SKINspectoR Technical Report

Rezky Mulia Kam[1], Aflaha Fathinah Fatahillah[2], Rafi Hazel Tafara[3], Felix Young[4], Marcellino Asanuddin[5]

School of Computer Science, Universitas Bina Nusantara

Abstract

Indonesia's healthcare system faces challenges in the equitable distribution of specialist personnel such as dermatologists, with the majority of specialists concentrated in urban hospitals. National projections indicate a sustained shortage in the overall number of dermatology specialists, which implies even greater scarcity in remote regions. Primary healthcare services in Indonesia's remote islands have experimented with teledermatology models due to limited access to dermatologists in person. These services demonstrate favorable response times and improved local clinical confidence, yet still reflect the constraints of severely limited specialist supply. To address these challenges, SKINspectoR offers a solution based on a combination of Computer Vision and Linguistics utilizing a pretrained Vision-Language Model (VLM), specifically Qwen2.5VL 3B, which is fine-tuned using the DoRA (Parameter Efficient Fine Tuning) technique on the multimodal SkinCAP dataset. This system is designed to provide dermatologist-level diagnosis with a focus on edge deployment that enables flexible software operation in environments without access to cloud computing (Embedded AI), thereby reaching remote areas with limited infrastructure.

Source Code: https://github.com/RezkyKam50/SKINspector.git
*Model Weights:*https://huggingface.co/azzenn4/Qwen2.5-3B-SkinCAP-DoRA

1. Introduction

SKINspectoR offers software based on a combination of Computer Vision & Linguistics using a pretrained VLM (Vision-Language Model), specifically Qwen2.5VL 3B, which is fine-tuned using the DoRA (Parameter Efficient Fine Tuning) technique on the multimodal SkinCAP dataset to provide diagnosis in the manner of a Dermatology expert, with a focus on edge deployment that enables flexible software operation in environments without access to cloud computing (Embedded AI).

2. Related Work

Our work builds upon several lines of research at the intersection of computer vision and natural language diagnosis that gained prominence with *(Esteva et al.)*, who demonstrated CNN performance comparable to board-certified dermatologists in skin cancer classification.

[1]*2702260773*
[2]*2702272350*
[3]*2702347263*
[4]*2702274513*
[5]*2702242650*

Recent advances in Vision-Language Models (VLMs) have enabled more nuanced multimodal reasoning in medical imaging. Qwen2.5-VL *(Bai et al.)* provides a state-of-the-art architecture that unifies visual and textual tokens within a single attention mechanism, supporting high-resolution image understanding and spatial reasoning capabilities critical for dermatological image analysis.

In the domain of medical multimodal datasets, SkinCaRe (Shen et al., 2024) offers a curated collection of dermatological image-caption pairs annotated by dermatologists, emphasizing descriptive clinical observation over rigid diagnostic labels. This dataset supports vision-language pretraining and fine-tuning for dermatology-specific tasks.

To address the challenges of efficient model adaptation, Parameter-Efficient Fine-Tuning (PEFT) methods such as LoRA (Low-Rank Adaptation) and its enhanced variant DoRA *(Liu et al., 2024)* has been developed. DoRA decomposes weight updates into magnitude and direction components, achieving performance closer to full fine-tuning while maintaining efficiency—a key consideration for medical applications where accuracy is paramount.

Further optimization for resource-constrained environments is enabled by QLoRA (Dettmers et al., 2023), which combines 4-bit quantization with LoRA adapters to enable fine-tuning of large models on consumer-grade hardware. Complementary to this, LigerKernel *(Hsu et al.)* provides optimized GPU kernels that reduce memory usage and improve training throughput, facilitating efficient VLM training on limited hardware.

For knowledge-enhanced generation, Retrieval-Augmented Generation (RAG) (Lewis et al., 2021) has been adopted to ground model predictions in external medical knowledge. Hybrid retrieval systems combining lexical (BM25) and semantic (PubMedBERT) search improve the relevance and accuracy of retrieved clinical context, addressing limitations in VLM parametric knowledge.

Finally, edge deployment frameworks such as LLaMA.cpp with GGML quantization enable efficient inference on devices without cloud connectivity, supporting teledermatology applications in low-resource settings. Together, these works provide the technical foundation for SKINspectoR's development as an embedded AI system for dermatological diagnosis in remote regions.

3. Dataset

SkinCAP is a multimodal dermatology dataset *(Shen et al.)* designed to support vision–language learning in the medical domain. The dataset provides paired visual and textual data, enabling models to learn the relationship between dermatological images and clinically accurate textual descriptions.

SkinCAP consists of approximately 4,000 image–caption pairs, where each caption is authored by board-certified dermatologists and passes through a multi-stage quality control process to ensure both clinical accuracy and linguistic consistency.

Composition	Explanation
Visual Modality	Dermatological skin images, including diverse skin conditions, lesions, and visual patterns.
Textual Modality	Medical captions that describe visual appearance, lesion characteristics, and dermatological context.

3.1 Dataset Composition

The SkinCAP dataset is composed of two primary modalities:

- Visual Modality
  Dermatological skin images covering diverse skin conditions, including lesions, variations in skin appearance, and visual patterns relevant to clinical observation.

- Textual Modality
  Medical captions that describe visual appearance, lesion characteristics, and dermatological context without enforcing a single definitive diagnosis.

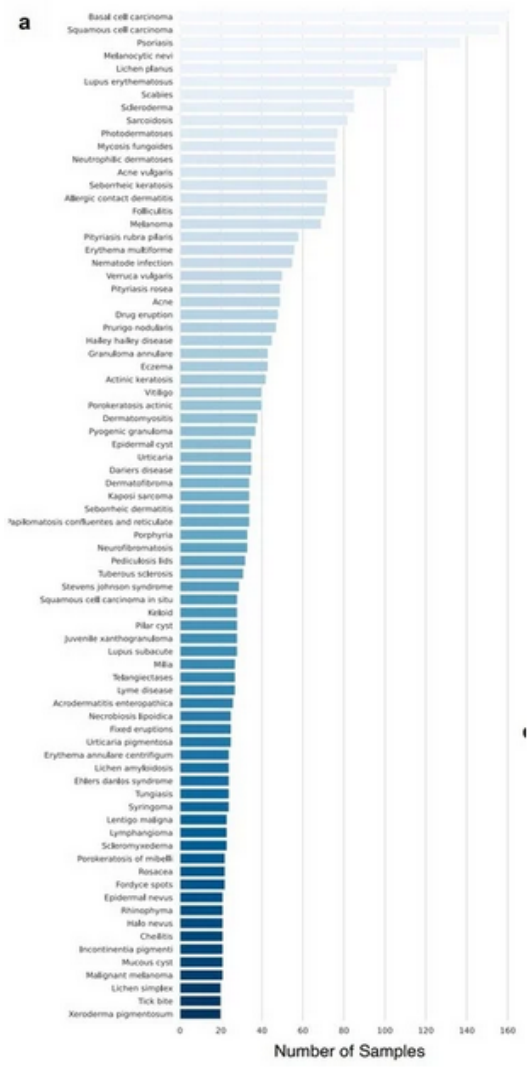


Figure a shows the **distribution of skin disease samples** in the SkinCAP dataset (sample size ≥ 20 per class), highlighting its **broad coverage across multiple disease categories.**

