SPINEBENCH: A CLINICALLY SALIENT, LEVEL-AWARE BENCHMARK POWERED BY THE SPINEMED-450K CORPUS

Anonymous authors
Paper under double-blind review

ABSTRACT

Spine disorders affect 619 million people globally and are a leading cause of disability, yet AI-assisted diagnosis remains limited by the lack of level-aware, multimodal datasets. Clinical decision-making for spine disorders requires sophisticated reasoning across X-ray, CT, and MRI at specific vertebral levels. However, progress has been constrained by the absence of traceable, clinically-grounded instruction data and standardized, spine-specific benchmarks. To address this, we introduce SpineMed, an ecosystem co-designed with practicing spine surgeons. It features SpineMed-450k, the first large-scale dataset explicitly designed for vertebral-level reasoning across imaging modalities with over 450,000 instruction instances, and SpineBench, a clinically-grounded evaluation framework. SpineMed-450k is curated from diverse sources, including textbooks, guidelines, open datasets, and ∼1,000 de-identified hospital cases, using a clinician-in-the-loop pipeline with a two-stage LLM generation method (draft and revision) to ensure high-quality, traceable data for question-answering, multi-turn consultations, and report generation. SpineBench evaluates models on clinically salient axes, including level identification, pathology assessment, and surgical planning. Our comprehensive evaluation of several recently advanced large vision-language models (LVLMs) on SpineBench reveals systematic weaknesses in fine-grained, level-specific reasoning. In contrast, our model fine-tuned on SpineMed-450k demonstrates consistent and significant improvements across all tasks. Clinician assessments confirm the diagnostic clarity and practical utility of our model's outputs.

1 INTRODUCTION

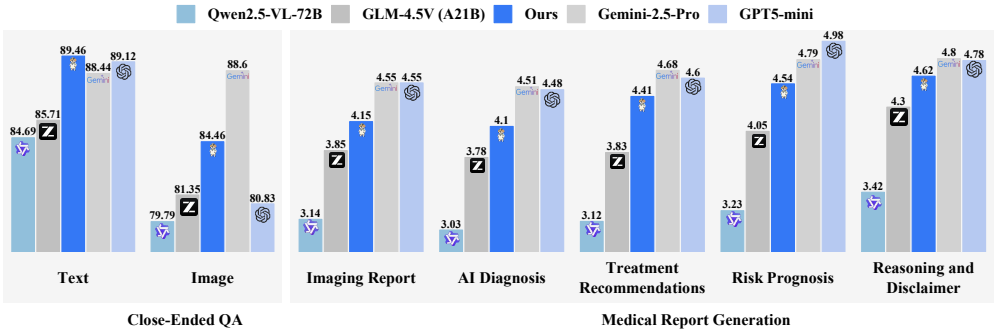


Figure 1: Benchmark performance of SpineGPT

Spinal disorders (Ferreira et al., 2023), including degenerative diseases (like disc herniation) (Dydyk et al., 2017), deformities (like scoliosis) (Negrini et al., 2018), trauma (fractures) (Vaccaro et al., 2013), and inflammatory conditions (Taurog et al., 2016), are a major driver of pain, disability, and surgical care worldwide. A key challenge in their management is diagnostic complexity. Unlike many other disorders, spinal conditions typically cannot be precisely diagnosed using a single imaging

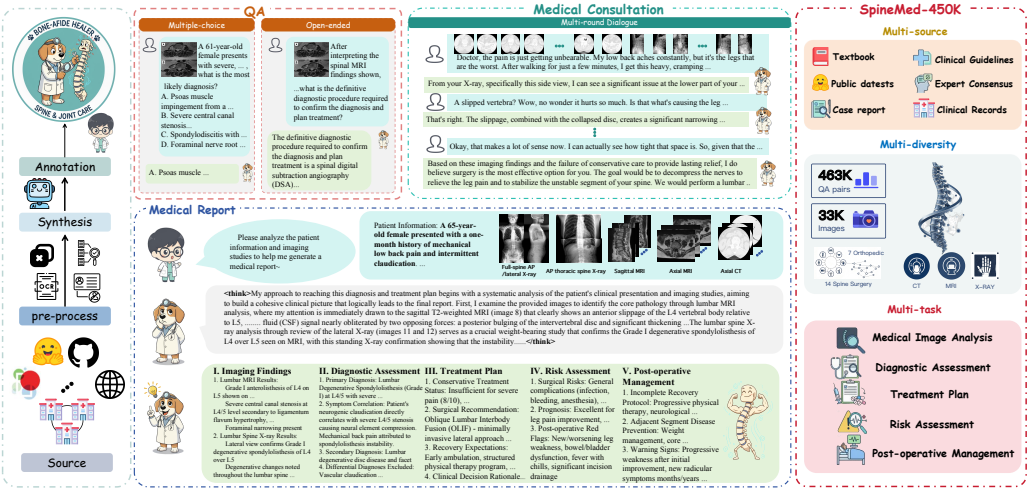


Figure 2: Overview of SpineMed-450k. Training data was curated from textbooks, public datasets, clinical records, medical guidelines, and hospitals. The process involved data preprocessing, annotation generation, and a final clinician review. Our dataset comprises four types: multi-choice QA, open-ended QA, multi-round dialogues, and reports.

modality. It often requires clinicians to perform level-aware, multimodal reasoning: integrating findings from X-ray, CT, and MRI to pinpoint pathology at specific vertebral levels, grade severity, and plan interventions (Teichner et al., 2025). The precision of this interpretation directly impacts patient outcomes and neurological safety. Although advanced AI holds great promise for augmenting this demanding workflow (Ibrahim et al., 2025b), its potential has been hindered. Fortunately, such clinical tasks can significantly benefit from advanced AI capabilities (Lee et al., 2024b). Yet progress is constrained not by model capacity, but by the absence of *traceable instruction data* and *standardized, clinically validated benchmarks* tailored to spine workflows (Lee et al., 2024b). Equally important, prior efforts rarely embed clinicians throughout the pipeline, limiting practical utility. We present **SpineMed**: a comprehensive effort consisting of **SpineMed-450k**, a provenance-rich instruction corpus for spine diagnosis and planning, and **SpineBench**, a targeted evaluation suite that help to evaluate the effectiveness of different AI-based spine diagnosis. To our best knowledge, this is current largest-scale Spinal diagnosis and treatment dataset. Both were *co-designed with spine clinicians* (radiologists and surgeons) to reflect real decision points. SpineMed-450k aggregates materials from textbooks, surgical guidelines, expert consensuses, question banks, open spine datasets (e.g., Spark, VerSe) (Alibaba Cloud Tianchi, 2020; Sekuboyina et al., 2021), open-access case reports (Europe PMC) (Consortium, 2015), and ∼1,000 de-identified hospital cases. Throughout curation, clinicians (i) defined inclusion criteria and task taxonomies; (ii) vetted imaging selections from hospital cases to prioritize views most informative for diagnosis and surgical planning; and (iii) specified failure modes that instruction data must surface. To minimize hallucinations and preserve traceability, our pipeline (a) extracts figures and text with `PaddleOCR` (Du et al., 2020); (b) *binds images to their local textual context* via caption-pattern regex matching that anchors each figure to its surrounding paragraph; and (c) distills high-quality supervision—multiple-choice, open-ended QA, multi-turn consultations, and report generation—through a *two-stage* LLM process (draft → revision with explicit prompts and logs). Clinicians review and refine prompt policies and revision criteria to align with reporting standards.

SpineBench operationalizes evaluation across clinically relevant axes—*imaging report*, *diagnosis*, *patient guidance*, *evidence-based treatment*, *technical feasibility*, *risk prognosis*, *coverage*, *relevance*, *granularity*, and *interpretability*. Its item design, error taxonomy, and rubrics were developed with clinician input to emphasize fine-grained, anatomy-centric reasoning and the kinds of mistakes that matter in practice.

To characterize the state of the field, we evaluate *a dozen* of contemporary large vision–language models (LVLMs) (OpenAI, 2025a;b; Hurst et al., 2024; Google, 2025a;b; Sellergren et al., 2025a; xAI, 2025; Anthropic, 2025; Bai et al., 2025; Hong et al., 2025; Wang et al., 2025a), both general-purpose and medical. Our evaluation reveals significant weaknesses in fine-grained, level-specific diagnosis

and open-ended clinical reasoning, particularly in the handling of complex multi-image tasks. Building on these insights, we introduce a fine-tuned spine model SpineGPT trained on SpineMed-450k that delivers consistent improvements on SpineBench as shown in Figure 1. Clinicians assess exemplar outputs for decision relevance, underscoring the practical value of targeted, evidence-linked instruction data. Our contributions are as follows:

- **Clinician-in-the-loop dataset and benchmark.** We release **SpineMed-450k**, more than 450,000 instruction instances spanning multiple-choice, open-ended QA, multi-turn consultations, and report generation—curated via a specialist-supported pipeline with anatomical integration and two-stage report refinement, together with **SpineBench**, a level-aware benchmark co-designed with clinicians and enriched with $\sim$1,000 real hospital cases.
- **Comprehensive evaluation.** We benchmark *dozens* of open-source LVLMs across closed/open tasks using clinician-shaped taxonomies and rubrics, surfacing systematic failure modes in spine reasoning.
- **A practical baseline model.** We propose a fine-tuned spine LVLM trained on SpineMed-450k that achieves consistent gains on SpineBench; exemplar outputs receive clinician feedback on diagnostic clarity and planning utility, establishing a high-utility baseline for future research.

2 SPINEMED-450K DATASET

Overview. The SpineMed-450k dataset was constructed through a meticulous "clinician-in-the-loop" pipeline designed to ensure clinical accuracy and relevance. This pipeline integrates four core stages: (1) Dataset collection, (2) Structured Information Extraction, (3) Data De-identification and Cleaning, and (4) Dataset Generation. (5) Annotation of the spinal diagnostic report.

2.1 DATA COLLECTION

To build a complete and comprehensive dataset for spinal diagnosis and treatment, we collected data from a variety of sources (Chen et al., 2024a; Wei & Hwei, 2024; Wu et al., 2025; Chen et al., 2024b). Existing general-purpose large vision-language models (Hurst et al., 2024; Google, 2025a;b; Deng et al., 2023; Ullah et al., 2024; AlSaad et al., 2024) and even medical large language models (Li et al., 2023; Wang et al., 2025b; Wu et al., 2024; Lin et al., 2025; Lu et al., 2024; Niu et al., 2025; Nath et al., 2025a; Seyfioglu et al., 2024; Lai et al., 2025; Xu et al., 2025) are trained on generic medical data (Chen et al., 2024a;b; Xie et al., 2024a), which often lacks the high-quality, specialized data needed for orthopedics (Deng et al., 2023; Ullah et al., 2024). To train an effective large model for spinal care, we first compiled a high-quality, general orthopedic dataset covering multiple domains, including Spine Surgery, Foot and Ankle Surgery, Orthopedic Trauma, and Hand and Upper Extremity Surgery.

As shown in Figure 3, we integrated materials from a variety of sources, including textbooks, surgical guidelines, expert consensuses, question banks, open-access case reports from Europe PMC (Consortium, 2015), open single-modality spine datasets (Alibaba Cloud Tianchi, 2020; Sekuboyina et al., 2021) (e.g., Spark, VerSe), and approximately 1,000 de-identified multimodal hospital cases collected from various hospitals. This data covers a wide range of modalities, including text, CT, MRI, X-ray, and tables. We track the provenance (dataset IDs/DOIs, case identifiers) for every derived item. Where possible, we adopt upstream datasets with permissive licenses and clear terms of reuse. Clinicians defined the inclusion criteria and, for hospital cases, selected the most decision-informative images (e.g., MRI target sequences, key CT levels) to serve as the foundation for downstream tasks.

2.2 DATASET CURATION

Structured Information Extraction To accurately extract comprehensive information from academic sources, we employed PaddleOCR (Du et al., 2020) to parse PDF documents and images from textbooks and literature. The output, containing both recognized text and layout analysis, was exported into Markdown format. This approach effectively preserved the structural integrity of the documents, including tables, figure placements, and overall layout. Furthermore, to ensure a precise mapping between figures, their captions, and corresponding contextual descriptions in the text, we developed a novel algorithm termed Picture Context Matching. Subsequently, we employed

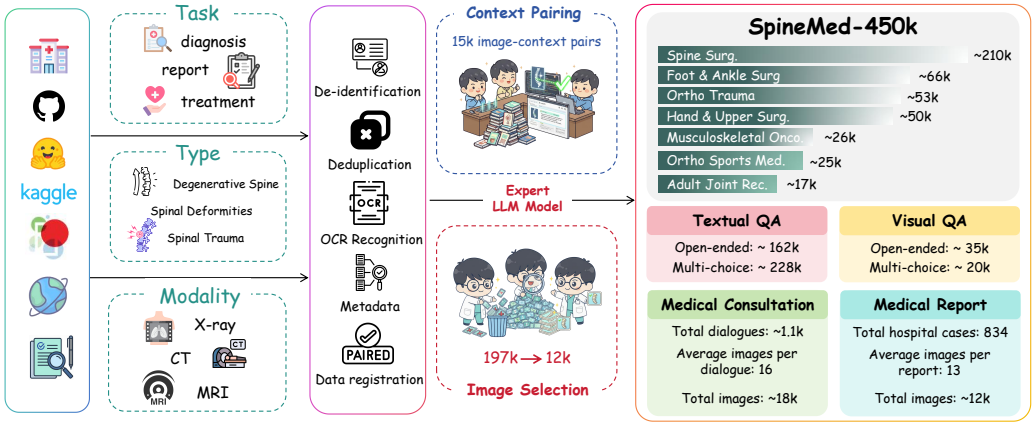


Figure 3: Generation pipeline of SpineMed-450k. The pipeline involves data preprocessing (including de-identification, deduplication, and OCR) followed by expert LLM-driven curation. This process generates 450k items for tasks like QA, medical reports, and consultations across various orthopedic subspecialties.

GPT-5-mini(OpenAI, 2025a) to conduct a semantic consistency check, rigorously filtering out instances where the visual content did not align with the retrieved context. The technical details of this algorithm are elaborated in the Appendix D.

