting treatment parameters across sources.
Left/right optic-radiation streamline counts are 20/496, 314/21, and 483/10 across scans. The final diffusion scan changes from 3 T to 7 T.	Left optic-radiation fiber count increased, whereas the right decreased.	A direction-only atom may omit acquisition differences between scans.

F.2 LLM–Clinician Agreement

Paired calibration. We report agreement with clinical experts separately from the LLM-judged comparison of baseline systems. The alignment statistics below come from a calibration on 10 hospital cases: five clinical experts each reviewed six cases, giving 30 clinician–case assessments, with three independent reviews per case. Each assessment was paired with an LLM-judge assessment of the same output under a shared rubric. These 30 paired assessments are distinct from the aggregate results of the expanded 30-case, 14-expert hospital evaluation in Appendix F.1.

Agreement criteria. Factuality, clinical coherence, omission, and citation accuracy were scored on 10-point scales. A pair agreed when the absolute difference between the LLM and clinician scores was at most two points. Hallucination used a binary rule: any hallucinated content constituted a failure, and no hallucination constituted a pass. Table 8 reports agreement counts and the correlation summaries for the scored dimensions; the hallucination row reports exact agreement.

Chance-corrected agreement. Table 8 adds quadratic-weighted Cohen's κ and two-way, absolute-agreement, single-measure ICC(A,1), computed between the paired LLM and clinician score columns. These are descriptive point estimates over the 30 assessments; repeated cases and clinicians are not modeled, and no cluster-adjusted intervals are reported. The previously reported agreement counts and correlation summaries are retained. Both added coefficients are undefined for hallucination because all ratings occupy the same category.

Agreement was observed in 23/30 pairs for factuality, 24/30 for clinical coherence, 26/30 for omission, and 24/30 for citation accuracy. For hallucination, both raters assigned a pass in all 30 pairs, giving exact agreement of 1.00. This alignment supports using the LLM judge ensemble for large-scale comparison across baselines, complemented by clinician evaluation.

Table 8: LLM–clinician agreement in the initial paired calibration (30 clinician–case assessments). Agreement on scored dimensions allows an absolute difference of at most two points on a 10-point scale. Hallucination uses binary pass/fail ratings and exact agreement.

Evaluation dimension	Agreement	Reported correlation / exact agreement	Weighted κ	ICC
Factuality	23/30	0.76 ± 0.06	0.74	0.74
Clinical coherence	24/30	0.78 ± 0.05	0.76	0.77
Omission	26/30	0.71 ± 0.06	0.69	0.69
Hallucination	30/30	1.00	N/A	N/A
Citation accuracy	24/30	0.74 ± 0.06	0.73	0.74

ICC: two-way absolute agreement, single measure, ICC(A,1). N/A: undefined because both rating columns are constant.

F.3 Expert-Reviewer Blind A/B Test

Because claims about medical AI value require evidence that goes beyond technical performance alone ("Show us the evidence for the value of medical AI", 2026), we tested whether automatic gains translate into real manuscript-review value through a blinded A/B study. In this study, 32 expert clinicians (with at least 5 years of review experience in clinical case reports) each reviewed five paired cases, yielding 160 pairwise judgments. In each pair, the clinician saw the same de-identified case context and images, a MedCaseAgent-Max report and the corresponding published reference in randomized order, then judged authorship and selected the stronger case report. We summarize source-identification consistency at the reviewer level by counting clinicians who correctly identify the generated report in at most three of five pairs. Under this criterion, 26 of 32 clinicians (81.3%) did not consistently identify report sources across their five assigned pairs; reviewers also more often selected MedCaseAgent as the stronger manuscript.

Study design. The paired cases are drawn from the post-January-2026 evaluation split, and report source labels are hidden from reviewers. The interface asks two primary forced-choice questions—which report appears to have been generated by an LLM, and which report the clinician prefers overall—to separate authorship indistinguishability from practical clinical-writing preference. Reviewers may also enter optional 1–5 rubric scores and free-text comments on hallucination, image interpretation, and overall quality. We report only aggregate statistics and omit reviewer identifiers, raw comments, and patient-level content.

Endpoint definition. The primary human endpoint is a descriptive measure of source-identification consistency among experienced clinicians across the five-case blinded round. Independent random guessing meets the same threshold with probability 81.25%. Under this criterion, most experienced clinicians did not consistently distinguish generated reports from published references across their five assigned pairs. Manuscript preference is used as complementary evidence of practical utility, and optional rubric scores provide additional descriptive detail.

Interface. Figure 13 documents the reviewer-facing interface used to collect the blinded comparisons, showing how the shared case context, paired reports, and medical images are presented before clinicians choose the more likely generated report, select the preferred report, and optionally enter rubric-aligned scores.

F.4 Web-Based User Studies and Feedback

We developed a web-based platform for user recruitment, user studies, and feedback collection, engaging more than 100 users. The platform supports study participation and collection of feedback on the system's usefulness for case-report preparation. For the expert A/B study, the reviewer-facing interface shown in Figure 13 records authorship judgments, report preferences, optional rubric scores, and free-text comments. User feedback has been largely positive on the system's usefulness for preparing case reports.

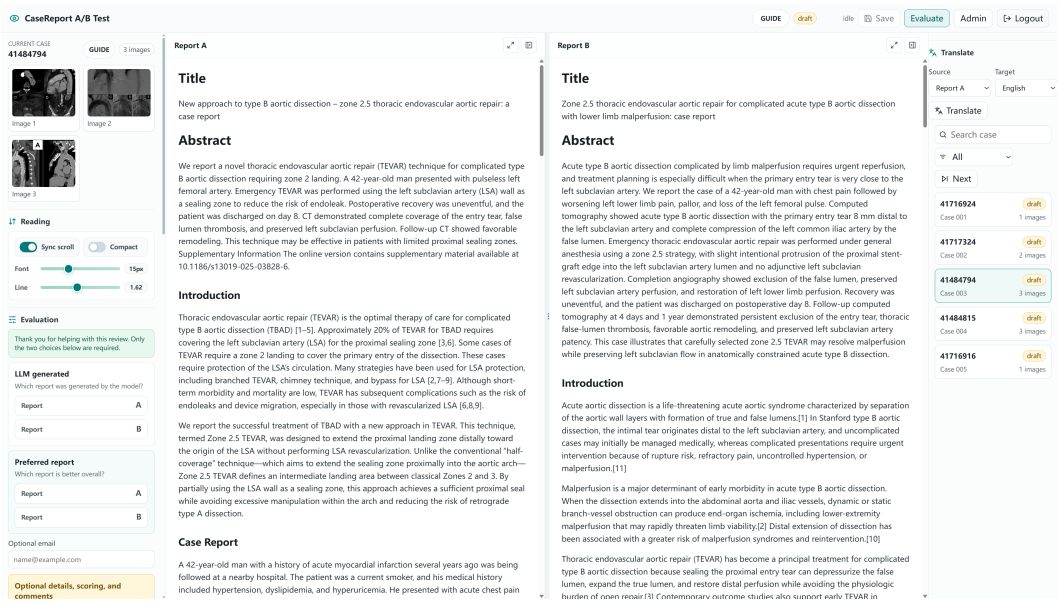


Figure 13: Expert web interface for the blinded A/B user study. Clinical reviewers inspect paired reports with shared medical images, guess which report was generated by an LLM, choose the stronger report overall, and may provide rubric-aligned scores and comments.

