MULTIKNEE: A MULTI-SEQUENCE MULTI-VIEW MRI BENCHMARK FOR EVIDENCE-CHAIN KNEE DIAGNOSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Medical imaging benchmarks for multimodal large language models (MLLMs) predominantly rely on single-image inputs or coarse-grained labels, as fine-grained annotations aligned across sequences and views are prohibitively expensive. We introduce **MultiKnee**, a hierarchically structured multi-sequence, multi-view knee MRI benchmark with fine-grained annotations at scale: 11,575 studies, 104,175 sequence-view units, and 821,823 VQAs across six anatomical systems, 71 clinical attributes, and 34 diagnostic categories. VQAs are organized as **explicitly linked Observation-Evidence-Diagnosis chains** that trace a clinical reasoning pathway: from perceiving images across sequences and views, to interpreting them within anatomical context, and reaching the final diagnosis. Across 20 model configurations, controlled-observation analysis reveals substantial differences in how models utilize complementary sequence and view information. Stage-wise analysis further reveals that models' reasoning bottlenecks lie primarily in evidence interpretation and diagnosis rather than observation. MultiKnee thus provides a large-scale infrastructure for multi-sequence, multi-view MRI research and a framework for understanding how MLLMs reason from observations to diagnoses.

1 INTRODUCTION

Multimodal large language models (MLLMs) are increasingly applied to medical image interpretation (Wang et al., 2026; Tanno et al., 2025; Busch et al., 2025). Knee MRI is a particularly demanding setting for such models: radiologists integrate signal, morphology, structural continuity, and lesion extent across complementary sequences and anatomical views to reach a diagnosis (Qiu et al., 2024). Multi-sequence multi-view MLLMs could assist radiologists in routine reporting, support decision-making in complex cases where findings are subtle or distributed across sequences and views, and facilitate education and quality assurance in musculoskeletal imaging Xie et al. (2025); Qiu et al. (2024); Gertz et al. (2024).

Medical VQA benchmarks have introduced semantic, localization, and evidence-related annotations (Liu et al., 2021; Nguyen et al., 2025), while knee MRI resources span classification, reconstruction, and detection (Zhao et al., 2022; Desai et al., 2021), as well as 3D grounded diagnostic reasoning (Sambara et al., 2025). But these datasets either operate on single images or coarse-grained inputs, lacking annotations aligned across sequences and views, due to the prohibitive annotation cost. Consequently, it remains limited for assessing MLLMs on multiple sequences and views (Fig. 1).

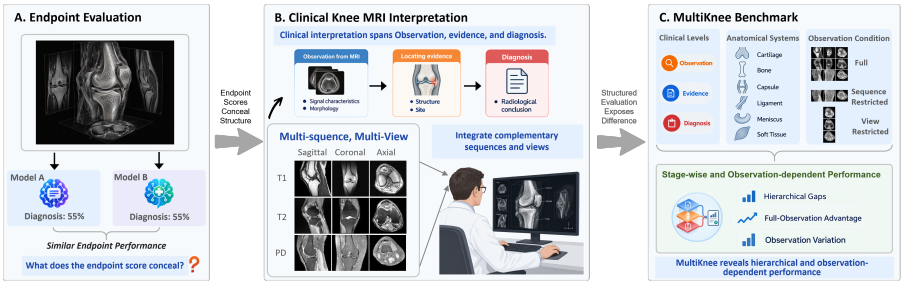

Figure 1: **From endpoint scores to clinical reasoning.** Examples show how MultiKnee reveals hidden reasoning gaps beyond accuracies via multi-sequence, multi-view, and stage-wise evaluation.

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.

To tackle these, we introduce a hierarchical **O**bservation–**E**vidence–**D**iagnosis **Chain** (**OEDChain**) for evaluating MLLMs on multi-sequence, multi-view knee MRI interpretation. It decomposes clinical reasoning into three linked stages: *Observation*, which assesses whether a model perceives clinically relevant image features across sequences and views; *Evidence*, which assesses whether the model interprets these features within anatomical context; and *Diagnosis*, which assesses whether the model reaches the correct clinical conclusion. These stages form a chain: each targets the same anatomical structure within an MRI study, tracing a complete reasoning pathway from image perception to diagnostic conclusion. The framework further treats the MRI input itself as a controlled variable: *full*, *sequence-restricted*, or *view-restricted*, enabling attribution of model failures to specific reasoning stages and to specific input dependencies. This dual decomposition provides a level of analytical granularity beyond existing knee MRI benchmarks.

To operationalize this framework, we construct the **MultiKnee** dataset, which comprises 11,575 studies, 104,175 sequence–view units, and 821,823 VQA items across six anatomical systems. MultiKnee is the first knee MRI benchmark to provide **cross-sequence and cross-view aligned fine-grained annotations at scale** across complementary sequences (T1, T2, PD) and anatomical views (axial, coronal, sagittal), constructed from paired MRI studies and radiology reports through a rigorous pipeline of report analysis, MRI grounding, question generation, independent audit, and sampled radiologist review. This multi-sequence, multi-view structure is not merely a data attribute but a design: each VQA item is grounded in sequence–view units, enabling **controlled-observation analysis** at the core of OEDChain: evaluating the same clinical question under full, sequence-restricted, and view-restricted inputs to attribute model performance to MRI information.

We evaluate 20 model configurations spanning general-purpose, medical-specialized, open-source, and fine-tuned models. Results validate both dimensions of the framework. First, **stage-resolved** O–E–D chain analysis shows that models with similar diagnosis-level scores can exhibit qualitatively different failure patterns: some degrade at the Observation-to-Evidence transition, others at the Evidence-to-Diagnosis step, with failure locations varying by anatomical system. Second, controlled-observation analysis shows that the value of multi-sequence, multi-view input varies systematically across reasoning stages and models: higher-level reasoning depends more heavily on complementary information, and different models rely on different sequences or views to reach the same diagnosis.

Our contributions are threefold: 1) We introduce OEDChain, a hierarchical framework with **explicitly linked** Observation–Evidence–Diagnosis stages over the same anatomical target, with MRI input as a controlled variable, enabling attribution of model failures to specific reasoning stages and input dependencies. 2) We construct MultiKnee, the first large-scale fine-grained multi-sequence, multi-view knee MRI benchmark grounded in specific sequence–view units to support **systematic controlled-observation analysis**. 3) Evaluation of 20 model configurations reveals that models with similar diagnosis-level scores exhibit distinct failure patterns along the reasoning chain, and that the value of multi-sequence, multi-view input varies across reasoning stages and models.

2 Related Work

Medical VQA Benchmarks. Medical MLLMs have been evaluated on visual question answering, diagnostic assistance, and report generation (Li et al., 2023; Chen et al., 2024a; Wang et al., 2026). Early VQA datasets established the foundation for this paradigm: VQA-RAD introduced clinician-generated radiology QA (Lau et al., 2018); PathVQA extended this to pathology (He et al., 2021); SLAKE added structured semantic annotations (Liu et al., 2021); and PMC-VQA scaled up with automatically generated QA pairs from biomedical images (Zhang et al., 2024). Larger benchmarks such as OmniMedVQA (Hu et al., 2024) and GMAI-MMBench (Chen et al., 2024b) further broadened task and modality coverage, with the latter spanning radiology, pathology, and dermatology. These resources provide broad measures of medical multimodal capability.

Recent benchmarks increasingly associate model outputs with semantic annotations, anatomical localization, or visual evidence. SLAKE provides structured semantic annotations (Liu et al., 2021), HEAL-MedVQA emphasizes evidence-aligned question answering (Nguyen et al., 2025), and X-PCR organizes evaluation around progressive clinical interpretation (Wang et al., 2026). However, these resources generally do not organize evaluation around a shared reasoning chain that links observation, evidence, and diagnosis for the same anatomical target, and imaging inputs are typically evaluated as fixed observations without comparing performance as available sequences or views change.

Table 1: **Comparison of benchmarks.** Multi-Seq./View denote multiple MRI sequence types/acquisition planes per examination. Stage Eval.: ✓, linked-stage evaluation; ∼, structured reasoning with localization and final diagnosis evaluation only; ✗, neither.

	Dataset	Multi-Seq.	Multi-View	Stage Eval.	#QA	#Images/Scans
Medical VQA	VQA-RAD (Lau et al., 2018)	✗	✗	✗	3,515	315 I
	PathVQA (He et al., 2021)	✗	✗	✗	32,795	4,998 I
	SLAKE (Liu et al., 2021)	✗	✗	✗	14,028[a]	642 I
	PMC-VQA (Zhang et al., 2024)	✗	✗	✗	226,946	149,075 I
	OmniMedVQA (Hu et al., 2024)	✗	✗	✗	127,995	118,010 I
	GMAI-MMBench (Chen et al., 2024b)	✗	✗	✗	25,831	NR
	HEAL-MedVQA (Nguyen et al., 2025)	✗	✗	✗	67,904	≈34K I
	X-PCR (Wang et al., 2026)	✗[b]	✗[b]	✓	177,868	26,415 I
Knee MRI	MRNet (Bien et al., 2018)	✓	✓	✗	–	1,370 S
	fastMRI+ (Knee) (Zhao et al., 2022)	✗	✗	✗	–	1,172 S
	SKM-TEA (Desai et al., 2021)	✗[c]	✗[d]	✗	–	155 S
	3DReasonKnee (Sambara et al., 2025)	✗	✗[d]	∼	≈494K[e]	7,970 V
	MultiKnee (Ours)	✓	✓	✓	**821,823**	**11,575 S**

I: images; S: scans/exams/studies; V: volumes. NR: unique-image count unverified. [a] English and Chinese. [b] Cross-modal ophthalmic inputs, not MRI sequences/planes. [c] Dual-echo qDESS, not distinct sequences. [d] 3D reformats, not separate acquisition planes. [e] Quintuples: MRI, question, box, reasoning, diagnosis.

Knee MRI Datasets. Knee MRI resources span classification, reconstruction, dense annotation, and spatially grounded diagnosis. MRNet supports abnormality, anterior cruciate ligament tear, and meniscal tear classification (Bien et al., 2018); fastMRI+ augments reconstruction data with pathology annotations (Zhao et al., 2022); and SKM-TEA provides quantitative MRI with dense image labels (Desai et al., 2021). More recently, 3DReasonKnee combines volumetric MRI with diagnostic questions, 3D bounding boxes, and reasoning chains, advancing spatially grounded vision–language evaluation (Sambara et al., 2025). However, these resources are limited in two respects: they are either not VQA-oriented or not multi-sequence, multi-view, and none provides fine-grained annotations at scale. As summarized in Table 1, **MultiKnee** is the first large-scale knee MRI benchmark that uniquely combines fine-grained multi-sequence, multi-view annotations with OEDChain hierarchy and controlled-observation evaluation, enabling model behavior to be characterized across clinical reasoning stages, anatomical systems, and MRI observation conditions.