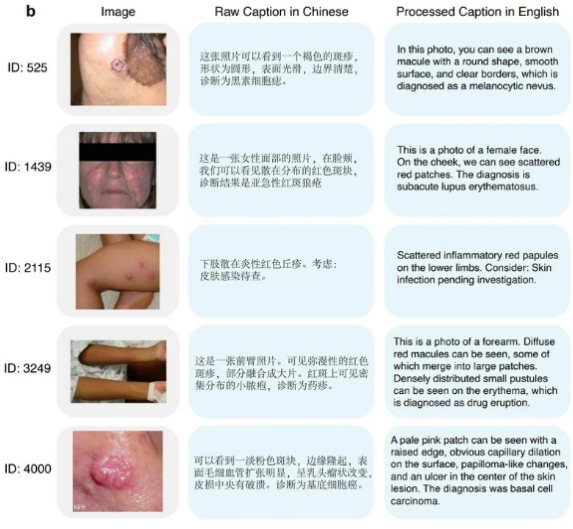


Figure b shows **representative examples of image–caption pairs** from the SkinCAP dataset. Each example consists of a clinical dermatology image and dermatologist-written captions (the original medical caption annotated by dermatologists in Chinese, and the corresponding processed English caption). The captions **emphasize observable clinical features**, such as lesion morphology, color, and anatomical location, illustrating the **observation-first annotation paradigm adopted in SkinCAP.**

Source: Adapted from Shen et al., "SkinCaRe: A Multimodal

Dermatology Dataset Annotated with Medical Caption and Chain-of-Thought Reasoning," 2024.

3.2 Purpose of the Dataset

The SkinCAP dataset is designed to:

- Emphasize descriptive clinical observation rather than rigid diagnostic classification.
- Provide supervised learning signals for vision–language models.
- Serve as a resource for model adaptation, fine-tuning, and system development in dermatological computer vision tasks.

3.3 Dataset Usage Pipeline

The SkinCAP dataset is utilized through an offline training and evaluation workflow, as illustrated below:

1. SkinCAP Dataset
   Used as input data for model preparation.
2. Model Preparation
   The dataset supports both pretraining and fine-tuning of vision–language models to improve dermatological understanding.
3. Offline Evaluation
   Behavioral evaluation is conducted to validate whether the model correctly understands dermatological images and descriptions.
4. Deployed Model (Inference-only)
   Only models that pass offline validation are deployed for real-world inference.

3.4 Deployment Constraints

- The SkinCAP dataset is only used offline during training and evaluation.
- The dataset is not accessed during inference or deployment.
- It does not function as a runtime knowledge base.
- Offline evaluation serves as a validation gate, ensuring only validated models are deployed.

4. Architecture

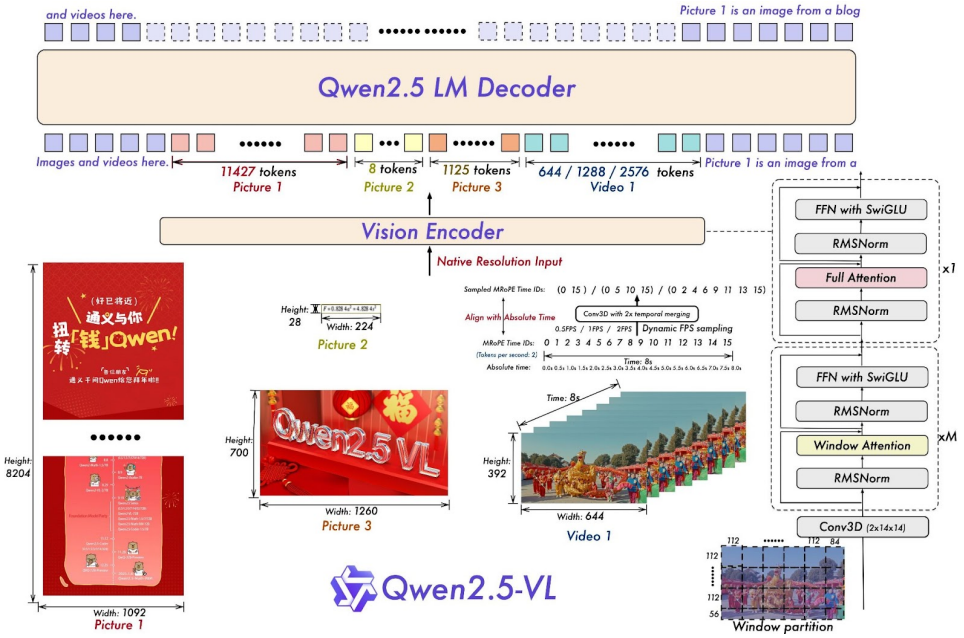


Fig 3.1: Qwen2.5 Architecture (Shuai et al.)

The model fuses Vision Tokens and Text Tokens within the same Attention Layer for a Unified Architecture *(Shuai et al.).* A projection layer maps dynamic visual embeddinginto the same dimensional space as text tokens and images go through a Vision Transformer (ViT-style) backbone *(Dosovitskiy et al.)*.

- The image is split into patches *(eg. 224x224 for 16x16 patches , 224/16 = 14 , 14x14 = 196 patches = 1 Vision Tokens)* - each patch becomes a visual token.
- Position information is added for each Vision Tokens so the model knows left from right.

Encodes spatial information explicitly, which enables:

- Bounding box reasoning.
- Referring expressions like “the button on the top right” which is important for strategic medical reasoning.

This encoder supports high-resolution and variable-resolution images which is very flexible, though higher HxW = Higher compute power requirement.

5. Training

The training is designed to perform **behavioral adaptation** of a pretrained multimodal model for dermatological explanation tasks while maintaining computational efficiency and deployment stability. It does not train a model from scratch, it uses a **pretrained vision–language model (Qwen2.5-VL)** *(Shuai et al.)* that already encodes general visual and linguistic knowledge.