Table 9: LLM-judged blind A/B testing averaged across Gemini 3.1 Pro, GPT-5.4, Qwen3.5-397B-A17B, and Claude Opus 4.6 judges. All entries are reported as percentages.

Method	Identified as Generated (%) ↓	Identified as Human (%) ↑	Preference vs. Ground Truth (%)
Qwen3.5-9B	86.53	13.47	24.84
Gemma-4-31B	82.31	17.69	30.58
Claude Opus 4.6	72.18	27.82	45.28
GPT-5.4	70.24	29.76	53.70
Gemini 3.1 Pro	71.07	28.93	50.48
MedCaseAgent-Base	79.67	20.33	36.29
MedCaseAgent-Tuned	75.54	24.46	43.87
MedCaseAgent-Max	63.91	36.09	64.26

G LLM-Judged Blind A/B Comparison Across Baselines

Table 9 extends the MedCaseAgent-Max-vs.-reference comparison reported with human judges in Appendix F.3 to all baseline systems, using an ensemble of four LLM judges (Gemini 3.1 Pro, GPT-5.4, Qwen3.5-397B-A17B, Claude Opus 4.6) under the same blinded-A/B protocol with hidden model names, randomized A/B order, and evidence-grounded rationales (Liu et al., 2023; Zheng et al., 2023). The ensemble enables coverage across all baselines at a scale that would be infeasible with clinicians. Paired LLM–clinician agreement is reported separately in Appendix F.2. MedCaseAgent-Max is the least often identified as generated (63.91%) and the most often preferred over the reference (64.26%), with MedCaseAgent-Base and MedCaseAgent-Tuned showing stepwise gains over the non-agentic baselines. The clinician study complements this automatic table: most expert reviewers fail to recover MedCaseAgent's source labels across a full blinded review round while still finding the generated manuscripts useful, supporting the claim that the system is strong under automatic evaluation and persuasive in clinical review.

G.1 JUDGE ROBUSTNESS AND SELF-PREFERENCE

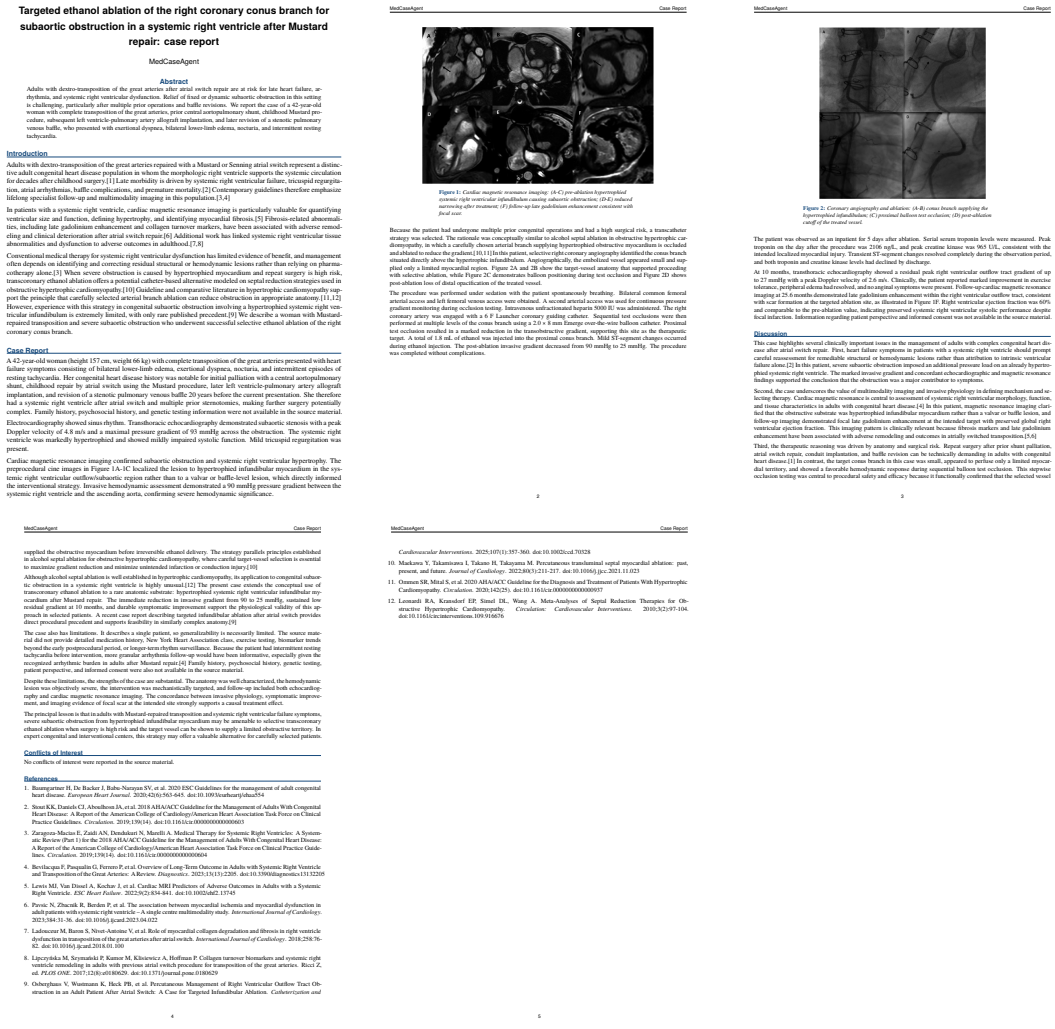
Three judges (GPT-5.4, Gemini 3.1 Pro, and Claude Opus 4.6) also serve as backbones in the configuration comparison. Table 10 reports general-quality scores under individual judges and the four-judge average.

Table 10: General-quality scores under individual judges and the four-judge average.