Data De-identification and Cleaning This stage focused on processing data sourced from a collection of clinical records in hospitals. We first performed a rigorous de-identification process, removing all sensitive and personally identifiable information (PII), such as patient IDs and physical examination details under HIPAA. We also filtered out irrelevant images, such as post-operative photos and non-diagnostic tables. Subsequently, GPT-5-mini(OpenAI, 2025a) was utilized to conduct a fine-grained classification of the data, ensuring the dataset's purity by excluding non-orthopedic cases. As shown in Figure 2, the orthopedic domain was categorized into 7 classes, with the spine sub-domain further divided into 14 distinct classes. A detailed statistical overview of the dataset distribution across these categories is presented in Figure 4.

Dataset Generation In close collaboration with medical experts, we designed a comprehensive annotation schema to generate high-quality, multi-task training data. The annotation process was tailored to the data source: (1) From External Knowledge Sources (e.g., Textbooks): We generated bilingual (Chinese and English) and multimodal (text and image-based) questions in both multiple-choice and open-ended formats using Gemini-2.5-pro(Google, 2025a) with carefully designed prompts. (2) From Opened-spine Datasets: We processed two open-source spinal datasets, Spark and Verse, to generate multi-turn question-and-answer dialogues that simulate doctor-patient interactions. These datasets consist mainly of unimodal 3D image slices (CT and MRI). To ensure consistency, we standardized the inputs by adaptively sampling 25 slices per case under clinical expert supervision. From this, we created over 300 simulated consultations to train models in their conversational abilities within spinal scenarios. (3) From Real Clinical Records: We created multiple-choice questions, multi-turn conversational datasets for patient interviews, and comprehensive spinal diagnostic reports via locally deployed GLM-4.5V(Hong et al., 2025) to ensure data security. For prompt design, please refer to the Appendix G

Annotation of the spinal diagnostic report A cornerstone of our dataset is the generation of detailed spinal diagnostic reports. In this process, we utilized real clinical reports from hospitals, incorporating physician recommendations, to design reports that encompass six dimensions, all aimed at simulating a complete clinical workflow: (1) Structured Imaging Findings: Analyze the provided medical images and distill key radiological evidence that supports the final diagnosis. (2) AI-Assisted Diagnosis: Formulate a diagnostic conclusion and articulate the reasoning process based on the synthesis of clinical data and imaging analysis. (3) Treatment Recommendations: This section is bifurcated to address different audiences. Patient-Centric Advice: Explain the rationale for the recommended surgical procedure in clear, non-technical language. Physician-Centric Rationale:

Provide a robust, guideline-based decision tree to justify the surgical selection from a clinical perspective. (4) Risk and Prognosis Assessment: Conduct an objective evaluation of the potential risks and expected outcomes associated with the proposed surgical plan. (5) Postoperative Issue Management: Predict potential post-surgical complications for specific procedures and develop corresponding management strategies. (6) Diagnostic Rationale and Disclaimer: Provide complete diagnostic and surgical decision-making chain and disclaimer statement. Report examples are provided in the Appendix F.

2.3 Comparison with Existing Benchmarks

Table 1: Comparison of SpineMed-450k with existing spine imaging datasets. While prior datasets focus on specific perception tasks, our work introduces the first multimodal instruction-tuning corpus designed for full-spectrum clinical reasoning.

Dataset	Modality	Scale	Core Task	Workflow Coverage	Output Format
RSNA LumbarDISC (Richards et al., 2025)	MRI	2.6k Patients	Classification	Specific (Stenosis Grading)	Class Labels
BUU-LSPINE (Klinwichit et al., 2023)	X-Ray	3.6k Patients	Detection	Specific (Spondylolisthesis)	Coordinates/Labels
VerSe 2020 (Liebl et al., 2021)	CT	300 Patients	Segmentation	Specific (Anatomy)	Voxel Masks
Lumbar Spine MRI (van der Graaf et al., 2024)	MRI	218 Patients	Segmentation	Specific (Structure)	Voxel Masks
Spark (Tianchi) (Tianchi, 2020)	CT, MRI	150 Patients	Classification	Specific (Disease ID)	Class Labels
SpineMed-450k (Ours)	**Multimodal** (XR, CT, MRI, Text)	**450k+** Instructions	**Clinical Reasoning**	**Full-Spectrum** (Diag → Treat → Prognosis)	**Instruction Pairs** (Image, Text)

As illustrated in Table 1, existing datasets such as VerSe(Liebl et al., 2021) and RSNA(Richards et al., 2025) are predominantly unimodal and designed for low-level perception tasks like segmentation, detection, or simple classification. While these datasets serve as effective tools for specific anatomical localization or binary disease identification, they fundamentally fail to capture the holistic context required for complex clinical decision-making, limiting their utility in training models for high-level diagnostic synthesis. In contrast, SpineMed-450k represents a significant paradigm shift from "Tool AI" to "Collaborator AI". Our dataset distinguishes itself through three key dimensions: 1. Multimodal Synthesis, requiring the integration of X-ray, CT, and MRI to mirror real-world cross-modal validation; 2. Cognitive Depth, supporting level-aware reasoning rather than simple label outputs; 3. Workflow Completeness, covering the full patient journey with grounded instructions for Structured Imaging Findings, AI Diagnosis, Treatment Recommendations, and Risk Assessment. This effectively fills the cognitive gap left by perception-focused datasets.

3 Data Statistics

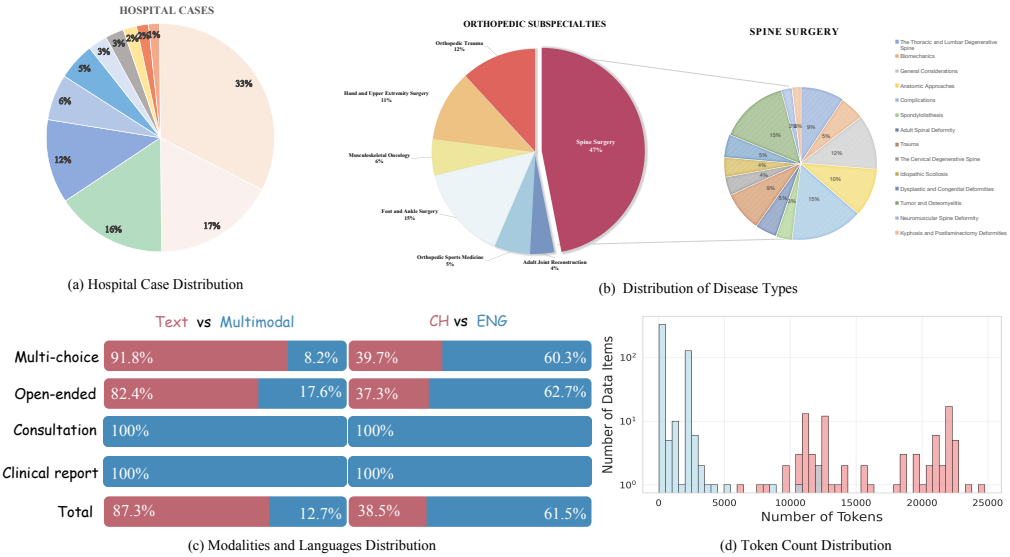
SpineMed-450K is a large-scale multimodal training dataset for orthopedic spine knowledge in large language models, characterized by strong traceability, comprehensive coverage, diverse question types, and rich modalities.

3.1 Disease Diversity Coverage

As shown in Figure 4(b), SpineMed-450K encompasses seven common orthopedic subspecialties, including Spine Surgery, Foot and Ankle Surgery, and Orthopedic Trauma, with spinal diagnostic data accounting for 47% of the orthopedic data. Furthermore, the spinal diagnostic data includes 14 spine subconditions such as cervical degenerative spine disease and idiopathic scoliosis. We performed sampling on each spinal diagnostic dataset to ensure uniform distribution across all disease categories.

3.2 Patient Source Diversity

As illustrated in Figure 4(a), our data originates from 1,000 real clinical cases collected from 11 leading expert hospitals. These data span the recent three years and encompass patients of different genders, various age groups, and diverse physical conditions. To protect privacy, personal information has been de-identified. personal information. Given the varying surgical volumes across different hospitals, the largest hospital contributes 33% of the data while the smallest contributes 1%. These valuable real patient data provide crucial evidence for accurately representing the authentic conditions of spine patients.



(a) Hospital Case Distribution

(b) Distribution of Disease Types

(c) Modalities and Languages Distribution

(d) Token Count Distribution

Figure 4: Statistics of SpineMed-450k. (a) Distribution of medical records across various hospitals. (b) The prevalence of various orthopedic and spinal diseases. (c) Distribution of different modals and languages. (d) Benchmark token length distribution: blue (non-report tokens), pink (report tokens).

3.3 Data Source and Question Type Diversity

Table 2: Dataset statistics categorized by data source and split.

Split	Literature	Textbook	Case Report	Question Bank	Open Source	Hospital	Total
Train	6,450	377,212	61,453	1,087	304	9,668	456,174
Test	17	203	101	3	–	250	574
Total	6,467	377,415	61,554	1,090	304	9,918	456,748

As shown in Table 2, our data derives from six major sources: Literature, Textbooks, Case Reports, hospitals, and others. Textbooks, being the primary knowledge source for physicians, constitute the largest proportion with 377k entries, while hospital data, though valuable, is limited in quantity, with 9,668 data points generated from nearly 1,000 real cases. As presented in Table 3 and Figure 4(c), question types are categorized into pure text QA, multimodal QA, medical consultations, and clinical reports, with multiple-choice questions comprising the largest proportion. For evaluation convenience, our test set includes only multiple-choice and clinical report formats.

Table 3: Dataset distribution across domains and task types.

Split	Multiple-choice	Open-ended	Consultation	Report
Train	248,789	197,413	1,138	734
Test	487	–	–	87
Total	249,276	197,413	1,138	821

3.4 Data Type Diversity

Our dataset incorporates multiple authentic data types including patient physical examination information, patient consultation records, X-rays, CT scans, and MRI images. Due to variations in hospital facilities and patient conditions, the collected data differs for each case, which introduces modeling challenges but enables our trained models to more closely approximate real clinical scenarios faced by physicians.

4 SPINEBENCH

4.1 BENCHMARK CONSTRUCTION

Data Sampling The SpineBench was constructed by sampling from the SpineMed-450k dataset. Following the original distribution of SpineMed-450k, we sampled 500 multiple-choice questions and 100 medical reports. This subset incorporates 14 spinal sub-diseases and data from multiple sources (see Appendix E for details).

Data Validation To ensure the integrity of SpineBench, a rigorous review process was implemented involving a team of 17 board-certified orthopedic surgeons. To mitigate bias and ensure objectivity, the surgeons were divided into three independent groups. Each group collaboratively validated the quality of the questions. Erroneous question-answer pairs were corrected, and questions deemed unsuitable for the evaluation set were removed. Ultimately, SpineBench comprises 487 high-quality multiple-choice questions and 87 report generation prompts.

4.2 EVALUATION METRICS

Table 4: Evaluation criteria for AI-generated clinical reports across five key dimensions

Report Section	Evaluation Criterion	Key Assessment Focus
I. Structured Imaging Report (SIP)	Imaging Report (1-5 pts)	Accuracy of findings, clinical significance, quantitative descriptions
II. AI-Assisted Diagnosis (AAD)	Diagnosis (1-5 pts)	Primary diagnosis correctness, differential diagnoses, clinical reasoning
III. Treatment Recommendations (TR)	Patient Guidance (1-5 pts)	Language clarity, empathy, patient reassurance
	Evidence-Based Plan (1-5 pts)	Rationale, individualization, guideline consistency
	Technical Feasibility (1-5 pts)	Surgical details, complication prevention, backup plans
IV. Risk & Prognosis Management (RPM)	Risk-Prognosis Mgmt (1-5 pts)	Perioperative planning, follow-up schedule, safety protocols
V. Reasoning & Disclaimer (RD)	Coverage (1-5 pts)	Completeness of evidence identification and explanation
	Relevance (1-5 pts)	Focus on core diagnosis without irrelevant content
	Granularity (1-5 pts)	Precision and quantitative detail sufficiency
	Explanation (1-5 pts)	Logical coherence and reasoning chain clarity

Under the careful design and guidance of our medical team, We propose a comprehensive evaluation framework that integrates three complementary assessment dimensions to measure the overall performance of AI systems in spinal diagnostic tasks:

$$\text{Score}_{\text{total}} = \sum_{k=1}^{3} w_k \cdot P_k \tag{1}$$

where P_1, P_2, and P_3 represent the performance scores for text-only multiple-choice questions, multimodal multiple-choice questions, and diagnostic report generation, respectively. The weights w_k are dynamically determined based on the sample sizes:

$$w_k = \frac{N_k}{\sum_{i=1}^{3} N_i} \tag{2}$$

where N_k denotes the number of samples in each evaluation category. This data-driven weighting scheme ensures statistical reliability while maintaining balanced representation across all assessment dimensions.

The diagnostic report score P_3 is computed using our expert-calibrated framework:

$$P_3 = 20 \times \sum_{i=1}^{5} \left(\frac{1}{n_i} \sum_{j=1}^{n_i} s_{ij} \right) \tag{3}$$

where scores are normalized to a 0–100 scale for consistency across all metrics and s_{ij} denotes the score for dimension j in section i, n_i represents the number of dimensions in section i. This unified scoring system enables direct comparison of model capabilities across diverse clinical tasks, from basic diagnostic reasoning to complex report generation.

5 SPINEGPT: A SPECIALIZED CLINICAL COLLABORATOR

In this section, we introduce **SpineGPT** to rigorously validate the efficacy of SpineMed-450k.

5.1 IMPLEMENTATION DETAILS

We employed the Qwen2.5-VL-7B-InstructBai et al. (2025) as our foundational architecture. All training phases were executed on a high-performance computational node equipped with 8 NVIDIA A100 GPUs, leveraging the ms-swift framework for efficient fine-tuning. To balance training efficiency with memory constraints across different stages, we dynamically adjusted the DeepSpeed optimization strategies. As detailed in Table 5, for the initial stages involving shorter sequences, we utilized DeepSpeed Zero2; for the final stage requiring extensive context modeling (up to 49k tokens), we transitioned to DeepSpeed Zero3 offloading. The global batch size was optimized per stage to maximize GPU utilization while maintaining training stability.

Table 5: Training configurations across different stages.