3 OEDCHAIN: HIERARCHICAL EVALUATION FRAMEWORK

3.1 OVERVIEW OF OEDCHAIN FRAMEWORK

Knee MRI interpretation is a multi-stage reasoning process: a radiologist perceives image features, interprets them in anatomical context, and reaches a diagnosis. These stages must be evaluated on the *same* anatomical target; otherwise, performance differences across stages are confounded by target difficulty variation rather than reflecting reasoning capability. The proposed **OEDChain** addresses this by decomposing interpretation into three linked stages, as illustrated in Fig. 2.

Observation assesses whether a model perceives clinically relevant image features across sequences and views. Knee MRI interpretation requires integrating complementary tissue contrasts from T1, T2, and PD sequences, as well as spatial information from axial, coronal, and sagittal views. Existing benchmarks typically present a single image or a fixed set of images and ask models to answer directly, without isolating whether the model has actually perceived the relevant features. Observation questions address this by grounding models' perception in the image data, requiring models to identify the bounding box, slice and sequence, and visual characteristics of features.

Evidence assesses whether the model interprets observed features within anatomical context. A signal abnormality in the medial meniscus, for example, must be recognized not merely as a visual pattern but as evidence located within a specific anatomical structure. Evidence builds upon Observation: a model may perceive a hyperintense region on a T2-weighted image yet fail to localize it to the posterior horn of the medial meniscus. Evidence questions target standardized imaging evidence (*e.g.*, tear, degeneration, joint effusion) for the same anatomical target as Observation.

Diagnosis assesses whether the model reaches the correct clinical conclusion by integrating the evidence into a radiological diagnosis, such as distinguishing a meniscal tear from degeneration. Diagnosis questions target clinical interpretations (*e.g.*, meniscus tear, osteoarthritis, ligament sprain) and require the diagnosis to be explicitly grounded in the evidence established at the previous stage.

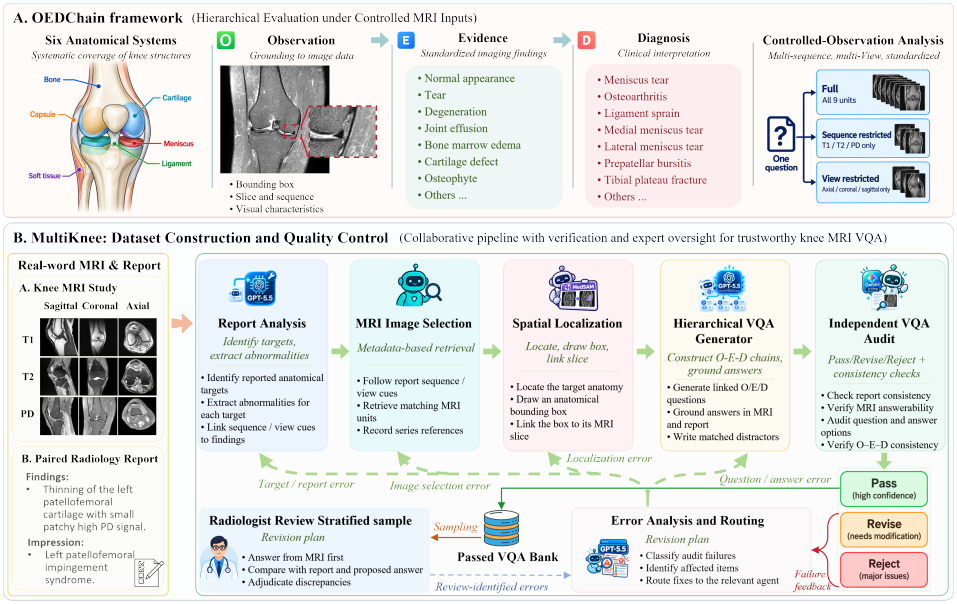

Figure 2: (A) **OEDChain framework** evaluates linked Observation–Evidence–Diagnosis chains across six anatomical systems under full, sequence-restricted, and view-restricted observation. (B) **MultiKnee construction pipeline** builds these chains through *Report Analysis*, *MRI Image Selection*, *Spatial Localization*, and *Hierarchical VQA Generation*.

These three stages form a chain: each targets the same anatomical structure within an MRI study, tracing a complete reasoning pathway from image perception to diagnostic conclusion. This chain structure makes model reasoning traceable: errors can be attributed to specific stages, revealing where and why the reasoning process breaks down.

3.2 Controlled-Observation Analysis

Knee MRI interpretation relies on complementary sequences and views, each providing distinct diagnostic information. To enable such analysis, we treat the MRI input as a controlled variable: seven observation conditions are evaluated while keeping the question, answer options, and reference answer fixed. **Full observation** provides all nine sequence–view units as the baseline. **Sequence-restricted observation** provides T1-only, T2-only, or PD-only images across the three views, revealing reliance on specific tissue contrasts. **View-restricted observation** provides axial-only, coronal-only, or sagittal-only images across the three sequences, revealing dependence on specific spatial orientations. Formally, for study i, the restricted inputs are

$$\mathcal{I}_i^s = \{\boldsymbol{o}_{ik} \in \mathcal{I}_i : s_{ik} = s\}, \qquad \mathcal{I}_i^v = \{\boldsymbol{o}_{ik} \in \mathcal{I}_i : v_{ik} = v\}, \tag{1}$$

where $s \in \{\mathrm{T1}, \mathrm{T2}, \mathrm{PD}\}$ and $v \in \{\mathrm{Axial}, \mathrm{Coronal}, \mathrm{Sagittal}\}$. By comparing a model's performance under full observation against each restricted condition, we can determine whether the complementary information across sequences and views is essential for accurate diagnostic reasoning.

3.3 OEDChain Evaluation Metrics

We report multiple-choice accuracy for each clinical level and anatomical system. The global score at each level is the unweighted mean of its six system-specific accuracies. Let $S_{m,l,c}$ denote the global score of model m at clinical level $l \in \{\mathrm{O}, \mathrm{E}, \mathrm{D}\}$ under observation condition c. To compare clinical levels under full observation, we define the hierarchical performance gap as

$$G_m^{x-y} = S_{m,x,\mathrm{full}} - S_{m,y,\mathrm{full}}, \qquad (x, y) \in \{(\mathrm{O}, \mathrm{E}), (\mathrm{E}, \mathrm{D}), (\mathrm{O}, \mathrm{D})\}. \tag{2}$$

$G_m^{\mathrm{O-E}}$ and $G_m^{\mathrm{E-D}}$ compare adjacent levels, while $G_m^{\mathrm{O-D}}$ compares observation and diagnosis. By definition, $G_m^{\mathrm{O-D}} = G_m^{\mathrm{O-E}} + G_m^{\mathrm{E-D}}$. A positive gap indicates higher accuracy at the first level. For restriction type $r \in \{\mathrm{seq}, \mathrm{view}\}$, we define full-observation advantage as

$$A_{m,l}^r = S_{m,l,\mathrm{full}} - \frac{1}{|\mathcal{C}_r|} \sum_{c \in \mathcal{C}_r} S_{m,l,c}, \tag{3}$$

and observation variation as

$$V^r_{m,l} = \max_{c \in \mathcal{C}_r} S_{m,l,c} - \min_{c \in \mathcal{C}_r} S_{m,l,c}. \tag{4}$$

Positive $A^r_{m,l}$ indicates higher accuracy under full observation than the mean of the restricted conditions. $V^r_{m,l}$ measures the accuracy range across those conditions, and a small value indicates consistent performance across sequences or views, but does not by itself imply high accuracy.

4 MULTIKNEE: DATASET CONSTRUCTION

MultiKnee is a knee MRI VQA dataset built to support the evaluation protocol defined in Sec. 3. Its construction requires meeting two conditions that existing resources do not satisfy: (i) each question must be linked into an Observation–Evidence–Diagnosis chain targeting the same anatomical structure, and (ii) each question must be grounded in fine-grained MRI annotations aligned across sequences and views, so that the available imaging information can be systematically varied during evaluation. Importantly, we establish and validate questions under full observation; we then use sequence- and view-restricted inputs as controlled perturbations to assess how model performance changes when we withhold complementary MRI information. Meeting these conditions is non-trivial: each item must be grounded in a specific MRI observation, the evidence-level interpretation must follow from the observation-level description, and the diagnosis must be supported by the evidence, while preserving alignment across the nine sequence–view combinations.

Implemented with LangChain (LangChain Inc.), the construction pipeline comprises five stages: report analysis, MRI unit selection, spatial localization, hierarchical VQA generation, and quality control. Quality control uses Gemini 3.1 Pro (Google DeepMind, 2026a) for independent auditing and GPT-5.5 (OpenAI, 2026) for error analysis and routing, supplemented by sampled radiologist review. Each stage records outputs and source references to enable targeted revision and re-auditing.

4.1 DATA COLLECTION

We collected 15,908 knee MRI studies from 14,557 patients at a collaborating hospital between 2016 and 2025. After expert checking to exclude incomplete or non-standard studies, we retained 11,575 studies from 10,773 patients, each paired with a radiology report written by radiologists as part of routine clinical practice. The retained studies contain T1-, T2-, and proton-density (PD)-weighted image series acquired in axial, coronal, or sagittal views, reflecting the standard clinical protocol for knee MRI examinations. This composition provides complementary information across sequences and views, and the diversity of anatomical targets and pathological conditions forms the basis for constructing O–E–D chains across a wide range of clinical scenarios.

To ensure data quality, four reviewers checked the MRI studies using automated scripts and manual review over six months. Each retained study was verified to contain MRI series that could be mapped to all nine sequence-view combinations and had no artifacts that compromised interpretation. Overall screening yielded a retention rate of 72.8%. The resulting dataset supports controlled-observation analysis, in which the available sequence–view units vary while the question, answer options, and reference answer remain fixed. We define each sequence–view combination as one observational unit. For study i, the MRI input is

$$\mathcal{I}_i = \left\{ \boldsymbol{o}_{ik} = (s_{ik},\, v_{ik},\, \boldsymbol{I}_{ik}) \right\}_{k=1}^{K_i}, \tag{5}$$

where s_{ik} is the sequence, v_{ik} is the view, and $\boldsymbol{I}_{ik}$ is the corresponding images. Given the three sequences and three views described above, each study contains $K_i = 9$ observational units, summing to 104,175 in total. These form the basis for controlled-observation analysis.

