

MasonNLP at MEDIQA-WV 2025: Multimodal Retrieval-Augmented Generation with Large Language Models for Medical VQA

A H M Rezaul Karim

George Mason University, VA, USA
akarim9@gmu.edu

Özlem Uzuner

George Mason University, VA, USA
ouzuner@gmu.edu

Abstract

Medical Visual Question Answering (MedVQA) enables natural language queries over medical images to support clinical decision-making and patient care. The MEDIQA-WV 2025 shared task addressed wound-care VQA, requiring systems to generate free-text responses and structured wound attributes from images and patient queries. We present the MasonNLP system, which employs a general-domain, instruction-tuned large language model with a retrieval-augmented generation (RAG) framework that incorporates textual and visual examples from in-domain data. This approach grounds outputs in clinically relevant exemplars, improving reasoning, schema adherence, and response quality across dBLEU, ROUGE, BERTScore, and LLM-based metrics. Our best-performing system ranked 3rd among 19 teams and 51 submissions with an average score of 41.37%, demonstrating that lightweight RAG with general-purpose LLMs—a minimal inference-time layer that adds a few relevant exemplars via simple indexing and fusion, with no extra training or complex re-ranking—provides a simple and effective baseline for multimodal clinical NLP tasks.¹

1 Introduction

Generating accurate answers to clinically relevant questions about medical images, known as Medical Visual Question Answering (MedVQA), requires integrating visual perception with domain-specific reasoning (Lin et al., 2023; Lau et al., 2018). Such systems can enhance diagnostics, support clinical training, and provide accessible, question-driven insights for clinicians and patients.

Compared to general VQA, MedVQA faces unique challenges, such as subtle anatomical or

pathological features that must be interpreted precisely, and questions often demanding specialized knowledge and logical inference (Lin et al., 2023; Liu et al., 2021). General VQA datasets lack this depth, motivating the creation of tailored medical benchmarks (Lin et al., 2023). Key resources include VQA-RAD for radiology (Lau et al., 2018), SLAKE with bilingual semantic annotations (Liu et al., 2021), and ImageCLEF’s VQA-Med series (Ben Abacha et al., 2019, 2021). PathVQA extends to pathology images (He et al., 2020b), PMC-VQA scales to over 227k Question Answer pairs for pretraining (Zhang et al., 2023), and MedFrameQA introduces multi-image reasoning for clinical scenarios (Yu et al., 2025). While these datasets drive progress, many methods still rely on resource-intensive fine-tuning and large domain corpora, limiting scalability.

Wound-care is a crucial MedVQA application, where image-based assessment guides treatment, monitors healing, and detects complications. Remote wound monitoring and telemedicine reduce costs, hospital visits, and infection risks (Sood et al., 2016; Chen et al., 2020), but variability in interpretation highlights the need for automated QA tools to support clinicians and empower patients.

The MEDIQA-WV shared task (Wound-care Visual Question Answering), part of ClinicalNLP 2025, addresses this challenge by generating free-text answers to patient-oriented wound-care questions using one or more images with annotations (Yim et al., 2025b). The shared task dataset includes bilingual (English/Chinese) queries, meta-data such as wound type and anatomic site, and systems are evaluated on fluency, relevance, and clinical accuracy.

We study an instruction-tuned general-domain LLM (Meta LLaMA-4 Scout 17B) (Meta, 2025) in a few-shot setup. It performs well on cases with small image details and short, generic question types, but degrades on images with sub-

¹Implementation can be found here: <https://github.com/AHMRezaul/MEDIQA-WV-2025>

tle or mixed findings, multi-part questions, and requests that require expert-level interpretation. To improve grounding and reasoning, we add a lightweight retrieval-augmented generation (RAG) (Lewis et al., 2020) layer by retrieving top-2 relevant text and image exemplars from the task corpus and appending them to the prompt. Since the dataset is not large enough for reliable fine-tuning and would add substantial compute and operational cost, a lightweight RAG setup was chosen.

Our contributions include:

- Demonstrating that a general-domain LLM with lightweight RAG can handle complex multimodal clinical tasks without domain-specific training.
- Showing that exemplar retrieval at inference improves reasoning and interpretability on clinical data.
- Providing a systematic analysis of how retrieval modality (text-only vs. multimodal) and prompting choices affect performance in medical visual question answering.

These results illustrate the promise of general-purpose LLMs, augmented with lightweight RAG, for transparent, flexible, and efficient solutions in clinical NLP and multimodal AI.

2 Related Work

Early VQA systems in both general and clinical domains relied on rule-based pipelines and small answer vocabularies, mapping hand-crafted cues or shallow features to fixed slots. These approaches lacked robustness to negation, uncertainty, and paraphrase (Malinowski et al., 2015). In the general domain, although VQA was framed as open-ended, many methods treated it as classification over restricted answer sets (Antol et al., 2015). Similar patterns appeared in early medical benchmarks, where evaluation emphasized exact match or lexical overlap, reinforcing closed-set, short-answer formats (Hasan et al., 2018; Ben Abacha et al., 2019, 2021). Such formulations constrained clinical expressivity and hindered nuanced responses.