Once deployed, the system operates strictly in inference-only mode, ensuring reproducibility, stability, and privacy preservation.

5.1 Training Objective

- Align visual understanding of skin images with medical textual descriptions.
- Improve image-to-text reasoning for dermatology-related cases.
- Resemble dermatologist-style explanations, use appropriate medical terminology.
- Avoid overly confident or speculative language.

5.2 Training Strategy

Base model: Qwen2.5-VL (Vision-Language Transformer).
Adaptation method:

- Parameter-Efficient Fine-Tuning (PEFT).
- QDoRA / QLoRA-style adapters.

Training Optimization:

- LigerKernel (Efficient CUDA Kernels for VLM Training)

5.3 Weight-Decomposed Low-Rank Adaptation

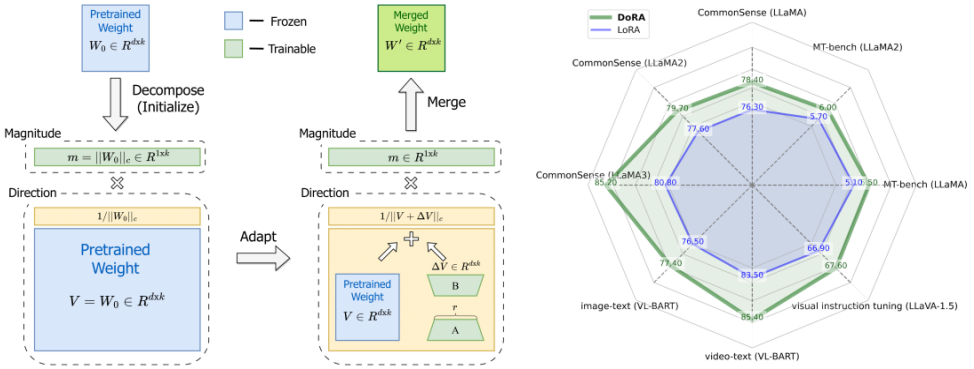


Fig 4.1: DoRA Mechanism (Liu et al.)

DoRA (Weight-Decomposed Low-Rank Adaptation) *(Liu et al.)* a part of Parameter-Efficient Fine Tuning (PEFT) method, enhances LoRA by decomposing pre-trained weights into two components: magnitude and direction. While standard LoRA applies low-rank updates directly to weights $W' = W_0 + BA$, DoRA separates this process by keeping the magnitude trainable as a vector while using LoRA specifically for directional updates:

$$W' = m\frac{(W_0 + BA)}{||W_0 + BA||}.$$

The key insight from the paper *(Liu et al.)* is that full fine-tuning exhibits a negative correlation between magnitude and direction changes, meaning substantial changes in one dimension require only minimal changes in the other—a pattern LoRA fails to replicate due to its proportional updates. DoRA overcomes this by decoupling these adaptations, achieving learning behavior closer to full fine-tuning. For medical applications like reasoning or image captioning, DoRA's superior performance is particularly valuable: it consistently outperforms LoRA by 2-4% across language tasks and nearly 1% on multimodal vision-language tasks like VL-BART, while maintaining the same inference efficiency and requiring similar or even fewer trainable parameters. This makes DoRA especially suitable for medical domains where accuracy improvements are critical, training data may be limited, and the nuanced understanding required for diagnostic reasoning or precise medical image description benefits from DoRA's more sophisticated weight adaptation strategy.

5.4 LigerKernel

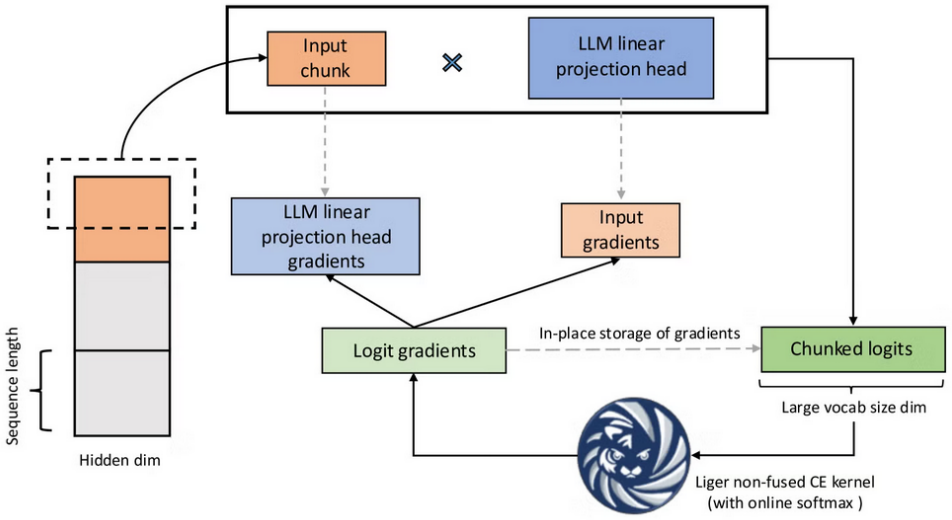


Fig 4.2: LigerKernel (Linkedin, PyTorch, Meta)

Liger Kernel *(Hsu et al.)* is an open-source library of optimized Triton kernels designed to significantly improve Large Language Model training efficiency through kernel operation fusing and input chunking techniques. Unlike standard PyTorch's eager mode execution, which executes operations step-by-step with overhead from function calls, dispatching, and kernel launches while materializing all intermediate activations in memory, Liger Kernel fuses multiple operations into single GPU kernels that minimize data movement between high-bandwidth memory (HBM) and faster on-chip SRAM. This approach achieves approximately 20% higher training throughput and 60% reduction in GPU memory usage compared to HuggingFace implementations by avoiding redundant memory copies and reducing the materialization of large intermediate tensors. For resource-constrained hardware, Liger Kernel is particularly valuable as it enables training with larger batch sizes, longer sequences, or on smaller GPUs that would otherwise run out of memory—for example, the Fused Linear Cross Entropy kernel uses chunked computation to prevent materializing massive logit tensors (which can be 16.8 GB for models with 256k vocabulary), making previously infeasible training configurations possible on limited hardware while maintaining numerical correctness and model convergence.