System	Qwen	Gemini	Claude	GPT	Average
Human Publication	76.74	78.35	74.85	72.13	75.52
GPT-5.4	77.65	78.32	75.84	74.78	76.65
Gemini 3.1 Pro	76.14	78.83	74.96	72.75	75.67
Claude Opus 4.6	75.60	77.02	75.10	70.84	74.64
MedCaseAgent-Base	66.24	67.70	64.36	62.42	65.18
MedCaseAgent-Max	84.10	85.48	82.34	80.76	83.17

H EXAMPLE CASE REPORTS GENERATED BY MEDCASEAGENT

Figures 14–16 show three complete example case reports generated by MedCaseAgent. Each generated output is included as a full-page sheet to show the manuscript structure, medical-image placement, citation formatting, and clinical timeline organization. These examples are qualitative outputs from the planner–writer–editor pipeline and complement the quantitative results in the main paper.



Targeted ethanol ablation of the right coronary conus branch for subaortic obstruction in a systemic right ventricle after Mustard repair: case report

MedCaseAgent

Abstract

Adults with dextro-transposition of the great arteries after atrial switch repair are at risk for late heart failure, arrhythmia, and systemic right ventricular dysfunction. Relief of fixed or dynamic subaortic obstruction in this setting is challenging, particularly after multiple prior operations and baffle revisions. We report the case of a 42-year-old woman with complete transposition of the great arteries, prior central aortopulmonary shunt, childhood Mustard procedure, subsequent left ventricle-pulmonary artery allograft implantation, and later revision of a stenotic pulmonary venous baffle, who presented with exertional dyspnea, bilateral lower-limb edema, nocturia, and intermittent resting tachycardia.

Introduction

Adults with dextro-transposition of the great arteries repaired with a Mustard or Senning atrial switch represent a distinctive adult congenital heart disease population in whom the morphologic right ventricle supports the systemic circulation for decades after childhood surgery.[1] Late morbidity is driven by systemic right ventricular failure, tricuspid regurgitation, atrial arrhythmias, baffle complications, and premature mortality.[2] Contemporary guidelines therefore emphasize lifelong specialist follow-up and multimodality imaging in this population.[3,4]

In patients with a systemic right ventricle, cardiac magnetic resonance imaging is particularly valuable for quantifying ventricular size and function, defining hypertrophy, and identifying myocardial fibrosis.[5] Fibrosis-related abnormalities, including late gadolinium enhancement and collagen turnover markers, have been associated with adverse remodeling and clinical deterioration after atrial switch repair.[6] Additional work has linked systemic right ventricular tissue abnormalities and dysfunction to adverse outcomes in adulthood.[7,8]

Conventional medical therapy for systemic right ventricular dysfunction has limited evidence of benefit, and management often depends on identifying and correcting residual structural or hemodynamic lesions rather than relying on pharmacotherapy alone.[3] When severe obstruction is caused by hypertrophied myocardium and repeat surgery is high risk, transcoronary ethanol ablation offers a potential catheter-based alternative modelled on septal reduction strategies used in obstructive hypertrophic cardiomyopathy.[10] Guideline and comparative literature in hypertrophic cardiomyopathy support the principle that carefully selected arterial branch ablation can reduce obstruction in appropriate anatomy.[11,12] However, experience with this strategy in congenital subaortic obstruction involving a hypertrophied systemic right ventricular infundibulum is extremely limited, with only rare published precedent.[9] We describe a woman with Mustard-repaired transposition and severe subaortic obstruction who underwent successful selective ethanol ablation of the right coronary conus branch.

Case Report

A 42-year-old woman (height 157 cm, weight 66 kg) with complete transposition of the great arteries presented with heart failure symptoms consisting of bilateral lower-limb edema, exertional dyspnea, nocturia, and intermittent episodes of resting tachycardia. Her congenital heart disease history was notable for initial palliation with a central aortopulmonary shunt, childhood repair by atrial switch using the Mustard procedure, later left ventricle-pulmonary artery allograft implantation, and revision of a stenotic pulmonary venous baffle 20 years before the current presentation. She therefore had a systemic right ventricle after atrial switch and multiple prior sternotomies, making further surgery potentially complex. Family history, psychosocial history, and genetic testing information were not available in the source material.

Electrocardiography showed sinus rhythm. Transthoracic echocardiography demonstrated subaortic stenosis with a peak Doppler velocity of 4.8 m/s and a maximal pressure gradient of 93 mmHg across the obstruction. The systemic right ventricle was markedly hypertrophied and showed mildly impaired systolic function. Mild tricuspid regurgitation was present.

Cardiac magnetic resonance imaging confirmed subaortic obstruction and systemic right ventricular hypertrophy. The preprocedural cine images in Figure 1A-1C localized the lesion to hypertrophied infundibular myocardium in the systemic right ventricular outflow/subaortic region rather than to a valvar or baffle-level lesion, which directly informed the interventional strategy. Invasive hemodynamic assessment demonstrated a 90 mmHg pressure gradient between the systemic right ventricle and the ascending aorta, confirming severe hemodynamic significance.

supplied the obstructive myocardium before irreversible ethanol delivery. The strategy parallels principles established in alcohol septal ablation for obstructive hypertrophic cardiomyopathy, where careful target-vessel selection is essential to maximize gradient reduction and minimize unintended infarction or conduction injury.[10]

Although alcohol septal ablation is well established in hypertrophic cardiomyopathy, its application to congenital subaortic obstruction in a systemic right ventricle is highly unusual.[12] The present case extends the conceptual use of transcoronary ethanol ablation to a rare anatomic substrate: hypertrophied systemic right ventricular infundibular myocardium after Mustard repair. The immediate reduction in invasive gradient from 90 to 25 mmHg, sustained low residual gradient at 10 months, and durable symptomatic improvement support the physiological validity of this approach in selected patients. A recent case report describing targeted infundibular ablation after atrial switch provides direct procedural precedent and supports feasibility in similarly complex anatomy.[9]

The case also has limitations. It describes a single patient, so generalizability is necessarily limited. The source material did not provide detailed medication history, New York Heart Association class, exercise testing, biomarker trends beyond the early postprocedural period, or longer-term rhythm surveillance. Because the patient had intermittent resting tachycardia before intervention, more granular arrhythmia follow-up would have been informative, especially given the recognized arrhythmic burden in adults after Mustard repair.[4] Family history, psychosocial history, genetic testing, patient perspective, and informed consent were also not available in the source material.