With deep learning, convolutional image encoders combined with recurrent or simple text encoders became standard, later enhanced by attention (Talafha and Al-Ayyoub, 2018; Lin et al., 2023). In the general domain, bottom-up/top-down attention over regions (Anderson et al., 2017) and

modular co-attention (Yu et al., 2019) set strong baselines, influencing medical adaptations (Lin et al., 2023). New datasets supported this shift: VQA-RAD (Lau et al., 2018) introduced clinically authored radiology questions; SLAKE (Liu et al., 2021) added bilingual annotations with semantic labels; PathVQA (He et al., 2020b) scaled pathology QA with textbook images but faced noise and coverage issues; Medical-Diff-VQA (Hu et al., 2023) introduced difference-based paired-image questions for comparative reasoning.

Transformer-based vision–language pretraining further reshaped the field. ViLBERT (Lu et al., 2019) and LXMERT (Tan and Bansal, 2019) learned joint cross-modal representations and adapted effectively to VQA. In medicine, MM-BERT (Khare et al., 2021) showed multimodal BERT (Devlin et al., 2019) pretraining improves MedVQA under data scarcity, and M2I2 (Li et al., 2023c) leveraged self-supervised masked modeling and contrastive alignment to advance results across VQA-RAD, PathVQA, and SLAKE. Hybrids also emerged: BPI-MVQA (Liu et al., 2022) combined transformers with retrieval signals for improved multimodal fusion. These approaches improved accuracy but generally required domain-specific pretraining or fine-tuning.

Large vision–language models (VLMs) and LLM–vision hybrids enabled open-ended generation. BLIP-2 (Li et al., 2023b) efficiently bridged frozen encoders and LLMs. LLaVA (Liu et al., 2023) introduced visual instruction tuning, while LLaVA-Med (Li et al., 2023a) adapted this strategy to biomedical content. Domain-specific conversational VLMs such as XrayGPT (Thawakar et al., 2024) aligned MedCLIP (Wang et al., 2022) encoders with Vicuna (Chiang et al., 2023) for chest X-ray QA and summarization, and R-LLaVA (Chen et al., 2024) enhanced MedVQA via ROI annotations. Generative perspectives also gained traction: PMC-VQA scaled to 227k QA pairs, training MedVInT for effective fine-tuning on VQA-RAD, SLAKE, and ImageCLEF (Zhang et al., 2024). Evaluation evolved from strict accuracy toward BLEU and other text-generation metrics to capture partial correctness and phrasing variability (Ben Abacha et al., 2019, 2021; Hasan et al., 2018).

RAG (Lewis et al., 2020) has emerged to mitigate hallucinations and data scarcity by grounding answers in evidence. RAMM (Yuan et al., 2023) combined retrieval with dedicated attention modules to set state-of-the-art results on multiple