For our Qwen2.5VL, we patch the kernels for optimizing the Rotary-Positional Embedding (RoPE), fusing Linear Cross Entropy, RMS Norm and Swi-Glu which results in 8.2GB VRAM usage down to 6.2GB VRAM usage. Paired with DoRA + LigerKernel we achieved 40% faster loss convergence (until step 300) and higher GPU utilization (from 60% to 100%) with batch size sets to 4 both for Training and Evaluation.

5.5 Quantized-LoRA (QLoRA/QDoRA)

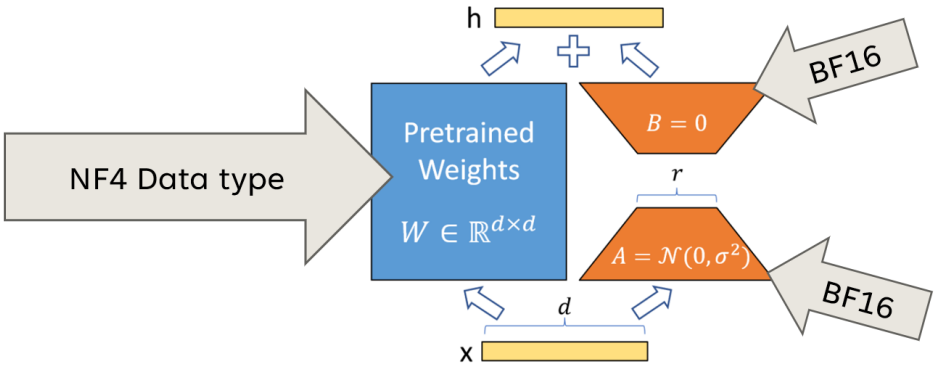


Fig 4.3: QLoRA (NVIDIA)

QLORA *(Dettmers et al.)* is an efficient finetuning method that enables training of very large language models (up to 65B parameters) on a single consumer GPU by combining 4-bit quantization with Low-Rank Adapters (LoRA). While standard LoRA reduces trainable parameters by adding small adapter weights to a frozen 16-bit model, QLORA goes further by quantizing the base model to 4-bit precision using a novel NormalFloat4 (NF4) data type that's optimized for normally distributed weights, along with Double Quantization to compress quantization constants and Paged Optimizers to handle memory spikes. The key innovation is that QLORA stores weights in 4-bit but dequantizes them to 16-bit (BFloat16) only during forward and backward passes, computing gradients solely for the LoRA adapters rather than the base model. This approach reduces memory requirements dramatically—from over 780GB to under 48GB for a 65B model—while maintaining full 16-bit finetuning performance, making state-of-the-art model training accessible on consumer hardware like a single 48GB GPU instead of requiring expensive multi-GPU server setups. This principle applies the same to QDoRA since it fundamentally relies upon LoRA.

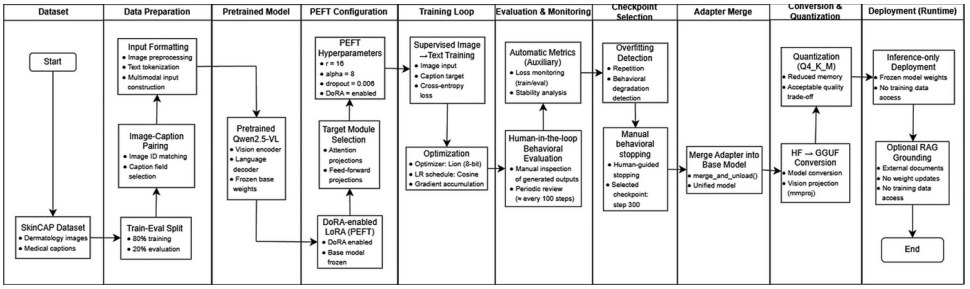
Combining these three, we're able to optimize both the ViT and the LLM counterpart of Qwen2.5VL-3B that enables us to fine-tune both model simultaneously on NVIDIA 40-series Ada Lovelace GPU architecture (8GB VRAM) which breaks the naive assumption that its not possible to train huge AI models on student budget.

Training focuses on:
- Visual–text alignment.
- Medical caption consistency.

5.6 Computational Consideration:

Training conducted using:
- HuggingFace Transformers.
- PEFT + BitsAndBytes + LigerKernel

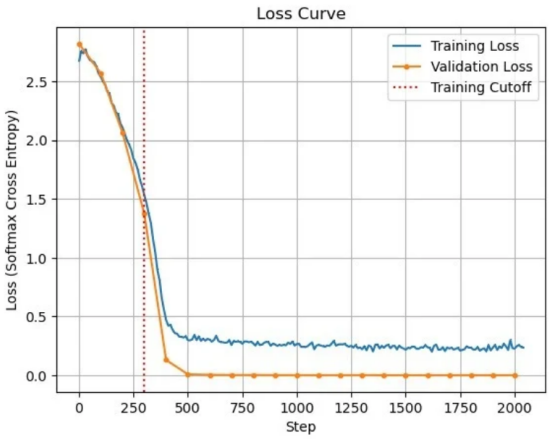
Fine-tuned weights are:
- Merged into base model.
- Converted and quantized for efficient local inference.



The model is adapted offline using **LoRA adapters with DoRA enabled** on a **pretrained Qwen2.5-VL** backbone. Training is monitored via **human-in-the-loop behavioral evaluation**, and the final checkpoint (**step 300**) is selected through manual behavioral stopping. Deployment is **inference-only**, with frozen weights and optional RAG grounding that does not affect model parameters.

The training mainly **improves how the model explains** so that its explanations are more aligned with how dermatologists describe skin conditions, which **directly improves the quality and safety of the SKINspector system.**

6. Evaluation



During fine-tuning, we monitor training and validation loss primarily as a **stability signal rather than a direct indicator of output quality**. Although validation loss converges early, we observe that **continued training leads to signs of behavioral overfitting**, such as repetitive and overconfident language in generated descriptions. Since our goal is behavioral adaptation rather than loss minimization, therefore, model selection relies on **periodic human-in-the-loop behavioral evaluation**, and training is **manually stopped at step 300 to prevent overfitting in descriptive behavior.**

