MVVQ-RAD: Medical Voice Vision Question-Reason Answer Dataset: A Comprehensive Multimodal Medical AI Dataset with Speech, Visual Localization, and Explainable Reasoning

Hsiang-Wei Hu[1,2][0009−0005−7297−2217]⋆, Pei-Shan Wang[1][0009−0003−2210−0176]⋆, Ren-Di Wu[1][0009−0007−4564−8790]⋆, Li-Ju Chen[1][0009−0006−4582−2018], and Zih-Jia Luo[1][0009−0001−3954−6022]

[1] International Academia of Biomedical Innovation Technology (B.I.T.), USA
hw.hsiang.wei@gmail.com
Corresponding author: Hsiang-Wei Hu
[2] Taiwan Artificial Intelligence Association, Taiwan

Abstract. We introduce the MVVQ-RAD dataset, the first comprehensive multimodal medical AI dataset that integrates visual grounding, speech synthesis, and interpretable reasoning components with uncertainty quantification. Built upon the widely used VQA-RAD dataset [1], MVVQ-RAD comprises 300 carefully curated samples spanning multiple imaging modalities—CT, MRI, and X-ray—with detailed diagnostic reasoning traces and validation from human experts. Our pipeline employs an innovative five-stage process executed through the LangGraph framework [2], enabling self-evaluation via LLM-as-a-judge [18]. The evaluation results achieve strong performance across key metrics: visual grounding quality (0.909), speech quality (0.718), reasoning quality (0.515), consistency score (0.542), and an overall quality score of 0.534. By enabling unprecedented multimodal integration and incorporating a rigorous quality assurance mechanism, this dataset effectively addresses a critical gap in explainable medical AI [3]. As an open-source resource, MVVQ-RAD facilitates the development of trustworthy medical AI systems capable of multimodal, interpretable outputs, thereby improving physician acceptance and enhancing clinical decision support [4]. The dataset will be fully released following institutional review by partner medical centers. Code available at `https://github.com/whats2000/MedVoiceQAReasonDataset`

Keywords: Medical AI · Multimodal Learning · Explainable AI · Visual Question Answering · Speech Processing · Medical Imaging

⋆ These authors contributed equally to this work.

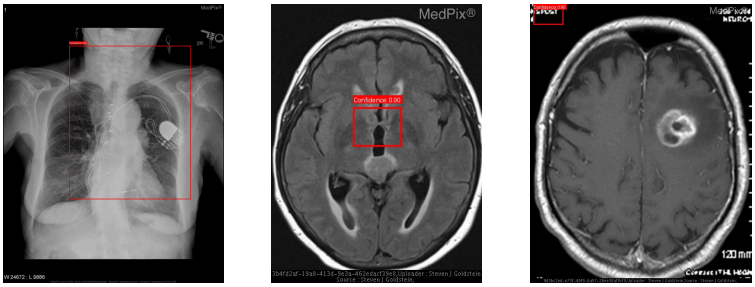

Fig. 1. We introduce an interface that helps medical experts evaluate and correct model inferences by revealing their reasoning basis. Qualitative analysis showed a clear gap between model reasoning and that of physicians in pure visual question answering tasks. **Left:** Bounding box is too large and exceeds the target area—model identified the region but needs refinement. **Middle:** Bounding box is slightly off—key features were detected, but alignment is imprecise. **Right:** When asked about modality, the model focused on blurred text in the upper-left corner, showing reliance on non-image features despite label obfuscation.

1 Introduction

Artificial intelligence (AI) has demonstrated high accuracy in medical image diagnosis. However, its clinical adoption remains limited due to a lack of explainability and trustworthiness. A core issue lies in the "black box" nature of most AI systems, which contrasts sharply with the evidence-based, interpretable reasoning required in medical decision-making. Clinicians not only need diagnostic results but also insight into the AI's reasoning process, its uncertainty estimates, and visual justification [3,5].

Existing medical visual question answering (VQA) datasets such as VQA-RAD [1] and PathVQA [6] provide aligned image-text data. However, they lack key components, including diagnostic reasoning traces, uncertainty annotations, multimodal integration, and visual grounding capabilities—features essential for supporting real-world clinical workflows.

Surprisingly, despite the impressive performance of many existing models on VQA benchmarks, we observed a substantial discrepancy between the model's reasoning process and that of human experts. As illustrated in Figure 1, while the model is capable of answering questions correctly, its underlying reasoning often deviates from clinical expectations. This highlights the need for interpretable inference mechanisms that allow medical professionals to examine and correct the model's decision-making process.