Despite these limitations, the strengths of the case are substantial. The anatomy was well characterized, the hemodynamic lesion was objectively severe, the intervention was mechanistically targeted, and follow-up included both echocardiography and cardiac magnetic resonance imaging. The concordance between invasive physiology, symptomatic improvement, and imaging evidence of focal scar at the intended site strongly supports a causal treatment effect.

The principal lesson is that in adults with Mustard-repaired transposition and systemic right ventricular failure symptoms, severe subaortic obstruction from hypertrophied infundibular myocardium may be amenable to selective transcoronary ethanol ablation when surgery is high risk and the target vessel can be shown to supply a limited obstructive territory. In expert congenital and interventional centers, this strategy may offer a valuable alternative for carefully selected patients.

Conflicts of Interest

No conflicts of interest were reported in the source material.

References

1. Baumgartner H, De Backer J, Babu-Narayan SV, et al. 2020 ESC Guidelines for the management of adult congenital heart disease. *European Heart Journal*. 2020;42(6):563-645. doi:10.1093/eurheartj/ehaa554

2. Stout KK, Daniels CJ, Aboulhosn JA, et al. 2018 AHA/ACC Guideline for the Management of Adults With Congenital Heart Disease: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. *Circulation*. 2019;139(14). doi:10.1161/cir.0000000000000603

3. Zaragoza-Macias E, Zaidi AN, Dendukuri N, Marelli A. Medical Therapy for Systemic Right Ventricles: A Systematic Review (Part 1) for the 2018 AHA/ACC Guideline for the Management of Adults With Congenital Heart Disease: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. *Circulation*. 2019;139(14). doi:10.1161/cir.0000000000000604

4. Bevilacqua F, Pasqualin G, Ferrero P, et al. Overview of Long-Term Outcome in Adults with Systemic Right Ventricle and Transposition of the Great Arteries: A Review. *Diagnostics*. 2023;13(13):2205. doi:10.3390/diagnostics13132205

5. Lewis MJ, Van Dissel A, Kocher J, et al. Cardiac MRI Predictors of Adverse Outcomes in Adults with a Systemic Right Ventricle. *JSC Heart Failure*. 2022;9(2):834-841. doi:10.1002/ehf2.13745

6. Pavoic N, Zhacnik R, Berden P, et al. The association between myocardial ischemia and myocardial dysfunction in adult patients with systemic right ventricle – A single centre multimodality study. *International Journal of Cardiology*. 2023;384:31-36. doi:10.1016/j.ijcard.2023.04.022

7. Ladouceur M, Baron S, Nivet-Antoine V, et al. Role of myocardial collagen degradation and fibrosis in right ventricle dysfunction in transposition of the great arteries after atrial switch. *International Journal of Cardiology*. 2018;258:78-82. doi:10.1016/j.ijcard.2018.01.100

8. Lipczyńska M, Szymański P, Kamor M, Kłosiewicz A, Hoffman P. Collagen turnover biomarkers and systemic right ventricle remodeling in adults with previous atrial switch procedure for transposition of the great arteries. Ricci Z, *ed. PLOS ONE*. 2017;12(8):e0180629. doi:10.1371/journal.pone.0180629

9. Ozbarbaus V, Wustmann K, Heck PB, et al. Percutaneous Management of Right Ventricular Outflow Tract Obstruction in an Adult Patient After Atrial Switch: A Case for Targeted Infundibular Ablation. *Catheterization and Cardiovascular Interventions*. 2025;107(1):357-360. doi:10.1002/ccd.70328

10. Maskowa Y, Takamisawa I, Takano H, Takeyama M. Percutaneous transluminal septal myocardial ablation: past, present, and future. *Journal of Cardiology*. 2022;80(3):211-217. doi:10.1016/j.jjcc.2021.11.023

11. Ommen SR, Mital S, et al. 2020 AHA/ACC Guideline for the Diagnosis and Treatment of Patients With Hypertrophic Cardiomyopathy. *Circulation*. 2020;142(25). doi:10.1161/cir.0000000000000937

12. Leonardi RA, Kransdorf EP, Simel DL, Wang A. Meta-Analyses of Septal Reduction Therapies for Obstructive Hypertrophic Cardiomyopathy. *Circulation: Cardiovascular Interventions*. 2010;3(2):97-104. doi:10.1161/circinterventions.109.916076

Figure 1: *Cardiac magnetic resonance imaging. (A–C) pre-ablation hypertrophied systemic right ventricular infundibulum causing subaortic obstruction; (D-E) reduced narrowing after treatment; (F) follow-up late gadolinium enhancement consistent with focal scar.*

Because the patient had undergone multiple prior congenital operations and had a high surgical risk, a transcatheter strategy was selected. The rationale was conceptually similar to alcohol septal ablation in obstructive hypertrophic cardiomyopathy, in which a carefully chosen arterial branch supplying hypertrophied obstructive myocardium is occluded and ablated to reduce the gradient.[10,11] In this patient, selective right coronary angiography identified the conus branch situated directly above the hypertrophic infundibulum. Angiographically, the embolized vessel appeared small and supplied only a limited myocardial region. Figure 2A and 2B show the target-vessel anatomy that supported proceeding with selective ablation, while Figure 2C demonstrates balloon positioning during test occlusion and Figure 2D shows post-ablation loss of distal opacification of the treated vessel.

The procedure was performed under sedation with the patient spontaneously breathing. Bilateral common femoral arterial access and left femoral venous access were obtained. A second arterial access was used for continuous pressure gradient monitoring during occlusion testing. Intravenous unfractionated heparin 5000 IU was administered. The right coronary artery was engaged with a 6 F Launcher coronary guiding catheter. Sequential test occlusions were then performed at multiple levels of the conus branch using a 2.0 × 8 mm Emerge over-the-wire balloon catheter. Proximal test occlusion resulted in a marked reduction in the transobstructive gradient, supporting this site as the therapeutic target. A total of 1.8 mL of ethanol was injected into the proximal conus branch. Mild ST-segment changes occurred during ethanol injection. The post-ablation invasive gradient decreased from 90 mmHg to 25 mmHg. The procedure was completed without complications.

Figure 2: *Coronary angiography and ablation: (A-B) conus branch supplying the hypertrophied infundibulum; (C) proximal balloon test occlusion; (D) post-ablation cutoff of the treated vessel.*

The patient was observed as an inpatient for 5 days after ablation. Serial serum troponin levels were measured. Peak troponin on the day after the procedure was 2106 ng/L, and peak creatine kinase was 965 U/L, consistent with the intended localized myocardial injury. Transient ST-segment changes resolved completely during the observation period, and both troponin and creatine kinase levels had declined by discharge.