4.2 CONSTRUCTION PIPELINE

Report Analysis Each MRI study in MultiKnee is paired with a radiology report written during routine clinical practice. GPT-5.5 (OpenAI, 2026) parses the report to extract anatomical targets, reported abnormalities, and sequence- or view-specific cues, while retaining references to the source text. These free-text findings are normalized into structured targets (*e.g.*, anatomical structure, evidence type, and location) for consistent use across the O–E–D chain. Because reports vary

in phrasing and granularity, source references are preserved to support downstream grounding, verification, and error tracing. The resulting targets and imaging cues guide subsequent MRI grounding but are not treated as image-verified evidence.

MRI Image Selection MRI image selection retrieves the sequence-view units needed to support subsequent question generation and answering. A custom program matches report-derived sequence and view cues to standardized MRI metadata and records the source references of the retrieved series, linking subsequent questions to their corresponding MRI inputs.

Spatial Localization Spatial localization links report-derived anatomical targets to candidate visual evidence in the retrieved MRI series. Since reported findings may be subtle or not clearly visible, metadata-based retrieval alone cannot establish their presence in the images. Gemini 3.1 Pro (Google DeepMind, 2026a) selects relevant slices and proposes an initial bounding box for each target. Each box prompts MedSAM (Ma et al., 2024) to generate a candidate mask. The boxes, masks, and source slices provide spatial references for subsequent VQA generation and audit. Separating series selection from spatial localization allows retrieval and localization errors to be traced to their respective steps.

Hierarchical VQA Generation Hierarchical VQA constructs MultiKnee's core artifacts: linked Observation–Evidence–Diagnosis chains targeting the same anatomical structure within one MRI study. Each chain contains an observation question on image characteristics, an evidence question interpreting those characteristics in anatomical context, and a diagnosis question on the supported radiological conclusion.

We use GPT-5.5 OpenAI (2026) as the generator, but independent generation can produce plausible yet inconsistent cross-level answers. To maintain the cross-level consistency of question, We use three complementary constraints. First, the radiology report provides a reference clinical interpretation for the O–E–D progression. Second, MRI grounding anchors each level to image evidence. Third, GPT-5.5 generates the chain sequentially from O to E to D, conditioning each later level on the preceding answer together with the report-derived target, selected MRI observations, and candidate localization. Each question is paired with three plausible distractors from the same anatomical region or condition family, and both individual questions and cross-level links are audited before retention.

4.3 Quality Control

Dataset construction involved 22 rounds of independent auditing, error-directed revision, and radiologist review. Per-round generation, expert review counts and pass rates are in the Appendix.

Independent audit. Gemini 3.1 Pro (Google DeepMind, 2026a) checks report consistency, MRI answerability, question and option quality, and O–E–D consistency: each chain must address the same anatomical target in the same study, with observation supporting evidence interpretation and evidence supporting diagnosis. Items receive pass, revise, or reject decisions with reason codes. Items requiring revision are corrected and re-audited; unsupported items are rejected. Chains are retained only when all questions and cross-level links pass.

Error analysis and routing. GPT-5.5 (OpenAI, 2026) traces audit or review identified failures to report analysis, MRI unit selection, spatial localization, or VQA generation and routes items for correction. It flags related items sharing potentially erroneous targets or rules and refines prompts for next round. Corrected and flagged items are re-audited before retention, including cross-level checks for affected chains.

Table 2: **MultiKnee statistics.** System counts are illustrative chain counts (one question each for O/E/D). Splits are patient-disjoint.

Split	#Studies	#Chains by Anatomical System						#Chains	#QA
		Cart.	Bone	Cap.	Lig.	Men.	Soft.		
Train	9,260	70,181	52,696	15,336	19,655	26,205	35,080	219,153	657,459
Validation	1,157	8,680	6,640	1,880	2,440	3,372	4,370	27,382	82,146
Test	1,158	8,800	6,410	1,960	2,560	3,296	4,380	27,406	82,218
Total	**11,575**	**87,661**	**65,746**	**19,176**	**24,655**	**32,873**	**43,830**	**273,941**	**821,823**

Cart./Cap./Lig./Men./Soft.: articular cartilage, capsule/bursae, ligament, meniscus, periarticular soft tissue.

Radiologist review. Four Radiologists randomly sample audit-passed items, including revised items, from each construction round: around 10% initially and 100 items per round as generation volume grows. They first answer using MRI inputs blinded to reports, then inspect and record disagreements and causes. Identified errors enter the routing module for correction and re-audit.

Table 3: Full-observation accuracy (%) across O–E–D stages and anatomical systems. Global scores are macro-averaged over the six systems within each stage. **Bold**: best; underlined: second-best.

	Model	Observation						Evidence						Diagnosis						Global		
		Cart.	Bone	Cap.	Lig.	Men.	Soft.	Cart.	Bone	Cap.	Lig.	Men.	Soft.	Cart.	Bone	Cap.	Lig.	Men.	Soft.	Obs.	Evid.	Diag.
Comm.	GPT-5.5 (OpenAI, 2026)	87.64	86.21	89.38	91.24	84.95	87.76	80.76	79.82	83.54	**86.27**	77.63	83.92	60.84	57.96	70.70	64.18	59.73	67.42	87.86	81.99	63.47
	Gemini 3.1-Pro (Google DeepMind, 2026a)	87.26	82.94	**90.18**	**94.37**	78.65	**97.01**	81.74	82.36	83.95	86.12	78.46	**84.74**	**64.82**	**67.34**	72.16	69.21	**66.03**	69.85	88.40	82.90	68.24
	Claude Opus 4.7 (Anthropic, 2026a)	86.48	83.02	88.16	92.84	78.04	90.54	74.83	72.91	69.44	78.26	67.85	71.41	40.27	37.66	52.40	45.76	35.44	43.81	86.51	72.45	42.56
	Claude Sonnet 4.6 (Anthropic, 2026b)	80.12	75.84	86.73	93.42	69.46	92.18	71.86	69.74	82.31	77.42	66.95	77.86	51.49	46.58	69.20	60.88	47.47	54.91	82.96	74.36	55.09
M.S.	MedGemma-1.5-4B-IT (Sellergren et al., 2025)	39.87	44.48	53.56	54.92	39.51	41.36	33.25	34.12	34.07	51.48	32.30	36.09	28.69	24.25	29.83	29.02	28.61	35.38	45.61	36.89	29.30
	Hulu-Med-7B (Jiang et al., 2025)	45.67	38.87	47.67	49.51	31.68	49.35	39.17	38.78	43.25	44.03	31.63	40.38	26.19	24.23	41.35	34.16	27.15	38.45	43.79	39.54	31.92
	Lingshu-7B (Xu et al., 2025)	30.42	31.20	35.10	46.80	29.40	44.87	30.20	34.80	38.40	29.10	33.20	27.14	27.54	31.36	23.75	26.28	28.91	27.00	36.30	32.14	27.47
	HuatuoGPT-Vision-7B (Chen et al., 2024a)	44.15	47.41	53.71	38.33	43.17	41.54	31.97	33.84	35.43	36.76	38.40	41.04	24.33	23.70	28.19	24.77	30.41	31.32	44.72	36.24	27.12
Open-Source	Qwen3.8-27B (Qwen Team, 2026a)	68.32	69.75	72.41	73.63	65.78	76.73	50.95	55.68	63.72	76.41	49.86	64.78	40.35	44.07	55.22	67.79	46.99	57.31	71.10	60.23	51.96
	Qwen3.5-9B (Qwen Team, 2026b)	54.28	61.73	67.45	70.16	64.82	74.90	53.24	35.28	33.55	59.84	38.29	52.21	28.64	35.72	31.56	39.56	26.34	38.90	65.56	45.40	33.45
	Qwen3.5-4B (Qwen Team, 2026b)	58.42	49.36	46.18	38.27	31.85	32.75	27.85	24.18	32.75	26.42	28.31	27.96	24.85	27.91	22.69	26.87	23.30	25.55	42.81	27.91	25.20
	InternVL3.5-38B (Wang et al., 2025)	58.20	63.40	68.50	71.80	70.60	72.26	46.20	52.10	59.30	63.40	49.80	61.72	41.86	48.58	59.39	51.70	44.50	54.61	67.46	55.42	50.11
	InternVL3.5-8B (Wang et al., 2025)	59.42	56.18	66.75	69.59	54.63	71.84	45.80	51.60	58.90	61.70	48.90	52.83	42.60	37.80	31.40	49.20	34.10	48.93	63.07	53.29	40.67
	InternVL3.5-4B (Wang et al., 2025)	48.72	50.15	60.84	69.63	57.26	75.74	38.24	43.67	43.77	48.92	35.46	49.32	35.40	31.80	31.51	49.80	35.60	32.55	60.39	43.23	36.11
	Gemma 4-12B (Google DeepMind, 2026b)	61.84	66.72	69.48	70.36	64.15	70.76	55.48	57.36	63.14	68.90	60.27	65.54	40.62	35.47	30.18	45.83	36.74	43.44	67.22	61.78	38.71
	Gemma-4-4B (Google DeepMind, 2026b)	49.85	54.26	58.47	63.75	47.92	68.27	42.36	48.61	55.52	63.10	43.08	62.55	39.61	34.37	35.86	48.59	30.73	49.26	57.09	52.54	39.74
FT	InternVL3.5-8B-FT	74.26	79.18	81.73	82.64	83.92	80.94	61.42	65.87	72.36	68.54	59.73	72.10	52.56	49.42	59.84	53.18	47.76	61.71	80.45	66.67	54.08
	Gemma-4-4B-FT	72.84	77.36	80.91	83.47	85.62	86.76	58.42	62.17	65.83	68.56	70.91	73.88	39.48	43.62	46.75	49.36	51.28	53.37	81.16	66.63	47.31
	Qwen3.5-9B-FT	86.91	84.67	80.53	90.14	**87.79**	87.88	74.84	67.87	68.27	77.56	64.91	70.13	42.18	36.74	43.18	46.83	46.91	48.12	86.32	70.60	44.00
	Qwen3.8-27B-FT	**88.27**	**86.94**	89.61	91.84	84.58	90.73	**82.14**	**82.75**	**84.18**	86.03	**79.65**	83.93	**64.82**	67.13	**72.86**	**69.45**	65.97	**70.46**	**88.66**	**83.11**	**68.45**

4.4 Dataset Statistics

Table 2 summarizes MultiKnee by split and anatomical system: 11,575 MRI studies, 273,941 complete O–E–D chains, and 821,823 VQA items. Each chain contains one question per level, yielding 273,941 questions at each of O, E, and D. Train/validation/test splits are patient-disjoint.