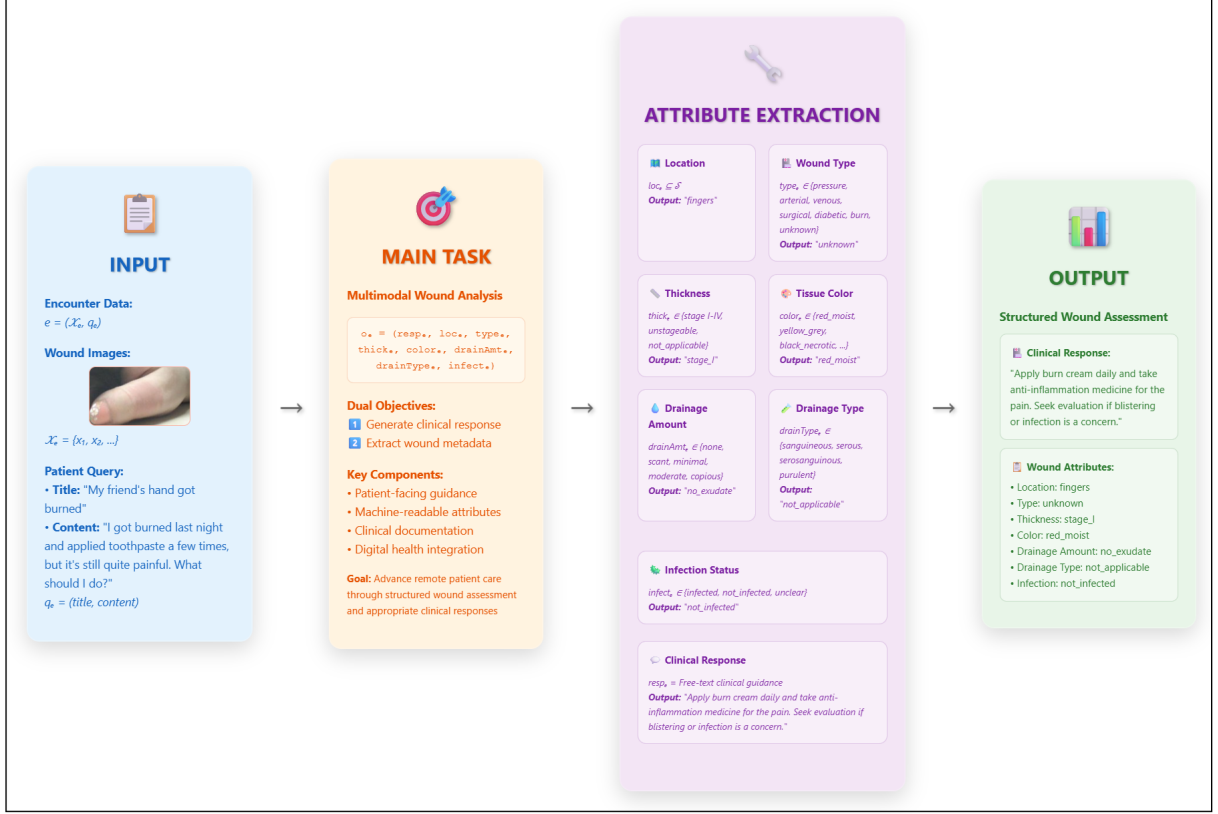


Figure 1: Task overview for MEDIQA-WV 2025. Inputs: wound images and a patient query. Outputs: free-text answer with structured wound attributes

MedVQA datasets. Fine-grained retrieval fusion with re-weighting further improved benchmarks like PathVQA and VQA-RAD without direct data access (Liang et al., 2025). Broader studies show retrieval strategies, granularity, and fusion strongly affect factuality, though best practices remain unsettled (Xiong et al., 2024).

Despite progress, challenges remain. Many systems rely on costly pretraining, curated corpora, or complex fusion stacks that limit transferability. Closed-set classification constrains answer diversity, while generative models risk hallucination if ungrounded. Our work addresses these issues with a general-domain, instruction-tuned LLM and lightweight RAG, which is a minimal, inference-time retrieval layer that adds a few relevant snippets via simple indexing and fusion, without extra training or complex re-ranking, to reduce hallucinations, respect data limits, and keep the system easy to reproduce. This approach of LLMs with RAG-based textual and visual exemplars preserves generative flexibility while improving interpretability and reproducibility by grounding answers in retrieved evidence, aligning with pragmatic, evidence-grounded MedVQA.

3 Task Description

The MEDIQA-WV shared task (Yim et al., 2025b) extends prior efforts in MedVQA to the wound-care domain. The objective is to advance remote patient care by generating clinically appropriate free-text responses to patient queries, while at the same time producing structured wound-related metadata that capture essential clinical details. This dual requirement reflects the need for both patient-facing guidance and machine-readable data that can be integrated into electronic health records (EHR).

Formally, each data instance corresponds to an *encounter* e , defined as a pair (\mathcal{X}_e, q_e) . The image set $\mathcal{X}_e = \{x_e^{(1)}, \dots, x_e^{(n)}\}$ contains one or more wound photographs, and the textual query q_e is bilingual, consisting of an English and a Chinese title and content.

The system must predict an output tuple with a response and the following metadata.

$$o_e = (resp_e, loc_e, type_e, thick_e, color_e, drainAmt_e, drainType_e, infect_e),$$

Where $resp_e$ is a free-text response and the remaining fields represent structured wound meta-

data. The anatomic location $loc_e \subseteq \mathcal{L}$ may include one or more sites (e.g., arm, chest, foot). The wound type $type_e \in \{pressure, arterial, venous, surgical, diabetic, \dots\}$ covers common etiologies. The wound thickness $thick_e \in \{stage\ I-IV, unstageable, not_applicable\}$. The tissue color $color_e$ is drawn from a finite set describing visual appearance (e.g., *red/moist, yellow/grey, black/necrotic*). Drainage is captured both in amount, $drainAmt_e \in \{none, scant, minimal, moderate, copious\}$, and in type, $drainType_e \in \{sanguineous, serous, serosanguinous, purulent\}$. Finally, the infection status $infect_e \in \{infected, not_infected, unclear\}$.

Training data provide full tuples o_e for each encounter, while in the test phase, only (\mathcal{X}_e, q_e) are given and systems must predict \hat{o}_e . Success in this task requires models to jointly reason over multi-modal inputs, differentiate clinically meaningful features, and generate outputs that are both fluent and structured for downstream clinical use.

4 Dataset

The MEDIQA-WV dataset (Yim et al., 2025a) was created to support wound assessment and patient counseling tasks. Each encounter consists of a unique identifier, one or more wound images, a bilingual query in English and Chinese, and a set of expert-generated responses in both languages. In addition to the free-text components, the training and validation splits contain structured gold-standard metadata covering the following attributes: wound_type, wound_thickness, tissue_color, drainage_amount, drainage_type, infection_status, and one or more anatomic_locations. All categorical values are drawn from a closed dictionary of medically valid terms, such as wound types $\{traumatic, surgical, pressure\}$, tissue colors $\{red\ moist, necrotic\ black\}$, drainage categories specifying both **amount** and **type**, and anatomic sites like *arm, knee, foot*. Figure 1 demonstrates an example data instance.

Split	Encounters	Responses	Images
Train	279	279	449
Validation	105	210	147
Test	93	279	152

Table 1: Dataset statistics: encounters, responses, and images per split.

4.1 Dataset Analysis

Table 1 summarizes the distribution of encounters, responses, and images across splits. The training set provides a single expert response per encounter, while validation is double-annotated, offering complementary perspectives. The test set is input-only and triple-annotated by medical professionals, though the gold-standard labels remain unpublished.