At 10 months, transthoracic echocardiography showed a residual peak right ventricular outflow tract gradient of up to 27 mmHg with a peak Doppler velocity of 2.6 m/s. Clinically, the patient reported marked improvement in exercise tolerance, peripheral edema had resolved, and no anginal symptoms were present. Follow-up cardiac magnetic resonance imaging at 25.6 months demonstrated late gadolinium enhancement within the right ventricular outflow tract, consistent with scar formation at the targeted ablation site, as illustrated in Figure 1F. Right ventricular ejection fraction was 60% and comparable to the pre-ablation value, indicating preserved systemic right ventricular systolic performance despite focal infarction. Information regarding patient perspective and informed consent was not available in the source material.

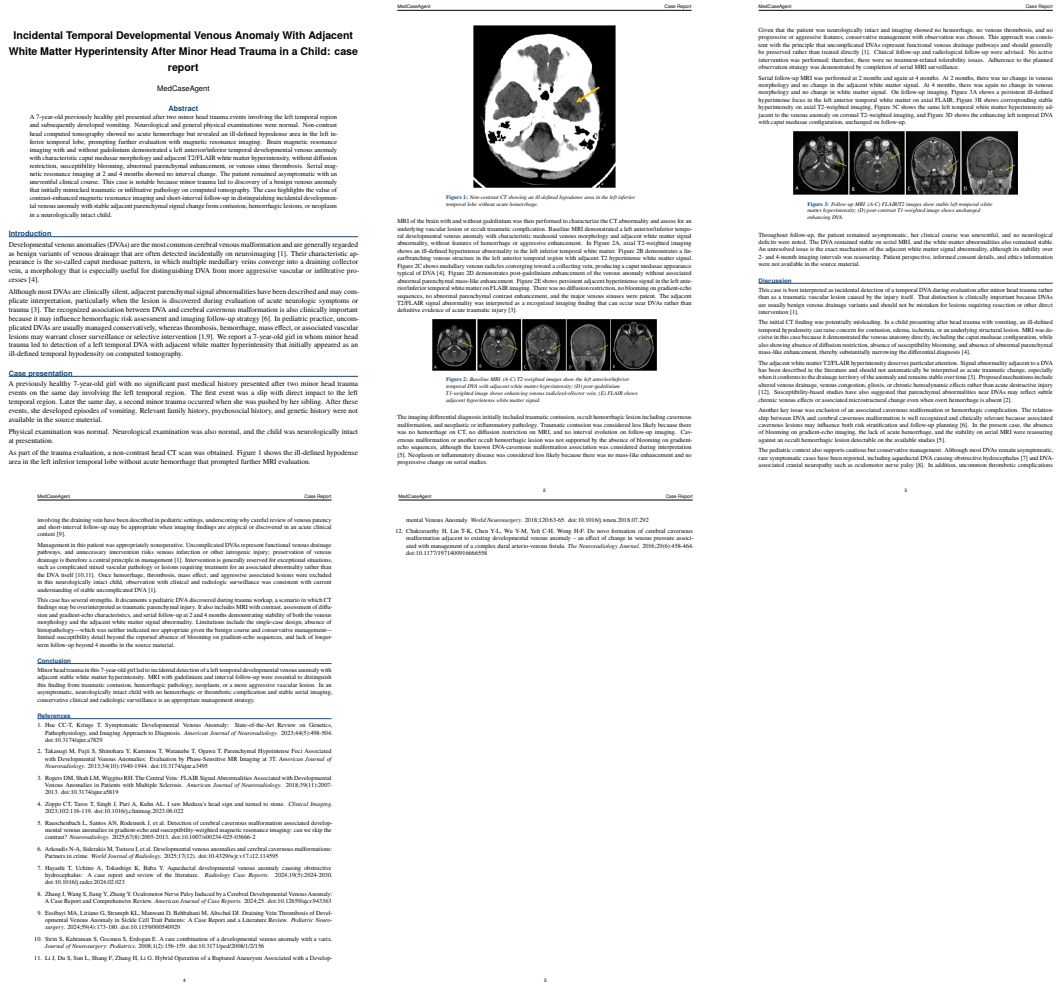
Discussion

This case highlights several clinically important issues in the management of adults with complex congenital heart disease after atrial switch repair. First, heart failure symptoms in patients with a systemic right ventricle should prompt careful reassessment for remediable structural or hemodynamic lesions rather than attribution to intrinsic ventricular failure alone.[2] In this patient, severe subaortic obstruction imposed an additional pressure load on an already hypertrophied systemic right ventricle. The marked invasive gradient and concordant echocardiographic and magnetic resonance findings supported the conclusion that the obstruction was a major contributor to symptoms.

Second, the case underscores the value of multimodality imaging and invasive physiology in defining mechanism and selecting therapy. Cardiac magnetic resonance is central to assessment of systemic right ventricular morphology, function, and tissue characteristics in adults with congenital heart disease.[4] In this patient, magnetic resonance imaging clarified that the obstructive substrate was hypertrophied infundibular myocardium rather than a valvar or baffle lesion, and follow-up imaging demonstrated focal late gadolinium enhancement at the intended target with preserved global right ventricular ejection fraction. This imaging pattern is clinically relevant because fibrosis markers and late gadolinium enhancement have been associated with adverse remodeling and outcomes in atrially switched transposition.[5,6]

Third, the therapeutic reasoning was driven by anatomy and surgical risk. Repeat surgery after prior shunt palliation, atrial switch repair, conduit implantation, and baffle revision can be technically demanding in adults with congenital heart disease.[1] In contrast, the target conus branch in this case was small, appeared to perfuse only a limited myocardial territory, and showed a favorable hemodynamic response during sequential balloon test occlusion. This stepwise occlusion testing was central to procedural safety and efficacy because it functionally confirmed that the selected vessel

Figure 14: Example 1 of a case report generated by MedCaseAgent. The full-page sheet shows the generated manuscript structure, image placement, citation formatting, and clinical timeline organization.



Incidental Temporal Developmental Venous Anomaly With Adjacent White Matter Hyperintensity After Minor Head Trauma in a Child: case report

MedCaseAgent

Abstract

A 7-year-old previously healthy girl presented after two minor head trauma events involving the left temporal region and subsequently developed vomiting. Neurological and general physical examinations were normal. Non-contrast head computed tomography showed no acute hemorrhage but revealed an ill-defined hypodense area in the left inferior temporal lobe, prompting further evaluation with magnetic resonance imaging. Brain magnetic resonance imaging with and without gadolinium demonstrated a left anterior/inferior temporal developmental venous anomaly with characteristic caput medusae morphology and adjacent T2/FLAIR white matter hyperintensity, without diffusion restriction, susceptibility blooming, abnormal parenchymal enhancement, or venous sinus thrombosis. Serial magnetic resonance imaging at 2 and 4 months showed no interval change. The patient remained asymptomatic with an uneventful clinical course. This case is notable because minor trauma led to discovery of a benign venous anomaly that initially mimicked traumatic or infiltrative pathology on computed tomography. The case highlights the value of contrast-enhanced magnetic resonance imaging and short-interval follow-up in distinguishing incidental developmental venous anomaly with stable adjacent parenchymal signal change from contusion, hemorrhagic lesions, or neoplasm in a neurologically intact child.