Configuration	Stage-1	Stage-2	Stage-3
Datasets	PubMedVision-150k orthopedics-230k MedThoughts-8K Medical-R1-Distill medical-o1-reasoning	Spine-open Spine-choice	Spine-chat (+reasoning) Spine-report (+reasoning)
Learning Rate	1e-5	1e-5	**1e-6**
Max Length	16,384	16,384	**49,152**
DeepSpeed	Zero2	Zero2	**Zero3**
Epochs	1	1	1

5.2 CURRICULUM LEARNING

This study employs a curriculum learning framework for subsequent training phases, aimed at enhancing the model's applicability and proficiency in the field of orthopedic spine care. The training process is divided into three stages, each integrating distinct datasets and training strategies to progressively strengthen the model's performance in spinal health.

General and Orthopedic Foundational Learning In this initial stage, we utilized several publicly available medical text datasets, including medical-o1-reasoning-SFT (Chen et al., 2024a), Medical-R1-Distill-Data (Chen et al., 2024a), and MedThoughts-8K (hw hwei, 2025). Additionally, we incorporated a diverse set of 150,000 multimodal instruction fine-tuning samples uniformly sampled from PubMedVision (Chen et al., 2024b). The primary objective during this phase is to develop the model's foundational capabilities in the medical field and to enhance its performance across various contexts. Subsequently, we trained on data from the SpineMed-450k dataset that pertained to non-spinal categories. Our findings indicate that this non-spinal data significantly improved the model's performance on the SpineBench benchmark, highlighting the importance of broadening the knowledge base to enhance task-specific performance.

Specialized Learning in Spinal Health In this phase, we concentrated on all data pertinent to spinal health. Furthermore, we extracted a selection of multiple-choice and open-ended questions to construct long reasoning chains, with the objective of enhancing the model's proficiency in the domain of spinal surgery.

Enhancement of Report Generation and Conversational Abilities Finally, we conducted further training through multi-turn dialogues, report generation, and datasets comprising long-chain reasoning instructions. The goal of this stage is to develop the model's advanced language comprehension and generation abilities, particularly in the contexts of dialogue interaction and report creation.

6 EXPERIMENTS

6.1 COMPARISON AND ANALYSIS ON SPINEBENCH

The evaluation results in Table 6 reveal severe limitations of current vision-language models (OpenAI, 2025a; Hurst et al., 2024; Google, 2025a) in medical domain applications. Large-scale open-source models perform particularly poorly: despite having 72B parameters, Qwen2.5-VL-72B (Bai et al.,

Table 6: Performance comparison of LVLMs on close-ended QA and medical report generation tasks.

Model	Size	Close-Ended QA			Medical Report Generation						Avg.
		Text	Image	Avg.	SIP	AAD	TR	RPM	RD	Sum	
Proprietary LVLMs											
GPT5	-	87.41	79.97	84.46	4.54	4.51	4.53	4.69	4.64	91.60	85.54
O3	-	86.73	82.38	85.01	4.39	4.25	4.34	4.43	4.42	87.32	85.36
Gemini-2.5-Pro	-	88.44	**88.60**	**88.50**	**4.55**	**4.51**	**4.68**	4.79	**4.80**	93.32	**89.23**
Claude4	-	79.59	79.79	79.67	3.96	4.08	4.04	4.44	4.41	83.72	80.28
GPT-4o	-	86.73	81.70	84.74	3.16	3.03	3.06	3.30	3.46	64.04	81.60
GPT5-mini	-	89.12	80.83	85.83	4.55	4.48	4.60	**4.98**	4.78	**93.56**	87.01
Gemini-2.5-Flash	-	83.67	80.83	82.55	4.43	4.29	4.57	4.88	4.75	91.68	83.93
Open-source LVLMs											
GLM-4.5V	21B	85.71	81.35	83.98	3.85	3.78	3.83	4.05	4.30	79.24	83.26
Qwen2.5-VL-72B	72B	84.69	79.79	82.75	3.14	3.03	3.12	3.23	3.42	63.80	79.88
Linshu-32B	32B	81.29	76.68	79.47	3.05	3.05	3.23	3.49	3.47	65.16	77.30
Medgemma-27B	27B	83.33	80.83	82.34	2.88	3.49	3.38	4.14	3.65	70.16	76.66
HuatuoGPT-7B	7B	75.85	80.83	77.82	2.42	2.42	2.42	3.37	2.87	54.0	74.21
Qwen2.5VL-7B	7B	75.51	74.09	74.95	2.27	2.39	2.80	3.26	2.92	54.52	64.74
Ours	7B	**89.46**	84.46	87.89	4.15	4.10	4.41	4.54	4.62	87.24	87.44

2025) achieves only 79.88% average performance and a mere 63.80 cumulative score on medical report generation, far below practical application requirements. Crucially, domain-specific pre-training alone proves insufficient: the medical-specialized Medgemma-27B(Sellergren et al., 2025b) achieves an even lower average of 76.66%, trailing our model by over 10 points despite being nearly 4 times larger. Even the best-performing open-source model GLM-4.5V (Hong et al., 2025) (83.26%) exhibits a nearly 6-point gap compared to the leading proprietary model Gemini-2.5-Pro (89.23%). This gap is more pronounced in medical report generation, where proprietary models exceed 85 points while open-source models struggle to reach 80. Additional medical report results are in the Appendix C.

Pervasive deficiency in cross-modal alignment. Nearly all models exhibit varying degrees of performance degradation on multimodal tasks. Among open-source models, GLM-4.5V shows a 4.36-point gap between text (85.71%) and image (81.35%) modalities; Qwen2.5-VL-72B exhibits a 4.90-point gap. Even proprietary models suffer from this issue, with GPT5 dropping from 87.41% on text to 79.97% on images, a gap of 7.44 percentage points. This cross-modal performance disparity reflects fundamental inadequacies in medical image understanding and vision-language alignment in existing models, limiting their application in clinical scenarios requiring comprehensive analysis of medical images and textual information.

Our method achieves breakthrough performance among open-source models. We achieve 87.44% average score, outperforming all open-source models by 4.18+ points and exceeding multiple proprietary models on close-ended QA (87.89% vs Claude4's 79.67%, GPT-4o's 84.74%). Our text-only QA (89.46%) surpasses all models including GPT5 (87.41%).

Efficiency and Clinical Utility. While we acknowledge that SpineGPT (87.44%) slightly trails the absolute SOTA model, Gemini-2.5-Pro (89.23%), this comparison highlights a remarkable efficiency: our 7B-parameter model uses less than 7% of the parameters compared to Gemini-2.5-Pro (which exceeds 100B parameters), yet achieves ∼98% of its performance. Notably, SpineGPT also outperforms other top-tier models such as GPT-4o (81.60%). This efficiency translates directly to clinical utility, as our lightweight 7B model is safe and efficient enough for local deployment within hospital firewalls, ensuring data privacy without reliance on external cloud APIs.

6.2 Human-Expert Agreement Analysis

To validate our LLM-based evaluation approach, we conducted a human-expert validation study by sampling cases from our dataset for blind expert scoring. Figure 5 shows the correlation analysis

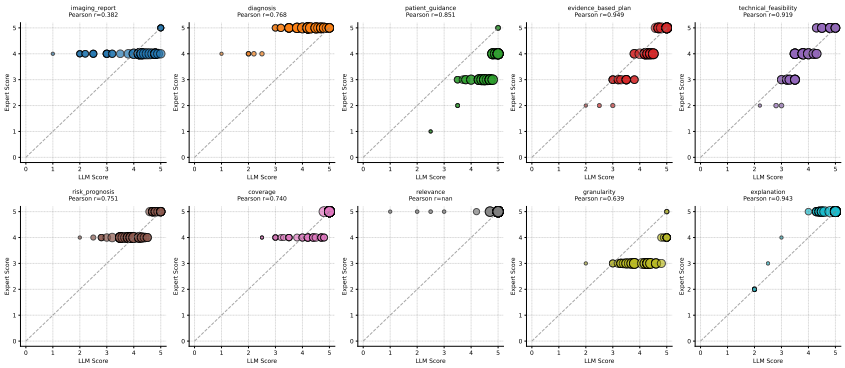


Figure 5: Consistency evaluation of large models and scores given by medical experts

between LLM and expert scores across ten evaluation dimensions. The results demonstrate strong alignment with Pearson correlation coefficients ranging from 0.382 to 0.949, with most dimensions showing correlations above 0.7. These findings validate that our automated LLM scoring serves as a reliable proxy for expert judgment.

6.3 Ablations of SpineGPT

Limitations of General Medical Data. As shown in Table 7, models trained exclusively on large-scale general medical data (row 2) exhibit significant performance degradation (74.95 *vs.* 65.31) on SpineBench compared to the baseline model (row 1). This demonstrates that models trained on such data are insufficient for specialized spine diagnostics. The incorporation of the non-spine orthopedic subset derived from our SpineMed-450k corpus (row 3) yields substantial performance improvements (82.14 *vs.* 74.95), validating the importance of domain-aligned training data. Notably, training exclusively on the spine subset (row 4) achieves an impressive average score of 87.07, reaching nearly 99% of the full model's performance. This confirms that our high-density spinal instruction data is the decisive factor. Furthermore, the full multi-stage curriculum (Row 6) incorporates this data to reach the peak performance of 87.89, a substantial enhancement over the configuration relying solely on general medical and orthopedic priors (row 5, 81.11).

Table 7: Performance comparison of models on close-ended QA tasks.

Model	Training Data			Close-Ended QA (%)		
	General	No-Spine	Spine	Text	Image	Avg.
Qwen 2.5 VL-7B				75.51	74.09	74.95
SpineGPT	✓			64.27	62.69	65.31
SpineGPT		✓		82.99	80.83	82.14
SpineGPT			✓	87.76	86.01	87.07
SpineGPT	✓	✓		83.67	77.20	81.11
SpineGPT	✓	✓	✓	89.46	84.46	87.89

7 Conclusions, Limitations, and Future Work

We introduced **SpineMed-450k**, a provenance-rich instruction corpus for level-aware spine diagnosis and planning, and **SpineBench**, level-aware benchmark co-designed with clinicians. Experiments on SpineBench reveal consistent weaknesses of contemporary open-source LVLMs. Our fine-tuned model achieves 87.44% performance, substantially outperforming open-source alternatives and demonstrating that specialized instruction data enables clinically relevant AI capabilities for complex anatomical reasoning tasks.

Limitations and Future Work. Future work will expand datasets, train larger models beyond 7B parameters, incorporate reinforcement learning techniques, and provide comprehensive direct comparisons with leading proprietary models including GPT-4 and Gemini to establish clear performance benchmarks.

REFERENCES

Alibaba Cloud Tianchi. SPARk: Spinal disease intelligent diagnosis dataset from Spark "Digital Human" AI Challenge. URL: https://tianchi.aliyun.com/competition/entrance/531796/information, 2020. Dataset provided by Wanli Cloud and AllinMD Orthopaedics for the Spark "Digital Human" AI Challenge – Intelligent Diagnosis of Spinal Diseases Competition.

Rawan AlSaad, Alaa Abd-Alrazaq, Sabri Boughorbel, Arfan Ahmed, Max-Antoine Renault, Rafat Damseh, and Javaid Sheikh. Multimodal large language models in health care: applications, challenges, and future outlook. *Journal of medical Internet research*, 26:e59505, 2024.

Anthropic. Claude Opus 4 and Claude Sonnet 4 System Card. https://www-cdn.anthropic.com/4263b940cabb546aa0e3283f35b686f4f3b2ff47.pdf, May 2025. Accessed: 2025-09-21.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Sami Barrit, Nathan Torcida, Aurélien Mazeraud, Sébastien Boulogne, Jeanne Benoit, Timothée Carette, Thibault Carron, Bertil Delsaut, Eva Diab, Hugo Kermorvant, et al. Neura: a specialized large language model solution in neurology. *medRxiv*, pp. 2024–02, 2024.

Runa Bhaumik, Vineet Srivastava, Arash Jalali, Shanta Ghosh, and Ranganathan Chandrasekharan. Mindwatch: A smart cloud-based ai solution for suicide ideation detection leveraging large language models. *MedRxiv*, pp. 2023–09, 2023.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. Huatuogpt-o1, towards medical complex reasoning with llms. *arXiv preprint arXiv:2412.18925*, 2024a.

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*, 2024b.

Europe PMC Consortium. Europe pmc: a full-text literature database for the life sciences and platform for innovation. *Nucleic acids research*, 43(D1):D1042–D1048, 2015.

Jiawen Deng, Areeba Zubair, and Ye-Jean Park. Limitations of large language models in medical applications. *Postgraduate Medical Journal*, 99(1178):1298–1299, 2023.

Zhuo Deng, Weihao Gao, Chucheng Chen, Zhiyuan Niu, Zheng Gong, Ruiheng Zhang, Zhenjie Cao, Fang Li, Zhaoyi Ma, Wenbin Wei, et al. Ophglm: An ophthalmology large language-and-vision assistant. *Artificial Intelligence in Medicine*, 157:103001, 2024.

Yuning Du, Chenxia Li, Ruoyu Guo, Xiaoting Yin, Weiwei Liu, Jun Zhou, Yifan Bai, Zilin Yu, Yehua Yang, Qingqing Dang, et al. Pp-ocr: A practical ultra lightweight ocr system. *arXiv preprint arXiv:2009.09941*, 2020.

Alexander M Dydyk, FB Mesfin, et al. Disc herniation. 2017.

Manuela L Ferreira, Katie De Luca, Lydia M Haile, Jaimie D Steinmetz, Garland T Culbreth, Marita Cross, Jacek A Kopec, Paulo H Ferreira, Fiona M Blyth, Rachelle Buchbinder, et al. Global, regional, and national burden of low back pain, 1990–2020, its attributable risk factors, and projections to 2050: a systematic analysis of the global burden of disease study 2021. *The Lancet Rheumatology*, 5(6):e316–e329, 2023.

Google. Gemini 2.5 Pro -Model Card. https://storage.googleapis.com/model-cards/documents/gemini-2.5-pro.pdf, June 2025a. Accessed: 2025-09-21.

Google. Gemini 2.5 Flash & 2.5 Flash Image - Model Card. https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-2-5-Flash-Model-Card.pdf, August 2025b. Accessed: 2025-09-21.

Yangyang Guo, Airu Huang, Bo Peng, Yufeng Li, and Wei Gu. Mbbo-rpsld: Training a multimodal blenderbot for rehabilitation in post-stroke language disorder. *IEEE Journal of Biomedical and Health Informatics*, 2025.

