[TÍTULO DA DISSERTAÇÃO]

[Sub-Título (se exisitir)]

[Nome Completo do(a) Candidato(a)]

Student No.: [Número do Aluno]

A dissertation submitted in partial fulfillment of the requirements for the degree of Master of Science, Specialisation Area of

Supervisor: [Nome do Orientador]
Co-Supervisor: [Nome do Co-orientador (caso exista)]

Evaluation Committee:
President:
[Nome do Presidente, Categoria, Escola]

Members:
[Nome do Vogal1, Categoria, Escola]
[Nome do Vogal2, Categoria, Escola] (até 4 vogais)

Porto, January 27, 2026

Statement Of Integrity

[Maintain only the version corresponding to the main language of the work]

I hereby declare that I have conducted this academic work with integrity.

I have not plagiarised or applied any form of undue use of information or falsification of results along the process leading to its elaboration.

Therefore, the work presented in this document is original and was authored by me, having not been previously used for any other purpose. The exceptions [REMOVE THIS CLAUSE IF IT DOES NOT APPLY - REMOVE THIS COMMENT] are explicitly acknowledged in the section that addresses ethical considerations. This section also states how Artificial Intelligence tools were used and for what purpose.

I further declare that I have fully acknowledged the Code of Ethical Conduct of P.PORTO.

ISEP, Porto, [Month] [Day], [Year]

Declaro ter conduzido este trabalho académico com integridade.

Não plagiei ou apliquei qualquer forma de uso indevido de informações ou falsificação de resultados ao longo do processo que levou à sua elaboração.

Portanto, o trabalho apresentado neste documento é original e de minha autoria, não tendo sido utilizado anteriormente para nenhum outro fim. As exceções [REMOVER ESTE PERÍODO NO CASO DE NÃO SE APLICAR - APAGAR ESTE COMENTÁRIO] estão explicitamente reconhecidas na secção onde são abordadas as considerações éticas. Esta secção também declara como as ferramentas de Inteligência Artificial foram utilizadas e para que finalidade.

Declaro ainda que tenho pleno conhecimento do Código de Conduta Ética do P.PORTO.

ISEP, Porto, [Dia] de [Mês] de [Ano]

Dedicatory

The dedicatory is optional. Below is an example of a humorous dedication.

"To my wife Marganit and my children Ella Rose and Daniel Adam without whom this book would have been completed two years earlier." in "An Introduction To Algebraic Topology" by Joseph J. Rotman.

Abstract

This document explains the main formatting rules to apply to a Master Dissertation work for the MSc in Artificial Intelligence Engineering of the Computer Engineering Department (DEI) of the School of Engineering (ISEP) of the Polytechnic of Porto (IPP).

The rules here presented are a set of recommended good practices for formatting the disseration work. Please note that this document does not have definite hard rules, and the discussion of these and other aspects of the development of the work should be discussed with the respective supervisor(s).

This document is based on a previous document prepared by Dr. Fátima Rodrigues (DEI/ISEP).

The abstract should usually not exceed 200 words, or one page. When the work is written in Portuguese, it should have an abstract in English.

Please define up to 6 keywords that better describe your work, in the *THESIS INFORMATION* block of the `main.tex` file.

Keywords: Keyword1, ..., Keyword6

Resumo

Após o resumo/abstract é obrigatório colocar as principais palavras-chave/keywords do tema em que se insere o trabalho desenvolvido, sendo permitido um máximo de 6 palavras-chave/keywords, estas devem ser caraterizadoras do trabalho desenvolvido e surgirem com frequência no documento escrito.

Para alterar a língua basta ir às configurações do documento no ficheiro `main.tex` e alterar para a língua desejada ('english' ou 'portuguese')[1]. Isto fará com que os cabeçalhos incluídos no template sejam traduzidos para a respetiva língua.

Palavras-chave: Keyword1, ..., Keyword6

[1] Alterar a língua requer apagar alguns ficheiros temporários; O target **clean** do **Makefile** incluído pode ser utilizado para este propósito.

Acknowledgement

The optional Acknowledgment goes here... Below is an example of a humorous acknowledgment.

"I'd also like to thank the Van Allen belts for protecting us from the harmful solar wind, and the earth for being just the right distance from the sun for being conducive to life, and for the ability for water atoms to clump so efficiently, for pretty much the same reason. Finally, I'd like to thank every single one of my forebears for surviving long enough in this hostile world to procreate. Without any one of you, this book would not have been possible." in "The Woman Who Died a Lot" by Jasper Fforde.