Step	Reference	Generated
400	This is a picture showing multiple red nodules on the shoulder skin, with smooth surfaces and varying sizes ranging from mung bean to peanut, consistent with the clinical presentation of hypertrophic scars (occurring after acne). It is recommended to seek dermatological consultation for possible local treatment and inflammation control.	The most common types of skin cancer are basal cell carcinoma, squamous cell carcinoma, and melanoma. Basal cell carcinoma is the most common type of skin cancer and is typically slow-growing. Squamous cell carcinoma is the second most common type of skin cancer and is usually more aggressive than basal cell carcinoma. Melanoma is the most serious type of skin cancer and is often associated with a higher risk of metastasis. The treatment options for these three types of skin cancer are different. For example, the treatment for basal cell carcinoma is usually surgery, while the treatment for squamous cell carcinoma is usually surgery or radiation therapy. The treatment for melanoma is usually surgery, chemotherapy, and immunotherapy. Therefore, the treatment plan should be determined according to the type of skin cancer.
...	...	...
2000	This is an image of skin lesions on both lower extremities, showing multiple papules and nodules scattered on the skin, not merging with each other, with clear borders, consistent with the clinical characteristics of nodular prurigo. Nodular prurigo is a chronic inflammatory skin disease with unclear etiology, characterized by the presence of papules and nodules on the skin, accompanied by itching. It is commonly seen on the lower extremities, buttocks, and other areas. Treatment mainly involves topical steroids, oral antihistamines, and in severe cases, phototherapy or immunosuppressive therapy may be considered.	The photo shows scattered red papules on both upper limbs, some of which merge into plaques, consistent with the clinical presentation of nodular erythema. Nodular erythema is a chronic inflammatory disease characterized by the appearance of nodules or plaques on the skin and mucous membranes, often accompanied by itching and pain. The etiology of nodular erythema is not fully understood, but it may be related to immune system abnormalities. Treatment typically involves anti-inflammatory medications and corticosteroids. Early diagnosis and treatment are crucial for controlling the progression of the disease.

This comparison illustrates the critical distinction between appropriate medical generalization and overconfident hallucination in vision-language model outputs during human-in-the-loop (HITL) evaluation.

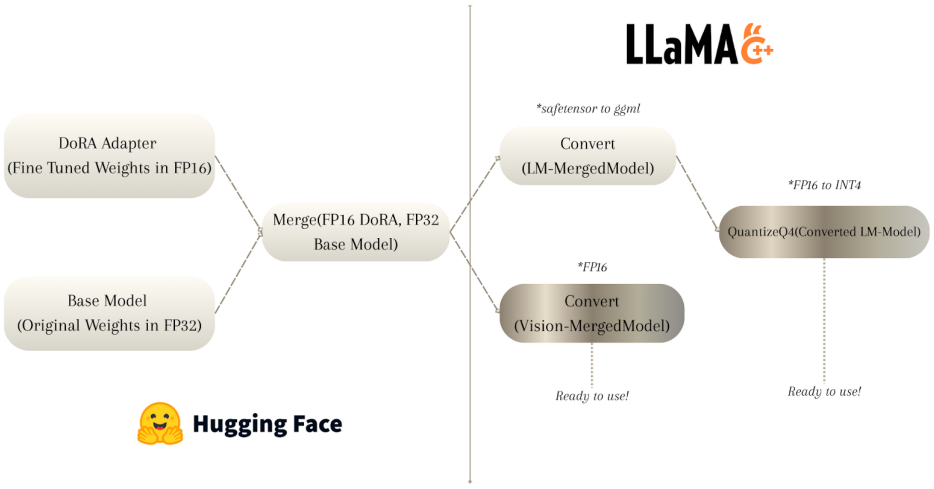
Step 400 demonstrates technically sound generalization by appropriately mapping the visual features observed in the reference image to a plausible dermatological diagnosis. The model correctly identifies the salient visual characteristics (multiple red nodules on shoulder skin with smooth surfaces and size variation ranging from mung bean to peanut) and generates a clinically reasonable differential diagnosis of hypertrophic scars secondary to acne. Importantly, the generated response maintains appropriate epistemic uncertainty by recommending dermatological consultation rather than providing definitive diagnosis or treatment, adhering to the principle that visual assessment alone is insufficient for conclusive medical determination. The response appropriately constrains its inference within the bounds of what can be reliably extracted from the image while acknowledging the limitations of image-based diagnosis.

In contrast, **Step 2000** exhibits problematic overconfidence manifested through diagnostic drift and unjustified medical assertions. While the reference description indicates nodular prurigo (a chronic inflammatory condition), the generated output pivots to nodular erythema, a distinct clinical entity, without adequate justification from the visual evidence. More critically, the generated text produces an extensive medical discourse on skin cancer classification, treatment modalities, and therapeutic protocols that bears no relationship to either the reference diagnosis or the visual presentation. This represents a severe hallucination event where the model confabulates detailed medical information ungrounded in the input, potentially introducing dangerous misinformation. The response

demonstrates overconfidence by presenting speculative differential diagnoses (basal cell carcinoma, squamous cell carcinoma, melanoma) and associated treatment algorithms with authoritative certainty, despite these conditions being visually and clinically inconsistent with the presented image.

This failure mode is particularly concerning in medical applications, as it conflates unrelated conditions and generates spurious clinical recommendations that could mislead healthcare providers or patients, highlighting the critical importance of calibrated confidence and faithful grounding in medical AI systems.

7. Deployment



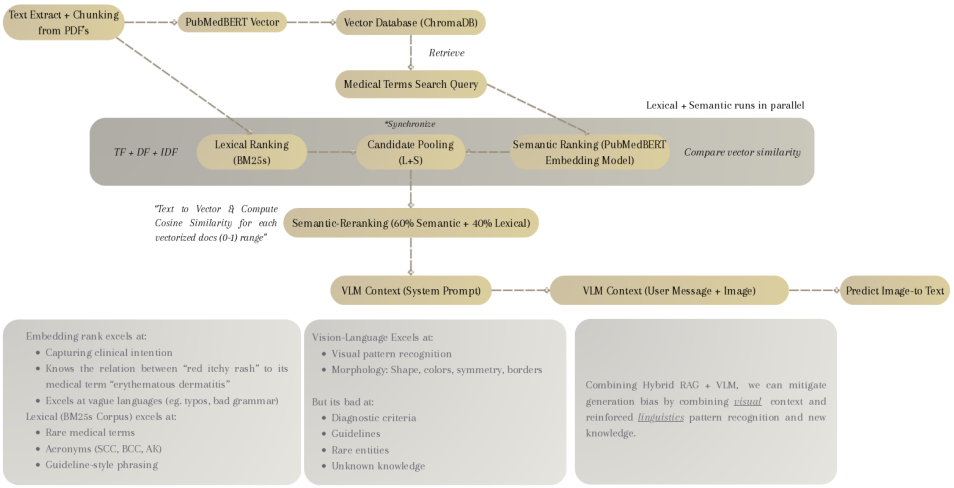
This figure delineates two distinct deployment pipelines for fine-tuned vision-language models, optimized for cloud-based inference via Hugging Face and resource-constrained edge deployment via LLaMA.cpp (ggml) ecosystems respectively.