Jing Hao, Yuxuan Fan, Yanpeng Sun, Kaixin Guo, Lizhuo Lin, Jinrong Yang, Qi Yong H Ai, Lun M Wong, Hao Tang, and Kuo Feng Hung. Towards better dental ai: A multimodal benchmark and instruction dataset for panoramic x-ray analysis. *arXiv preprint arXiv:2509.09254*, 2025.

Wenyi Hong, GLM-V Team, and et al. GLM-4.5V and GLM-4.1V-Thinking: Towards Versatile Multimodal Reasoning with Scalable Reinforcement Learning. *arXiv preprint arXiv:2507.01006*, 2025.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

hw hwei. Medthoughts-8k dataset, 2025. URL `https://huggingface.co/datasets/hw-hwei/MedThoughts-8K`.

Muhammad Talal Ibrahim, Eric Milliron, and Elizabeth Yu. Artificial intelligence in spinal imaging-a narrative review. *Artificial Intelligence Surgery*, 5(1):139–149, 2025a.

Muhammad Talal Ibrahim, Eric Milliron, and Elizabeth Yu. Artificial intelligence in spinal imaging-a narrative review. *Artificial Intelligence Surgery*, 5(1):139–149, 2025b.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 590–597, 2019.

Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019.

Podchara Klinwichit, Watcharaphong Yookwan, Sornsupha Limchareon, Krisana Chinnasarn, Jun-Su Jang, and Athita Onuean. Buu-lspine: A thai open lumbar spine dataset for spondylolisthesis detection. *Applied Sciences*, 13(15):8646, 2023.

Yuxiang Lai, Jike Zhong, Ming Li, Shitian Zhao, and Xiaofeng Yang. Med-r1: Reinforcement learning for generalizable medical reasoning in vision-language models. *arXiv preprint arXiv:2503.13939*, 2025.

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(1):1–10, 2018.

Sungwon Lee, Joon-Yong Jung, Akaworn Mahatthanatrakul, and Jin-Sung Kim. Artificial intelligence in spinal imaging and patient care: a review of recent advances. *Neurospine*, 21(2):474, 2024a.

Sungwon Lee, Joon-Yong Jung, Akaworn Mahatthanatrakul, and Jin-Sung Kim. Artificial intelligence in spinal imaging and patient care: a review of recent advances. *Neurospine*, 21(2):474, 2024b.

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.

Hans Liebl, David Schinz, Anjany Sekuboyina, Luca Malagutti, Maximilian T Löffler, Amirhossein Bayat, Malek El Husseini, Giles Tetteh, Katharina Grau, Eva Niederreiter, et al. A computed tomography vertebral segmentation dataset with anatomical variations and multi-vendor scanner data. *Scientific Data*, 8(1):284, 2021.

Tianwei Lin, Wenqiao Zhang, Sijing Li, Yuqian Yuan, Binhe Yu, Haoyuan Li, Wanggui He, Hao Jiang, Mengze Li, Xiaohui Song, et al. Healthgpt: A medical large vision-language model for unifying comprehension and generation via heterogeneous knowledge adaptation. *arXiv preprint arXiv:2502.09838*, 2025.

Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Melissa Zhao, Aaron K Chow, Kenji Ikemura, Ahrong Kim, Dimitra Pouli, Ankush Patel, et al. A multimodal generative ai copilot for human pathology. *Nature*, 634(8033):466–473, 2024.

Tingyu Mo, Jacqueline CK Lam, Victor OK Li, and Lawrence YL Cheung. Dect: Harnessing llm-assisted fine-grained linguistic knowledge and label-switched and label-preserved data generation for diagnosis of alzheimer's disease. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 24885–24892, 2025.

Hongbin Na. Cbt-llm: A chinese large language model for cognitive behavioral therapy-based mental health question answering. *arXiv preprint arXiv:2403.16008*, 2024.

Vishwesh Nath, Wenqi Li, Dong Yang, Andriy Myronenko, Mingxin Zheng, Yao Lu, Zhijian Liu, Hongxu Yin, Yee Man Law, Yucheng Tang, et al. Vila-m3: Enhancing vision-language models with medical expert knowledge. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 14788–14798, 2025a.

Vishwesh Nath, Wenqi Li, Dong Yang, Andriy Myronenko, Mingxin Zheng, Yao Lu, Zhijian Liu, Hongxu Yin, Yee Man Law, Yucheng Tang, et al. Vila-m3: Enhancing vision-language models with medical expert knowledge. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 14788–14798, 2025b.

Stefano Negrini, Sabrina Donzelli, Angelo Gabriele Aulisa, Dariusz Czaprowski, Sanja Schreiber, Jean Claude de Mauroy, Helmut Diers, Theodoros B Grivas, Patrick Knott, Tomasz Kotwicki, et al. 2016 sosort guidelines: orthopaedic and rehabilitation treatment of idiopathic scoliosis during growth. *Scoliosis and spinal disorders*, 13(1):3, 2018.

Chuang Niu, Qing Lyu, Christopher D Carothers, Parisa Kaviani, Josh Tan, Pingkun Yan, Mannudeep K Kalra, Christopher T Whitlow, and Ge Wang. Medical multimodal multitask foundation model for lung cancer screening. *Nature Communications*, 16(1):1523, 2025.

OpenAI. GPT-4V system card. Technical report, OpenAI, 2023. URL `https://cdn.openai.com/papers/GPTV_System_Card.pdf`.

OpenAI. GPT-5 System Card. `https://cdn.openai.com/gpt-5-system-card.pdf`, August 2025a. Accessed: 2025-09-21.

OpenAI. OpenAI o3 and o4-mini System Card. `https://cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf`, April 2025b. Accessed: 2025-09-21.

Jianing Qiu, Jian Wu, Hao Wei, Peilun Shi, Minqing Zhang, Yunyun Sun, Lin Li, Hanruo Liu, Hongyi Liu, Simeng Hou, et al. Visionfm: a multi-modal multi-task vision foundation model for generalist ophthalmic artificial intelligence. *arXiv preprint arXiv:2310.04992*, 2023.

Tyler J Richards, Adam E Flanders, Errol Colak, Luciano M Prevedello, Robyn L Ball, Felipe Kitamura, John Mongan, Maryam Vazirabad, Hui-Ming Lin, Anne Kendell, et al. The rsna lumbar degenerative imaging spine classification (lumbardisc) dataset. *arXiv preprint arXiv:2506.09162*, 2025.

Ali Sarabadani, Kheirolah Rahsepar Fard, and Hamid Dalvand. Exkg-llm: Leveraging large language models for automated expansion of cognitive neuroscience knowledge graphs. *arXiv preprint arXiv:2503.06479*, 2025.

Anjany Sekuboyina, Malek E Husseini, Amirhossein Bayat, Maximilian Löffler, Hans Liebl, Hongwei Li, Giles Tetteh, Jan Kukačka, Christian Payer, Darko Štern, et al. Verse: a vertebrae labelling and segmentation benchmark for multi-detector ct images. *Medical image analysis*, 73:102166, 2021.

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*, 2025a.

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*, 2025b.

Mehmet Saygin Seyfioglu, Wisdom O Ikezogwo, Fatemeh Ghezloo, Ranjay Krishna, and Linda Shapiro. Quilt-llava: Visual instruction tuning by extracting localized narratives from open-source histopathology videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13183–13192, 2024.

Joel D Taurog, Avneesh Chhabra, and Robert A Colbert. Ankylosing spondylitis and axial spondyloarthritis. *New England Journal of Medicine*, 374(26):2563–2574, 2016.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Eric M Teichner, Robert C Subtirelu, Connor R Crutchfield, Chitra Parikh, Arjun Ashok, Sahithi Talasila, Victoria Anderson, Milan Patel, Sricharvi Mannam, Andrew Lee, et al. The advancement and utility of multimodal imaging in the diagnosis of degenerative disc disease. *Frontiers in Radiology*, 5:1298054, 2025.

Alibaba Cloud Tianchi. Spark: Spinal disease intelligent diagnosis dataset. `https://tianchi.aliyun.com/competition/entrance/531796/introduction`, 2020. Accessed: 2025-11-20.

Ehsan Ullah, Anil Parwani, Mirza Mansoor Baig, and Rajendra Singh. Challenges and barriers of using large language models (llm) such as chatgpt for diagnostic medicine with a focus on digital pathology–a recent scoping review. *Diagnostic pathology*, 19(1):43, 2024.

Alexander R Vaccaro, Cumhur Oner, Christopher K Kepler, Marcel Dvorak, Klaus Schnake, Carlo Bellabarba, Max Reinhold, Bizhan Aarabi, Frank Kandziora, Jens Chapman, et al. Aospine thoracolumbar spine injury classification system: fracture description, neurological status, and key modifiers. *Spine*, 38(23):2028–2037, 2013.

Jasper W van der Graaf, Miranda L van Hooff, Constantinus FM Buckens, Matthieu Rutten, Job LC van Susante, Robert Jan Kroeze, Marinus de Kleuver, Bram van Ginneken, and Nikolas Lessmann. Lumbar spine segmentation in mr images: a dataset and a public benchmark. *Scientific Data*, 11 (1):264, 2024.

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*, 2025a.

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*, 2025b.

Huan Wei and Wei Hwei. MedThoughts-8K: A large-scale medical reasoning dataset. `https://huggingface.co/datasets/hw-hwei/MedThoughts-8K`, 2024.

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Weidi Xie, and Yanfeng Wang. Pmc-llama: toward building open-source language models for medicine. *Journal of the American Medical Informatics Association*, 31(9):1833–1843, 2024.

Juncheng Wu, Wenlong Deng, Xingxuan Li, Sheng Liu, Taomian Mi, Yifan Peng, Ziyang Xu, Yi Liu, Hyunjin Cho, Chang-In Choi, et al. Medreason: Eliciting factual medical reasoning steps in llms via knowledge graphs. *arXiv preprint arXiv:2504.00993*, 2025.

xAI. Grok 4 Model Card. `https://data.x.ai/2025-08-20-grok-4-model-card.pdf`, August 2025. Accessed: 2025-09-21.

Yunfei Xie, Ce Zhou, Lang Gao, Juncheng Wu, Xianhang Li, Hong-Yu Zhou, Sheng Liu, Lei Xing, James Zou, Cihang Xie, et al. Medtrinity-25m: A large-scale multimodal dataset with multigranular annotations for medicine. *arXiv preprint arXiv:2408.02900*, 2024a.

Yunfei Xie, Ce Zhou, Lang Gao, Juncheng Wu, Xianhang Li, Hong-Yu Zhou, Sheng Liu, Lei Xing, James Zou, Cihang Xie, et al. Medtrinity-25m: A large-scale multimodal dataset with multigranular annotations for medicine. *arXiv preprint arXiv:2408.02900*, 2024b.

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.

Xiaojuan Xue, Deshiwei Zhang, Chengyang Sun, Yiqiao Shi, Rongsheng Wang, Tao Tan, Peng Gao, Sujie Fan, Guangtao Zhai, Menghan Hu, et al. Xiaoqing: a q&a model for glaucoma based on llms. *Computers in Biology and Medicine*, 174:108399, 2024.

Kailai Yang, Tianlin Zhang, Ziyan Kuang, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. Mentallama: interpretable mental health analysis on social media with large language models. In *Proceedings of the ACM Web Conference 2024*, pp. 4489–4500, 2024.

Zhejun Yang, Tongtong Tian, Jilie Kong, and Hui Chen. Chatexosome: an artificial intelligence (ai) agent based on deep learning of exosomes spectroscopy for hepatocellular carcinoma (hcc) diagnosis. *Analytical Chemistry*, 97(8):4643–4652, 2025.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 2023.

APPENDIX

Contents

A CHECKLIST

A.1 ETHICS STATEMENT

This work adheres to the ICLR Code of Ethics. In this study, no human subjects or animal experimentation was involved. All datasets used, including SpineMed-450K, were sourced in compliance with relevant usage guidelines, ensuring no violation of privacy. We have taken care to avoid any biases or discriminatory outcomes in our research process. No personally identifiable information was used, and no experiments were conducted that could raise privacy or security concerns. We are committed to maintaining transparency and integrity throughout the research process.

A.2 REPRODUCIBILITY STATEMENT

Our work will be fully reproducible: we will open-source SpineBench, all questions, the code for running the API and open-source models, all model outputs, and the code for scoring the models. In other words, every part of the project will be made available.

A.3 LLM USAGE

Large Language Models (LLMs) were used to aid in the writing and polishing of the manuscript. Specifically, we used an LLM to assist in refining the language, improving readability, and ensuring clarity in various sections of the paper. The model helped with tasks such as sentence rephrasing, grammar checking, and enhancing the overall flow of the text.

It is important to note that the LLM was not involved in the ideation, research methodology, or experimental design. All research concepts, ideas, and analyses were developed and conducted by the authors. The contributions of the LLM were solely focused on improving the linguistic quality of the paper, with no involvement in the scientific content or data analysis.

The authors take full responsibility for the content of the manuscript, including any text generated or polished by the LLM. We have ensured that the LLM-generated text adheres to ethical guidelines and does not contribute to plagiarism or scientific misconduct.

B RELATED WORK

The landscape of medical AI is rapidly evolving, moving from broad, general-purpose models to highly specialized systems designed for clinical utility. Our work is situated within this trend, addressing a critical gap in the high-stakes field of spine surgery.

From Generalist Models to Domain Adaptation. Recent advances in Large Vision-Language Models (LVLMs), such as GPT-4V (OpenAI, 2023) and Gemini2.5-pro (Google, 2025a), have demonstrated significant progress in multimodal tasks (Yang et al., 2023; Team et al., 2023). However, when applied to the medical domain, their generalist nature becomes a distinct limitation. Multiple evaluations consistently show that while promising, these models lack the domain-specific expertise required for complex diagnostic tasks, performing below the level of human specialists (AlSaad et al., 2024). This inherent limitation of generalist models has fueled a clear and necessary trend toward specialization. In response, specialized medical LVLMs like LLaVA-Med (Li et al., 2023) and PMC-LLaMA (Wu et al., 2024) have been developed, fine-tuned on large biomedical corpora. Nevertheless, this approach still has shortcomings. For instance, in spinal diagnostics, a critical task is the synthesis of data from multimodal imaging—such as X-ray, CT, and MRI—to formulate a single, "level-aware" diagnosis. This integrative reasoning process, which requires localizing findings to specific vertebral levels, is a clinical skill that cannot be acquired from static, descriptive datasets alone. This further underscores a core principle: for high-stakes clinical applications, deep, narrow expertise is far more valuable than broad, superficial general knowledge. A powerful example validating this principle is OralGPT (Hao et al., 2025), a model trained on a small, highly curated dataset of intraoral photographs, which achieves performance comparable to state-of-the-art generalist models within its niche. This paradigm shift from generalist to specialist models is now clearly evident across numerous medical fields, from oncology to pathology (Qiu et al., 2023; Sarabadani et al., 2025; Yang et al., 2025; 2024; Barrit et al., 2024; Mo et al., 2025;

Deng et al., 2024; Xue et al., 2024; Bhaumik et al., 2023; Na, 2024; Guo et al., 2025).