Contents

List of Figures

List of Tables

List of Algorithms

List of Source Code

List of Symbols

a	distance	m
P	power	W ($\mathrm{J\,s^{-1}}$)
ω	angular frequency	rad

List of Acronyms

GPU	Graphics Processing Unit.
LLM	Large Language Model.
LoRA	Low-Rank Adaptation.
RAG	Retrieval-Augmented Generation.
SLM	Small Language Model.
VLM	Vision-Language Model.
VRAM	Video Random Access Memory.

Chapter 1

Introduction

1.1 Contextualization

The field of Natural Language Processing (NLP) has undergone a radical transformation over the last decade . Historically, early NLP relied heavily on statistical methods and Recurrent Neural Networks (RNNs) or Long Short-Term Memory (LSTM) networks to process text (Chu et al. 2024; Raeini et al. 2025). While effective for short sequences, these architectures struggled with long-range dependencies and lacked parallelization capabilities. The paradigm shifted fundamentally with the introduction of the Transformer architecture in 2017 (Vaswani et al. 2017). The Transformer mechanism introduced "self-attention," allowing models to weigh the importance of different words in a sentence regardless of their positional distance. This architecture laid the foundation for the current generation of generative AI, enabling models to be trained on massive datasets to learn the statistical structure of language itself, giving birth to the new term LLMs.

The impact of the Transformer architecture was further amplified by the systematic scaling of model parameters, training data, and computational resources. Empirical evidence has shown that such increase in scale lead to predictable and often non-linear improvements in performance across a wide range of linguistic tasks, a phenomenon formalized through "Scaling Laws"(Pearce et al. 2024). As the models grew, new capabilities also began to emerge, enabling them to perform tasks without explicit task-specific training (Brown et al. 2020). These abilities were compared to the ones of human-like mind such as Chain of Thought, Multi-Step Reasoning which led LLMs to be positioned as the general-purpose language understanding and generation systems (Wei et al. 2023).

This generalization capability has caused a paradigm switch in Human-Computer Interaction (HCI), moving the digital interface from rigid, command-based interactions to natural language conversations. In the present day, LLMs are seen as "Foundation Models", a term describing systems trained on broad data that can be adapted to a vast range of downstream tasks. Beyond the initial applications in software engineering, these models are now driving transformative shifts in high-stakes domains, ranging from personalized education Kasneci et al. 2023; W. Zhang et al. 2025a and financial forecasting S. Wu et al. 2024 to complex legal reasoning Lei Chen et al. 2025a and scientific discovery H. Wang et al. 2024. This shift enabled users to retrieve and reason over complex information using simple natural language prompts. Due to these changes toward language-centered interaction, LLMs have begun to cause an impact across multiple sectors of society significantly increasing human productivity by automating routine cognitive tasks and offering decision support (Brynjolfsson, D. Li, and Raymond 2023; Garg et al. 2025; Al-Naqbi et al. 2024).

Despite the dominance of these massive Foundation Models, the research trajectory has recently been divided. While one path continues to pursue larger parameters for generalist capabilities, a parallel body of research has emerged challenging the assumption that "bigger is always better" for domain-specific performance (Lepagnol et al. 2024). This perspective was validated by studies on Compute-Optimal training, most notably DeepMind's "Chinchilla" research, which demonstrated that model performance depends less on sheer parameter count and more on the quality and quantity of tokens seen during training (J. Hoffmann et al. 2022), meaning that when models are exposed to a vast and high quality data, they could achieve performance levels comparable to larger models.

This realization gave rise to the class of Small Language Models (SLMs). Typically defined as models with fewer than 10 billion parameters Belcak et al. 2025a, SLMs prioritize training efficiency and data quality. However, to compete with the capabilities of larger models, SLMs are frequently deployed in conjunction with augmentation strategies, specifically Fine-Tuning and Retrieval-Augmented Generation (RAG).

To understand this architectural shift, it is necessary to define the concept of Agentic Systems. Unlike traditional conversational models that passively generate text in response to a prompt, an Agentic System utilizes the Language Model as a cognitive controller capable of autonomous reasoning and planning L. Wang et al. 2024; Xi et al. 2023. In this paradigm, the model is equipped with a feedback loop (often termed the Perception-Action loop) that allows it to decompose abstract user goals into executable sub-tasks, invoke external tools (such as APIs or databases), and iteratively refine its output based on environmental feedback.