5 Experimental and Results

5.1 Experimental Setup

Evaluated Models. We evaluate 20 model configurations on MultiKnee: four commercial MLLMs (GPT-5.5 (OpenAI, 2026), Gemini 3.1-Pro (Google DeepMind, 2026a), Claude Opus 4.7 (Anthropic, 2026a), and Claude Sonnet 4.6 (Anthropic, 2026b)); four medical-specialized models (MedGemma-1.5-4B-IT (Sellergren et al., 2025), Hulu-Med-7B (Jiang et al., 2025), Lingshu-7B (Xu et al., 2025), and HuatuoGPT-Vision-7B (Chen et al., 2024a)); eight general-purpose open-source models from the Qwen3.8 (Qwen Team, 2026a), Qwen3.5 (Qwen Team, 2026b), InternVL3.5 (Wang et al., 2025), and Gemma-4 (Google DeepMind, 2026b) families; and four fine-tuned variants, Qwen3.8-27B-FT, Qwen3.5-9B-FT, InternVL3.5-8B-FT, and Gemma-4-4B-FT, to assess the benefits of structured supervision across clinical stages and observation conditions.

Implementation details. We use a patient-level 8:1:1 split, with 9,260 training, 1,157 validation, and 1,158 test studies; all models share the same test split. Controlled-observation evaluation varies only the available sequence–view units, keeping the question, options, and target fixed. Open-source models use greedy decoding (`do_sample=False`, `max_new_tokens=1024`). Commercial APIs use temperature $= 0$ when supported, except Claude Opus 4.7 (Anthropic, 2026a), which uses its default sampling configuration. Fine-tuning uses NVIDIA H100 GPUs: one GPU for models up to 4B parameters and multiple GPUs for larger models. Additional details are provided in the appendix.

5.2 Hierarchical and Anatomical Performance

Table 3 and Fig. 3 reveal three patterns. First, performance declines across O–E–D, but drop locations and magnitudes vary across models. Second, medical specialization alone does not ensure transfer to multi-sequence, multi-view knee MRI. Third, models show shared anatomical patterns that shift across clinical stages. Together, these results show that diagnosis accuracy alone is insufficient to characterize model behavior.

Stage-wise performance profiles. Commercial models generally lead in observation and evidence, but this advantage is less consistent at diagnosis. For example, Claude Opus 4.7 achieves 86.51% at O and 72.45% at E, but only 42.56% at D, below Qwen3.8-27B at 51.96%. This shows that strong early-stage performance does not guarantee strong diagnosis. Models with similar O–D gaps can also fail at different stages. Qwen3.8-27B and Gemini 3.1-Pro show similar O–D gaps (19.14 vs. 20.16), but Qwen3.8-27B drops by 10.87/8.27 across O–E/E–D, whereas Gemini drops by 5.50/14.66.

Stage-dependent anatomical patterns. Despite differences in overall accuracy, models exhibit shared anatomical performance patterns within each clinical stage. Among the four commercial models, ligament and periarticular soft tissue are consistently among the strongest categories at observation, whereas capsule-related questions achieve the highest diagnosis accuracy. These patterns indicate that anatomical difficulty is stage-dependent rather than uniform across the O–E–D chain.

Medical task alignment. Medical-specialized models achieve 27.12–31.92% diagnosis accuracy, below Qwen3.8-27B at 51.96%. This suggests that medical specialization alone does not ensure transfer to multi-sequence, multi-view knee MRI. However, backbone differences confound this comparison: Lingshu-7B is built on Qwen2.5-VL-7B-Instruct (Xu et al., 2025), whereas HuatuoGPT-Vision-7B uses Qwen2-7B within a LLaVA-v1.5 architecture (Chen et al., 2024a). Performance therefore reflects both backbone capability and task alignment.

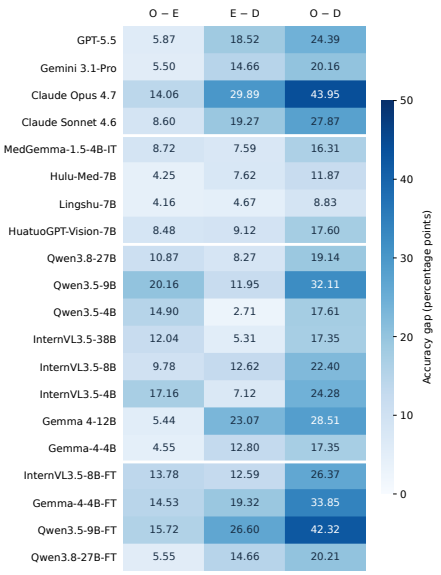

Figure 3: Hierarchical performance gaps under full observation.

5.3 Controlled Observation

Fig. 4 show that performance depends on both MRI input composition and clinical stage, with full observation not consistently outperforming restricted inputs.

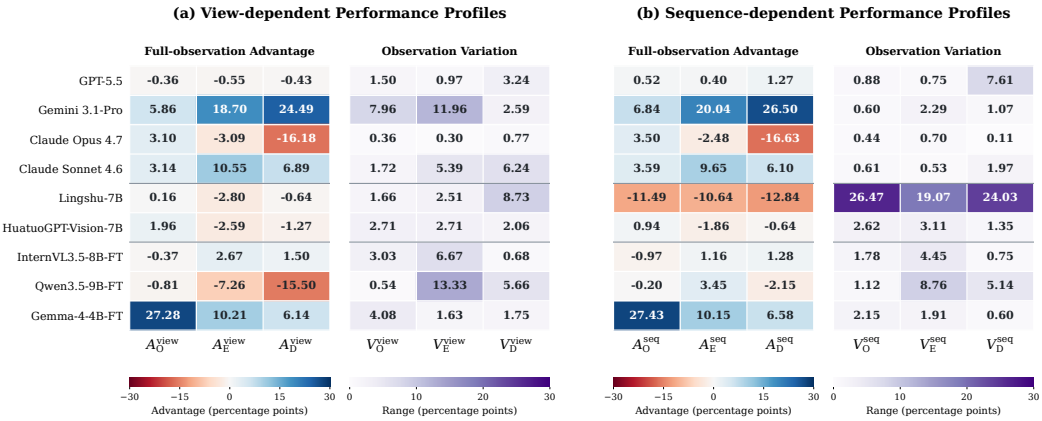

Figure 4: View- and sequence-dependent performance profiles.

More available information does not guarantee better diagnosis. Only four of the nine configurations show positive view-based diagnosis advantages. Gemini 3.1-Pro benefits strongly from full observation ($A_D^{view} = 24.49$, $A_D^{seq} = 26.50$), whereas Claude Opus 4.7 shows $-16.18/-16.63$ and Qwen3.5-9B-FT $-15.50/-2.15$ under view/sequence restriction. These contrasts show that access to complementary MRI information does not guarantee effective utilization. Moreover, A reflects an average benefit: GPT-5.5 has $A_D^{seq} = 1.27$, although its PD-only diagnosis accuracy (65.31%) exceeds full observation (63.47%). Thus, negative or positive A should be interpreted as differences in input utilization rather than intrinsic benefit or harm of additional MRI information.

Sequence and view dependence is model-specific. Restricted inputs reveal distinct dependence profiles across models. For Lingshu-7B, diagnosis variation is 8.73 points across views but 24.03 across sequences, with accuracies of 30.98%, 55.01%, and 34.95% under PD-, T1-, and T2-only inputs. In contrast, Gemini 3.1-Pro shows greater evidence-level variation across views ($V_E^{view} = 11.96$) than sequences ($V_E^{seq} = 2.29$). Claude Opus 4.7 further shows that low variation does not imply strong performance: $V_D^{view} = 0.77$ and $V_D^{seq} = 0.11$, despite only 42.56% full-observation diagnosis accuracy. These results show that MRI input dependence is strongly model-specific.

Table 4: **MRI dependence.** Accuracy (%) on the same questions with and without MRI images.

	Observation		Evidence		Diagnosis	
Model	**MRI**	**No MRI**	**MRI**	**No MRI**	**MRI**	**No MRI**
GPT-5.5 (OpenAI, 2026)	87.86	33.45	81.99	26.68	63.47	22.96
Gemini 3.1-Pro (Google DeepMind, 2026a)	88.40	31.87	82.90	24.56	68.24	21.22
Claude Opus 4.7 (Anthropic, 2026a)	86.51	29.61	72.45	22.47	42.56	24.01
Qwen3.8-27B (Qwen Team, 2026a)	71.10	25.42	60.23	22.36	51.96	23.64
Qwen3.5-9B-FT	86.32	38.95	70.60	30.09	44.00	25.77

Input dependence changes across clinical stages. Gemma-4-4B-FT shows a view-based advantage of 27.28 points at O, decreasing to 10.21 at E and 6.14 at D, whereas Gemini 3.1-Pro shows the opposite trend (5.86, 18.70, and 24.49). For Claude Opus 4.7, sequence-based advantage even reverses from 3.50 at O to -2.48 at E and -16.63 at D. Thus, the benefit of additional MRI information depends not only on the model and input composition, but also on the clinical reasoning stage.

5.4 Stage-wise Effects of Structured Fine-Tuning

Fine-tuning improves performance unevenly across the reasoning chain. All four matched model pairs improve across Observation, Evidence, and Diagnosis, demonstrating effective task-specific supervision from MultiKnee. Qwen3.8-27B gains 17.56/22.88/16.49 points across O/E/D, InternVL3.5-8B 17.38/13.38/13.41, and Gemma-4-4B 24.07/14.09/7.57. These stage-specific gains alter model reasoning profiles rather than uniformly improving the chain.

Fine-tuned models retain distinct sequence- and view-dependent profiles. At diagnosis, Gemma-4-4B-FT shows view/sequence advantages of 6.14/6.58 points, whereas Qwen3.5-9B-FT shows -15.50/-2.15; their corresponding variations are 1.75/0.60 and 5.66/5.14. Thus, MultiKnee reveals not only whether fine-tuning improves performance, but also how it alters stage-wise and observation-dependent behavior.