To address these gaps, we propose MVVQ-RAD, a novel multimodal medical AI dataset that integrates image, speech, and textual data, structured around a five-stage processing pipeline with a comprehensive quality assurance mechanism. MVVQ-RAD builds upon VQA-RAD, selecting 300 representative sam-

ples, and augments them with bounding box annotations, speech processing modules, structured diagnostic reasoning, and human expert validation.

The development of MVVQ-RAD not only fills a critical dataset void but also lays a foundation for the advancement of trustworthy, explainable AI systems in medicine. By delivering multimodal explanations and uncertainty estimates, MVVQ-RAD aims to foster clinician trust and support clinical decision-making. The dataset will be fully open-sourced following medical institutional review, contributing to the standardization and transparency of medical AI research [4].

2 Related Work

Medical Visual Question Answering (Med-VQA) has become a key area of research in AI, especially as the need for intelligent systems that can interpret medical images and respond to clinical questions continues to grow. Unlike general-purpose VQA tasks, building robust datasets for Med-VQA is much more difficult. The challenges stem largely from concerns around patient privacy and the fact that medical data often requires annotation by trained professionals. Despite these hurdles, researchers have made steady progress over the years, developing several influential datasets to help move the field forward.

VQA-RAD VQA-RAD [1] is a widely used, manually curated dataset designed specifically for visual question answering in the radiology domain. It features naturally occurring questions and answers authored and validated by clinicians, a design choice that overcomes the limitations of automatically generated data. The dataset consists of 315 radiology images and 3,515 associated visual questions, 1,515 of which are open-ended. Questions and answers are generally concise. VQA-RAD serves as a foundational benchmark for Med-VQA systems, with evaluation metrics such as simple accuracy, mean accuracy, and BLEU. Results from VQA-RAD show that models trained only on general VQA data perform poorly, largely due to difficulties with specialized terminology and medical phrasing. The dataset's primary goal is to encourage the development of clinically relevant and robust VQA systems.

PathVQA PathVQA [6] is the first dataset tailored specifically for VQA in pathology. It supports the development of AI systems aimed at interpreting pathology images and answering related questions, with a long-term vision of approaching board-exam-level understanding. It contains 4,998 pathology images and 32,799 question-answer pairs. Unlike earlier Med-VQA datasets that focus on multiple-choice formats, PathVQA emphasizes open-ended questions, significantly increasing its difficulty. The dataset was constructed via a semi-automated process using images and captions sourced from pathology textbooks and online repositories. Questions were then automatically generated and manually verified. While the dataset explores a range of visual features—such as location, shape, and color—its reliance on rule-based question generation limits linguistic diversity. Moreover, its design does not fully replicate the contextual

depth of real clinical cases, which typically include patient history and diagnostic context.

SLAKE SLAKE (Semantically-Labeled Knowledge-Enhanced) [11] is a bilingual Med-VQA dataset available in English and Chinese. It contains 642 radiology images and over 14,000 question-answer pairs. A unique aspect of SLAKE is its rich semantic annotation, including segmentation masks and bounding boxes, created by experienced physicians. It also includes a structured medical knowledge graph consisting of organ- and disease-related triples, enabling complex, knowledge-intensive queries. SLAKE spans multiple imaging modalities such as CT, MRI, and X-ray, and covers various anatomical regions including the brain, chest, and abdomen. The dataset encourages diverse question types that require both visual understanding and external knowledge reasoning. Despite its high-quality annotations and structured knowledge, SLAKE still falls short of achieving clinical-level benchmarks. Moreover, the dataset remains relatively small in image volume compared to recent large-scale alternatives.

PMC-VQA PMC-VQA [12] is a large-scale dataset designed to support generative Med-VQA tasks, with a focus on free-form answer generation. It comprises 149,000 medical images and 227,000 question-answer pairs, making it one of the largest datasets in the domain. The dataset includes a wide variety of image types, from radiology scans to signal charts. It was created through an automated pipeline using large language models (e.g., ChatGPT) to generate Q&A pairs from figure captions in biomedical literature. A series of filtering steps ensures visual relevance, and a manually curated test set was constructed to support robust evaluation. Benchmark results show that text-only models struggle with PMC-VQA, underscoring the importance of visual reasoning. Pretrained models like MedVInT, trained on PMC-VQA, have achieved state-of-the-art results across various Med-VQA benchmarks. Nonetheless, the dataset still faces challenges related to content consistency and the difficulty of objectively evaluating generative outputs.