The Hugging Face deployment pipeline commences with a base model parameterized in FP32 precision and a DoRA (Weight-Decomposed Low-Rank Adaptation) adapter module containing fine-tuned weights in FP16 precision. These components undergo a merging operation that integrates the adapter weights into the base model architecture, yielding a consolidated FP32 model ready for inference within the Hugging Face framework. While suitable for server-side deployment with abundant computational resources, this pipeline maintains high precision at the cost of increased memory requirements and computational overhead.

The LLaMA.cpp deployment pipeline, specifically engineered for edge computing scenarios with limited computational resources, employs a sophisticated conversion methodology optimized for on-device inference via the ggml backend. Leveraging the merged FP32 model as input, the pipeline bifurcates into two parallel processing streams designed to minimize resource consumption while maintaining clinical utility. The primary stream processes the language model component (LM-MergedModel) through a safetensor-to-ggml conversion utility, followed by Q4 quantization that reduces numerical precision from FP16 to INT4, thereby achieving substantial reductions in memory footprint (approximately 75% compression) and inference latency while preserving model efficacy. This aggressive quantization enables deployment on edge devices with constrained memory bandwidth, such as mobile processors, embedded systems, or IoT devices commonly found in remote healthcare facilities. Concurrently, the secondary stream handles the vision encoder (Vision-MergedModel) through an independent conversion process that maintains FP16 precision to ensure visual feature extraction fidelity for dermatological image analysis. This dual-stream architecture produces deployment-ready artifacts optimized for the LLaMA.cpp runtime, demonstrating a principled approach to model compression that applies differential quantization strategies based on modality-specific performance requirements and edge deployment constraints. The resulting model architecture enables autonomous

operation in resource-limited environments without cloud connectivity, addressing the infrastructural challenges characteristic of teledermatology deployment in remote Indonesian islands where network reliability and computational resources are severely constrained.

7.1 Retrieval Augmented Generation



We leverage a hybrid retrieval-augmented generation (RAG) *(Lewis et al.)* systems architecture that synergistically combines lexical and semantic search methodologies to optimize medical information retrieval for vision-language model inference.

The pipeline initiates with document preprocessing, wherein medical literature PDFs undergo text extraction and chunking operations. These text segments are subsequently encoded via PubMedBERT *(Gu et al.)*, a domain-specific transformer model pretrained on biomedical corpora, generating dense vector representations that are indexed in a ChromaDB vector database. Upon receiving a medical terminology search query, the system executes parallel retrieval operations through two complementary pathways. The lexical ranking module employs traditional information retrieval metrics (Term Frequency, Document Frequency, and Inverse Document Frequency) implemented via BM25s algorithm to identify syntactically relevant documents. Concurrently, the semantic ranking module leverages the PubMedBERT embedding space to compute vector similarity between query and document representations, capturing conceptual relevance beyond surface-form matching. These dual retrieval streams synchronize at a candidate pooling stage, which aggregates results from both approaches.

The system subsequently applies a weighted semantic reranking strategy, combining 60% semantic similarity scores with 40% lexical matching scores, effectively balancing the precision of embedding-based retrieval with the recall advantages of traditional keyword matching. This hybrid scoring mechanism produces a unified ranking of candidate documents, which are then vectorized and filtered based on cosine similarity thresholds to ensure relevance quality. The retrieved context is finally integrated into the vision-language model's inference pipeline, where it augments both the system prompt and user message components alongside the input medical image, enabling grounded image-to-text prediction.

The architectural design addresses complementary strengths and limitations of each retrieval paradigm. The embedding-based semantic search excels at capturing clinical intent, establishing conceptual relationships between colloquial descriptions and formal medical terminology (e.g., mapping "red itchy rash" to "erythematous dermatitis"), and demonstrating robustness to linguistic variations and grammatical imperfections. The lexical component (BM25s) provides superior performance for rare medical terminology, standardized acronyms (e.g., SCC, BCC, AK), and guideline-specific phraseology that may be underrepresented in embedding spaces. Conversely, vision-language models demonstrate proficiency in visual pattern recognition and morphological feature extraction (shape, color, symmetry, border characteristics) but exhibit limitations in diagnostic criteria interpretation, clinical guideline adherence, identification of rare pathological entities, and

knowledge domains absent from pretraining corpora. By synthesizing visual context from the VLM with linguistically grounded retrieval from the hybrid RAG system, this architecture mitigates generation bias through the integration of visual perception and reinforced linguistic pattern recognition, ultimately enhancing both factual accuracy and clinical utility of the generated medical descriptions.

7.2 RAG Ablation Study

The experiments were conducted with the following VLM generation parameters:

Context Size (8192 Tokens)

Context Size defines the maximum number of tokens the model can process in a single prompt, including both input and output. It represents the model's memory window.
$C = 8192 \text{ tokens}$
where C is the maximum sequence length. Any text beyond this limit cannot be processed in a single inference pass.

Repeat Penalty (1.0)

Repeat Penalty discourages the model from repeating tokens that have already appeared. The modified probability for a token is:
$P'(x_t) = \frac{P(x_t)}{\alpha^{n(x_t)}}$
where $P(x_t)$ is the original probability of token x_t, $\alpha = 1.0$ is the penalty coefficient, and $n(x_t)$ is the count of how many times token x_t has appeared. At $\alpha = 1.0$, no penalty is applied.

Temperature (0.7)

Temperature controls the randomness of predictions by scaling the logits before applying softmax:
$P(x_t \mid x_{<t}) = \frac{\exp(z_t/\tau)}{\sum_{i=1}^{V} \exp(z_i/\tau)}$
where z_t are the logits, $\tau = 0.7$ is the temperature, and V is the vocabulary size. Lower values ($\tau < 1$) make the model more deterministic and confident.