5.5 Benchmark Validation

MRI dependence. We evaluate six configurations on the same held-out questions with and without MRI, keeping options and answers fixed. Removing MRI reduces O/E/D accuracy by 45.68–56.90, 37.87–58.34, and 18.23–47.02 percentage points, respectively (Table 4). These declines confirm that MultiKnee performance depends strongly on imaging input rather than textual cues alone.

Construction workflow validation. We compare multi-stage and single-pass VQA generation on the same 50 MRI studies and task targets, with feedback and prompt updates disabled. Each method produces 2,700 questions in 900 O–E–D chains across six systems. Two blinded musculoskeletal experts independently review each question, which passes only with agreement on MRI answerability, answer correctness, and question/option quality. The multi-stage workflow achieves a 79.68% pass rate versus 41.52% for single-pass generation, indicating higher expert-validated question quality.

6 Conclusion

We introduced OEDChain, a hierarchical framework that decomposes knee MRI interpretation into linked Observation–Evidence–Diagnosis stages over the same anatomical target, with MRI input treated as a controlled variable. Built on this framework, MultiKnee is the first large-scale fine-grained knee MRI benchmark grounded in sequence–view units, comprising 821,823 VQAs. Evaluation of 20 model configurations shows that similar diagnosis-level scores can mask distinct reasoning-chain failure patterns, while the utility of multi-sequence, multi-view information varies substantially across models and clinical stages. Fine-tuning consistently improves performance but produces uneven gains across the reasoning chain, further highlighting the difficulty of translating image evidence into accurate diagnosis. MultiKnee thus provides task-specific supervision and a stage-resolved benchmark for understanding and improving medical MLLMs.

MRI-based diagnostic reasoning still requires further investigation. Future work will explore whether stage-specific supervision can further reduce the gap between evidence interpretation and diagnosis.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, code development, and figure refinement. The authors reviewed and verified the resulting text, code, and figures. Language models were also used as part of the benchmark construction pipeline to generate candidate structured annotations and VQA items from MRI studies and paired radiology reports. The generated questions and their cross-level links underwent the quality-control procedure described in Sec. 4.3, including independent auditing and sampled review by musculoskeletal radiologists. The authors are responsible for the benchmark design, quality-control decisions, experimental analyses, and conclusions.

ETHICS STATEMENT

This study uses retrospective knee MRI examinations and paired radiology reports from a collaborating hospital. Institutional ethics approval was obtained, and all imaging and report data were de-identified before benchmark construction. No participants were recruited and no new clinical data were collected specifically for this study. MultiKnee is intended for research and evaluation and not for direct clinical diagnosis or real-time decision-making.

REPRODUCIBILITY STATEMENT

Sections 3–5 describe the OEDChain framework, MultiKnee construction and quality control, evaluation protocol, dataset splits, model settings, and fine-tuning configuration. Appendices A–E provide additional details on dataset composition, MRI inputs, evaluation and training procedures, construction prompts and auditing, radiologist review, and complete controlled-observation results. Code and benchmark resources will be released upon acceptance, subject to applicable institutional and data-governance requirements.

REFERENCES

Gui Wang, Zehao Zhong, YongSong Zhou, Yudong Li, Ende Wu, Wooi Ping Cheah, Rong Qu, Jianfeng Ren, and Linlin Shen. X-pcr: A benchmark for cross-modality progressive clinical reasoning in ophthalmic diagnosis. *arXiv preprint arXiv:2604.20350*, 2026.

Ryutaro Tanno et al. Collaboration between clinicians and vision–language models in radiology report generation. *Nature Medicine*, 31:599–608, 2025. doi: 10.1038/s41591-024-03302-1.

Felix Busch, Lena Hoffmann, Daniel Pinto Dos Santos, Marcus R Makowski, Luca Saba, Philipp Prucker, Martin Hadamitzky, Nassir Navab, Jakob Nikolas Kather, Daniel Truhn, et al. Large language models for structured reporting in radiology: past, present, and future. *European Radiology*, 35(5):2589–2602, 2025.

Zelin Qiu, Zhuoyao Xie, Huangjing Lin, Yanwen Li, Qiang Ye, Menghong Wang, Shisi Li, Yinghua Zhao, and Hao Chen. Learning co-plane attention across mri sequences for diagnosing twelve types of knee abnormalities. *Nature Communications*, 15(1):7637, 2024.

Yuxue Xie, Zhonghua Hu, Hongyue Tao, Yiwen Hu, Haoyu Liang, Xinmin Lu, Lei Wang, Xiangwen Li, and Shuang Chen. Large language models for efficient whole-organ mri score-based reports and categorization in knee osteoarthritis. *Insights into Imaging*, 16(1):100, 2025.

Roman Johannes Gertz, Thomas Dratsch, Alexander Christian Bunck, Simon Lennartz, Andra-Iza Iuga, Martin Gunnar Hellmich, Thorsten Persigehl, Lenhard Pennig, Carsten Herbert Gietzen, Philipp Fervers, et al. Potential of gpt-4 for detecting errors in radiology reports: implications for reporting accuracy. *Radiology*, 311(1):e232714, 2024.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. SLAKE: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1650–1654. IEEE, 2021. doi: 10.1109/ISBI48211.2021.9434010.

Dung Nguyen, Minh Khoi Ho, Huy Ta, Thanh Tam Nguyen, Qi Chen, Kumar Rav, Quy Duong Dang, Satwik Ramchandre, Son Lam Phung, Zhibin Liao, Minh-Son To, Johan Verjans, Phi Le Nguyen, and Vu Minh Hieu Phan. Localizing before answering: A benchmark for grounded medical visual question answering. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, 2025. doi: 10.24963/ijcai.2025/853.

Ruiyang Zhao, Burhaneddin Yaman, Yuxin Zhang, Russell Stewart, Austin Dixon, Florian Knoll, Zhengnan Huang, Yvonne W. Lui, Michael S. Hansen, and Matthew P. Lungren. fastmri+, clinical pathology annotations for knee and brain fully sampled magnetic resonance imaging data. *Scientific Data*, 9:152, 2022. doi: 10.1038/s41597-022-01255-z.

Arjun D. Desai, Andrew M. Schmidt, Elka B. Rubin, Christopher M. Sandino, Marianne S. Black, Valentina Mazzoli, Kathryn J. Stevens, Robert Boutin, Christopher Ré, Garry E. Gold, Brian A. Hargreaves, and Akshay S. Chaudhari. Skm-tea: A dataset for accelerated mri reconstruction with dense image labels for quantitative clinical evaluation. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2021. doi: 10.48550/arXiv.2203.06823. arXiv:2203.06823.

Sraavya Sambara, Sung Eun Kim, Xiaoman Zhang, Luyang Luo, Shreya Johri, Mohammed Baharoon, Du Hyun Ro, and Pranav Rajpurkar. 3dreasonknee: Advancing grounded reasoning in medical vision language models. In *Biocomputing 2026: Proceedings of the Pacific Symposium*, pages 99–113. World Scientific, 2025.

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. In *Advances in Neural Information Processing Systems*, 2023.

Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Ruifei Zhang, Zhenyang Cai, Ke Ji, et al. HuatuoGPT-Vision, towards injecting medical visual knowledge into multimodal LLMs at scale. *arXiv preprint arXiv:2406.19280*, 2024a.

Jason J. Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific Data*, 5:180251, 2018. doi: 10.1038/sdata.2018.251.

Xuehai He, Zhuo Cai, Wenlan Wei, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Towards visual question answering on pathology images. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Short Papers*, pages 708–718. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-short.90.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Development of a large-scale medical visual question-answering dataset. *Communications Medicine*, 4(1):277, 2024. doi: 10.1038/s43856-024-00709-2.

Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22170–22183, 2024.

Pengcheng Chen, Jin Ye, Guoan Wang, Yanjun Li, Zhongying Deng, Wei Li, Tianbin Li, Haodong Duan, Ziyan Huang, Yanzhou Su, et al. Gmai-mmbench: A comprehensive multimodal evaluation benchmark towards general medical ai. *Advances in Neural Information Processing Systems*, 37: 94327–94427, 2024b.

Nicholas Bien, Pranav Rajpurkar, Robyn L. Ball, Jeremy Irvin, Allison Park, Erik Jones, Michael Bereket, Bhavik N. Patel, Kristen W. Yeom, Katie Shpanskaya, Safwan Halabi, Evan Zucker, Gary Fanton, Derek F. Amanatullah, Christopher F. Beaulieu, Geoffrey M. Riley, Roger J. Stewart, Francis G. Blankenberg, David B. Larson, Richard H. Jones, Curtis P. Langlotz, Andrew Y. Ng, and Matthew P. Lungren. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of mrnet. *PLOS Medicine*, 15(11):e1002699, 2018. doi: 10.1371/journal.pmed.1002699.

LangChain Inc. LangGraph. https://github.com/langchain-ai/langgraph. Software repository. Accessed: 2026-09-24.

Google DeepMind. Gemini 3.1 pro - model card. https://deepmind.google/models/model-cards/gemini-3-1-pro/, February 2026a. Published 19 February 2026. Accessed: 2026-04-26.

OpenAI. Introducing gpt-5.5. https://openai.com/index/introducing-gpt-5-5/, April 2026. OpenAI product release page. Accessed: 2026-04-26.

Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature communications*, 15(1):654, 2024.

Anthropic. What's new in claude opus 4.7. https://platform.claude.com/docs/en/about-claude/models/whats-new-claude-4-7, 2026a. Claude API Docs. Accessed: 2026-04-26.

Anthropic. Introducing claude sonnet 4.6. https://www.anthropic.com/news/claude-sonnet-4-6, 2026b. Accessed: 2026-05-06.

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.

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*, 2025.

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*, 2025.

Qwen Team. Qwen3.8-max: A new bar for coding and cowork, August 2026a. URL https://qwen.ai/blog?id=qwen3.8.

Qwen Team. Qwen3.5-9b. https://huggingface.co/Qwen, March 2026b. Accessed: 2026-05-02.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Google DeepMind. Gemma 4. https://deepmind.google/models/gemma/gemma-4/, 2026b. Accessed: 2026-05-04.

A Dataset Composition and MRI Inputs

A.1 Dataset Composition

Data source and ethics. The MRI studies and corresponding radiology reports were obtained from a collaborating hospital and cover examinations performed between 2016 and 2025. The study received institutional ethics approval, and all imaging and report data were de-identified before benchmark construction. The hospital name and identifying approval details are withheld for double-blind review.