3 Methods

Figure 2 presents the complete MVVQ-RAD pipeline, which consists of two primary components: (1) an AI-driven annotation and verification system that integrates multi-step reasoning and validation modules, and (2) a human-in-the-loop interface designed for interactive correction. The following sections provide a detailed breakdown of each component.

3.1 AI-driven Annotation and Verification Nodes

The MVVQ-RAD pipeline starts with a preprocessing stage in which each image is converted into a standardized PNG format. Additionally, metadata from

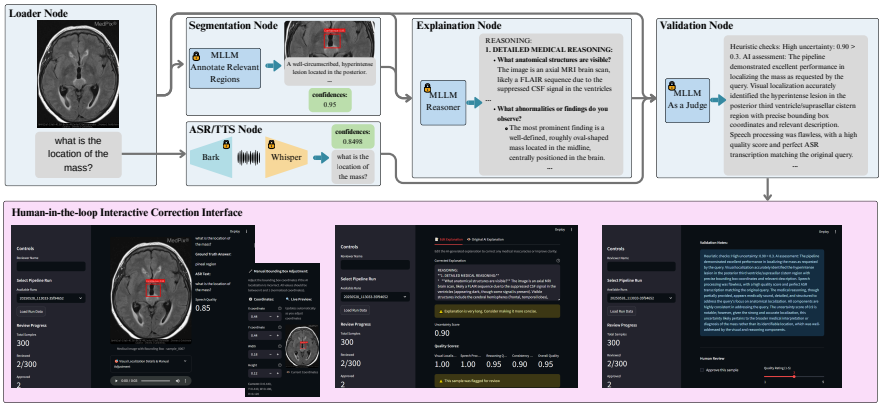

Fig. 2. MVVQ-RAD dataset creation pipeline. The system includes a multi-stage AI annotation and verification framework and a human-interactive correction interface.

the original VQA-RAD dataset is loaded to support subsequent annotation and verification steps. The processed data is denoted as:

$$\mathcal{D}_0 = \{I, Q, M\} \tag{1}$$

where I is the input medical image, Q is the original clinical question, and M is the associated metadata.

The data $\mathcal{D}_0$ is then forwarded to two parallel processing nodes:

Segmentation Node This node leverages a multimodal large language model (MLLM) to identify and annotate relevant regions within the image, generating bounding boxes along with corresponding semantic information with the prompt (see Fig. 3).

Input:

$$(I, Q) \tag{2}$$

Output:

$$\begin{cases} \mathbf{b} = (x, y, w, h), & \text{normalized bounding box coordinates} \\ c \in [0, 1], & \text{confidence score} \\ r, & \text{region description} \\ \rho, & \text{relevance reasoning} \end{cases} \tag{3}$$

ASR/TTS Node This node synthesizes speech from the input question and re-transcribes it via ASR to assess the speech quality and accuracy.

Input:

$$Q \tag{4}$$

Output:

$$\begin{cases} s, & \text{path to synthesized speech file} \\ Q', & \text{transcribed question via ASR} \\ q_s \in [0,1], & \text{speech quality score} \\ \mu, & \text{ASR metadata} \end{cases} \tag{5}$$

Chain-of-Thought Reasoning Node This node integrates visual and linguistic features to generate a clinical reasoning explanation and associated uncertainty, guided by a specifically designed prompt (see Fig. 4).

Input:

$$(Q, \mathbf{b}, I) \tag{6}$$

Output:

$$\begin{cases} \epsilon, & \text{generated reasoning explanation} \\ u \in [0,1], & \text{uncertainty score} \end{cases} \tag{7}$$

Verification Node The final node aggregates all outputs and computes overall quality metrics and review decisions, using a structured evaluation prompt (see Fig. 5).

Input:

$$\mathcal{X} = \{I, Q, \mathbf{b}, c, r, \rho, s, Q', q_s, \mu, \epsilon, u\} \tag{8}$$

Output:

$$\begin{cases} \delta \in \{\text{True}, \text{False}\}, & \text{whether human review is needed} \\ \kappa, & \text{critic notes or missing components} \\ \mathbf{q} = (q_v, q_a, q_r, q_{\text{cons}}, q_{\text{total}}), & \text{quality scores} \end{cases} \tag{9}$$

where q_v is the visual localization quality, q_a is the speech processing quality, q_r is the reasoning quality, q_{cons} is the consistency score, and q_{total} is the overall quality score. Each quality score is normalized within the range $[0, 1]$.