Encounters contain varying numbers of images, reflecting the clinical setting where multiple photos capture different wound angles or progress. In the training split, 170 encounters include a single image, while 109 (39%) contain multiple (up to nine) images. Validation includes 72 encounters with single images and 33 encounters with multiple images, and the test set has 55 single-image and 38 multiple-image encounters. Both the validation and test sets contain up to four images for a single encounter. Queries and responses also differ across splits. English queries average 46 words in training, 44 in validation, and 52 in test. Responses are 29 words on average for training, but become longer in validation (41 words) and test (47 words).

The metadata distribution is highly skewed. Traumatic wounds dominate with 330 cases (85.9%), while arterial and venous ulcers appear only once each (0.3%). Infection status is similarly imbalanced: 325 encounters (84.6%) are labeled as not infected, 39 as unclear (10.2%), and only 20 as infected (5.2%). Wound thickness is concentrated in stage I and stage II, and common anatomical sites include the lower leg, fingers, and hand. Although annotations generally follow the predefined dictionary, occasional inconsistencies appear, such as “*sole*” instead of “*foot-sole*” or drainage mismatches like “*no exudate*” paired with a specific drainage type. These rare cases highlight the need for normalization.

Overall, the dataset integrates structured wound metadata, bilingual queries, and expert responses into a challenging benchmark. The skewed label distributions and queries with multiple images, and the small size of training data, make fine-tuning difficult. These properties motivate using an LLM with RAG to retrieve similar examples from the training data, so answers stay close to the data, avoid generic responses, and follow the required output format.

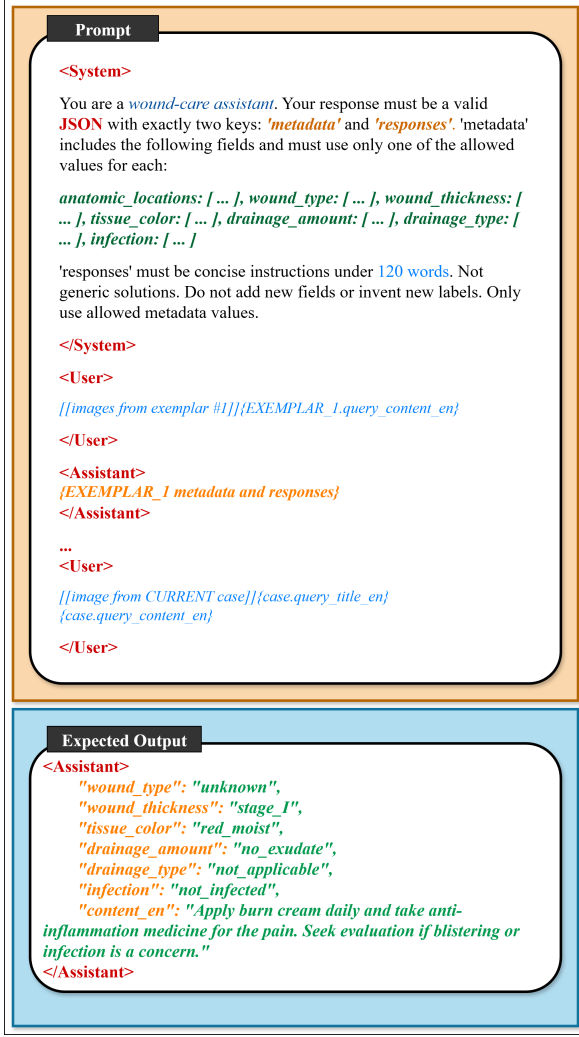


Figure 2: Structured prompt with retrieved exemplars and the expected output schema.

5 Methodology

To test how a general-domain LLM performs on a MedVQA task without domain-specific training, the *meta-llama/Llama-4-Scout-17B-16E-Instruct* (Meta, 2025) model was chosen. It follows instructions well, has open weights for reproducible research, and is a strong multimodal variant in the Meta-LLaMA (Touvron et al., 2023) family, offering a long context window and reliable vision-language support.

5.1 Model Configuration

We used the 17B instruction-tuned LLaMA-4 model, implemented via Hugging Face transformers with automatic GPU mapping. Inference ran in bfloat16 for efficiency, with a maximum generation length of 4096 tokens, temperature 0.2, and top-p 0.9. For multimodal inputs, the model was paired with the LLaMA-4

processor to jointly encode text prompts and wound images.

5.2 Prompt Design

We explored three prompting strategies: zero-shot, few-shot, and RAG. An example prompt is provided in Figure 2.

Zero-shot prompting. The model received only a system instruction defining its role as a wound-care assistant. Outputs were constrained to valid JSON by dividing the output tuple into two top-level keys: metadata and responses. Metadata used categorical labels from a wound-care data dictionary (e.g., wound type, tissue color, drainage, infection status), while responses provided short patient-facing instructions (≤ 120 words). This setting tested schema adherence without exemplars.

Few-shot prompting. We added two exemplar encounters from the training set, chosen after evaluating on the validation set, to reduce schema violations and improve metadata consistency. Each exemplar included wound image(s) and query text as a user turn, followed by the reference response as an assistant turn, guiding the model to emulate JSON structure and style. We limit exemplars to two because adding more, together with images, metadata, and the current prompt, exceeds the model’s context window.