Consequently, the true potential of SLMs is not impactful when they operate in isolation, but when they are integrated into Agentic Systems. As argued by Belcak et al. 2025a, the integration of SLMs with augmenting tools, such as Retrieval-Augmented Generation (RAG) and being able to fetch data with external tools such as API calls, fundamentally transforms the role of the model. In this configuration, the SLM is a true competitor comparable to a LLM, achieving better high-quality and precision answers to that of a larger model (Hsieh et al. 2023; Lepagnol et al. 2024; Pingua et al. 2025; Soudani et al. 2024), not only that but their lightweight architecture makes them an appealing choice for scenarios where computational efficiency is ciritcal and data privacy is needed since they are able to be hosted locally (Dettmers et al. 2023a; J. Kim et al. 2025; Lepagnol et al. 2024). While a generalist LLM attempts to handle all tasks via internal parametric memory, a Agentic System decomposes complex workflows into modular sub-tasks. By specializing and equipping an SLM with RAG, the system creates a "Specialized Agent" capable of retrieving verifiably accurate medical context without needing the massive overhead of a trillion-parameter model (Belcak et al. 2025a).

Nevertheless, a limitation remains when relying on a single model, regardless of its size, to handle a diverse array of tasks. When a solitary agent is tasked with perceiving complex and visual inputs, retrieving context, reasoning, and generating patient-centric responses simultaneously, it faces "cognitive overload," often leading to a degradation in performance known as the "lost-in-the-middle" phenomenon or context dilution (N. F. Liu et al. 2023). To mitigate this, the frontier of Agentic AI has shifted toward Multi-Agent Systems (MAS) (Hua Chen et al. 2025). The core philosophy of MAS is "decomposition": breaking down a complex, multifaceted problem into smaller, manageable sub-tasks, each handled by a specialized agent (Fu et al. 2025; Q. Wu et al. 2023). Rather than relying on a single

monolithic model to act as a universal expert, multi-agent architectures adopt a divide-and-conquer strategy in which coordination and specialization are treated as first-class design principles. Within such architectures, a particular importance is placed on the routing or orchestration component, which is responsible for analyzing incoming inputs and delegating tasks to appropriate downstream agents.

This direction aligns closely with the concept of heterogeneous agentic systems proposed by Belcak et al. 2025a, which argues that effective agentic architectures are inherently composite, combining models of different modalities and scales according to the demands of each sub-task. By leveraging modality-specific inductive biases and maintaining clear functional boundaries between agents, these modular architectures aim to improve robustness, interpretability, and scalability, while enabling flexible system evolution through the replacement or refinement of individual components without retraining the entire system (Belcak et al. 2025a; Bubeck et al. 2023).

1.2 Problem Description

The prevailing approach to Generative AI has relied heavily on scaling laws (Pearce et al. 2024), assuming that larger parameters equate to superior performance. However, this monolithic, scale-centric model has revealed significant structural limitations. The computational and energy costs associated with the training and inference of these models have become prohibitive for the majority of organizations, particularly when considering the substantial hardware requirements for deployment (Avinash et al. 2025; Dettmers et al. 2023a). For real-time applications, the latency introduced by routing data to massive cloud-based models creates a bottleneck that degrades user experience, rendering them impractical for responsive, edge-based diagnostic tools.

More critically, within highly regulated domains such as healthcare, the reliance on centralized cloud infrastructures or external AI services conflicts with imperative requirements for data privacy and sovereignty. The transmission of sensitive patient data, specifically dermatological imagery and personal medical history, to external API endpoints introduces unacceptable risks regarding data residency and confidentiality. As noted by Petrick et al. 2023 and Khalid et al. 2023, the "black box" nature of commercial LLM APIs often prevents granular control over data retention policies, creating a compliance gap that necessitates the use of local, controllable architectures.

Beyond infrastructure, the fundamental architecture of generalist LLMs poses a safety risk in diagnostic scenarios. Large models trained on the open internet function as probabilistic engines rather than knowledge bases; they are prone to "hallucinations," generating plausible but factually incorrect medical advice with high confidence (Ji et al. 2022). In a monolithic setup, it is difficult to constrain the model to strictly medical facts. This lack of determinism and explainability creates a "trust gap." A generalist model cannot easily point to the specific medical journal it used to derive a diagnosis, making verification impossible for the user. This necessitates a shift toward systems that decouple reasoning (the model) from knowledge (the data), a feature inherent to RAG-augmented architectures but inefficient to implement at the scale of LLMs.