3.2 Human-in-the-loop Interactive Correction Interface

Our system includes a human-in-the-loop (HITL) review interface designed to assist medical professionals in evaluating and appropriately correcting AI-generated outputs. The core functions of this interface are as follows:

- **Annotation Preview**: Provides toggles to show or hide bounding boxes and the corresponding clinical question for each sample.
- **Speech Playback Panel**: Allows users to play back the synthesized speech, view the transcribed text, and inspect the associated confidence score.
- **Interactive Bounding Box Editor**: Offers an interactive interface that enables real-time adjustment of bounding box position and size, with immediate visual feedback.
- **Reasoning Editor**: Enables domain experts to review and revise the AI-generated reasoning process.

- **Agent Suggestion Panel**: Displays evaluation scores along with the Agent's generated feedback and correction suggestions.
- **Final Review Section**: Allows reviewers to assign quality scores, apply predefined quality tags, and write structured comments summarizing their review.

3.3 Experimental Setup

Models and Baselines The models used in our pipeline are as follows. For text-to-speech (TTS) synthesis, we adopted the `suno/bark` model [15]. The corresponding speech-to-text (STT) verification was conducted using the `Whisper large-v3` model [16]. For bounding box annotation, reasoning generation, and final verification, we employed the `Gemini 2.5 Flash` model [17] as the large language model (LLM) agent for semantic evaluation (LLM-as-a-judge [18]).

All experiments were conducted using samples from the VQA-RAD dataset [1], which contains 3,515 visual question–answer (VQA) pairs. Of these, 1,515 (43.1%) are free-form questions, and 733 of them have paraphrased counterparts, meaning that 48.3% of the free-form questions have corresponding rephrased versions.

Experiment Details The entire experimental pipeline was executed on a single NVIDIA RTX 3090 GPU. The full annotation process took a total of 5 hours and 37 minutes.

4 Results

4.1 Agent Self-Evaluation Results Analysis and Discussion:

Table 1 presents the self-evaluation results from the Validate Agent across 300 annotated samples. While the Visual Quality Agent reported high confidence levels, subsequent manual inspection revealed that several bounding boxes were spatially misaligned. This highlights the critical need for an expert-driven correction interface, as illustrated in Figure 1.

We also observed that the agent systematically assigned lower scores for reasoning quality, suggesting that clinical reasoning generation remains a key area for improvement. Particularly concerning is the Yes/No question category: although it accounts for a large portion of the dataset, the model's performance on these items remains consistently poor across quality metrics.

4.2 Comparative Analysis with Existing Datasets

As shown in Table 2, current publicly available medical VQA datasets and models largely lack support for multimodal inputs and reasoning capabilities. In contrast, our dataset introduces not only text, image, and speech modalities, but also explicitly links visual regions to textual explanations via bounding box annotations, thereby enhancing interpretability.

Table 1. Statistics by Question Type

Question Type	Count	Uncertainty	Overall Quality	Speech Quality	Visual Quality	Reasoning Quality
Abnormality Detection	12	0.84	0.55	0.68	0.94	0.53
Counting	3	0.87	0.42	0.78	0.92	0.27
Long Answer	5	0.85	0.59	0.78	0.94	0.52
Modality	19	0.96	0.81	0.86	0.99	0.88
Organ Identification	29	0.88	0.60	0.74	0.88	0.60
Short Phrase	26	0.79	0.54	0.70	0.86	0.47
Single Word	40	0.87	0.55	0.74	0.94	0.53
Yes/No	166	0.89	0.49	0.70	0.90	0.47

Furthermore, the integration of uncertainty estimation helps reduce the burden on medical reviewers by identifying low-confidence cases that require closer inspection.

Table 2. Comprehensive comparison with existing medical VQA datasets and models.