Retrieval-augmented prompting. To improve grounding and reduce hallucinations (Lewis et al., 2020), we designed a multimodal RAG pipeline combining dense similarity search with exemplar-driven prompting, where we encoded questions and images into vectors, then retrieved the nearest training examples for that encounter and placed those exemplars in the prompt. Two indices were built with FAISS (Douze et al., 2024): semantic text embeddings from sentence-transformers/all-MiniLM-L6-v2 and vision-language embeddings from CLIP (openai/clip-vit-base-patch32)². We tested both the text-only and multimodal (text+image) retrieval setup.

At inference, we retrieve training encounters most similar to the inference-case using combined text and image similarity with equal weight ($\alpha = 0.5$). We evaluated other α values that placed more weight on images, but performance declined with more weight for the image, and so an approach

²sentence-transformers, openai-clip

with image-only retrieval was not explored. We select the top two exemplars because validation runs gave the best overall metrics, and adding more with images and metadata caused the prompt to exceed the model’s context window. This setup reduced schema violations, improved metadata predictions, and outperformed zero- / few-shot prompting.

5.3 Experimental Setup

Images were resized to 224×224 and passed with text. Decoding used nucleus sampling without beam search to balance diversity and format compliance. All runs were performed on NVIDIA A100 GPUs (80 GB), enabling full 17B model inference with multimodal inputs. We logged raw generations to audit both successful and erroneous outputs.

5.4 Post-processing

LLMs often generate extraneous text or malformed JSON, so we implemented a normalization pipeline. We first stripped any Markdown code fences or leading text before the opening brace, then parsed outputs to enforce exactly two keys: metadata and responses. Metadata entries were validated against the wound-care dictionary, discarding invalid fields. Responses were mapped to the English patient instruction. The cleaned output was merged into each case under its `encounter_id`, producing the final structured predictions for evaluation.

This layered design enabled systematic comparison of zero-shot, few-shot, and retrieval-augmented prompting, quantifying the benefits of contextual grounding and exemplar retrieval on schema adherence, metadata accuracy, and response validity.

6 Evaluation

The MEDIQA-WV 2025 shared task employs a multi-dimensional evaluation protocol that combines surface overlap, semantic similarity, and clinical plausibility.

For lexical similarity, the task uses `deltaBLEU` (Galley et al., 2015), which extends BLEU (Papineni et al., 2002) by rewarding partial matches across multiple references. Complementary recall-oriented measures include ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-Lsum (Lin, 2004), capturing different levels of n -gram and sequence overlap.

Semantic similarity is evaluated with BERTScore (Zhang et al., 2019), using two variants:

BERT-mn, which averages over references, and BERT-mx, which takes the maximum score to reward alignment with at least one gold annotation. English responses are scored with `microsoft/deberta-xlarge-mnli` (He et al., 2020a), while Chinese responses are scored with `lang=zh` for multilingual alignment.

To assess plausibility and instructional quality beyond surface metrics, three large multimodal language models (LMLMs) act as automatic judges: (i) DeepSeek-V3-0324 (Azure AI Foundry), (ii) Gemini-1.5-pro-002 (Google GenAI), and (iii) GPT-4o (Azure AI Foundry)³. Using standardized prompts in English and Chinese, these models independently score outputs for usefulness, contextuality, and clinical appropriateness, reducing model-specific bias.

A final `average_score` (Avg) aggregates results across all metrics, combining fidelity, semantic alignment, and plausibility into a robust benchmark for multimodal clinical generation systems.

7 Results and Discussion

7.1 Leaderboard Performance

The MEDIQA-WV 2025 shared task attracted participation from **19** teams, producing a total of **51** submissions. Our **MasonNLP** system ranked competitively, achieving an average score of **41.37%** on its best run. As shown in Table 2, both of our submissions placed in the top five overall, underscoring the robustness of our general-domain LLM pipeline against more specialized approaches. Notably, while the leading system achieved the highest overall performance (47.30%), our systems demonstrated comparable strength across multiple metrics, reflecting effective phrasing and semantic alignment. This suggests that our lightweight retrieval and prompting strategies can yield results close to top-level systems.

7.2 Ablation Study

To better understand the contribution of the retrieval and prompting strategy, we conducted an ablation across four configurations: (1) LLaMA-4 + RAG with *image+text* retrieval, (2) LLaMA-4 + RAG with *text-only* retrieval, (3) LLaMA-4 in *few-shot*, and (4) LLaMA-4 in *zero-shot*. Results in Table 3 demonstrate three key effects. First, retrieval markedly improves all evaluation metrics, confirming its role in grounding predictions. Second,

³DeepSeek, Gemini-1.5-pro, GPT-4o

Team	dBLEU	R1	R2	RL	RLsum	BERT-mn	BERT-mx	DeepSeekV3	Gemini	GPT-4o	Avg
MasonNLP	8.89	70.99	48.62	42.19	42.27	59.01	63.27	53.55	55.38	55.38	41.37
MasonNLP	7.31	72.79	48.44	43.31	43.25	60.42	64.55	58.92	56.45	53.23	41.07
EXL Services–Health	9.92	<i>79.09</i>	<i>56.13</i>	<i>45.61</i>	45.60	62.18	66.90	68.23	<i>64.52</i>	<i>71.51</i>	<i>47.30</i>
EXL Services–Health	<i>13.04</i>	71.18	51.28	45.17	45.72	61.88	<i>67.43</i>	63.49	59.14	62.90	45.75
DermaVQA	7.65	78.99	53.91	45.49	45.48	60.62	63.68	42.74	45.70	37.10	37.71

Table 2: Leaderboard results on MEDIQA-WV 2025. MasonNLP best runs in bold; best per column in italics.