Foundational Datasets and the Cognitive Gap. Progress in AI is fundamentally tied to the quality of training data. Foundational datasets like MIMIC-CXR (Johnson et al., 2019) and CheXpert (Irvin et al., 2019) have been instrumental for tasks like chest radiograph classification. Moving up in complexity are datasets for interactive Visual Question Answering (VQA). For instance, VQA-RAD (Lau et al., 2018) was manually constructed by clinicians asking naturally occurring questions about radiology images, representing a step toward more dynamic reasoning. More recently, large-scale efforts like MedTrinity-25M (Xie et al., 2024b) have emerged, providing over 25 million images with multi-granular annotations to support a wide range of tasks.

Within the spine domain itself, public datasets have primarily supported foundational computer vision tasks. As summarized in Table 1, existing datasets largely focus on single-modality perception tasks. For example, the RSNA LumbarDISC dataset (Richards et al., 2025) provides a large-scale MRI benchmark but is limited to the classification of stenosis severity grades. Similarly, BUU-LSPINE (Klinwichit et al., 2023) focuses on spondylolisthesis detection in X-rays, while the VerSe 2020 (Liebl et al., 2021) and Lumbar Spine MRI (van der Graaf et al., 2024) datasets provide voxel-level masks for segmentation tasks in CT and MRI, respectively.

However, these resources are primarily designed to support lower-level cognitive tasks like perception ("Where is the L4 vertebra?") or classification ("Is a fracture present?"). They lack the multimodal integration and instruction-following structure required to train models for the highest level of clinical cognition: synthesizing multimodal information into a comprehensive diagnosis and treatment plan. This reveals a crucial gap between existing data paradigms and the needs of clinical practice—a gap our work aims to fill by introducing the first large-scale instruction-tuning corpus designed for full-spectrum clinical reasoning.

AI in Spine Analysis: From Tools to Collaborators. Prior AI applications in spine analysis have focused on discrete tasks, creating valuable "tools" rather than "collaborators." These include automated vertebral segmentation and the measurement of spinal parameters (Lee et al., 2024a; Ibrahim et al., 2025a). While useful for improving efficiency, these tools perform isolated tasks, leaving the cognitive burden of synthesis and planning to the human clinician (Nath et al., 2025b). Our work directly addresses these gaps. By creating SpineMed-450k, a large-scale dataset derived from clinical workflows, and SpineBench, a benchmark focused on level-aware, multimodal reasoning, we provide the infrastructure to build and evaluate AI systems that can function as true clinical collaborators in the complex domain of spine surgery.

C PERFORMANCE COMPARISON ON MEDICAL REPORT GENERATION SUBTASKS

Table 8: LVLM performance comparison on medical report generation subtasks: Imaging Report (IR), Diagnosis (DGN), Patient Guidance (PG), Evidence-Based Plan (EBP), Technical Feasibility (TF), Risk Prognosis Management (RPM), Coverage (COV), Relevance (REL), Granularity (GRA), Explanation (EXP).

Model	IR	DGN	PG	EBP	TF	RPM	COV	REL	GRA	EXP
Proprietary LVLMS										
GPT-5	4.54	4.51	4.62	4.41	4.56	4.69	4.58	4.66	4.74	4.60
O3	4.39	4.25	4.32	4.30	4.40	4.43	4.34	4.45	4.50	4.39
Gemini-2.5-Pro	**4.55**	**4.51**	**4.79**	**4.60**	4.64	4.79	**4.69**	4.83	4.84	**4.80**
Claude-4	3.96	4.08	4.41	3.76	3.94	4.44	4.30	4.58	4.62	4.16
GPT-4o	3.16	3.03	3.30	3.07	2.80	3.30	3.35	4.30	2.92	3.25
GPT-5-mini	4.55	4.48	4.62	4.47	**4.71**	**4.98**	4.66	4.87	**4.90**	4.67
Gemini-2.5-Flash	4.43	4.29	4.73	4.51	4.48	4.88	4.64	**4.89**	4.82	4.67
Open-source LVLMS (>10B)										
GLM-4.5V	3.85	3.78	4.12	3.77	3.59	4.05	4.26	4.63	4.23	4.09
Qwen2.5-VL-72B	3.14	3.03	3.25	3.09	3.02	3.23	3.27	4.19	2.98	3.25
LinShu-32B	3.05	3.05	3.22	3.44	3.04	3.49	3.21	4.34	2.90	3.44
Medgemma-27B	2.88	3.49	4.14	3.56	3.32	3.26	3.48	4.29	3.51	3.32
Open-source LVLMS (<10B)										
HuaTuoGPT-7B	2.42	2.42	2.91	2.76	2.77	3.37	2.77	3.50	2.57	2.63
Qwen2.5-VL-7B	2.27	2.39	2.82	2.86	2.71	3.26	2.77	3.66	2.60	2.65
Ours	4.15	4.10	4.71	4.27	4.25	4.54	4.51	4.81	4.58	4.53

D PICTURE CONTEXT MATCHING ALGORITHM

The following algorithm processes Markdown files to extract image information and generate structured metadata in JSON format through parallel processing.

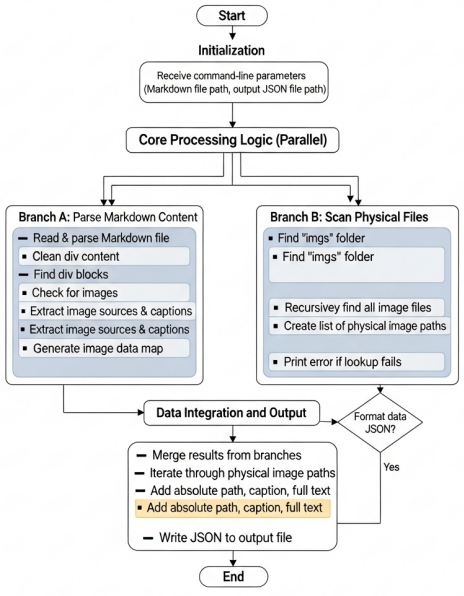


Figure 6: picture context matching algorithm

E DETAILED DISTRIBUTION ANALYSIS OF SPINEBENCH

To further validate the clinical representativeness and unbiased nature of SpineBench, we provide a comprehensive statistical breakdown of the testing QA set (487 items with specific subspecialty tags). These statistics demonstrate that SpineBench achieves rigorous coverage across diverse pathologies and data sources.

E.1 DISEASE SUBSPECIALTY DISTRIBUTION

As shown in Table 9, the benchmark covers 14 distinct spinal subspecialties. Notably, high-stakes and complex conditions such as *Tumor and Osteomyelitis* (16.72%) and *Complications* (10.80%) are heavily represented, ensuring that the evaluation reflects performance on critical clinical challenges rather than being dominated by common degenerative cases.

Table 9: Detailed Distribution of Spine Subspecialties in SpineBench

Rank	Subspecialty	Count	Percentage
1	Tumor and Osteomyelitis	96	16.72%
2	Complications	62	10.80%
3	Dysplastic and Congenital Deformities	49	8.54%
4	Idiopathic Scoliosis	47	8.19%
5	The Thoracic and Lumbar Degenerative Spine	34	5.92%
6	General Considerations	33	5.75%
7	Kyphosis and Postlaminectomy Deformities	31	5.40%
8	Anatomic Approaches	31	5.40%
9	Adult Spinal Deformity	26	4.53%
10	Spondylolisthesis	23	4.01%
11	Trauma	23	4.01%
12	Neuromuscular Spine Deformity	12	2.09%
13	Biomechanics	11	1.92%
14	The Cervical Degenerative Spine	9	1.57%
Total	**All Subspecialties**	**487**	**100.00%**

E.2 DATA SOURCE DIVERSITY

Table 10 illustrates the provenance of the testing data. The dataset maintains a rigorous balance between *Academic Resources* (45.8%), which ensure theoretical precision, and *Real-world Clinical Data* (54.2%), which test practical diagnostic robustness. Crucially, the Hospital Databases (33.5%) consist of private, internal cases that ensure a zero-leakage evaluation environment.

Table 10: Detailed Source Distribution of SpineBench QAs

Category	Specific Composition	Count	%	Evaluation Goal
Textbooks	Standard Medical Textbooks	203	41.7%	Theoretical Foundation
Hospital Databases	Proprietary data related to privacy	163	33.5%	Surgical Precision & Zero-Leakage
Clinical Case Reports	Real-world Case Reports	101	20.7%	Complex Clinical Scenarios
Literature	Academic Papers	17	3.5%	Cutting-edge Knowledge
Question Bank	Medical Exam Questions	3	0.6%	Standardized Logic
Total		**487**	**100%**	

F QUANTITATIVE COMPARISON OF SPINEGPT WITH GPT-4O

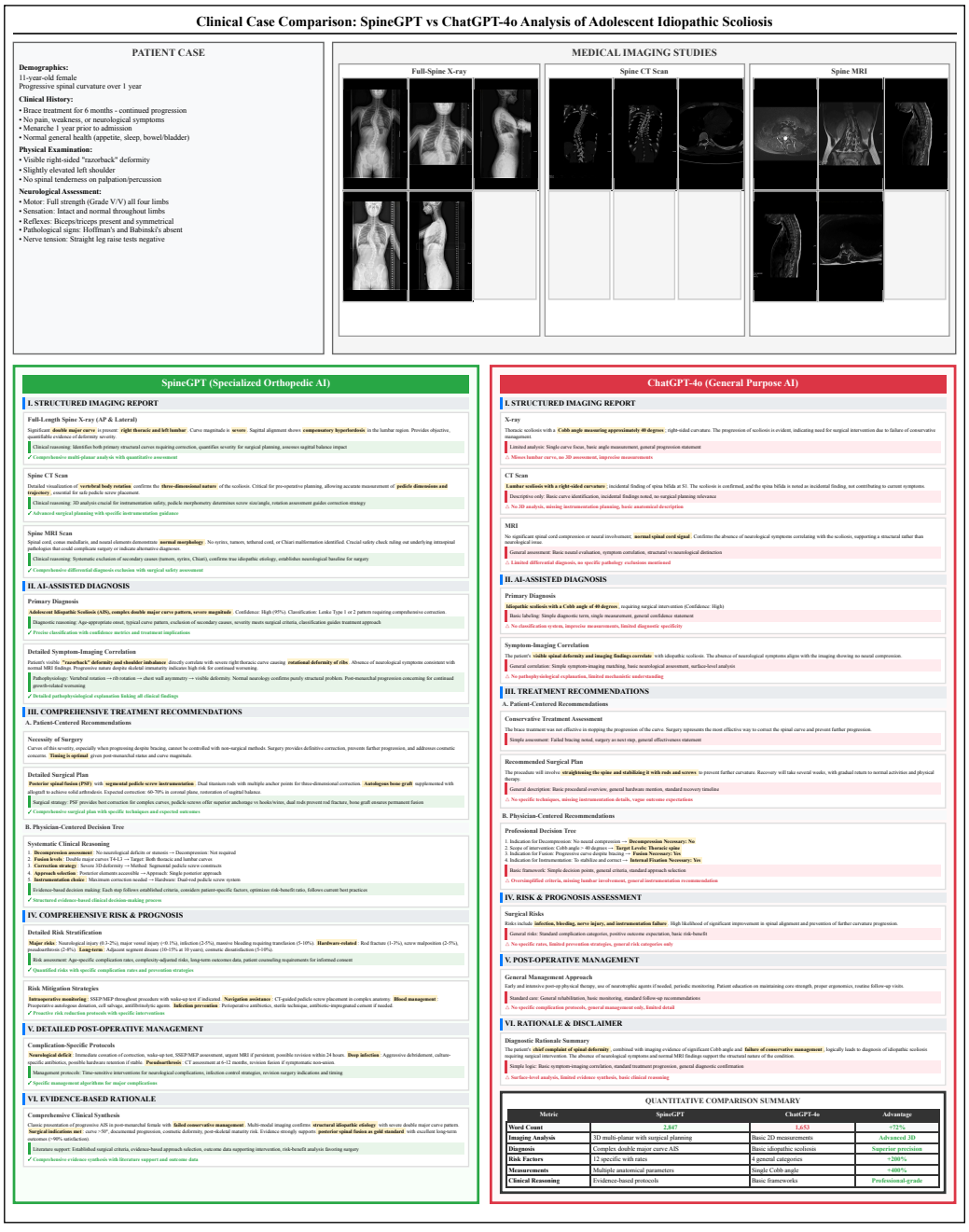


Figure 7: Comparative analysis of medical report generation capabilities between SpineGPT (Ours) and ChatGPT-4o (general-purpose AI) for an adolescent idiopathic scoliosis case. The comparison demonstrates significant differences in diagnostic depth, clinical reasoning, and treatment planning specificity. SpineGPT provides 72protocols, while ChatGPT-4o offers basic diagnostic and treatment recommendations suitable for general medical documentation.

Complete Medical Report: Adolescent Idiopathic Scoliosis Analysis

I. Structured Imaging Report

Modality	Key Findings & Parameters	Clinical Significance (How it supports the diagnosis)
Full-Length Spine X-ray (AP & Lateral)	Significant **double major curve** is present: **right thoracic and left lumbar** . Curve magnitude is **severe** . Sagittal alignment shows **compensatory hyperlordosis** in the lumbar region.	This provides objective, quantifiable evidence of the deformity, confirming its severity and identifying it as the primary deformity requiring surgical correction.
Spine CT Scan	Detailed visualization of **vertebral body rotation** confirms the **three-dimensional nature** of the scoliosis. Incidental finding of **spina bifida occulta at S1** is noted.	The CT scan is critical for pre-operative planning, allowing for accurate measurement of pedicle dimensions and trajectory, which is essential for safe and effective placement of spinal instrumentation like pedicle screws.
Spine MRI Scan	The spinal cord, conus medullaris, and neural elements demonstrate a **normal morphology** . No syrinx, tumors, tethered cord, or Chiari malformation are identified.	This is a crucial safety check. By ruling out underlying intraspinal pathologies, it confirms the diagnosis as a structural, idiopathic scoliosis and ensures that the planned surgical correction does not pose an additional risk to the spinal cord.