Dataset/Model	Samples	Modalities	Reasoning	Speech	Localization
VQA-RAD [1]	3,515	Image+Text	x	x	x
PathVQA [6]	32,799	Image+Text	x	x	x
SLAKE [11]	14,028	Image+Text	x	x	x
PMC-VQA [12]	227,194	Image+Text	x	x	x
LLaVA-Med [7]	–	Image+Text	Partial	x	x
VILA-M3 [8]	–	Image+Text	Partial	x	v
Me-LLaMA [9]	–	Text	v	x	x
Medical-mT5 [10]	–	Text	v	x	x
Ours	**300**	**Image+Text+Speech**	**v**	**v**	**v**

5 Discussion

The MVVQ-RAD dataset represents a paradigm shift in the design of medical AI datasets by systematically integrating multimodal components and agent-based validation workflows. We address limitations in existing datasets by improving clinical applicability and aligning data more closely with real-world diagnostic scenarios.

Moreover, the introduction of AI-driven self-evaluation significantly mitigates the instability commonly associated with manual annotation, while also reducing the time burden for human experts. Through our interactive user interface, clinical researchers can focus on validating domain-specific reasoning rather than low-level annotation tasks.

Importantly, the modular LangGraph-based design enables seamless replacement of processing nodes as newer models emerge, ensuring compatibility with

state-of-the-art architectures. It also allows us to design task-specific expert agents tailored to a variety of medical applications.

MVVQ-RAD is built to support the long-term development of multimodal large language model (LLM) agents in the healthcare domain. By integrating speech, text, and visual data with higher-level annotation features—such as emotion, role, intent, and task—the dataset aims to advance explainability and multimodal reasoning capabilities in medical LLM systems.

6 Future Work

1. **Improving Reasoning Capabilities.** While the system performs stably on direct image-to-description tasks, it struggles with complex reasoning or arithmetic-based questions such as counting (see Table 1). Enhancing multi-step and causal reasoning can improve both diagnostic accuracy and interpretability in medical imaging systems [19,20].
2. **Optimizing Counting Tasks.** Although modality-based questions exhibit the highest model uncertainty, they result in the best output quality when answered. This suggests that higher-threshold automation strategies can be designed to prioritize machine-handled review, reducing human annotation load. Integrating dedicated counting modules and advanced attention mechanisms may significantly improve accuracy and confidence in VQA systems [21,22].
3. **Collaboration with Clinical Experts and Dataset Refinement.** Future work will involve deeper collaboration with medical professionals to refine workflows and clean the dataset further. This includes introducing segmentation mask outputs, expanding to joint prediction tasks, and extending support for multilingual medical QA scenarios.

7 Conclusion

Innovation in Multimodal Integration This work presents the first integration of image, text, and speech modalities into a single VQA dataset and annotation workflow (Image+Text+Speech). Compared to existing datasets that only support image and text, our approach more accurately reflects the complex and multimodal nature of real-world clinical environments.

Integration of Localization, Chain-of-Thought Reasoning, and Uncertainty Estimation The annotation pipeline simultaneously incorporates bounding box localization, chain-of-thought (CoT) reasoning, and uncertainty estimation. This integrated design not only enhances the explainability of AI outputs but also enables dynamic filtering of low-confidence cases, allowing human reviewers to focus their attention more efficiently and effectively.