System	dBLEU	R1	R2	RL	RLsum	BERT-mn	BERT-mx	DeepSeekV3	Gemini	GPT-4o	Avg
LLaMA-4 + RAG (image+text)	8.89	70.99	48.62	42.19	42.27	59.01	63.27	53.55	55.38	55.38	41.37
LLaMA-4 + RAG (text only)	7.31	72.79	48.44	43.31	43.25	60.42	64.55	58.92	56.45	53.23	41.07
LLaMA-4 (few-shot)	4.67	41.50	27.30	23.50	24.10	41.60	44.20	35.00	33.90	33.90	23.63
LLaMA-4 (zero-shot)	1.73	25.00	17.00	14.00	14.50	29.00	30.00	20.00	21.60	21.60	14.10

Table 3: Ablation of prompting and retrieval strategies. Best per column in bold.

the inclusion of images supplied visual evidence for image-dependent details, as shown by higher dBLEU and GPT-4o scores. Third, even without retrieval, moving from zero-shot to few-shot reduces hallucinations and yields more consistent phrasing, though the gap to retrieval-based models remains large. Together, these trends highlight that retrieval complements prompting and that multimodal retrieval is particularly effective for wound-specific guidance. This systematic progression from zero-shot to multimodal RAG reveals clear patterns in how different retrieval modalities and prompting approaches affect MedVQA performance.

7.3 Discussion and Implications

Our results show a clear progression in performance from zero-shot prompting to multimodal RAG. In the *zero-shot* setting with the **LLaMA-4 17B** model, scores were very low (dBLEU 1.73), largely due to the model’s failure to produce the required structured JSON output despite explicit instructions.

Adding a few in-context exemplars improved formatting and raised dBLEU to 4.67, but responses remained generic and lacked clinically specific detail. Retrieval with *textual exemplars* addressed this issue more effectively. By grounding outputs in semantically similar queries and solutions, the model produced more structured and concrete recommendations, with Rouge-L increasing from 23.50 (few-shot) to 43.31, and GPT-4o judgments rising substantially.

Extending retrieval to include *images* further boosted contextual grounding, particularly for wound-site descriptions and infection cues, lifting dBLEU to 8.89. However, gains were not universal. Visual neighbors sometimes introduced noise when image relevance was weak, slightly trailing

text-only retrieval in a few metrics.

Overall, the ablation confirms that moving from zero-shot to exemplar-based and multimodal retrieval progressively improves structure and specificity. A lightweight RAG pipeline combining textual and visual evidence provides a strong, reproducible baseline for multimodal clinical tasks without domain-specific fine-tuning.

8 Error Analysis

In the absence of gold-standard labels, we evaluate model behavior along four axes: (i) schema conformance against an allowed-value dictionary, (ii) content form and genericness (length, template reuse, lexical alignment to the query), (iii) intent coverage for common asks (healing time, stitches/sutures, tetanus), and (iv) hallucination/ over-claim heuristics (e.g., asserting infection without cues).

8.1 Zero-shot LLaMA-4

On 93 queries, the model produced 93 answers with one empty reply (1.1%). Answers are short (mean 18.1 words with max 53) and frequently reuse stock advice, like “cover with a bandage” (25/93), “monitor for signs of infection” (23/93), “apply antibiotic ointment” (22/93), with additional phrases such as “seek medical attention” (9/93), “consult a doctor” (6/93), and “keep the area clean and dry” (5/93). Although 90 outputs are unique (only two duplicates and one missing), query–answer lexical overlap is low, indicating a generic style that often under-engages the user’s ask. Intent coverage lacks precision as well. For healing-time questions, only 1/16 answers include a numeric time frame; for stitches/sutures, 4/13 mention suture care or removal timing; for tetanus, 4/7 mention vaccination/ booster guidance. Hallucination screening flags 31/93 answers that assert infection without any infection

Improvement Type	Zeroshot Prediction	RAG Prediction
Hallucination Reduction	<i>Infection: infected</i> Instruction: “Antibiotics may be needed.”	<i>Infection: not_infected</i> Instruction: “No signs of infection; continue saline cleaning and dry dressing.”
Specificity of Response	<i>Location: finger</i> Instruction: “Keep the area clean and avoid movement.”	<i>Location: fingertip</i> Instruction: “Clean fingertip wound twice daily, apply antibiotic ointment, and avoid immersion in water.”
Vocabulary Normalization	<i>Type: trauma</i> Instruction: “Healing depends on care.”	<i>Type: traumatic</i> Instruction: “Traumatic wound; healing time approx. 2–3 weeks with proper care.”