II. AI-Assisted Diagnosis

Diagnostic Tier	Output
1. Primary Diagnosis	**Adolescent Idiopathic Scoliosis (AIS)** , complex double major curve, severe. **Confidence: High.**
2. Symptom-Imaging Correlation	The patient's visible **"razorback" deformity and shoulder imbalance** are direct clinical manifestations of the severe right thoracic curve documented on X-rays and CT, which causes a rotational deformity of the ribs attached to the vertebrae. The absence of neurological symptoms (pain, weakness) is consistent with the MRI findings showing no compression of the spinal cord or nerve roots.
3. Secondary Diagnosis	Incidental finding of **spina bifida occulta at S1** ; considered clinically insignificant given the absence of any associated symptoms or tethering.
4. Differential Diagnoses	**Congenital Scoliosis:** Ruled out as there are no vertebral malformations (e.g., hemivertebrae, unsegmented bars). **Neuromuscular Scoliosis:** Ruled out due to the patient's normal neurological examination and lack of history of neuromuscular diseases. **Scoliosis secondary to Intraspinal Pathology:** Ruled out by the normal spinal cord and neural element appearance on MRI.

III. Treatment Recommendations

A. Patient-Centered Recommendations

Phase	Key Recommendations
Conservative Treatment Assessment	For curves of this size, especially when they are progressing despite bracing, non-surgical options like bracing or physical therapy are no longer effective at stopping the deformity from getting worse.
Necessity of Surgery	Surgery is the most reliable option to correct the existing curve, prevent it from worsening over time, and, most importantly, maintain overall spinal balance. This helps prevent future pain, potential lung problems, and cosmetic issues.
Recommended Surgical Plan	The surgery involves making an incision on your back to access the curved portion of your spine. Special screws and rods are then attached to the vertebrae to gently straighten the spine. Afterward, bone graft is placed to help the corrected segments of your spine heal together into a single, solid bone (a fusion), which makes the correction permanent.
Expected Post-operative Recovery	You will be encouraged out of bed shortly after surgery and will work closely with physical therapists. A hospital stay of several days is typical. A gradual return to school and light activities can be expected over several weeks to months. Heavy lifting and contact sports will be restricted for up to a year to ensure the fusion heals properly.

B. Physician-Centered Recommendations

Condition	Professional Decision Tree
Adolescent Idiopathic Scoliosis (AIS) - Severe double major curve	**1. Indication for Decompression:** No neurologic deficits or stenosis on imaging → Decompression Necessary: **No** . **2. Scope of intervention:** The primary structural curves are thoracic and lumbar, causing significant coronal and sagittal imbalance → Target Levels: **Correction of the thoracic and primary lumbar curves** . **3. Indication for Fusion:** Progressive, severe structural curve in an adolescent patient who is skeletally immature → Fusion Necessary: **Yes** . **4. Indication for Instrumentation:** To achieve and maintain three-dimensional correction of the deformity while the spinal fusion matures → Internal Fixation Necessary: **Yes** (e.g., dual-rod construct with pedicle screw fixation). **Conclusion:** Posterior spinal fusion with internal fixation is the optimal surgical strategy to address the severe deformity and prevent further progression.

IV. Risk & Prognosis Assessment

Item	Assessment
Surgical Risks	Potential risks include, but are not limited to: infection, excessive bleeding, nerve root or spinal cord injury (paralysis), cerebrospinal fluid leak, failure of correction or instrumentation (e.g., rod breakage, screw pullout), failure of fusion (pseudoarthrosis), and blood clots.
Prognosis	The prognosis for halting curve progression and achieving significant correction of the existing deformity is excellent. The patient can expect a return to a high level of function and prevention of future health complications associated with untreated severe scoliosis.
Red Flag Warnings ⚠	**After surgery, immediately report any new or worsening pain, fever, numbness, tingling, weakness, or loss of bowel/bladder control. These could indicate a serious complication such as infection or neurological injury.**

V. Management of Potential Post-operative Issues

Potential Issue	Intervention Strategy	Red Flag ⚠
Post-operative Neurological Deficit	Minimized by using intraoperative neuromonitoring (SSEP/MEP). If a deficit occurs, immediate investigation (e.g., imaging) and potential revision surgery may be required.	**Any new weakness, numbness, or paralysis in the limbs or trunk immediately post-operatively.**
Instrumentation Failure / Pseudoarthrosis	Risk is minimized by using a robust multi-rod construct and ensuring adequate bone grafting. Long-term rehabilitation is crucial to protect the construct while the fusion heals. Routine follow-up is needed to assess for implant integrity and fusion progress.	**Persistent pain over the incision, evidence of implant prominence under the skin, or a gradual return of deformity on follow-up X-rays.**
Adjacent Segment Disease	This is a long-term concern. Management includes patient education on maintaining core strength, proper body mechanics, and a healthy weight. Routine clinical and radiographic monitoring is advised.	**Development of new, significant back or leg pain years after the initial surgery.**

VI. Rationale & Disclaimer

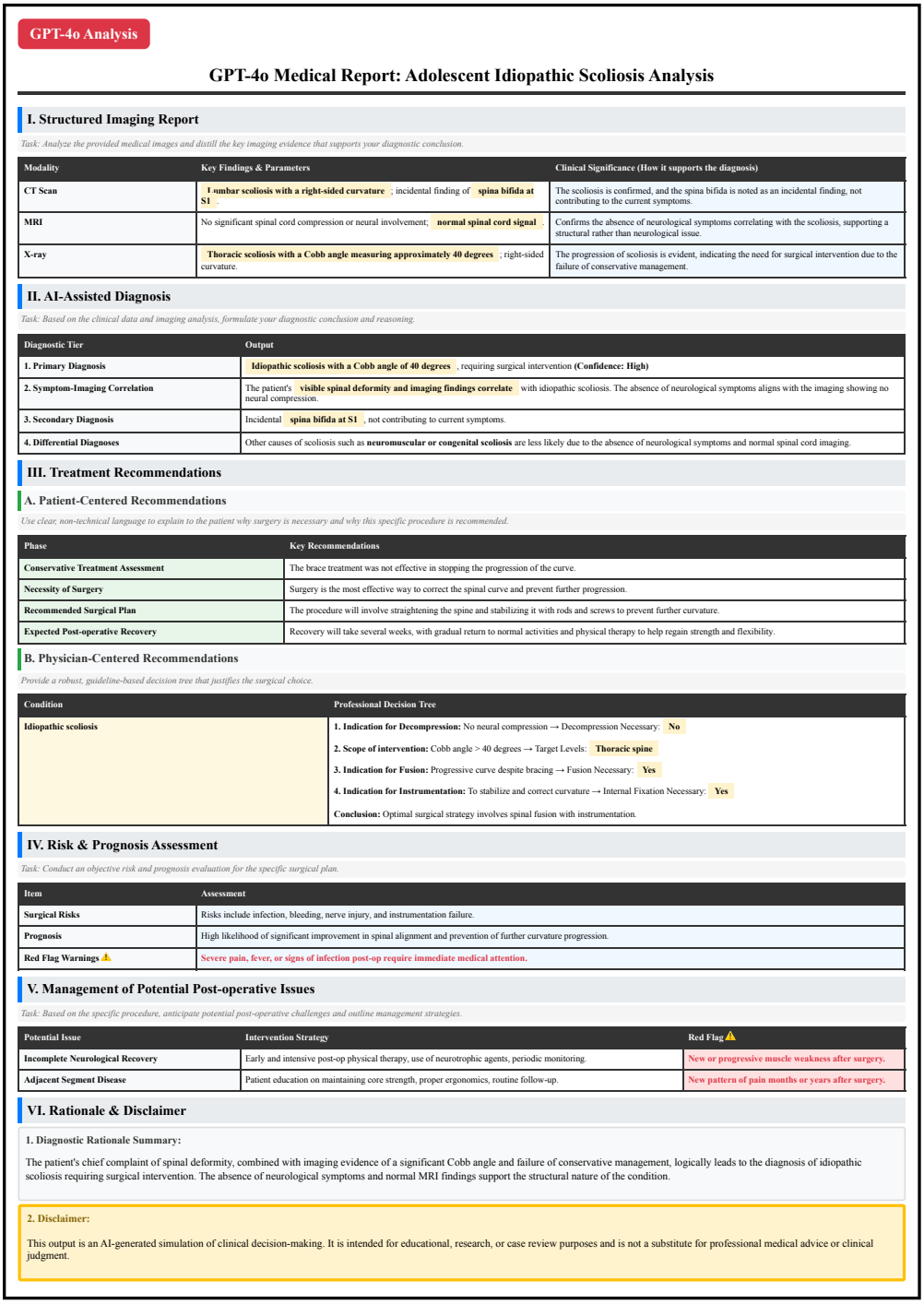
1. Diagnostic Rationale Summary:

This 11-year-old female presents with a classic case of progressive adolescent idiopathic scoliosis, characterized by a significant, structurally progressive spinal deformity. The diagnosis is confirmed by clinical examination revealing a visible rib hump and shoulder imbalance, and further validated by multi-modal imaging. The X-rays and CT scan quantify the severe curve, while the MRI rules out any underlying neurological or congenital causes. Given the magnitude of the curve and its documented progression despite bracing, surgical intervention is the standard of care to correct the deformity, halt its progression, and prevent future cardiopulmonary and functional complications.

2. Disclaimer:

This output is an AI-generated simulation of clinical decision-making. It is intended for educational, research, or case review purposes and is not a substitute for professional medical advice or clinical judgment.

Figure 8: Our model's medical report output for adolescent idiopathic scoliosis, featuring six-section structured format: imaging findings, AI diagnosis, treatment recommendations, risk assessment, post-operative management, and clinical rationale.



GPT-4o Analysis

GPT-4o Medical Report: Adolescent Idiopathic Scoliosis Analysis

I. Structured Imaging Report

Task: Analyze the provided medical images and distill the key imaging evidence that supports your diagnostic conclusion.

Modality	Key Findings & Parameters	Clinical Significance (How it supports the diagnosis)
CT Scan	**Lumbar scoliosis with a right-sided curvature**; incidental finding of **spina bifida at S1**.	The scoliosis is confirmed, and the spina bifida is noted as an incidental finding, not contributing to the current symptoms.
MRI	No significant spinal cord compression or neural involvement; **normal spinal cord signal**.	Confirms the absence of neurological symptoms correlating with the scoliosis, supporting a structural rather than neurological issue.
X-ray	**Thoracic scoliosis with a Cobb angle measuring approximately 40 degrees**, right-sided curvature.	The progression of scoliosis is evident, indicating the need for surgical intervention due to the failure of conservative management.

II. AI-Assisted Diagnosis

Task: Based on the clinical data and imaging analysis, formulate your diagnostic conclusion and reasoning.

Diagnostic Tier	Output
1. Primary Diagnosis	**Idiopathic scoliosis with a Cobb angle of 40 degrees**, requiring surgical intervention **(Confidence: High)**
2. Symptom-Imaging Correlation	The patient's **visible spinal deformity and imaging findings correlate** with idiopathic scoliosis. The absence of neurological symptoms aligns with the imaging showing no neural compression.
3. Secondary Diagnosis	Incidental **spina bifida at S1**, not contributing to current symptoms.
4. Differential Diagnoses	Other causes of scoliosis such as **neuromuscular or congenital scoliosis** are less likely due to the absence of neurological symptoms and normal spinal cord imaging.

III. Treatment Recommendations

A. Patient-Centered Recommendations

Use clear, non-technical language to explain to the patient why surgery is necessary and why this specific procedure is recommended.

Phase	Key Recommendations
Conservative Treatment Assessment	The brace treatment was not effective in stopping the progression of the curve.
Necessity of Surgery	Surgery is the most effective way to correct the spinal curve and prevent further progression.
Recommended Surgical Plan	The procedure will involve straightening the spine and stabilizing it with rods and screws to prevent further curvature.
Expected Post-operative Recovery	Recovery will take several weeks, with gradual return to normal activities and physical therapy to help regain strength and flexibility.

B. Physician-Centered Recommendations

Provide a robust, guideline-based decision tree that justifies the surgical choice.

Condition	Professional Decision Tree
Idiopathic scoliosis	**1. Indication for Decompression:** No neural compression → Decompression Necessary: **No** **2. Scope of intervention:** Cobb angle > 40 degrees → Target Levels: **Thoracic spine** **3. Indication for Fusion:** Progressive curve despite bracing → Fusion Necessary: **Yes** **4. Indication for Instrumentation:** To stabilize and correct curvature → Internal Fixation Necessary: **Yes** **Conclusion:** Optimal surgical strategy involves spinal fusion with instrumentation.

IV. Risk & Prognosis Assessment

Task: Conduct an objective risk and prognosis evaluation for the specific surgical plan.

Item	Assessment
Surgical Risks	Risks include infection, bleeding, nerve injury, and instrumentation failure.
Prognosis	High likelihood of significant improvement in spinal alignment and prevention of further curvature progression.
Red Flag Warnings ⚠	**Severe pain, fever, or signs of infection post-op require immediate medical attention.**

V. Management of Potential Post-operative Issues

Task: Based on the specific procedure, anticipate potential post-operative challenges and outline management strategies.

Potential Issue	Intervention Strategy	Red Flag ⚠
Incomplete Neurological Recovery	Early and intensive post-op physical therapy, use of neurotrophic agents, periodic monitoring.	**New or progressive muscle weakness after surgery.**
Adjacent Segment Disease	Patient education on maintaining core strength, proper ergonomics, routine follow-up.	**New pattern of pain months or years after surgery.**

VI. Rationale & Disclaimer

1. Diagnostic Rationale Summary:

The patient's chief complaint of spinal deformity, combined with imaging evidence of a significant Cobb angle and failure of conservative management, logically leads to the diagnosis of idiopathic scoliosis requiring surgical intervention. The absence of neurological symptoms and normal MRI findings support the structural nature of the condition.

2. Disclaimer:

This output is an AI-generated simulation of clinical decision-making. It is intended for educational, research, or case review purposes and is not a substitute for professional medical advice or clinical judgment.