Cohort selection. The initial collection comprised 15,908 knee MRI studies from 14,557 patients. Screening excluded 4,333 studies, leaving 11,575 studies from 10,773 patients in the final cohort. Table A summarizes the exclusion reasons and counts.

Studies were required to have a paired radiology report and usable MRI data that could be mapped to the predefined sequence–view schema. Missing required components led to exclusion unless an equivalent series could be unambiguously mapped to the required component. Non-target or non-standard acquisitions included localizer-only acquisitions and protocols that could not be reliably mapped to the schema. Multiple studies from the same patient were retained if they passed screening.

Table A: Cohort selection and exclusion counts.

Stage or exclusion reason	**Studies**
Initial knee MRI collection	15,908
Missing paired report	1,245
Missing or unmappable required MRI series	688
Non-target or non-standard acquisition	971
Corrupted or non-diagnostic images	1,429
Retained studies	11,575

Linked clinical questions. The retained studies contain 273,941 linked Observation–Evidence–Diagnosis (O–E–D) chains and 821,823 VQA items, with one question at each level of every chain. Each chain concerns the same anatomical target within the same MRI study. Observation questions assess visible MRI characteristics and their spatial references; Evidence questions interpret those observations within a specific anatomical structure; and Diagnosis questions assess the radiological conclusion supported by that evidence.

The training, validation, and test splits are patient-disjoint: all retained studies from the same patient are assigned to one split.

A.2 MRI Inputs

Sequence–view observational units. An observational unit is indexed by an MRI sequence category and an anatomical view. The sequence categories are T1-weighted, T2-weighted, and proton-density-weighted (PD); the views are axial, coronal, and sagittal. Each retained study is represented by the nine sequence–view combinations, yielding 104,175 observational units across 11,575 studies. This count refers to sequence–view units, not individual slices or VQA items.

Controlled observation conditions. Table B defines the seven observation conditions. Full observation includes all nine units. Sequence-restricted conditions retain one sequence across all three views, whereas view-restricted conditions retain one view across all three sequences. The question, answer options, and reference answer remain fixed across conditions; only the available observational units change.

A.3 Clinical Vocabulary

Clinical levels. The ontology follows the clinical reasoning process from observing MRI characteristics, to interpreting them as anatomically grounded evidence, and then to reaching a diagnosis. Table C summarizes the three levels.

Table B: Observation conditions used in controlled evaluation. Ax, Cor, and Sag denote axial, coronal, and sagittal views, respectively. The unit count specifies the number of sequence–view combinations, not the number of input images.

Condition	Included sequence–view units	Units
Full	All combinations of T1, T2, and PD with Ax, Cor, and Sag	9
T1-only	Ax–T1, Cor–T1, Sag–T1	3
T2-only	Ax–T2, Cor–T2, Sag–T2	3
PD-only	Ax–PD, Cor–PD, Sag–PD	3
Axial-only	Ax–T1, Ax–T2, Ax–PD	3
Coronal-only	Cor–T1, Cor–T2, Cor–PD	3
Sagittal-only	Sag–T1, Sag–T2, Sag–PD	3

Table C: **Three reasoning levels in the MultiKnee ontology.**

Level	Definition	Examples
Observation	Visible MRI characteristics and spatial references identified from the images.	Signal intensity, cartilage thickness, ligament continuity, slice and sequence, candidate image region.
Evidence	Interpretation of observed characteristics within a specific anatomical structure or site.	Abnormal signal in the posterior horn of the medial meniscus, ACL fiber disruption, patellofemoral cartilage defect.
Diagnosis	Radiological conclusion supported by the anatomical evidence.	Meniscal tear, ACL rupture, cartilage degeneration, Baker cyst.

Representative Ontology Examples. Table D provides examples of Observation attributes, anatomically grounded Evidence, and Diagnosis categories across the six anatomical systems.

Table D: **Representative ontology examples across O–E–D levels.** Observation describes MRI characteristics; Evidence interprets them within an anatomical structure; Diagnosis gives the supported radiological conclusion.

System	Observation attributes	Anatomically grounded evidence	Diagnosis categories
Articular cartilage	Cartilage thickness; surface regularity; signal alteration; defect depth; defect extent	Patellofemoral cartilage thinning; medial or lateral femorotibial cartilage defect	Patellofemoral cartilage degeneration; focal chondral defect; advanced chondral loss
Bone	T1 marrow signal; T2/PD marrow signal; fracture line; marrow signal pattern	Tibial plateau marrow signal abnormality; subchondral cyst; osteophyte at a specified site	Bone marrow edema; bone contusion; fracture; osteochondral lesion
Capsule/Bursae	Fluid volume; synovial thickening; fluid signal; capsular integrity	Fluid in the joint cavity or suprapatellar bursa; fluid in the popliteal cyst region	Knee joint effusion; suprapatellar effusion; Baker cyst; synovitis
Ligament	Fiber continuity; morphology; thickness; signal intensity; periligamentous edema	ACL fiber disruption; abnormal signal and thickening of the MCL; abnormal signal in the PCL or LCL	ACL partial tear; ACL complete rupture; PCL injury; MCL injury; LCL injury
Meniscus	Intrameniscal signal; linear PD hyperintensity; surface-reaching signal; morphology	Abnormal signal in a specified meniscal horn or root; meniscal extrusion; displaced fragment	Meniscal degeneration; medial meniscus tear; lateral meniscus tear; meniscal root tear
Periarticular soft tissue	Subcutaneous signal; tendon signal; tendon continuity; tendon thickening; fat pad signal	Edema in the prepatellar soft tissue; abnormal signal in the patellar tendon or infrapatellar fat pad	Prepatellar soft-tissue edema; fat pad impingement; tendon injury; periarticular soft-tissue injury

B Evaluation and Fine-Tuning Details

B.1 Question Presentation and Scoring

Each Observation, Evidence, and Diagnosis question is evaluated independently. The model receives only the MRI observations available under the specified observation condition, the question, and four answer options. Radiology reports, answers to other questions in the O–E–D chain, and images outside the specified condition are not provided.

The following instruction is used for MRI-based multiple-choice evaluation:

```
You are interpreting a knee MRI examination.
Answer the multiple-choice question using the
MRI images provided in this request. The images
are labeled by sequence and view. Select the
single best option based on the available images.
Do not use information from a radiology report
or assume that images from other sequences or
views are available.

Output only one capital letter: A, B, C, or D.
Do not provide an explanation.
```

The question is presented in the following format:

```
MRI images:
{labeled_MRI_images}

Question:
{question}

Options:
A. {option_A}
B. {option_B}
C. {option_C}
D. {option_D}

Answer:
```

Here, {`labeled_MRI_images`} denotes the supplied MRI images with their sequence–view labels, such as `[Sagittal--PD]` and `[Coronal--T2]`. The same question, option texts, and reference answer are used under full and restricted observation. Only the available MRI images change: full observation includes all nine sequence–view units, whereas each restricted condition includes three units from one sequence or one view.

We report multiple-choice accuracy for each clinical level and anatomical system. The global accuracy at each level is the unweighted mean of the six anatomical-system accuracies.

B.2 Model Settings

We evaluate 20 model configurations: four commercial models, four medically specialized open-source models, eight general-purpose open-source models, and four fine-tuned variants. Table E lists their API model IDs or checkpoint repositories and the evaluation periods recorded for this study.

For locally evaluated models, decoding uses `do_sample=False` and `max_new_tokens=1024`. For commercial APIs, temperature is set to zero when supported. Claude Opus 4.7 uses its default sampling configuration.

C Fine-Tuning Details

We fine-tune three open-source models on the MultiKnee training split: Qwen3.5-9B, Qwen3.8-27B, InternVL3.5-8B, and Gemma-4-4B. We evaluate their fine-tuned variants against the corresponding original models on the patient-disjoint test split.

C.1 Training Data

The training split contains 9,260 MRI studies and 657,459 VQA items. These items form 219,153 linked Observation–Evidence–Diagnosis (O–E–D) chains. Each chain concerns one anatomical target

Table E: Model identifiers and reported evaluation periods for the 20 evaluated configurations. Repository names identify public model repositories; checkpoint revisions and evaluation periods must be checked against the experimental records.

Reported name	Checkpoint repository or API model ID	Evaluation period
GPT-5.5	`gpt-5.5`	May–Aug 2026
Gemini 3.1-Pro	`gemini-3.1-pro-preview`	May–Aug 2026
Claude Opus 4.7	`claude-opus-4-7`	May–Aug 2026
Claude Sonnet 4.6	`claude-sonnet-4-6`	May–Aug 2026
MedGemma-1.5-4B-IT	`google/medgemma-1.5-4b-it`	May–Aug 2026
Hulu-Med-7B	`ZJU-AI4H/Hulu-Med-7B`	May–Aug 2026
Lingshu-7B	`lingshu-medical-mllm/Lingshu-7B`	May–Aug 2026
HuatuoGPT-Vision-7B	`FreedomIntelligence/HuatuoGPT-Vision-7B`	May–Aug 2026
Qwen3.8-27B	`Qwen/Qwen3.8-27B`	Sep 2026
Qwen3.5-9B	`Qwen/Qwen3.5-9B`	May–Aug 2026
Qwen3.5-4B	`Qwen/Qwen3.5-4B`	May–Aug 2026
InternVL3.5-38B	`OpenGVLab/InternVL3_5-38B`	May–Aug 2026
InternVL3.5-8B	`OpenGVLab/InternVL3_5-8B`	May–Aug 2026
InternVL3.5-4B	`OpenGVLab/InternVL3_5-4B`	May–Aug 2026
Gemma 4-12B	`google/gemma-4-12B-it`	Jun–Aug 2026
Gemma-4-4B	`google/gemma-4-E4B-it`	May–Aug 2026
InternVL3.5-8B-FT	[Actual local checkpoint ID / revision]	[Actual evaluation period]
Gemma-4-4B-FT	[Actual local checkpoint ID / revision]	[Actual evaluation period]
Qwen3.5-9B-FT	[Actual local checkpoint ID / revision]	[Actual evaluation period]
Qwen3.8-27B-FT	[Actual local checkpoint ID / revision]	[Actual evaluation period]

within one study and contains one question at each of the three levels. Thus, the training split contains 219,153 Observation questions, 219,153 Evidence questions, and 219,153 Diagnosis questions.

Dataset partitioning is performed at the patient level: studies from the same patient are assigned to the same split. The fine-tuned models are evaluated only on the held-out test split.