Table 4: Examples of improvements from zero-shot to RAG, grouped by improvement type.

cues in the corresponding queries; about a quarter of these are hedged (e.g., “*may be infected*”), and explicit speculative diagnosis terms (e.g., fracture, necrosis) are rare (4/93). Overall, zero-shot outputs are fluent and safety-oriented but frequently generic, under-answer explicit asks, and sometimes over-call infection in the absence of evidence.

8.2 LLAMA-4 + RAG (Image+Text)

We examined 93 predictions for schema conformance, value validity, and content quality. All seven fields were present for every item. True out-of-vocabulary (OOV) rates were low as *anatomic_locations* had 8 OOV entries driven by common synonyms (*leg*, *finger/fingertip*, *shin*), while single-valued fields each had at most one OOV instance (*wound_type* 1/93; *wound_thickness* 4/93 due to *partial/partial thickness*; *tissue_color*, *drainage_amount*, *drainage_type*, *infection* each 1/93). Label distributions reflected the training and development set analysis with *wound_type* mostly being *traumatic* (88.0%), *infection* favoring *not_infected* (52.2%) with mass on *infected* (27.2%) and *unclear* (20.7%), and *wound_thickness* was dominated by *stage_II* (50.6%). There was exactly one instance with no generated response. Responses were longer than the zero-shot system (mean 28.4 words with a max of 96) and remained largely unique (91/93) but still exhibited a generic tone. About 60% of answers had very low lexical overlap with their queries, and common advice tokens were frequent (e.g., “antibiotic” in 45.2%; “debridement” in 5.4%). Intent coverage improved but remained uneven. 7/44 (15.9%) healing-time questions received a concrete range; 4/13 (31%) stitches/ sutures were addressed; 4/7 (57%) tetanus was handled. Hallucination risk was limited (6/93, 6.5% infection assertions without cues), and safety-related replies were appropriately

cautious, though consistent crisis templates would be beneficial.

8.3 Observed Improvements from Zero-shot to RAG

Relative to zero-shot, RAG reduces over-assertion of infection substantially (31/93 → 6/93) and produces longer, more informative answers that better reflect the query context, particularly for time-to-heal questions (a larger share of timeline-bearing replies). RAG outputs also conform to a schema with low OOV rates, eliminating synonym-induced errors through canonicalization. Nonetheless, both systems retain some generic phrasing and leave room for stronger intent coverage on stitches and return-to-activity guidance. Taken together, RAG shifts the model from broadly safe, generic counseling toward more specific, schema-consistent, and less hallucinatory answers, as also reflected in the examples presented in Table 4.

9 Conclusion

We investigated wound-care VQA in the MEDIQA-WV 2025 shared task using a general-domain, instruction-tuned LLM combined with lightweight RAG. Our study shows that this approach can handle challenging multimodal questions without domain-specific training. The framework integrates textual and visual neighbors at inference time and is simple to reproduce. Results demonstrate clear gains from zero-shot to exemplar-driven prompting, with multimodal retrieval being the best-performing system. Error analysis confirmed that retrieval reduces hallucinations and improves metadata consistency, though challenges remain when neighbors are only partially relevant. Overall, our findings highlight retrieval-augmented generation as a transparent, efficient, and generalizable approach for advancing multimodal clinical NLP.

Limitations

Our generation is closely tied to the in-domain training data used for retrieval, so outputs can mirror its gaps and biases. Higher-quality and more diverse exemplars would likely yield more specific and reliable responses. Incorporating external knowledge (e.g., vetted clinical guidelines or curated web corpora) could broaden coverage and reduce omissions.

References

- Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2017. Bottom-up and top-down attention for image captioning and vqa. *arXiv preprint arXiv:1707.07998*, 2(4):8.
- Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.
- Asma Ben Abacha, Sadid A Hasan, Vivek V Datla, Dina Demner-Fushman, and Henning Müller. 2019. Vqa-med: Overview of the medical visual question answering task at imageclef 2019. In *Proceedings of CLEF (Conference and Labs of the Evaluation Forum) 2019 Working Notes*. 9-12 September 2019.
- Asma Ben Abacha, Mourad Sarrouiti, Dina Demner-Fushman, Sadid A Hasan, and Henning Müller. 2021. Overview of the vqa-med task at imageclef 2021: Visual question answering and generation in the medical domain. In *Proceedings of the CLEF 2021 Conference and Labs of the Evaluation Forum-working notes*. 21-24 September 2021.
- Lihong Chen, Lihui Cheng, Wei Gao, Dawei Chen, Chun Wang, and Xingwu Ran. 2020. Telemedicine in chronic wound management: systematic review and meta-analysis. *JMIR mHealth and uHealth*, 8(6):e15574.
- Xupeng Chen, Zhixin Lai, Kangrui Ruan, Shichu Chen, Jiayang Liu, and Zuozhu Liu. 2024. R-llava: Improving med-vqa understanding through visual region of interest. *arXiv preprint arXiv:2410.20327*.
- Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. <https://lmsys.org/blog/2023-03-30-vicuna/>. LMSYS Org Blog.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.
- Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library. *arXiv preprint arXiv:2401.08281*.
- Michel Galley, Chris Brockett, Alessandro Sordoni, Yangfeng Ji, Michael Auli, Chris Quirk, Margaret Mitchell, Jianfeng Gao, and Bill Dolan. 2015. deltableu: A discriminative metric for generation tasks with intrinsically diverse targets. *arXiv preprint arXiv:1506.06863*.
- Sadid A Hasan, Yuan Ling, Oladimeji Farri, Joey Liu, Henning Müller, and Matthew Lungren. 2018. Overview of imageclef 2018 medical domain visual question answering task. *Proceedings of CLEF 2018 Working Notes*.
- Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020a. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.
- Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. 2020b. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*.
- Xinyue Hu, L Gu, Q An, M Zhang, L Liu, K Kobayashi, T Harada, R Summers, and Y Zhu. 2023. Medical-diff-vqa: a large-scale medical dataset for difference visual question answering on chest x-ray images. *PhysioNet*, 12:13.
- Yash Khare, Viraj Bagal, Minesh Mathew, Adithi Devi, U Deva Priyakumar, and CV Jawahar. 2021. Mmbert: Multimodal bert pretraining for improved medical vqa. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, pages 1033–1036. IEEE.
- Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. 2018. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.
- Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023a. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564.

- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.
- Pengfei Li, Gang Liu, Lin Tan, Jinying Liao, and Shenjun Zhong. 2023c. Self-supervised vision-language pretraining for medical visual question answering. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE.
- Xiao Liang, Di Wang, Bin Jing, Zhicheng Jiao, Ronghan Li, Ruyi Liu, Qiguang Miao, and Quan Wang. 2025. Fine-grained knowledge fusion for retrieval-augmented medical visual question answering. *Information Fusion*, 120:103059.
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.
- Zhihong Lin, Donghao Zhang, Qingyi Tao, Danli Shi, Gholamreza Haffari, Qi Wu, Mingguang He, and Zongyuan Ge. 2023. Medical visual question answering: A survey. *Artificial Intelligence in Medicine*, 143:102611.
- Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. 2021. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, pages 1650–1654. IEEE.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.
- Shengyan Liu, Xuejie Zhang, Xiaobing Zhou, and Jian Yang. 2022. Bpi-mvqa: a bi-branch model for medical visual question answering. *BMC Medical Imaging*, 22(1):79.
- Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.
- Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz. 2015. Ask your neurons: A neural-based approach to answering questions about images. In *Proceedings of the IEEE international conference on computer vision*, pages 1–9.
- AI Meta. 2025. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation. <https://ai.meta.com/blog/llama-4-multimodal-intelligence/>, checked on, 4(7):2025.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.
- Aditya Sood, Mark S Granick, Chloé Triel, Julie Lano, Sylvie Palmier, Evelyne Ribal, and Luc Téot. 2016. The role of telemedicine in wound care: a review and analysis of a database of 5,795 patients from a mobile wound-healing center in languedoc-roussillon, france. *Plastic and reconstructive surgery*, 138(3S):248S–256S.
- Bashar Talafha and Mahmoud Al-Ayyoub. 2018. Just at vqa-med: A vgg-seq2seq model. In *CLEF (working notes)*.
- Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.
- Omkar Chakradhar Thawakar, Abdelrahman M Shaker, Sahal Shaji Mullappilly, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, Jorma Laaksonen, and Fahad Khan. 2024. Xraygpt: Chest radiographs summarization using large medical vision-language models. In *Proceedings of the 23rd workshop on biomedical natural language processing*, pages 440–448.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jiemeng Sun. 2022. Medclip: Contrastive learning from unpaired medical images and text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2022, page 3876.
- Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. [Benchmarking retrieval-augmented generation for medicine](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6233–6251, Bangkok, Thailand. Association for Computational Linguistics.
- Wen-wai Yim, Asma Ben Abacha, Robert Doerning, Chia-Yu Chen, Jiaying Xu, Anita Subbarao, Zixuan Yu, Fei Xia, M Kennedy Hall, and Meliha Yetisgen. 2025a. Woundcarevqa: A multilingual visual question answering benchmark dataset for wound care. *Journal of Biomedical Informatics*.
- Wen-wai Yim, Asma Ben Abacha, Meliha Yetisgen, and Fei Xia. 2025b. Overview of the mediqa-wv 2025 shared task on wound care visual question answering. In *Proceedings of the 7th Clinical Natural Language Processing Workshop*. Association for Computational Linguistics.
- Suhao Yu, Haojin Wang, Juncheng Wu, Cihang Xie, and Yuyin Zhou. 2025. Medframeqa: A multi-image medical vqa benchmark for clinical reasoning. *arXiv preprint arXiv:2505.16964*.

- Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. 2019. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6281–6290.
- Zheng Yuan, Qiao Jin, Chuanqi Tan, Zhengyun Zhao, Hongyi Yuan, Fei Huang, and Songfang Huang. 2023. Ramm: Retrieval-augmented biomedical visual question answering with multi-modal pre-training. In *Proceedings of the 31st ACM international conference on multimedia*, pages 547–556.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.
- Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*.
- Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024. Development of a large-scale medical visual question-answering dataset. *Communications Medicine*, 4(1):277.