Top P (0.95)

Top P (nucleus sampling) samples from the smallest set of tokens whose cumulative probability exceeds threshold p:

$$V_p = \min \left\{ V' \subseteq V : \sum_{x \in V'} P(x) \geq p \right\}$$

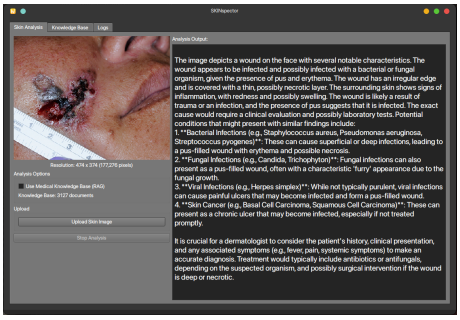
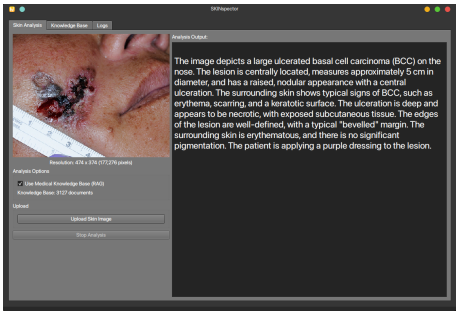
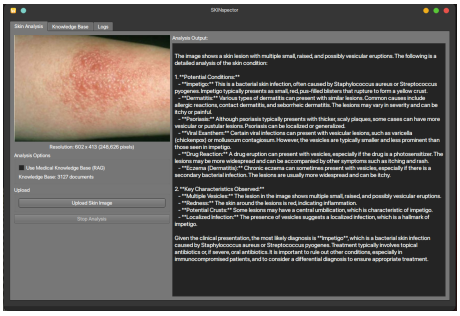
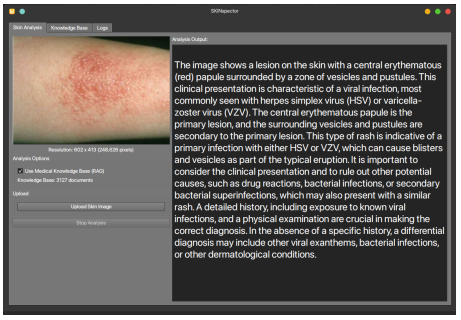
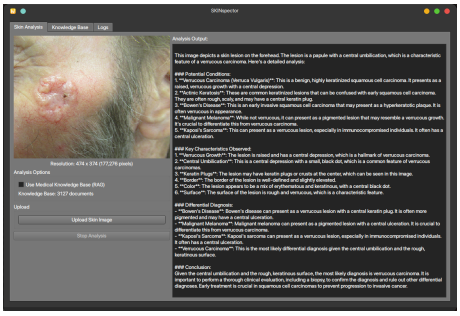
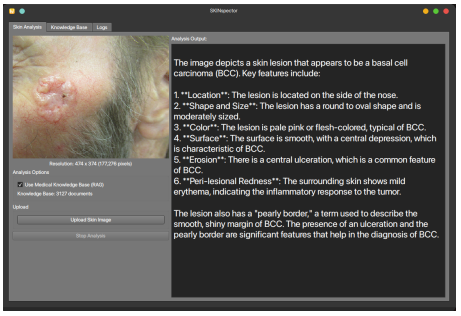
where $p = 0.95$. Only tokens in V_p are considered for sampling, filtering out low-probability tail tokens.

Top K (25)

Top K restricts sampling to only the k most probable tokens:

$$P'(x_t) = \begin{cases} P(x_t) & \text{if } x_t \in \text{top-}k \\ 0 & \text{otherwise} \end{cases}$$

where $k = 25$. The distribution is then renormalized over these k tokens before sampling.

Bare - Inference	*RAG Inference*
	
	
	

In **non-RAG** (Retrieval-Augmented Generation) inference, the model relies exclusively on parametric knowledge encoded during pre-training to generate diagnostic assessments. For instance, when analyzing a facial lesion with erythema and purulent discharge, a non-RAG system might broadly classify it as *"an infected wound, possibly bacterial or fungal"* based on visual pattern recognition alone. The model generates differential diagnoses by sampling from its learned probability distributions, often producing generic categories like *"bacterial infections (e.g., Staphylococcus aureus, Pseudomonas aeruginosa)"* without specificity regarding which organism is more likely given the particular morphological features present. The reasoning remains superficial because the model cannot access current antimicrobial resistance patterns, updated treatment guidelines, or specific clinical criteria that distinguish between similar presentations. Additionally, when encountering ambiguous cases—such as distinguishing between verrucous carcinoma and actinic keratosis—the non-RAG system may list both conditions with equal confidence without providing the clinical decision framework that practicing dermatologists use to differentiate them, resulting in diagnostically imprecise outputs that require significant expert interpretation.

In contrast, **RAG-based inference** fundamentally transforms this process by retrieving and integrating specific medical knowledge during generation. When analyzing the same infected facial wound, the RAG system queries its knowledge base and retrieves precise clinical definitions, then generates a more structured analysis: *"The wound appears to be infected and possibly infected with a bacterial or fungal organism, given the presence of pus and erythema."* Critically, the system can then specify that *"potential conditions include: 1. Bacterial Infections (e.g., Staphylococcus aureus, Pseudomonas aeruginosa, Streptococcus pyogenes): These can cause superficial or deep infections, leading to a pus-filled wound with erythema and possible necrosis."* This granularity stems from retrieved documents containing specific organism-symptom associations. For the verrucous carcinoma case, the

RAG system retrieves distinguishing features and generates detailed differential criteria: it identifies the *"central umbilication" and "rough, keratinous surface"* as key diagnostic features, then systematically compares these against retrieved descriptions of Bowen's Disease, malignant melanoma, and Kaposi's sarcoma. The system can state definitively that *"verrucous carcinoma is the most likely differential diagnosis given the central umbilication and the rough, keratinous surface,"* because it has accessed specific morphological criteria from authoritative sources rather than generating text purely from statistical correlations. This retrieval-grounded reasoning provides not only higher diagnostic accuracy but also explicit justification chains that align with evidence-based clinical practice, making the system's outputs auditable and trustworthy for medical decision support applications.

8. Contribution

- Rezky Mulia Kam: He's the first author and programmer that come up with the idea of building a software specific for Dermatologist by leveraging cutting edge methods from research papers at the intersection of Computer Vision & Linguistics and identifying present real-world problems as a reinforcement of Marcell's previous idea for CNN-based acne classification.
- Aflaha Fathinah Fatahillah: Afla helped this project by simplifying the abstraction of our complex end-to end training pipeline into a readable explanation for broad ranges of audiences by creating a sequential flow-diagram.
- Rafi Hazel Tafara: Hazel helped this project by creating an insightful data visualization and identifying key insights of our training and evaluation results for HITL.
- Felix Young: Felix helped this project by summarizing *"SkinCaRe: A Multimodal Dermatology Dataset Annotated with Medical Caption and Chain-of-Thought Reasoning." (Shen et al.)* paper into a concise technical explanation for our medical-oriented project.
- Marcellino Asanuddin: Marcell helped this project by coming up with an idea upon CNN-based acne classification and reinforcing our technical report writing for potential future research publication.