C.2 Training Input and Target

Each training example contains the supplied MRI images, one multiple-choice question, and four answer options. Images are labeled by sequence and view, such as `[Sagittal--PD]` and `[Coronal--T2]`. The question and option format follows the evaluation prompt in Section B.1.

For supervised fine-tuning, the input ends with `Answer:`, and the target completion is the reference option letter, A, B, C, or D. The training target does not include a chain-of-thought explanation or free-form diagnostic report. Questions from the same O–E–D chain have separate option-letter targets; reference answers from other levels are not supplied as input.

D Construction and Human Quality Assessment

D.1 Construction Details

The construction pipeline comprises report analysis, MRI unit selection, spatial localization, hierarchical VQA generation, and quality control. Table F summarizes the input and output of each stage. The prompt templates for stages that invoke a multimodal or language model are given below. All report-derived information is used for dataset construction and verification; radiology report text is not supplied to evaluated models in the MRI-based benchmark evaluation.

MRI unit selection. A custom program maps report-derived sequence and view cues to standardized MRI series metadata. Algorithm 1 summarizes this metadata-based retrieval. The nine available sequence–view units remain associated with the study; report cues identify candidate units for subsequent grounding. This step does not verify a reported abnormality or select a lesion slice.

Table F: Inputs and outputs of the MultiKnee construction stages. The MRI unit selection stage uses a custom program, and MedSAM receives a bounding-box prompt rather than a text prompt.

Stage / method	Input	Output
Report analysis / GPT-5.5	De-identified report and study identifier	Candidate targets, reported findings, anatomical locations, stated sequence/view cues, and supporting report spans
MRI unit selection / custom program	Candidate targets, report-derived cues, and MRI series metadata	Retrieved sequence–view units and source series references
Spatial localization / Gemini 3.1 Pro	Candidate target and retrieved MRI observations	Selected slice and initial anatomical bounding box, linked to the source MRI
Candidate mask / MedSAM	Selected MRI slice and initial bounding box	Candidate segmentation mask
O–E–D generation / GPT-5.5	Report-derived target, selected MRI observations, candidate localization, and preceding generated answers	Three linked questions, four options and one reference answer per level, with supporting MRI references
Independent audit / Gemini 3.1 Pro	Generated chain, MRI references, localization, and report reference	Pass/revise/reject decisions and reason codes for questions and cross-level links
Error analysis and routing / GPT-5.5	Stage outputs and audit or radiologist feedback	Likely error source, proposed revision stage, and items requiring re-audit

Algorithm 1 Metadata-based MRI unit selection

Require: Series metadata $\mathcal{M}_i$ for study i; optional sequence cue s_t and view cue v_t
Ensure: Available units $\mathcal{U}_i$ and candidate units $\mathcal{C}_t$, with source-series references

$\mathcal{U}_i \leftarrow$
for all series r in $\mathcal{M}_i$ **do**
  $s_r \leftarrow \textsc{MapSequence}(r)$
  $v_r \leftarrow \textsc{MapView}(r)$
  if $s_r \in \{\mathrm{T1}, \mathrm{T2}, \mathrm{PD}\}$ **and** $v_r \in \{\mathrm{Ax}, \mathrm{Cor}, \mathrm{Sag}\}$ **then**
    $\mathcal{U}_i \leftarrow \mathcal{U}_i \cup \{(s_r, v_r, \textsc{SeriesRef}(r))\}$
  end if
end for
$\mathcal{C}_t \leftarrow$
for all $u = (s_u, v_u, \mathrm{ref}_u)$ in $\mathcal{U}_i$ **do**
  if (s_t is unspecified **or** $s_u = s_t$) **and** (v_t is unspecified **or** $v_u = v_t$) **then**
    $\mathcal{C}_t \leftarrow \mathcal{C}_t \cup \{u\}$
  end if
end for
return $(\mathcal{C}_t, \mathcal{U}_i)$

Spatial localization. The input to spatial localization consists of a candidate anatomical target from report analysis and the MRI units retrieved for that target. Each supplied image is associated with its source series, sequence, view, and slice identifier. Gemini 3.1 Pro examines these images, selects a slice relevant to the target, and proposes an initial bounding box on that slice. Its requested output is the localization status, the selected image reference, and the box.

The Gemini 3.1 Pro prompt is:

```
You are locating an anatomical target in a
knee MRI study.

Candidate target:
{anatomical_structure_and_site}

Report-derived description:
{reported_observation_or_abnormality}

MRI images:
{MRI_images_with_series_sequence_view_slice_labels}

Examine the supplied MRI images and perform
the following steps:
```

```
1. Identify the image slice that best shows
   the candidate anatomical target.
2. On that slice, propose a bounding box
   enclosing the relevant anatomical region.
3. Report the source series, sequence, view,
   and slice identifier of the selected image.
4. If the target cannot be located reliably,
   return "uncertain" and do not invent a box.

Return the following fields:
localization_status:
source_series_id:
sequence:
view:
slice_id:
bounding_box:
brief_reason:

Use the MRI images to select the slice and
box. The report-derived description is a
candidate cue, not confirmation that the
reported abnormality is visible.
```

When Gemini 3.1 Pro returns a usable box, the selected MRI slice and box are supplied to MedSAM to generate a candidate segmentation mask. Thus, MedSAM takes an image and a spatial box prompt as input and produces a mask as output; no natural-language prompt is used at this step. The source slice, initial box, and candidate mask are retained together as spatial references for VQA generation and audit.

These outputs establish candidate anatomical localization, not the correctness of the report-derived diagnosis. Keeping the source series and slice references also allows errors in MRI unit retrieval to be distinguished from errors in slice selection or localization.

Hierarchical VQA generation. GPT-5.5 constructs one linked Observation–Evidence–Diagnosis (O–E–D) chain for an anatomical target within one MRI study. Generation consists of three successive calls. All calls receive the study and target identifiers, the report-derived clinical reference, the selected MRI observations, and candidate spatial references. The Evidence call additionally receives the generated Observation question and answer; the Diagnosis call receives the generated Observation and Evidence questions and answers.

The report supplies a reference clinical interpretation, whereas the MRI observations and spatial references supply image-based grounding. A question is not considered MRI-answerable solely because its proposed answer agrees with the report.

Observation generation. The input is the target, report reference, selected MRI observations, and candidate localization. The output is one Observation question with four options, one proposed reference answer, and the supporting MRI source reference. The prompt template is:

```
You are generating the Observation question
of a knee MRI O--E--D chain.

Study ID:
{study_id}

Anatomical target and site:
{target}

Report-derived reference and source span:
{report_reference}
```

```
MRI images, labeled by source series,
sequence, view, and slice:
{selected_mri_images}

Candidate localization:
{selected_slice_box_and_mask}

Write ONE multiple-choice question about
a visible MRI characteristic or its
image-based spatial reference for this
target. The question should test what can
be observed in the supplied MRI images;
do not ask for the final diagnosis.

Provide exactly four options labeled
A, B, C, and D. Select one reference
answer and write three plausible
distractors relevant to the same
anatomical region.

Return:
study_id
target
level: Observation
question
option_A
option_B
option_C
option_D
reference_option
supporting_mri_source

If the proposed answer cannot be supported
by the supplied MRI, flag the question
instead of treating the report as sufficient
image evidence.
```

Evidence generation. The input additionally includes the generated Observation question and its proposed answer. The output is an Evidence question interpreting the observation in the same anatomical structure or site:

```
You are generating the Evidence question
of the SAME knee MRI O--E--D chain.

Study ID:
{study_id}

Anatomical target and site:
{target}

Report-derived reference and source span:
{report_reference}

MRI images and candidate localization:
{selected_mri_images_and_localization}

Preceding Observation question:
{observation_question}
```

```
Preceding Observation reference answer:
{observation_answer}

Write ONE multiple-choice question asking
how the observed MRI characteristics
should be interpreted within the specified
anatomical structure or site.

Keep the study and anatomical target
unchanged. The intended Evidence answer
must follow from the Observation answer
and be assessable from the supplied MRI.
Do not ask for the final diagnosis.

Provide exactly four options labeled
A, B, C, and D: one proposed reference
answer and three plausible alternatives
from the same anatomical region or
finding family.

Return:
study_id
target
level: Evidence
question
option_A
option_B
option_C
option_D
reference_option
supporting_mri_source
link_to_observation

If the interpretation is not supported
by the images and the preceding
Observation answer, flag the question.
```

Diagnosis generation. The input includes both preceding questions and their proposed answers. The output is a Diagnosis question for the same target:

```
You are generating the Diagnosis question
of the SAME knee MRI O--E--D chain.

Study ID:
{study_id}

Anatomical target and site:
{target}

Report-derived clinical reference
and source span:
{report_reference}

MRI images and candidate localization:
{selected_mri_images_and_localization}

Observation question and reference answer:
{observation_question_and_answer}
```

```
Evidence question and reference answer:
{evidence_question_and_answer}

Write ONE multiple-choice question asking
for the radiological conclusion supported
by the established anatomical evidence.

Keep the study and anatomical target
unchanged. The intended Diagnosis answer
must be consistent with the Observation
and Evidence answers and supported by
the supplied MRI observations. Agreement
with the report alone is not sufficient.

Provide exactly four options labeled
A, B, C, and D: one proposed reference
answer and three plausible diagnoses
from the same condition family or
anatomical region.

Return:
study_id
target
level: Diagnosis
question
option_A
option_B
option_C
option_D
reference_option
supporting_mri_source
link_to_observation_and_evidence

If the diagnosis cannot be supported
by the available MRI evidence, flag
the question for review.
```

The three generated questions, proposed answers, options, MRI source references, and cross-level links form a candidate chain for independent audit. Preceding answers are used in *construction* to maintain cross-level consistency; they are not thereby supplied to evaluated models when individual questions are presented.

Independent audit. Gemini 3.1 Pro audits the generated O–E–D chain separately from the GPT-5.5 generation step. Its inputs are the three generated questions and proposed answers, their shared study and anatomical target, the supporting MRI observations and spatial references, and the relevant report reference.

The audit evaluates each question for MRI answerability, support for its proposed answer, consistency with the report reference, and single-answer question and option quality. It then evaluates whether all three questions concern the same target in the same study and whether Observation supports Evidence and Evidence supports Diagnosis.