This scenario reveals a saturation point in the strategy of pure scalability and motivates an urgent shift toward a new design philosophy centered on smaller, more efficient, and controllable models. Small Language Models (SLMs), when appropriately specialized, offer a viable

alternative by preserving performance while enabling privacy-preserving and resource-efficient deployment. Furthermore, when paired with external tools for context augmentation, SLMs offer a pathway to mitigate the "black box" nature of Natural Language Generation, enhancing interpretability.

This transition aligns with the prospective vision of the present industry, which posits that SLMs represent the future of Agentic Systems (Belcak et al. 2025a). In this new perspective, model size becomes secondary to specialization. Current research demonstrates that a specialized SLM, when augmented by context-enrichment tools such as Retrieval-Augmented Generation (RAG) and adaptation methods like Fine-Tuning (Oruganty et al. 2025; Pingua et al. 2025; Soudani et al. 2024), can achieve response quality superior to that of generalist LLMs, particularly when supported by external evidentiary verification (Hassan et al. 2025).

However, as previously noted, a single LLM Agent does not guarantee clinical intelligence, specifically when complex tasks or workflows are involved. The prevailing trend is evolving toward Multi-Agent Architectures, where system complexity resides not within a single neural network, but in the orchestration of multiple specialists. This evolution is driven by the demonstrated cognitive limitations of monolithic systems in multitasking environments. Recent evidence suggests that agents, when subjected to demanding workloads, suffer from severe performance degradation. Klang et al. (Klang et al. 2025) demonstrated that the accuracy of a monolithic agent collapses from 73.1% to 16.6% under high cognitive load. In contrast, a multi-agent system was able to sustain an accuracy of 65.3% under the same conditions, validating the superior efficacy and robustness of this modular architecture (Tian et al. 2025; M. Zhou et al. 2025).

1.3 Objectives

In response to the critical challenges identified in Problem Description section (1.2), specifically the prohibitive computational costs of monolithic LLM systems, the privacy risks inherent to cloud-based architectures, and the documented cognitive degradation of single agents under multitasking loads, this research aims to validate an architecture that prioritizes efficiency over scale and specialization over generality.

To mitigate the "black box" nature of generalist models and ensure compliance with healthcare data privacy, the proposed solution moves away from a "one-size-fits-all" cloud dependency. Instead, it implements a modular pipeline where the cognitive load is distributed across specialized local agents. Specifically, this work develops a system initiated by a Vision-Language Model (VLM) acting as a "Router." This router analyzes patient-uploaded imagery to classify dermatological conditions (e.g., Eczema, Melanoma) and dynamically directs the user's session to a dedicated Small Language Model (SLM).

These downstream SLM agents are engineered not as creative generators, but as specialized advisors augmented with Retrieval-Augmented Generation (RAG). By grounding the SLM's responses in a curated vector database of medical literature, the system aims to provide verifiable, context-aware medical guidance. This approach intends to demonstrate that a coordinated ensemble of lightweight models (8B parameters) can achieve diagnostic utility comparable to massive cloud-based models, while enabling local, privacy-preserving deployment on consumer-grade hardware.

1.3.1 Main Objective

Dermatology is a multimodal field that requires combining visual analysis with medical knowledge. While AI is currently good at isolated tasks, such as using CNNs for images or LLMs for text, there is a lack of integrated systems that can handle both steps seamlessly. Therefore, dermatology is the ideal domain to validate Small Language Models (SLMs). Unlike massive Cloud models, SLMs can run locally to guarantee data privacy, and they can be paired with Retrieval-Augmented Generation (RAG) to ensuring the medical advice is factually grounded.

Driven by this specific intersection of privacy, multimodal reasoning, and resource efficiency, the main objective of this dissertation is:

To design, implement, and validate a privacy-preserving Multi-Agent System that orchestrates a Vision-Language Model (VLM) for visual classification and specialized Small Language Models (SLMs) for advisory; aiming to demonstrate that a modular architecture can provide context-aware, verifiable dermatological support comparable to monolithic LLM Systems.