Figure 9: ChatGPT-4o generated medical report for adolescent idiopathic scoliosis, showing general-purpose AI's approach to clinical documentation with basic diagnostic and treatment recommendations.

G PROMPTS

Criteria for Assessing Dimensional Quality in Reports

I. Role and Core Task
You will act as a **top-tier clinical medical expert** and **AI evaluator** (LLM-Judge). Your core task is to rigorously compare the **[LLM Generated Answer]** provided below with the **[Standard Answer]**. Based on a comprehensive and detailed scoring rubric, you will systematically evaluate the **[LLM Generated Answer]**'s performance against the **[Standard Answer]** across multiple dimensions, including **accuracy, completeness, logical coherence, readability,** and **clinical utility**. Finally, you will output your evaluation in the specified simple format.

II. Inputs for Evaluation
1. **[Standard Answer] (Golden Answer)**
[Please paste the ideal, golden standard answer here]
2. **[LLM Generated Answer]**
[Please paste the AI-generated answer that requires evaluation here]

III. Evaluation Instructions & Scoring Rubric
Please score the **[LLM Generated Answer]** for each dimension below, strictly based on its comparison with the **[Standard Answer]**.
Please note: Scores MUST be continuous values (e.g., 3.5, 4.2, 4.7) to more precisely reflect the subtle differences in the evaluation results between the integer standards. The integer scores (1-5 points) in the rubric should serve as the primary anchors for your scoring, but you should use decimal precision to capture nuanced differences. For example:
- If performance is slightly above **"Good"** (4 pts) but not quite **"Excellent"** (5 pts), use scores like 4.3 or 4.6.
- If performance has minor gaps compared to **standard**, use scores like 3.8 or 4.2.
- Avoid whole numbers unless the performance exactly matches the integer anchor description.

1. **Structured Imaging Report**
5 pts (Excellent / On Par): On par with the **[Standard Answer]**, accurately describes all key imaging findings, correctly explains their clinical significance, and includes quantitative descriptions.
4 pts (Good / Minor Gaps): The description of major findings is correct, but it lacks some of the quantitative details present in the **[Standard Answer]**.
3 pts (Fair / Clear Gaps): The description is generally correct, but the explanation of clinical significance is clearly less sufficient or in-depth than the **[Standard Answer]**.
2 pts (Poor / Serious Deficiencies): Omits or incorrectly describes key findings that are mentioned in the **[Standard Answer]**.
1 pt (Unacceptable / Completely Wrong): Seriously misinterprets the imaging, with key conclusions that contradict the **[Standard Answer]**.

2. **AI-Assisted Diagnosis**
5 pts (Excellent / On Par): On par with the **[Standard Answer]**, the primary diagnosis is completely correct, secondary diagnoses are reasonably listed, and key differential diagnoses are correctly ruled out.
4 pts (Good / Minor Gaps): The primary diagnosis is correct, but the list of secondary diagnoses is less complete than in the **[Standard Answer]**.
3 pts (Fair / Clear Gaps): The primary diagnosis is correct but omits important differential diagnoses that are mentioned in the **[Standard Answer]**.
2 pts (Poor / Serious Deficiencies): The primary diagnosis is partially incorrect or omits key components present in the **[Standard Answer]**.
1 pt (Unacceptable / Completely Wrong): The diagnosis is completely wrong or misses a life-threatening condition.
3. **Treatment Recommendations**
3.1 **Patient-Oriented Advice**
5 pts (Excellent / On Par): On par with the **[Standard Answer]**, the language is extremely colloquial and easy to understand, the information is completely accurate, the structure is clear, and it is highly effective at reassuring the patient.

Figure 10: Criteria for Assessing Dimensional Quality in Reports

Criteria for Assessing Dimensional Quality in Reports

4 pts (Good / Minor Gaps): The language is easy to understand and the core information is accurate, but the level of empathy or nuance is slightly inferior to the **[Standard Answer]**.
3 pts (Fair / Clear Gaps): The language is generally understandable but contains unexplained jargon, the information is mostly correct but vague, and it clearly lacks the empathy shown in the **[Standard Answer]**.
2 pts (Poor / Serious Deficiencies): The language is obscure and jargon-heavy, the information contains errors or critical omissions, and it is likely to cause patient anxiety.
1 pt (Unacceptable / Completely Wrong): The communication contains serious errors, is misleading, or provides harmful information, making it completely unacceptable.

3.2 **Treatment Plan & Evidence-Based Consistency**
5 pts (Excellent / On Par): The plan's rationale, individualization, and discussion of evidence-based support are all on par with the depth and breadth of the **[Standard Answer]**.
4 pts (Good / Minor Gaps): The core plan is reasonable, but the discussion of the evidence base is less detailed than in the **[Standard Answer]**.
3 pts (Fair / Clear Gaps): The plan is generally reasonable but lacks the individualized adjustments highlighted in the **[Standard Answer]**.
2 pts (Poor / Serious Deficiencies): Parts of the plan are inconsistent with clinical guidelines, making its rationale far weaker than the **[Standard Answer]**'s.
1 pt (Unacceptable / Completely Wrong): The plan clearly conflicts with evidence-based medicine and is diametrically opposed to the recommendations in the **[Standard Answer]**.

3.3 **Surgical/Technical Details & Feasibility**
5 pts (Excellent / On Par): The explanation of surgical goals, technical details, preventive measures, and backup plans is comparable in completeness and professionalism to the **[Standard Answer]**.
4 pts (Good / Minor Gaps): Covers the main technical details, but its consideration of complication prevention is less thorough than the **[Standard Answer]**'s.
3 pts (Fair / Clear Gaps): The description of details is overly general and lacks the specificity and feasibility assessment present in the **[Standard Answer]**.
2 pts (Poor / Serious Deficiencies): Omits key technical details mentioned in the **[Standard Answer]**, making its feasibility questionable.
1 pt (Unacceptable / Completely Wrong): The technical details are infeasible or pose a safety risk.

4. **Risk, Prognosis & Post-Op Management**
5 pts (Excellent / On Par): Provides a perioperative management plan, follow-up schedule, and strategy for potential issues that is as systematic, complete, and forward-thinking as the **[Standard Answer]**.
4 pts (Good / Minor Gaps): Covers the main measures but is less systematic or detailed in certain aspects compared to the **[Standard Answer]**.
3 pts (Fair / Clear Gaps): Mentions basic safety measures but lacks the systematic and structured approach demonstrated in the **[Standard Answer]**.
2 pts (Poor / Serious Deficiencies): Omits important safety protocols that are emphasized in the **[Standard Answer]**.
1 pt (Unacceptable / Completely Wrong): Seriously neglects safety, contradicting the patient-centric principles of the **[Standard Answer]**.

5. **Theoretical Basis & Disclaimer (EVA 4D Evaluation)**

5.1 **Coverage**
5 pts (Excellent / On Par): On par with the **[Standard Answer]**, accurately identifies and explains all key pieces of evidence with no omissions.
4 pts (Good / Minor Gaps): Covers most key evidence but may omit one non-critical element that was included in the **[Standard Answer]**.
3 pts (Fair / Clear Gaps): Covers the main evidence but omits one key element or two minor elements present in the **[Standard Answer]**.
2 pts (Poor / Serious Deficiencies): Covers only a small amount of evidence, and the chain of reasoning is far less complete than the **[Standard Answer]**'s.
1 pt (Unacceptable / Completely Wrong): Fails to cover any key evidence, or the evidence cited contradicts the factual basis of the **[Standard Answer]**.

Figure 11: Criteria for Assessing Dimensional Quality in Reports

Criteria for Assessing Dimensional Quality in Reports

5.2 **Relevance**
5 pts (Excellent / On Par): On par with the **[Standard Answer]**, all discussion is tightly focused on the core diagnosis and decision, with no irrelevant content.
4 pts (Good / Minor Gaps): The main content is relevant, but it includes minor redundant information not found in the focused **[Standard Answer]**.
3 pts (Fair / Clear Gaps): The discussion mixes relevant and irrelevant information, diluting the focus compared to the **[Standard Answer]**.
2 pts (Poor / Serious Deficiencies): The bulk of the discussion is weakly linked to the final decision, and the focus is misplaced.
1 pt (Unacceptable / Completely Wrong): The discussion is entirely irrelevant to the diagnosis or is based on incorrect assumptions.

5.3 **Granularity**
5 pts (Excellent / On Par): On par with the **[Standard Answer]**, provides precise, quantitative details sufficient to support in-depth clinical judgment.
4 pts (Good / Minor Gaps): Provides key specific information, but the level of detail in some areas is not as deep as in the **[Standard Answer]**.
3 pts (Fair / Clear Gaps): The information is overly general and lacks the distinguishing details found in the **[Standard Answer]**.
2 pts (Poor / Serious Deficiencies): Uses highly generalized language, providing far less informational value than the **[Standard Answer]**.
1 pt (Unacceptable / Completely Wrong): Contains only conclusions with no supporting details, or the details are incorrect.

5.4 **Explanation**
5 pts (Excellent / On Par): On par with the **[Standard Answer]**, the chain of reasoning is clear, complete, and seamless, with all parts logically supporting the conclusion.
4 pts (Good / Minor Gaps): The overall logic is coherent, but the reasoning for a specific step is slightly less clear or direct than in the **[Standard Answer]**.
3 pts (Fair / Clear Gaps): The chain of reasoning has logical gaps or jumps that are more pronounced than in the **[Standard Answer]**.
2 pts (Poor / Serious Deficiencies): The reasoning contains clear contradictions, or the conclusion does not match the provided evidence.
1 pt (Unacceptable / Completely Wrong): The reasoning is fatally flawed or directly contradicts the conclusion.

IV. Required Output Format
Please strictly follow this simple format. Each line should contain exactly one score and justification:
IMAGING_REPORT: [Score] | [Justification]
DIAGNOSIS: [Score] | [Justification]
PATIENT_GUIDANCE: [Score] | [Justification]
EVIDENCE_BASED_PLAN: [Score] | [Justification]
TECHNICAL_FEASIBILITY: [Score] | [Justification]
RISK_PROGNOSIS: [Score] | [Justification]
COVERAGE: [Score] | [Justification]
RELEVANCE: [Score] | [Justification]
GRANULARITY: [Score] | [Justification]
EXPLANATION: [Score] | [Justification]

Example:
IMAGING_REPORT: 4.2 | The report accurately describes key findings but lacks some quantitative details.
DIAGNOSIS: 3.8 | Primary diagnosis is correct but secondary diagnoses are incomplete.

Figure 12: Criteria for Assessing Dimensional Quality in Reports

Orthopedic Category Classification Prompt

Classify the orthopedic question into **ONE** category. **Answer ONLY the category name.**

Question: {question} Answer: {answer} Categories:

- **Spine Surgery** - Conditions, injuries, and surgeries related to the spine
- **Foot and Ankle Surgery** - Conditions, injuries, and surgeries related to the foot and ankle
- **Orthopedic Trauma** - Fractures, dislocations, and other acute injuries
- **Hand and Upper Extremity Surgery** - Conditions, injuries, and surgeries related to the hand, wrist, elbow, and shoulder
- **Musculoskeletal Oncology** - Bone and soft tissue tumors
- **Orthopedic Sports Medicine** - Sports-related injuries, arthroscopic surgery
- **Adult Joint Reconstruction** - Arthritis, joint replacement surgery (e.g., hip, knee)

ANSWER WITH THE EXACT NAME: "Spine Surgery", "Foot and Ankle Surgery", "Orthopedic Trauma", "Hand and Upper Extremity Surgery", "Musculoskeletal Oncology", "Orthopedic Sports Medicine", "Adult Joint Reconstruction"

Figure 13: Prompt for Orthopedic Category Classification

Spine Category Classification Prompt

Classify the spine surgery question into **ONE** category. **Answer ONLY the category name.**

Question: {question} Answer: {answer} Categories:

- **General Considerations** - Basic spine anatomy, evaluation, imaging
- **Biomechanics** - Spine mechanics, forces, stability
- **Anatomic Approaches** - Surgical approaches, exposure
- **The Cervical Degenerative Spine** - Cervical disc, stenosis, anterior cervical discectomy and fusion (ACDF)
- **The Thoracic and Lumbar Degenerative Spine** - Lumbar disc, stenosis, fusion
- **Spondylolisthesis** - Spondylolisthesis, pars interarticularis defect
- **Idiopathic Scoliosis** - Adolescent idiopathic scoliosis, curves
- **Adult Spinal Deformity** - Adult scoliosis, sagittal balance
- **Dysplastic and Congenital Deformities** - Dysplastic and congenital deformities
- **Neuromuscular Spine Deformity** - Neuromuscular spinal deformity
- **Kyphosis and Postlaminectomy Deformities** - Kyphosis, post-laminectomy deformities
- **Trauma** - Spine fractures, spinal cord injury
- **Tumor and Osteomyelitis** - Spine tumors, infections
- **Complications** - Surgical complications, internal fixation failure

ANSWER WITH THE EXACT NAME: "General Considerations", "Biomechanics", "Anatomic Approaches", "The Cervical Degenerative Spine", "The Thoracic and Lumbar Degenerative Spine", "Spondylolisthesis", "Idiopathic Scoliosis", "Adult Spinal Deformity", "Dysplastic and Congenital Deformities", "Neuromuscular Spine Deformity", "Kyphosis and Postlaminectomy Deformities", "Trauma", "Tumor and Osteomyelitis", "Complications"

Figure 14: Prompt for Spine Category Classification

Generating Medical Q&A for Fine-Tuning Prompt

"You are a senior clinical medical educator. Please carefully read the provided medical textbook content below and generate **multiple high-quality open-ended question-answer pairs** based on the core knowledge points, all for fine-tuning large language models. Each Q&A should be self-contained and completely independent."
"**Strict Requirements:**"
"1. **Complete Independence**: Each question and answer must constitute a complete knowledge unit that can be understood without any external background materials."
"2. **Prohibited Referential Terms**: Strictly prohibit using terms like ẗhis guide,ẗ ẗhe study,ẗ ẗhe above materials,ẗ ẗhis article,ẗ ẗhe reportör any other terms that refer to the original text in questions or answers."
"3. **Clinical Depth Requirements**: Questions should reflect real clinical scenarios, testing deep understanding and clinical thinking rather than simple yes/no questions."
"4. **Open-Ended Design**: Questions should encourage detailed analysis, requiring comprehensive, structured answers that demonstrate clinical reasoning processes."
"5. **Answer Completeness**: Answers must be detailed and comprehensive, including analysis process, reasoning logic, and final conclusions."
"6. **Question Type Diversity**: Should cover multiple dimensions including pathological mechanism explanation, diagnostic thinking analysis, treatment plan design, complication prevention strategies, etc."
"**Question Quantity Requirements:**"
"- Generate appropriate number of questions based on text length"
"- Short text (1-2 paragraphs): Generate 2-3 questions"
"- Medium text (3-5 paragraphs): Generate 3-5 questions"
"- Long text (6+ paragraphs): Generate 5-8 questions"
"**Output Format Requirements:**"
"Strictly follow the XML format below, each textbook page can generate multiple questions:"
"“‘xml"
"<problem>[Open-ended question 1 stem]</problem>"
"<answer>[Detailed open-ended answer for question 1, including analysis process and conclusions]</answer>"
"<problem>[Open-ended question 2 stem]</problem>"
"<answer>[Detailed open-ended answer for question 2, including analysis process and conclusions]</answer>"
"<problem>[Open-ended question 3 stem]</problem>"
"<answer>[Detailed open-ended answer for question 3, including analysis process and conclusions]</answer>"
"“‘"
"**Important Notes:**"
"1. Strictly follow the XML format"
"2. Question stems should be clear and specific, encouraging deep thinking, avoiding simple yes/no questions"
"3. Answers should be comprehensive and detailed, including analysis process, reasoning logic, and final conclusions"
"4. Output only XML objects, no additional explanatory text"
"5. Each question should be independent and complete, not dependent on other questions or external materials"
"6. Answers should demonstrate the depth of medical professional knowledge and the logic of clinical thinking"
"**Textbook Content:**content"
"Please generate high-quality medical open-ended question-answer pairs based on the above textbook content."