The Gemini 3.1 Pro prompt template is:

```
You are independently auditing a
knee MRI multiple-choice O--E--D chain.
You did not generate these questions.

Study ID:
```

```
{study_id}

Shared anatomical target:
{anatomical_target_and_site}

Relevant report reference:
{report_text_or_supporting_spans}

MRI images, labeled by source series,
sequence, view, and slice:
{labeled_mri_images}

Candidate spatial references:
{selected_slices_boxes_and_masks}

Generated Observation question,
options, and proposed answer:
{observation_item}

Generated Evidence question,
options, and proposed answer:
{evidence_item}

Generated Diagnosis question,
options, and proposed answer:
{diagnosis_item}

Audit EACH question independently:

1. Can its intended answer be determined
   from the supplied MRI images?
2. Is the proposed reference answer
   supported by those images?
3. Is the answer consistent with the
   relevant report reference?
4. Are the question and four options
   clear and appropriate for a
   single-answer multiple-choice task?
5. Do the distractors avoid being
   equally correct based on the
   available MRI evidence?

Then audit the CHAIN:

6. Do all three questions concern the
   same anatomical target within the
   same MRI study?
7. Does the proposed Observation answer
   support the Evidence interpretation?
8. Does the proposed Evidence answer
   support the Diagnosis conclusion?

For each of the three questions and
each of the two cross-level links,
return:
- decision: pass, revise, or reject
- reason_code
- brief explanation
- affected question, option, answer,
```

```
  image reference, or chain link

Return an overall chain decision.
Mark the chain as pass only if all
three questions and both cross-level
links pass.

Do not approve a proposed answer
solely because it matches the report.
If the MRI evidence is insufficient,
state that explicitly. Do not silently
replace the proposed reference answer.
```

The audit output contains a decision and reason code for each question and cross-level link, together with an overall chain decision. Items marked for revision are routed to the affected construction stage and re-audited after correction. Unsupported items are rejected. A chain is retained only when its questions and cross-level links pass.

Error analysis and routing. When an item fails independent audit or a radiologist identifies a discrepancy, GPT-5.5 receives the failure feedback, available outputs from preceding construction stages, and their source references. It identifies which stage is most likely responsible: report analysis, MRI unit selection, spatial localization, or VQA generation.

The requested output is a proposed error category, the affected stage and fields, a correction plan, and the questions or cross-level links that require re-auditing. GPT-5.5 does not itself assign the final post-correction pass decision.

The GPT-5.5 prompt template is:

```
You are analyzing a failed knee MRI
O--E--D construction item. Identify
where the error likely originated
and what must be checked again.

Study ID:
{study_id}

Anatomical target:
{target_and_site}

Report analysis output and supporting
report spans:
{report_analysis_trace}

MRI unit selection output and source
series references:
{mri_selection_trace}

Spatial localization output, including
available slice, box, and mask references:
{localization_trace}

Generated O--E--D questions, options,
proposed answers, and supporting MRI
references:
{generated_chain}

Independent audit decision and
reason codes, if available:
{audit_feedback}
```

```
Radiologist feedback, if available:
{radiologist_feedback}

Determine whether the failure may
originate from one or more of:

1. Report analysis: an incorrect or
   unsupported extracted target,
   finding, location, or imaging cue.
2. MRI unit selection: an unsuitable
   sequence--view unit or source series.
3. Spatial localization: an unsuitable
   slice, bounding box, or candidate mask.
4. VQA generation: an unsupported,
   ambiguous, or inconsistent question,
   option, reference answer, or
   cross-level link.

For EACH suspected source, return:
error_stage
affected_item_or_field
evidence_for_the_assessment
proposed_action

Also return:
related_questions_to_recheck
cross_level_links_to_recheck
recommended_disposition:
  revise or reject

Use the stage outputs and source
references to support the assessment.
Do not treat agreement with the report
as sufficient MRI evidence. If the
available MRI cannot support the
target or answer, recommend rejection.
Do not mark the item as passed.
```

The proposed correction is applied at the affected construction stage. The corrected question and any related questions or links whose answers or MRI references may have changed are submitted for independent re-audit. An item unsupported by the available MRI is rejected. A chain is retained only after all of its questions and cross-level links pass audit.

Radiologist review. The construction process comprises 22 rounds of independent audit, error-directed revision, and sampled radiologist review. Four radiologists review randomly sampled items that have passed the automated audit, including items that passed after revision. The number reviewed in each round is reported in Table G.

For each sampled question, reviewers first answer using the MRI images while blinded to the corresponding radiology report. They then examine the proposed answer and relevant references, record discrepancies, and identify possible causes. Questions that meet the review criteria are counted as expert passes; the expert pass rate in Table G is calculated over the questions reviewed in that round.

Review-identified errors are returned to the error-analysis and routing stage for correction and re-audit. Expert pass rates describe the sampled questions; they are not applied to all newly generated questions to calculate the number finally retained. Across the 22 rounds, 997,200 questions were newly generated, 2,236 were sampled for expert review, and 821,823 were finally retained after construction and quality control.

Table G: **Per-round construction and expert review.** Expert pass rates are calculated over the questions reviewed in each round. The final column reports questions retained after the construction and quality-control process.

Round	Rule version	New questions	Expert-reviewed	Expert pass rate	Final retained
1	V1	1,200	100	76.00%	777
2	V2	1,200	108	69.44%	831
3	V3	1,200	112	84.82%	981
4	V4	1,200	116	77.59%	1,002
5	V5	1,200	100	84.00%	969
6	V6	2,400	100	88.00%	2,118
7	V7	2,400	100	87.00%	2,211
8	V8	2,400	100	93.00%	2,280
9	V9	4,800	100	81.00%	4,278
10	V10	9,600	100	91.00%	8,025
11	V11	9,600	100	96.00%	8,928
12	V12	19,200	100	84.00%	16,035
13	V13	19,200	100	90.00%	17,868
14	V14	38,400	100	88.00%	29,703
15	V15	38,400	100	96.00%	35,751
16	V16	38,400	100	91.00%	33,408
17	V17	76,800	100	89.00%	66,417
18	V18	38,400	100	97.00%	36,096
19	V19	76,800	100	80.00%	63,867
20	V20	153,600	100	87.00%	125,154
21	V21	307,200	100	85.00%	251,901
22	V22	153,600	100	78.00%	113,223
Total	—	**997,200**	**2,236**	—	**821,823**

Table H: **Accuracy under full and restricted observation.** Scores (%) are unweighted means over six anatomical systems at the observation (Obs.), evidence (Evid.), and diagnosis (Diag.) stages. Full uses all nine sequence–view combinations. Ax, Cor, and Sag denote axial, coronal, and sagittal views, respectively.

	Model	Full			View-restricted Input									Sequence-restricted Input								
					Ax (PD+T1+T2)			Cor (PD+T1+T2)			Sag (PD+T1+T2)			PD (Ax+Cor+Sag)			T1 (Ax+Cor+Sag)			T2 (Ax+Cor+Sag)		
		Obs.	Evid.	Diag.	Obs.	Evid.	Diag.	Obs.	Evid.	Diag.	Obs.	Evid.	Diag.	Obs.	Evid.	Diag.	Obs.	Evid.	Diag.	Obs.	Evid.	Diag.
Comm.	GPT-5.5 (OpenAI, 2026)	87.86	81.99	63.47	87.38	81.95	61.83	88.41	82.75	64.81	88.88	82.92	65.07	87.81	82.04	65.31	86.93	81.29	57.70	87.29	81.44	63.59
	Gemini 3.1-Pro (Google DeepMind, 2026a)	88.40	82.90	68.24	80.42	60.69	43.88	79.62	59.97	44.98	87.58	71.93	42.39	81.85	63.72	42.32	81.58	63.44	41.66	81.25	61.43	41.25
	Claude Opus 4.7 (Anthropic, 2026a)	86.51	72.45	42.56	83.28	75.38	58.26	83.31	75.55	58.93	83.64	75.68	59.03	83.18	75.14	59.13	83.12	75.17	59.24	82.74	74.47	59.21
	Claude Sonnet 4.6 (Anthropic, 2026b)	82.96	74.36	55.09	80.34	66.08	50.29	80.42	60.69	44.05	78.70	64.66	50.26	79.07	64.51	49.73	79.68	65.04	49.47	79.36	64.57	47.76
M.S.	Lingshu-7B (Xu et al., 2025)	36.30	32.14	27.47	36.41	36.02	32.57	36.84	35.30	27.91	35.18	33.51	23.84	39.13	35.75	30.98	65.35	54.82	55.01	38.88	37.77	34.95
	HuatuoGPT-Vision-7B (Chen et al., 2024a)	44.72	36.24	27.12	42.09	39.84	29.32	41.74	39.52	28.58	44.45	37.13	27.26	42.45	36.38	27.05	43.83	38.42	27.84	45.07	39.49	28.40
FT	InternVL3.5-8B-FT	80.45	66.67	54.08	79.88	62.40	52.22	82.81	68.13	52.90	79.78	61.46	52.62	80.37	62.68	52.37	81.74	67.13	53.12	82.15	66.71	52.92
	Qwen3.5-9B-FT	86.32	70.60	44.00	86.77	83.12	62.93	87.31	69.79	58.31	87.31	80.68	57.27	87.15	68.78	44.18	86.03	61.96	44.94	86.38	70.72	49.32
	Gemma-4-4B-FT	81.16	66.63	47.31	55.42	56.99	40.22	54.88	56.92	41.33	51.34	55.36	41.97	52.57	55.53	40.59	53.91	56.46	41.10	54.72	57.44	40.50

E Complete Controlled-Observation Results

Table H reports accuracy (%) for the nine model configurations shown in Figure 4. It includes full observation and all six restricted conditions at the Observation, Evidence, and Diagnosis levels. Each score is the unweighted mean of the six anatomical-system accuracies.

View-restricted conditions retain all three sequences in the specified view, whereas sequence-restricted conditions retain all three views for the specified sequence. The condition-specific scores are used to compute the full-observation advantage and observation variation shown in Figure 4.

The individual scores reveal patterns that are less apparent from the summary metrics. For HuatuoGPT-Vision-7B, T2-only scores exceed full-observation scores at all three levels: 45.07% versus 44.72% at Observation, 39.49% versus 36.24% at Evidence, and 28.40% versus 27.12% at Diagnosis. For InternVL3.5-8B-FT, coronal-only input yields higher Observation (82.81% versus 80.45%) and Evidence (68.13% versus 66.67%) scores than full observation, but a lower Diagnosis score (52.90% versus 54.08%). These results show that the effect of restricting MRI input is not uniform across models or clinical levels. By evaluating the same questions across sequence–view conditions and O–E–D levels, MultiKnee makes such differences observable and identifies where a model's performance changes along the path from image observation to diagnosis.