Introduction

Developmental venous anomalies (DVAs) are the most common cerebral venous malformation and are generally regarded as benign variants of venous drainage that are often detected incidentally on neuroimaging [1]. Their characteristic appearance is the so-called caput medusae pattern, in which multiple medullary veins converge into a draining collector vein, a morphology that is especially useful for distinguishing DVA from more aggressive vascular or infiltrative processes [4].

Although most DVAs are clinically silent, adjacent parenchymal signal abnormalities have been described and may complicate interpretation, particularly when the lesion is discovered during evaluation of acute neurologic symptoms or trauma [3]. The recognized association between DVA and cerebral cavernous malformation is also clinically important because it may influence hemorrhagic risk assessment and imaging follow-up strategy [6]. In pediatric practice, uncomplicated DVAs are usually managed conservatively, whereas thrombosis, hemorrhage, mass effect, or associated vascular lesions may warrant closer surveillance or selective intervention [1,9]. We report a 7-year-old girl in whom minor head trauma led to detection of a left temporal DVA with adjacent white matter hyperintensity that initially appeared as an ill-defined temporal hypodensity on computed tomography.

Case presentation

A previously healthy 7-year-old girl with no significant past medical history presented after two minor head trauma events on the same day involving the left temporal region. The first event was a slip with direct impact to the left temporal region. Later the same day, a second minor trauma occurred when she was pushed by her sibling. After these events, she developed episodes of vomiting. Relevant family history, psychosocial history, and genetic history were not available in the source material.

Physical examination was normal. Neurological examination was also normal, and the child was neurologically intact at presentation.

As part of the trauma evaluation, a non-contrast head CT scan was obtained. Figure 1 shows the ill-defined hypodense area in the left inferior temporal lobe without acute hemorrhage that prompted further MRI evaluation.

Figure 1: *Non-contrast CT showing an ill-defined hypodense area in the left inferior temporal lobe without acute hemorrhage.*

MRI of the brain with and without gadolinium was then performed to characterize the CT abnormality and assess for an underlying vascular lesion or occult traumatic complication. Baseline MRI demonstrated a left anterior/inferior temporal developmental venous anomaly with characteristic medullo venous morphology and adjacent white matter signal abnormality, without features of hemorrhage or aggressive enhancement. In Figure 2A, axial T2-weighted imaging shows an ill-defined hyperintense abnormality in the left inferior temporal white matter. Figure 2B demonstrates a linearbranching venous structure in the left anterior temporal region with adjacent T2 hyperintense white matter signal. Figure 2C shows medullary venous radicles converging toward a collecting vein, producing a caput medusae appearance typical of DVA [4]. Figure 2D demonstrates post-gadolinium enhancement of the venous anomaly without associated abnormal parenchymal mass-like enhancement. Figure 2E shows persistent adjacent hyperintense signal in the left anterior/inferior temporal white matter on FLAIR imaging. There was no diffusion restriction, no blooming on gradient-echo sequences, no abnormal parenchymal contrast enhancement, and the major venous sinuses were patent. The adjacent T2/FLAIR signal abnormality was interpreted as a recognized imaging finding that can occur near DVAs rather than definitive evidence of acute traumatic injury [3].

Figure 2: *Baseline MRI. (A–C) T2-weighted images show the left anterior/inferior temporal DVA with adjacent white matter hyperintensity; (D) post-gadolinium T1-weighted image shows enhancing venous radicles/collector vein; (E) FLAIR shows adjacent hyperintense white matter signal.*

The imaging differential diagnosis initially included traumatic contusion, occult hemorrhagic lesion including cavernous malformation, and neoplastic or inflammatory pathology. Traumatic contusion was considered less likely because there was no hemorrhage on CT, no diffusion restriction on MRI, and no interval evolution on follow-up imaging. Cavernous malformation or another occult hemorrhagic lesion was not supported by the absence of blooming on gradient-echo sequences, although the known DVA-cavernous malformation association was considered during interpretation [5]. Neoplasm or inflammatory disease was considered less likely because there was no mass-like enhancement and no progressive change on serial studies.

Given that the patient was neurologically intact and imaging showed no hemorrhage, no venous thrombosis, and no progressive or aggressive features, conservative management with observation was chosen. This approach was consistent with the principle that uncomplicated DVAs represent functional venous drainage pathways and should generally be preserved rather than treated directly [1]. Clinical follow-up and radiological follow-up were advised. No active intervention was performed; therefore, there were no treatment-related tolerability issues. Adherence to the planned observation strategy was demonstrated by completion of serial MRI surveillance.

Serial follow-up MRI was performed at 2 months and again at 4 months. At 2 months, there was no change in venous morphology and no change in the adjacent white matter signal. At 4 months, there was again no change in venous morphology and no change in white matter signal. On follow-up imaging, Figure 3A shows a persistent ill-defined hyperintense focus in the left anterior temporal white matter on axial FLAIR, Figure 3B shows corresponding stable hyperintensity on axial T2-weighted imaging, Figure 3C shows the same left temporal white matter hyperintensity adjacent to the venous anomaly on coronal T2-weighted imaging, and Figure 3D shows the enhancing left temporal DVA with caput medusae configuration, unchanged on follow-up.

Figure 3: *Follow-up MRI. (A-C) FLAIR/T2 images show stable left temporal white matter hyperintensity; (D) post-contrast T1-weighted image shows unchanged enhancing DVA.*

Throughout follow-up, the patient remained asymptomatic, her clinical course was uneventful, and no neurological deficits were noted. The DVA remained stable on serial MRI, and the white matter abnormalities also remained stable. An unresolved issue is the exact mechanism of the adjacent white matter signal abnormality, although its stability over 2- and 4-month imaging intervals was reassuring. Patient perspective, informed consent details, and ethics information were not available in the source material.

Discussion

This case is best interpreted as incidental detection of a temporal DVA during evaluation after minor head trauma rather than as a traumatic vascular lesion caused by the injury itself. That distinction is clinically important because DVAs are usually benign venous drainage variants and should not be mistaken for lesions requiring resection or other direct intervention [1].