References

1. Lau, J.J., Gayen, S., Ben Abacha, A., Demner-Fushman, D.: A dataset of clinically generated visual questions and answers about radiology images. Scientific Data **5**(1), 180251 (2018). https://doi.org/10.1038/sdata.2018.251
2. LangChain Inc.: LangGraph: Stateful orchestration framework for agentic AI workflows, Version 0.4.5. https://github.com/langchain-ai/langgraph, last accessed 2025/06/01
3. Holzinger, A., Biemann, C., Pattichis, C.S., Kell, D.B.: What do we need to build explainable AI systems for the medical domain? arXiv preprint arXiv:1712.09923 (2017) https://arxiv.org/abs/1712.09923
4. Topol, E.J.: High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine **25**(1), 44–56 (2019). https://doi.org/10.1038/s41591-018-0300-7
5. Lipton, Z.C.: The mythos of model interpretability. Commun. ACM **61**(10), 36–43 (2018). https://doi.org/10.1145/3233231
6. He, X., Zhang, Y., Mou, L., Xing, E.P., Xie, P.: PathVQA: 30,000+ questions for medical visual question answering. arXiv preprint arXiv:2003.10286 (2020). https://arxiv.org/abs/2003.10286
7. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: LLaVA-Med: training a large language-and-vision assistant for biomedicine in one day. In: Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS 2023), pp. 28541–28564. Curran Associates, Red Hook (2023). https://openreview.net/forum?id=GSuP99u2kR
8. Nath, V., Li, W., Yang, D., Myronenko, A., Zheng, M., Lu, Y., Liu, Z., Yin, H., Law, Y.M., Tang, Y., Guo, P., Zhao, C., Xu, Z., He, Y., Harmon, S., Simon, B., Heinrich, G., Aylward, S., Edgar, M., Zephyr, M., Molchanov, P., Turkbey, B., Roth, H., Xu, D.: VILA-M3: Enhancing vision-language models with medical expert knowledge. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 14788–14798. IEEE, New Orleans (2025)
9. Xie, Q., Chen, Q., Chen, A., Peng, C., Hu, Y., Lin, F., Peng, X., Huang, J., Zhang, J., Keloth, V., Zhou, X., He, H., Ohno-Machado, L., Wu, Y., Xu, H., Bian, J.: Me-LLaMA: Foundation large language models for medical applications. Research Square **rs.3.rs-4240043** (2024). https://doi.org/10.21203/rs.3.rs-4240043/v1
10. García-Ferrero, I., Agerri, R., Atutxa Salazar, A., Cabrio, E., de la Iglesia, I., Lavelli, A., Magnini, B., Molinet, B., Ramirez-Romero, J., Rigau, G., Villa-Gonzalez, J.M., Villata, S., Zaninello, A.: MedMT5: an open-source multilingual text-to-text LLM for the medical domain. In: Calzolari, N., Kan, M.-Y., Hoste, V., Lenci, A., Sakti, S., Xue, N. (eds.) Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pp. 11165–11177. ELRA and ICCL, Torino, Italy (2024). https://aclanthology.org/2024.lrec-main.974/
11. Liu, B., Zhan, L.M., Xu, L., Ma, L., Yang, Y., Wu, X.M.: SLAKE: A semantically labeled knowledge-enhanced dataset for medical visual question answering. In: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), pp. 1650–1654. IEEE, Nice, France (2021). https://doi.org/10.1109/ISBI48211.2021.9434010
12. Zhang, X., Wu, C., Zhao, Z., Lin, W., Zhang, Y., Wang, Y., Xie, W.: PMC-VQA: Visual instruction tuning for medical visual question answering. arXiv preprint arXiv:2305.10415 (2023). https://arxiv.org/abs/2305.10415

13. Ribeiro, M.T., Singh, S., Guestrin, C.: "Why should I trust you?" Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2016), pp. 1135–1144. ACM, New York (2016). https://doi.org/10.1145/2939672.2939778
14. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: BERTScore: evaluating text generation with BERT. In: International Conference on Learning Representations (ICLR) (2020). https://openreview.net/forum?id=SkeHuCVFDr
15. Suno AI: Bark: text-to-audio generation. https://huggingface.co/suno/bark, last accessed 2025/06/01
16. Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision. In: Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., Scarlett, J. (eds.) Proceedings of the 40th International Conference on Machine Learning (ICML 2023). Proceedings of Machine Learning Research, vol. 202, pp. 28492–28518. PMLR, Honolulu, Hawaii, USA (2023). https://proceedings.mlr.press/v202/radford23a.html
17. Google DeepMind: Gemini 2.5 Flash, https://deepmind.google/technologies/gemini/, last accessed 2025/06/01
18. Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E.P., Zhang, H., Gonzalez, J.E., Stoica, I.: Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In: Proceedings of the 37th International Conference on Neural Information Processing Systems (NeurIPS 2023), pp. 2020–2048. Curran Associates Inc., Red Hook, NY, USA (2023). https://openreview.net/forum?id=uccHPGDlao
19. Zhao, G., Feng, Q., Chen, C., Zhou, Z., Yu, Y.: Diagnose like a radiologist: hybrid neuro-probabilistic reasoning for attribute-based medical image diagnosis. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(11), 7400–7416 (2022). https://doi.org/10.1109/TPAMI.2021.3130759
20. Huang, S., Wang, L., Liao, J., Liu, L.: Multi-attentional causal intervention networks for medical image diagnosis. Knowledge-Based Systems **299**, 111993 (2024). https://doi.org/10.1016/j.knosys.2024.111993
21. Chen, M., Wang, Y., Chen, S., Wu, Y.: Counting attention based on classification confidence for visual question answering. In: 2019 IEEE International Conference on Parallel and Distributed Processing with Applications, Big Data and Cloud Computing, Sustainable Computing and Communications, Social Computing and Networking (ISPA/BDCloud/SustainCom/SocialCom), pp. 1173–1179. IEEE (2019). https://doi.org/10.1109/ISPA-BDCloud-SustainCom-SocialCom48970.2019.00167
22. Zhang, Y., Hare, J., Prügel-Bennett, A.: Learning to count objects in natural images for visual question answering. In: International Conference on Learning Representations (ICLR) (2018). https://openreview.net/forum?id=B12Js_yRb