1.3.2 Secondary Objectives

- **Gather data on different type of Skin Conditions Disease:** Aggregate and preprocess a verified set of skin condition images for visual classification, alongside a corpus of validated medical literature to construct the Knowledge Bases required for the RAG retrieval systems.
- **Implementation of a Visual Language Model:** Configure a Vision Language Model (VLM) capable of analyzing user-uploaded imagery to classify distinct skin pathologies (e.g., Eczema, Psoriasis, Melanoma) and dynamically route the session to the appropriate downstream agent.
- **Create RAG ingestion pipeline:** Design a robust retrieval architecture by evaluating specific data processing strategies, including semantic chunking, hybrid search algorithms (keyword + vector), and re-ranking mechanisms, to maximize information density within the constrained context windows of Small Language Models.
- **Develop specialized SLM-RAG Agents:** Develop lightweight agents using Small Language Models integrated with condition-specific vector databases, ensuring that responses are grounded in retrieved context rather than model weights alone to minimize hallucinations.
- **Orchestrate the Multi-Agent System workflow:** Design the control logic that enables seamless state transfer between the Visual Language Model (VLM) and the Small Language Model (SLM), maintaining context without the latency or overhead of a monolithic system.
- **Evaluation of Small Language Models againts Large Language Models:** Validate the architecture against Ground Truth benchmarks for diagnostic accuracy to demonstrate the precision anda accuracy of SLMs over generalist LLMs .

1.4 Ethical Considerations and Social Impact

The deployment of Agentic AI systems in healthcare, specifically within dermatological triage, introduces significant ethical challenges regarding data privacy, algorithmic fairness, and patient safety Khalid et al. 2023; Ziller et al. 2024. While the proposed architecture leverages Small Language Models (SLMs) to mitigate computational overhead, it must inherently address the risks associated with automated medical decision support and comply with emerging regulatory frameworks.

1.4.1 Medical and Diagnostic Ethics

Limitations and Disclaimers: The proposed system, while leveraging advanced Vision Language Models for skin disease classification, must operate within clear ethical boundaries regarding medical practice. The system should be positioned as a supplementary informational tool rather than a replacement for professional medical diagnosis. Users must receive explicit disclaimers that the classification results are preliminary and require confirmation by licensed dermatologists. This is particularly crucial given that skin diseases can present similarly across different conditions, and misclassification could lead to delayed treatment or inappropriate self-care measures (Taylor et al. 2025).

Accuracy and Reliability Concerns: Vision Language Models, despite their sophistication, may exhibit varying performance across different skin tones, disease severities, and image qualities. The training data's representation becomes ethically significant; if the model is predominantly trained on lighter skin tones (Fitzpatrick types I-III), it may perform poorly on darker complexions (types IV-VI), perpetuating healthcare disparities (M. Groh et al. 2021).

1.4.2 Regulatory Compliance: The EU AI Act

High-Risk Classification: Under the European Union Artificial Intelligence Act (EU AI Act), AI systems intended to be used for medical triage or as safety components of medical devices are classified as "High-Risk AI Systems" (Annex III) (European Parliament and Council 2024). Consequently, this architecture is designed with specific adherence to *Article 14 (Human Oversight)*, ensuring that the system acts as a Clinical Decision Support System (CDSS) rather than an autonomous prescriber.

Transparency and Data Governance: In compliance with *Article 13* (Transparency), the interface is designed to clearly explicitly inform the user that they are interacting with an automated agent. Furthermore, to satisfy data governance requirements regarding bias monitoring (*Article 10*), the proposed validation phase specifically evaluates the VLM's performance across diverse demographic groups to identify and document potential discriminatory outputs.

1.4.3 Privacy and Data Protection

Sensitive Health Information: User-uploaded images of skin conditions are highly sensitive Personally Identifiable Information (PII) and Protected Health Information (PHI). Traditional monolithic LLM architectures often require sending this data to centralized cloud API endpoints (e.g., OpenAI, Anthropic) Belcak et al. 2025a, creating risks of data interception and non-consensual training.

Edge AI and Sovereignty: The proposed SLM-based Multi-Agent System supports the paradigm of Edge AI Yuntao Wang et al. 2024. By utilizing smaller, resource-efficient models, the architecture enables the potential for local execution (on-device or on private servers). This ensures data privacy, as patient data does not necessarily need to traverse public cloud infrastructure to receive a high-quality inference, aligning with GDPR principles regarding data minimization.

1.4.4 Clinical Safety and Hallucination Mitigation

Liability and Accountability: Generative models are prone to "hallucinations", generating plausible but factually incorrect information Ji et al. 2022. In a medical context, such errors can be dangerous. Small Language Models, having fewer parameters, theoretically possess a narrower knowledge base than larger models, which could increase this risk.