The initial CT finding was potentially misleading. In a child presenting after head trauma with vomiting, an ill-defined temporal hypodensity can raise concern for contusion, edema, ischemia, or an underlying structural lesion. MRI was decisive in this case because it demonstrated the venous anatomy directly, including the caput medusae configuration, while also showing absence of diffusion restriction, absence of susceptibility blooming, and absence of abnormal parenchymal mass-like enhancement, thereby substantially narrowing the differential diagnosis [4].

The adjacent white matter T2/FLAIR hyperintensity deserves particular attention. Signal abnormality adjacent to a DVA has been described in the literature and should not automatically be interpreted as acute traumatic change, especially when it conforms to the drainage territory of the anomaly and remains stable over time [3]. Proposed mechanisms include altered venous drainage, venous congestion, gliosis, or chronic hemodynamic effects rather than acute destructive injury [12]. Susceptibility-based studies have also suggested that parenchymal abnormalities near DVAs may reflect subtle chronic venous effects or associated microstructural change even when overt hemorrhage is absent [2].

Another key issue was exclusion of an associated cavernous malformation or hemorrhagic complication. The relationship between DVA and cerebral cavernous malformation is well recognized and clinically relevant because associated cavernous lesions may influence both risk stratification and follow-up planning [6]. In the present case, the absence of blooming on gradient-echo imaging, the lack of acute hemorrhage, and the stability on serial MRI were reassuring against an occult hemorrhagic lesion detectable on the available studies [5].

The pediatric context also supports cautious but conservative management. Although most DVAs remain asymptomatic, rare symptomatic cases have been reported, including aqueductal DVA causing obstructive hydrocephalus [7] and DVA-associated cranial neuropathy such as oculomotor nerve palsy [8]. In addition, uncommon thrombotic complications

involving the draining vein have been described in pediatric settings, underscoring why careful review of venous patency and short-interval follow-up may be appropriate when imaging findings are atypical or discovered in an acute clinical context [9].

Management in this patient was appropriately nonoperative. Uncomplicated DVAs represent functional venous drainage pathways, and unnecessary intervention risks venous infarction or other iatrogenic injury; preservation of venous drainage is therefore a central principle in management [1]. Intervention is generally reserved for exceptional situations, such as complicated mixed vascular pathology or lesions requiring treatment for an associated abnormality rather than the DVA itself [10,11]. Once hemorrhage, thrombosis, mass effect, and aggressive associated lesions were excluded in this neurologically intact child, observation with clinical and radiologic surveillance was consistent with current understanding of stable uncomplicated DVA [1].

This case has several strengths. It documents a pediatric DVA discovered during trauma workup, a scenario in which CT findings may be overinterpreted as traumatic parenchymal injury. It also includes MRI with contrast, assessment of diffusion and gradient-echo characteristics, and serial follow-up at 2 and 4 months demonstrating stability of both the venous morphology and the adjacent white matter signal abnormality. Limitations include the single-case design, absence of histopathology—which was neither indicated nor appropriate given the benign course and conservative management—limited susceptibility detail beyond the reported absence of blooming on gradient-echo sequences, and lack of longer-term follow-up beyond 4 months in the source material.

Conclusion

Minor head trauma in this 7-year-old girl led to incidental detection of a left temporal developmental venous anomaly with adjacent stable white matter hyperintensity. MRI with gadolinium and interval follow-up were essential to distinguish this finding from traumatic contusion, hemorrhagic pathology, neoplasm, or a more aggressive vascular lesion. In an asymptomatic, neurologically intact child with no hemorrhagic or thrombotic complication and stable serial imaging, conservative clinical and radiologic surveillance is an appropriate management strategy.

References

1. Hsu CC-T, Krings T. Symptomatic Developmental Venous Anomaly: State-of-the-Art Review on Genetics, Pathophysiology, and Imaging Approach to Diagnosis. *American Journal of Neuroradiology*. 2023;44(5):498-504. doi:10.3174/ajnr.A7829
2. Takasugi M, Fujii S, Shinohara Y, Kaminou T, Watanabe T, Ogawa T. Parenchymal Hypointense Foci Associated with Developmental Venous Anomalies: Evaluation by Phase-Sensitive MR Imaging at 3T. *American Journal of Neuroradiology*. 2013;34(10):1940-1944. doi:10.3174/ajnr.a3495
3. Rogers DM, Shah LM, Wiggins RH. The Central Vein: FLAIR Signal Abnormalities Associated with Developmental Venous Anomalies in Patients with Multiple Sclerosis. *American Journal of Neuroradiology*. 2018;39(11):2007-2013. doi:10.3174/ajnr.a5819
4. Zoppo CT, Tarro T, Singh J, Puri A, Kuhn AL. I saw Medusa's head sign and turned to stone. *Clinical Imaging*. 2023;102:118-119. doi:10.1016/j.clinimag.2023.08.022
5. Rauschenbach L, Santos AN, Rodemerk J, et al. Detection of cerebral cavernous malformation associated developmental venous anomalies in gradient-echo and susceptibility-weighted magnetic resonance imaging: can we skip the contrast? *Neuroradiology*. 2025;67(8):2005-2013. doi:10.1007/s00234-025-03606-2
6. Arkudis N-A, Siderakis M, Tsetsou I, et al. Developmental venous anomalies and cerebral cavernous malformations: Partners in crime. *World Journal of Radiology*. 2025;17(12). doi:10.4329/wjr.v17.i12.114595
7. Hayashi T, Uchino A, Tokushige K, Baba Y. Aqueductal developmental venous anomaly causing obstructive hydrocephalus: A case report and review of the literature. *Radiology Case Reports*. 2024;19(5):2024-2030. doi:10.1016/j.radcr.2024.02.023
8. Zhang J, Wang S, Jiang Y, Zhang Y. Oculomotor Nerve Palsy Induced by a Cerebral Developmental Venous Anomaly: A Case Report and Comprehensive Review. *American Journal of Case Reports*. 2024;25. doi:10.12659/ajcr.943363
9. Essibayi MA, Liriano G, Strumph KL, Manswan D, Bebbahani M, Abichal DF. Draining Vein Thrombosis of Developmental Venous Anomaly in Sickle Cell Trait Patients: A Case Report and a Literature Review. *Pediatric Neurosurgery*. 2024;59(4):173-180. doi:10.1159/000540929
10. Sirin S, Kahraman S, Gocmen S, Erdogan E. A rare combination of a developmental venous anomaly with a varix. *Journal of Neurosurgery: Pediatrics*. 2008;1(2):158-159. doi:10.3171/ped/2008/1/2/158
11. Li J, Du S, Sun L, Shang F, Zhang H, Li G. Hybrid Operation of a Ruptured Aneurysm Associated with a Develop-

mental Venous Anomaly. *World Neurosurgery*. 2018;120:63-65. doi:10.1016/j.wneu.2018.07.292
12. Chakravarthy H, Lin T-K, Chen Y-L, Wu Y-M, Yeh C-H, Wong H-F. De novo formation of cerebral cavernous malformation adjacent to existing developmental venous anomaly – an effect of change in venous pressure associated with management of a complex dural arterio-venous fistula. *The Neuroradiology Journal*. 2016;29(6):458-464. doi:10.1177/1971400916666558

Figure 15: Example 2 of a case report generated by MedCaseAgent. The full-page sheet shows the generated manuscript structure, image placement, citation formatting, and clinical timeline organization.