Prompt for Segmentation Node

You are a medical imaging AI assistant. Your task is to identify the most relevant region in this medical image that relates to the following question:

Question: {text_query}

Please analyze the image and:
1. Identify the anatomical region or abnormality that is most relevant to answering this question
2. Provide a bounding box (x, y, width, height) in normalized coordinates (0-1) that encompasses this region
3. Explain your reasoning for selecting this region

Your response should be in the following JSON format:

```
{{
  "bounding_box": {{
    "x": <normalized x coordinate of top-left corner>,
    "y": <normalized y coordinate of top-left corner>,
    "width": <normalized width>,
    "height": <normalized height>
  }},
  "confidence": <confidence score 0-1>,
  "region_description": "<description of what this region contains>",
  "relevance_reasoning": "<explanation of why this region is relevant to the question>"
}}
```

If no specific region can be identified or the entire image is relevant, you may return a bounding box that covers most or all of the image.

Fig. 3. This figure illustrates the prompt used in the segmentation node.

Prompt for Explanation Node

You are an expert medical imaging specialist. Analyze this medical image and provide detailed reasoning.

MEDICAL IMAGE QUERY: {query}

VISUAL LOCALIZATION: The relevant region has been identified with the following bounding box:
- Coordinates: {visual_box}

Please provide:

1. DETAILED MEDICAL REASONING:
  - What anatomical structures are visible?
  - What abnormalities or findings do you observe?
  - How does the visual localization relate to the query?
  - What differential diagnoses should be considered?
  - What is your final assessment?

2. REASONING CHAIN:
  - Step 1: Initial observation
  - Step 2: Feature analysis
  - Step 3: Clinical correlation
  - Step 4: Conclusion

3. UNCERTAINTY ASSESSMENT:
  - How confident are you in this assessment? (0.0 = very uncertain, 1.0 = very certain)
  - What factors contribute to uncertainty?
  - What additional information would improve confidence?

Format your response as:
REASONING: [detailed medical reasoning]
UNCERTAINTY_SCORE: [float between 0.0 and 1.0]

Fig. 4. This figure illustrates how the explanation node employs chain-of-thought (CoT) reasoning to enhance the interpretability of answers.

Prompt for Validation Node

You are a medical AI quality assurance specialist. Evaluate this complete medical image analysis pipeline output.

ORIGINAL QUERY: {query}

PIPELINE OUTPUTS:
1. VISUAL LOCALIZATION: {visual_box}
2. SPEECH SYNTHESIS QUALITY: {speech_quality}
3. SPEECH RECOGNITION RESULT: "{asr_text}"
4. MEDICAL REASONING: {explanation}
5. UNCERTAINTY SCORE: {uncertainty}

EVALUATION CRITERIA:
1. VISUAL LOCALIZATION QUALITY:
- Are the bounding box coordinates reasonable?
- Does the localization seem relevant to the query?

2. SPEECH PROCESSING QUALITY:
- Is the speech quality score acceptable (>0.7)?
- Does the ASR text match the original query?

3. MEDICAL REASONING QUALITY:
- Is the reasoning medically sound and detailed?
- Does it properly address the query?
- Is the explanation clear and comprehensive?

4. CONSISTENCY CHECK:
- Are all components internally consistent?
- Do the outputs align with each other?

5. UNCERTAINTY ASSESSMENT:
- Is the uncertainty score reasonable?
- High uncertainty (>0.7) may indicate quality issues

Provide a comprehensive assessment with:
- needs_review: boolean indicating if human review is required
- Quality scores (0.0-1.0) for each component
- Detailed critic notes explaining your assessment

Consider flagging for review if:
- Visual localization seems irrelevant or has poor coordinates
- Speech processing quality is below 0.7
- Medical reasoning is unclear, too brief, or medically unsound
- High uncertainty (>0.7) without clear justification
- Inconsistencies between pipeline components

Fig. 5. This figure illustrates how the Evaluate node utilizes LLM-as-a-judge for semantic evaluation.