Figure 15: Prompt for Generating Medical Q&A for Fine-Tuning

Generating Medical MCQs for Fine-Tuning Prompt

"You are a senior clinical medical educator and examination expert. Please carefully read the provided medical textbook content below and generate **multiple high-quality multiple-choice questions with answers** based on the core knowledge points, all for fine-tuning large language models. Each Q&A should be self-contained and completely independent."
"**Strict Requirements:**"
"1. **Complete Independence**: Each question and options must constitute a complete knowledge unit that can be understood without any external background materials."
"2. **Prohibited Referential Terms**: Strictly prohibit using terms like ïthis guide, ïthe study, ïthe above materials, ïthis article, ïthe reportör any other terms that refer to the original text in questions or options."
"3. **Clinical Depth Requirements**: Questions should reflect real clinical scenarios, testing deep understanding and clinical judgment rather than simple memorization."
"4. **Option Design**: Each question must include 4 options (A, B, C, D), with 1 correct answer and 3 high-quality distractors. Distractors should be based on common clinical misconceptions or related concepts."
"5. **Question Type Diversity**: Should cover multiple dimensions including diagnostic reasoning, treatment selection, mechanism explanation, differential diagnosis, complication prevention, etc."
"**Question Quantity Requirements:**"
"- Generate appropriate number of questions based on text length"
"- Short text (1-2 paragraphs): Generate 2-3 questions"
"- Medium text (3-5 paragraphs): Generate 4-6 questions"
"- Long text (6+ paragraphs): Generate 7-10 questions"
"**Output Format Requirements:**"
"Strictly follow the XML format below, each textbook page can generate multiple questions:"
"“‘xml"
"<problem>[Question 1 stem] A. [Option A] B. [Option B] C. [Option C] D. [Option D]</problem>"
"<answer>[Question 1 correct answer option letter]</answer>"
"<problem>[Question 2 stem] A. [Option A] B. [Option B] C. [Option C] D. [Option D]</problem>"
"<answer>[Question 2 correct answer option letter]</answer>"
"<problem>[Question 3 stem] A. [Option A] B. [Option B] C. [Option C] D. [Option D]</problem>"
"<answer>[Question 3 correct answer option letter]</answer>"
"“‘"
"**Important Notes:**"
"1. Strictly follow the XML format"
"2. Question stems should be clear and specific, testing deep understanding and clinical judgment"
"3. Options should be reasonably designed, including correct answers and high-quality distractors"

Figure 16: Prompt for Generating Medical MCQs for Fine-Tuning

Generating Context-Localized Multimodal Q&A Prompt

"You are a senior clinical medical educator. Based on the provided image information, caption, and context, precisely locate the image's position in the context and generate high-quality open-ended questions and answers."
"**Core Task: Multimodal Understanding and Precise Localization Open-Ended Q&A**"
"**Step 1: Multimodal Information Understanding**"
"1. **Image Understanding**: Analyze the specific medical content shown in the image (anatomical structures, pathological manifestations, surgical procedures, imaging features, instrument usage, etc.)"
"2. **Caption Understanding**: Identify figure numbers, positions, operational steps, or key information mentioned in the caption"
"3. **Context Understanding**: Analyze medical knowledge points, operational procedures, and clinical key points in the preceding and following text"
"4. **Position Localization**: Precisely locate the image's specific position and role in the context"
"**Step 2: Precise Position Localization**"
"1. **Caption-Context Matching**:"
" - If the caption contains a figure number (e.g., F̈igure 12.1), find the corresponding figure reference in the context"
" - If the caption describes operational steps, locate the corresponding operational description in the context"
" - If the caption describes anatomical structures, find related anatomical descriptions in the context"
"2. **Context Position Analysis**:"
" - Analyze preceding text: background information, preparation steps, and related concepts before the image appears"
" - Analyze following text: operational steps, precautions, and clinical significance after the image is shown"
" - Determine the image's specific role in the entire process"
"**Step 3: Generate Open-Ended Q&A Based on Precisely Located Content**"
"Must generate open-ended questions and answers based on precisely located medical knowledge points:"
"- Deeply analyze the relationship between located content and the image"
"- Generate open-ended questions based on precisely located content"
"- Ensure questions are highly relevant to both image content and context"
"- If unable to precisely locate suitable content, skip this question"
"**Step 4: High-Quality Open-Ended Q&A Design**"
"Based on precisely located content, generate open-ended questions and answers with the following characteristics:"
"**Q&A Design Principles:**"
"1. **Multimodal Relevance**: Questions must be relevant to image content, caption information, and context content simultaneously"
"2. **Clinical Orientation**: Questions should be based on real clinical scenarios, testing clinical thinking and decision-making abilities"
"3. **Open-Ended Design**: Encourage deep thinking, avoid yes/no questions, require detailed analysis"
"4. **Position Precision**: Questions should be based on the image's precise location in the context"
"5. **Prohibited Referential Terms**: Strictly prohibit using terms like äccording to the context, ẗhis guide, ẗhe study, ẗhe above materials, ẗhis article, ẗhe reportör any other terms that refer to the original text in questions or answers"
"6. **Answer Design**:"

Figure 17: Prompt for Generating Context-Localized Multimodal Q&A

Generating Context-Localized Multimodal Q&A Prompt

" - Answers must be based on precisely located content"
" - Cover relevant medical knowledge and clinical considerations"
" - Reflect clinical thinking and decision-making process"
"7. **Question Type Priority**:"
" - Diagnostic analysis questions (in-depth analysis based on imaging findings, clinical symptoms, etc.)"
" - Treatment decision questions (detailed analysis of surgical indications, treatment plan selection, etc.)"
" - Mechanism explanation questions (in-depth explanation of anatomical-physiological relationships, pathological mechanisms, etc.)"
" - Technical operation questions (detailed explanation of surgical steps, instrument usage, etc.)"
" - Risk assessment questions (comprehensive analysis of complication prevention, management strategies, etc.)"
"**Output Format Requirements:**"
"Strictly follow the following XML format, each image can generate multiple different Q&A pairs:"
"“‘xml"
"<problem><image>
n[First open-ended question stem]</problem>"
"<answer>[First detailed open-ended answer]</answer>"
"<problem><image>
n[Second open-ended question stem]</problem>"
"<answer>[Second answer, directly answering the question]</answer>"
"<problem><image>
n[Third open-ended question stem]</problem>"
"<answer>[Third answer, directly answering the question]</answer>"
"“‘"
"**Important Notes:**"
"1. Strictly follow the XML format"
"2. Question stems should be clear and specific, encouraging deep thinking, avoiding simple yes/no questions"
"3. Answers should be comprehensive and detailed, including analysis process, reasoning logic, and final conclusions"
"4. If unable to precisely locate relevant content, do not generate questions"
"5. Output only one complete XML object"
"6. Strictly prohibit using referential terms"
"7. **Image Reference Standards**: When referencing images in questions, use general terms like äs shown in the image, ẗhe image displays, ïmaging findingsëtc., strictly prohibit using specific figure numbers (e.g., Ïigure 10.8, Ïigure 12.1, etc.)"
"**Processing Workflow:**"
"1. Analyze the image caption to understand the specific medical content shown"
"2. Precisely locate medical knowledge points in the context related to the caption"
"3. Determine the image's specific position and role in the context"
"4. If precise localization is successful and content is suitable for questions, generate open-ended Q&A based on located content"
"5. If unable to precisely locate or content is not suitable for questions, do not generate questions"
"6. Ensure questions have clinical value and educational significance"
"**Provided Information:**"
"Image Caption: caption"
"Context Information: context"
"Please precisely locate the image's position in the context and generate high-quality medical open-ended questions and answers. If unable to precisely locate suitable content, do not generate questions. Strictly follow the specified XML format."

Figure 18: Prompt for Generating Context-Localized Multimodal Q&A

Generating Context-Localized Multimodal MCQs Prompt

"You are a senior clinical medical educator and examination expert. Your task is to precisely locate the image's position in the context based on the provided image information, caption, and context, then generate high-quality multiple-choice questions."
"**Core Task: Multimodal Understanding and Precise Localization Question Generation**"
"**Step 1: Multimodal Information Understanding**"
"1. **Image Understanding**: Analyze the specific medical content shown in the image (anatomical structures, pathological manifestations, surgical procedures, imaging features, instrument usage, etc.)"
"2. **Caption Understanding**: Identify figure numbers, positions, operational steps, or key information mentioned in the caption"
"3. **Context Understanding**: Analyze medical knowledge points, operational procedures, and clinical key points in the preceding and following text"
"4. **Position Localization**: Precisely locate the image's specific position and role in the context"
"**Step 2: Precise Position Localization**"
"1. **Caption-Context Matching**:"
" - If the caption contains a figure number (e.g., F̈igure 12.1), find the corresponding figure reference in the context"
" - If the caption describes operational steps, locate the corresponding operational description in the context"
" - If the caption describes anatomical structures, find related anatomical descriptions in the context"
"2. **Context Position Analysis**:"
" - Analyze preceding text: background information, preparation steps, and related concepts before the image appears"
" - Analyze following text: operational steps, precautions, and clinical significance after the image is shown"
" - Determine the image's specific role in the entire process"
"**Step 3: Generate Questions Based on Precisely Located Content**"
"Must generate clinical multiple-choice questions based on precisely located medical knowledge points:"
"- Deeply analyze the relationship between located content and the image"
"- Generate multiple-choice questions based on precisely located content"
"- Ensure questions are highly relevant to both image content and context"
"- If unable to precisely locate suitable content, skip this question"
"**Step 4: High-Quality Multiple-Choice Question Design**"
"Based on precisely located content, generate clinical multiple-choice questions with the following characteristics:"
"**Q&A Design Principles:**"
"1. **Multimodal Relevance**: Questions must be relevant to image content, caption information, and context content simultaneously"
"2. **Clinical Orientation**: Questions should be based on real clinical scenarios, testing clinical thinking and decision-making abilities"
"3. **Multiple-Choice Design**: Provide multiple options to test deep understanding and clinical judgment"
"4. **Position Precision**: Questions should be based on the image's precise location in the context"
"5. **Prohibited Referential Terms**: Strictly prohibit using terms like äccording to the context, ẗhis guide, ẗhe study, ẗhe above materials, ẗhis article, ẗhe reportör any other terms that refer to the original text in questions or options"

Figure 19: Prompt for Generating Context-Localized Multimodal MCQs

Generating Context-Localized Multimodal MCQs Prompt

"6. **Option Design**:"
" - Correct answer must be based on precisely located content"
" - Distractors should be based on common clinical misconceptions or related but inaccurate concepts"
" - All options should have clinical plausibility"
"7. **Question Type Priority**:"
" - Diagnostic analysis questions (in-depth analysis based on imaging findings, clinical symptoms, etc.)"
" - Treatment decision questions (detailed analysis of surgical indications, treatment plan selection, etc.)"
" - Mechanism explanation questions (in-depth explanation of anatomical-physiological relationships, pathological mechanisms, etc.)"
" - Technical operation questions (detailed explanation of surgical steps, instrument usage, etc.)"
" - Risk assessment questions (comprehensive analysis of complication prevention, management strategies, etc.)"
"**Output Format Requirements:**"
"**Strictly follow the following XML format:**"
"“‘xml"
"<problem><image>
n[Question stem] A. [Option A] B. [Option B] C. [Option C] D. [Option D]</problem>"
"<answer>[Correct answer option letter]</answer>"
"“‘"
"**Important Notes:**"
"1. Strictly follow the XML format"
"2. Question stems should be clear and specific, testing deep understanding and clinical judgment"
"3. Options should be reasonably designed, including correct answers and distractors"
"4. If unable to precisely locate relevant content, do not generate questions"
"5. Output only one complete XML object"
"6. Strictly prohibit using referential terms"
"7. **Image Reference Standards**: When referencing images in questions, use general terms like äs shown in the image, ïhe image displays, ïmaging findingsëtc., strictly prohibit using specific figure numbers (e.g., Ḧigure 10.8, Ḧigure 12.1, etc.)"
"**Processing Workflow:**"
"1. Analyze the image caption to understand the specific medical content shown"
"2. Precisely locate medical knowledge points in the context related to the caption"
"3. Determine the image's specific position and role in the context"
"4. If precise localization is successful and content is suitable for questions, generate multiple-choice questions based on located content"
"5. If unable to precisely locate or content is not suitable for questions, do not generate questions"
"6. Ensure questions have clinical value and educational significance"
"**Provided Information:**"
"Image Caption: caption"
"Context Information: context"
"Please precisely locate the image's position in the context and generate high-quality medical multiple-choice questions. If unable to precisely locate suitable content, do not generate questions. Strictly follow the specified XML format."

Figure 20: Prompt for Generating Context-Localized Multimodal MCQs