9. Limitations

The deployment of vision-language models in clinical dermatology presents several critical considerations that warrant careful examination from both technical and practical perspectives.

Regulatory and Validation Constraints. Real-world medical applications necessitate stringent regulatory compliance frameworks and comprehensive expert validation protocols that fundamentally diverge from conventional model evaluation paradigms. While human-in-the-loop (HITL) methodologies and mathematical performance metrics (e.g., BLEU, ROUGE, perplexity) provide quantitative assessments of model behavior, they remain insufficient proxies for clinical utility and patient safety. Medical device regulations, such as those governed by FDA 510(k) clearance or CE marking under the Medical Device Regulation (MDR), mandate prospective clinical trials, inter-rater reliability studies with board-certified dermatologists, and rigorous post-market surveillance—requirements that substantially exceed the scope of standard machine learning validation protocols. This regulatory-technical gap represents a significant barrier to clinical translation, as models demonstrating strong performance on benchmark datasets may nonetheless fail to meet the evidentiary standards required for medical deployment.

Methodological Scope and Scientific Contribution. The present work primarily emphasizes technical implementation leveraging contemporary deep learning frameworks (PyTorch, Transformers, ChromaDB) and established architectural paradigms (LoRA adapters, hybrid retrieval systems, vision-language transformers). While demonstrating engineering proficiency, the research trajectory prioritizes systems integration and framework utilization over fundamental scientific innovation. The methodology does not introduce novel theoretical frameworks for multimodal reasoning, propose new architectures addressing dermatology-specific challenges, or advance domain-specific representation learning techniques. Consequently, the contribution resides primarily within the engineering domain rather than constituting a multidisciplinary scientific advancement

that bridges computer vision, medical informatics, and dermatopathology. Future iterations would benefit from deeper collaboration with clinical domain experts to identify and address substantive medical AI research questions beyond technical demonstration.

Misapplication Risk and Responsible Deployment. The system architecture presents considerable potential for misuse by non-expert users who may misinterpret model outputs as definitive medical diagnoses. Naive users lacking medical training may fail to recognize the probabilistic nature of model predictions, the limitations of image-based diagnosis without clinical context (patient history, symptom duration, systemic manifestations), and the necessity of differential diagnosis by qualified practitioners. This misapplication risk is amplified by the model's tendency toward confident generation (as evidenced in the HITL evaluation), which may mislead users into believing generated descriptions constitute authoritative medical opinions. Deployment safeguards must include explicit disclaimers, confidence calibration mechanisms, and user interface designs that emphasize the assistive rather than autonomous nature of the system. Without such protections, the technology risks contributing to diagnostic errors, delayed treatment seeking, or inappropriate self-medication.

Quantization-Induced Error Propagation in Multimodal Inference. The adoption of Q4 (4-bit) quantization utilizing integer representations rather than floating-point arithmetic introduces significant concerns regarding error accumulation in autoregressive generation *(Wu et al.)*. Unlike floating-point quantization schemes (e.g., FP16, BF16) that preserve numerical dynamic range, integer quantization maps continuous weight distributions to discrete bins, inducing quantization noise that propagates through sequential decoding operations. In vision-language architectures, this error accumulation manifests across both modalities: the vision encoder experiences degraded feature extraction precision, while the language decoder suffers from compounded prediction errors as each generated token conditions subsequent token probabilities based on increasingly noisy hidden states. For teledermatology applications requiring high-fidelity visual feature discrimination (subtle color variations, texture gradients, border irregularities) and precise medical terminology generation, Q4 quantization may introduce unacceptable error rates. The next-token prediction cascade inherent to autoregressive language models amplifies these effects, as early quantization errors bias the probability distribution over subsequent tokens, potentially causing semantic drift or the generation of clinically inappropriate terminology. This technical limitation necessitates comprehensive ablation studies quantifying the impact of quantization precision on both perceptual fidelity and clinical accuracy before the model can be responsibly considered for real-world teledermatology deployment.

10. Conclusions and Future work

Our paper demonstrated a promising potential for real-world deployment specifically in any remote regions that lacks access to experienced Dermatologist and datacenter-level compute power, besides the listed limitations that require a medical expert to validate which we have not yet found to collaborate due to logistical issues during the development of this project .

Future work will emphasize on filling the gap of current limitations explicitly listed at section 9.

References

Bai, Shuai. “[2502.13923] Qwen2.5-VL Technical Report.” *arXiv*, 19 February 2025, https://arxiv.org/abs/2502.13923. Accessed 27 December 2025.

Alexey, Dosovitskiy. “[2010.11929] An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale” *arXiv*, 3 Jun 2021, https://arxiv.org/abs/2010.11929. Accessed 27 December 2025.

Yuhao, Shen. “[2405.18004] SkinCaRe: A Multimodal Dermatology Dataset Annotated with Medical Caption and Chain-of-Thought Reasoning.” *arXiv*, 9 Nov 2025, https://arxiv.org/abs/2405.18004. Accessed 27 December 2025.

Patrick, Lewis. “[2005.11401] Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks.” *arXiv*, 12 Apr 2021, https://arxiv.org/abs/2005.11401. Accessed 27 December 2025.

Yu, Gu. “[2007.15779] Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing.” *arXiv*, 16 Sept 2021, https://arxiv.org/abs/2007.15779. Accessed 27 December 2025.

Xiaoxia, Wu. “[2301.12017] Understanding INT4 Quantization for Transformer Models: Latency Speedup, Composability, and Failure Cases.” *arXiv*, 30 May 2023, https://arxiv.org/abs/2301.12017. Accessed 27 December 2025.

Shih-Yang, Liu. “[2402.09353] DoRA: Weight-Decomposed Low-Rank Adaptation.” *arXiv*, 9 Jul 2024, https://arxiv.org/abs/2402.09353. Accessed 27 December 2025.

Pin-Lun, Hsu. “[2410.10989] Liger Kernel: Efficient Triton Kernels for LLM Training.” *arXiv*, 24 Jan 2025, https://arxiv.org/abs/2410.10989. Accessed 27 December 2025.

Tim, Dettmers. “[2305.14314] QLoRA: Efficient Finetuning of Quantized LLMs.” *arXiv*, 23 May 2023, https://arxiv.org/abs/2305.14314. Accessed 27 December 2025.

Esteva, A., Kuprel, B., Novoa, R. *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118 (2017). https://doi.org/10.1038/nature21056