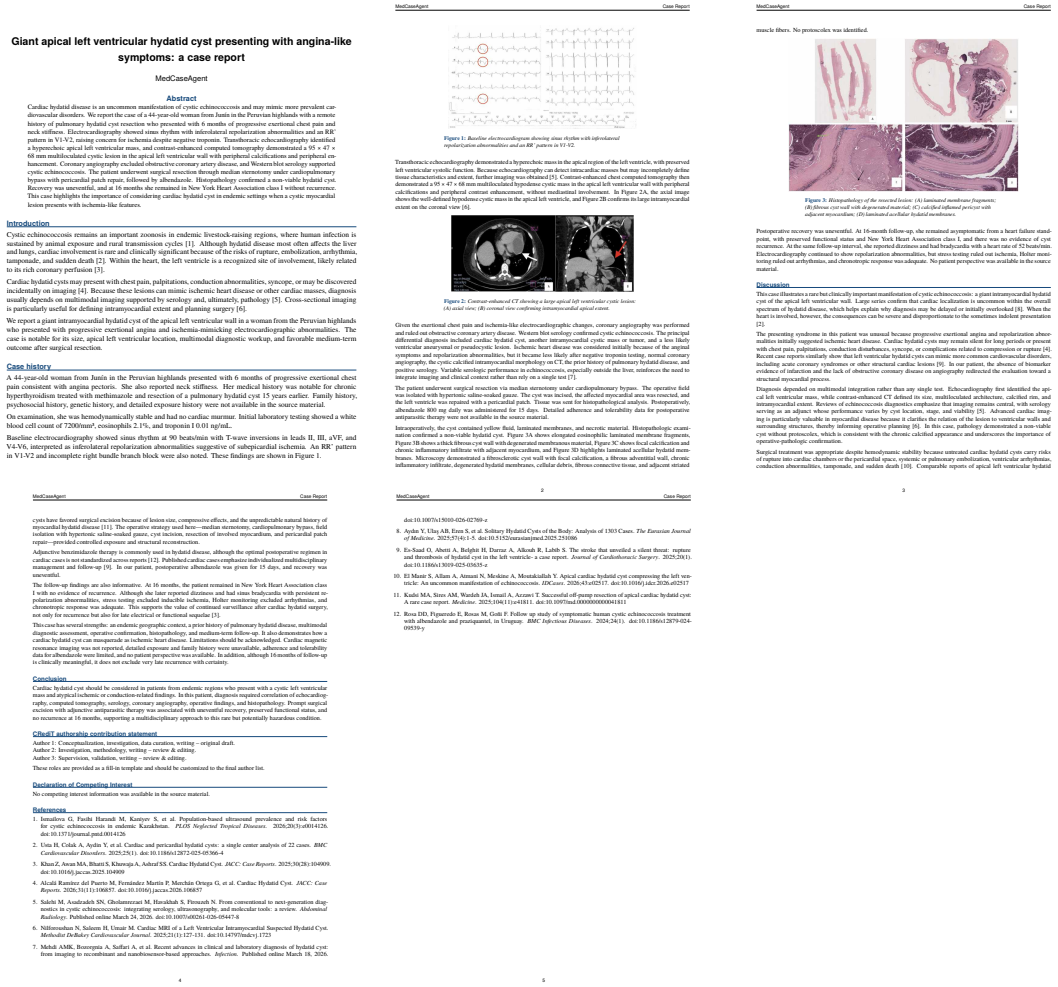


Figure 16: Example 3 of a case report generated by MedCaseAgent. The full-page sheet shows the generated manuscript structure, image placement, citation formatting, and clinical timeline organization.

I BROADER IMPACTS

MedCaseAgent is intended to support clinicians in writing case reports. Case-report writing is a time-consuming part of clinical knowledge sharing: clinicians must organize patient history, diagnostic evidence, images, treatment course, outcomes, and related literature before a report can be submitted. Unlike open-ended research-paper generation, this writing task is strongly evidence-based and grounded in an observed patient case, making it well suited to a clinician-controlled tool that structures evidence rather than inventing new claims.

The main positive impact is therefore to reduce the documentation and literature-review burden involved in manuscript preparation. MedCaseAgent helps organize fragmented notes, medical images, prior-case evidence, and citation candidates into a structured draft that remains available for expert review. This support is especially useful for rare, atypical, or unprecedented presentations, where clinicians must compare a complex patient course against sparse prior literature.

MedCaseAgent may also provide useful cues before writing begins. During the planning stage, retrieval over analogous cases, literature, images, and disease-level evidence can help surface diagnostic possibilities or comparative context that clinicians may want to consider. These planning outputs are not meant to replace expert diagnosis, but they can serve as organized signals for specialists who are already evaluating a difficult case. More broadly, the system may enable faster dissemination of rare-disease observations, more consistent case-report structure, and broader participation from clinical teams with limited time for manuscript drafting.

Responsible use should keep clinicians at the center of the workflow. Patient facts, image interpretations, and citation support should be verified before submission or dissemination, and any deployment on real clinical records should follow local governance, consent, de-identification, and access-control requirements.

J RELEASE AND LLM USAGE DISCLOSURE

We will release the MedCase-150K and MedCase-Bench datasets, together with the MedCaseAgent source code, data-construction scripts, evaluation prompts, metadata, and data splits, under CC-BY-NC 4.0.

This disclosure concerns LLMs as the modeling technology studied in this work, including the agentic planner, writer, editor, and LLM-judge evaluation components described in the main paper. Separate from the method, LLM-based tools were used only for manuscript writing and editing assistance. They were not used to fabricate patient records, experimental measurements, baselines, or evaluation outcomes; final reported numbers must be produced by the evaluation pipeline and be traceable to released scripts, prompts, and logs.