RAG as an Ethical Safeguard and Explainability: The implementation of Retrieval-Augmented Generation (RAG) is not merely a technical optimization but an ethical imperative. By constraining the SLM to answer only based on retrieved, validated medical chunks, the system shifts from "creative generation" to "summarization of ground truth," significantly reducing the risk of fabricating medical advice (Hassan et al. 2025).

1.4.5 Environmental Impact and Democratization

Carbon Footprint: The training and inference of massive Monolithic LLMs carry a substantial carbon footprint (P. Liu et al. 2024). Promoting a "bigger is better" approach restricts advanced AI medical tools to well-funded institutions with massive compute clusters.

Chapter 2

State of the Art: Agentic AI and Small Language Models in Healthcare

2.1 Research Methodology

The systematic literature review was conducted following the PRISMA 2020 statement Page et al. 2021, a framework designed to ensure transparency and reproducibility in systematic reviews. This methodology was selected for several compelling reasons that align with the nature of this research domain. First, the fields of Agentic AI, Small Language Models, and Vision-Language Models are characterized by rapid innovation cycles, with foundational architectures and benchmark results being superseded within months of publication. PRISMA's structured approach ensures that the evidence synthesis captures this dynamism while maintaining methodological rigor. Second, the framework's emphasis on explicit documentation of search strategies, inclusion criteria, and study selection processes enhances the reproducibility of findings—a critical consideration given the interdisciplinary nature of this work spanning computer science, medical informatics, and clinical dermatology.

The review process was structured into four distinct phases: (1) identification of relevant studies through systematic database searching; (2) screening of titles and abstracts based on predefined inclusion criteria; (3) eligibility assessment of full-text articles against methodological quality standards; and (4) qualitative synthesis of the selected studies to address the defined research questions. The temporal scope of the search spanned publications from January 2023 to January 2026, a period deliberately chosen to capture the rapid maturation of Small Language Models following the release of instruction-tuned models such as Llama-2, Mistral, and Phi-2. Studies published before 2023 were excluded from the systematic review but referenced as foundational works where necessary to establish theoretical context.

Quality assessment of the selected studies followed a structured appraisal protocol designed specifically for this research domain. Each study was evaluated against four criteria: (1) provision of direct quantitative comparisons between SLM-based systems and state-of-the-art LLMs; (2) explicit employment of multi-agent architectures or composite systems rather than single-model inference; (3) evaluation within specialized high-stakes domains requiring domain grounding; and (4) availability of reproducible artifacts such as open-source code, datasets, or detailed architectural specifications. This quality assessment framework ensures that the synthesized evidence directly addresses the research questions while maintaining standards appropriate for technical AI research.

2.2 Research Questions

The systematic review was guided by a hierarchical structure of research questions designed to comprehensively examine the viability of Small Language Models within agentic architectures for specialized domains. The Main Research Question (MRQ) establishes the central thesis under investigation, while the Sub-Research Questions (SRQs) decompose this inquiry into specific, addressable components that collectively inform the overarching question.

The Main Research Question driving this review is:

MRQ: To what extent can Small Language Models achieve performance equivalence with Large Language Models in specialized domains through Agentic Architectures?

This question emerges directly from the tension identified in Chapter 1 between the demonstrated capabilities of massive foundation models and the practical constraints of computational cost, latency, and data privacy that limit their deployment in resource-constrained or privacy-sensitive contexts such as healthcare. To systematically address this central question, four Sub-Research Questions were formulated, as presented in Table 2.1.

Table 2.1: Sub-Research Questions and Their Rationale

ID	Research Question	Rationale
SRQ1	What are the limitations of Large Language Models that justify the architectural shift toward specialized Small Language Models?	Establishes the motivation for investigating alternatives to monolithic LLM architectures by examining their inherent constraints in deployment scenarios requiring efficiency, privacy, or specialized domain performance.
SRQ2	What are the key characteristics of Multi-Agent Systems compared to Monolithic Single-Agent architectures?	Understanding the architectural principles that enable distributed cognitive systems is essential for evaluating whether task decomposition and agent specialization can compensate for reduced model scale.
SRQ3	What is the comparative efficacy of Retrieval-Augmented Generation versus Parameter-Efficient Fine-Tuning for specializing Small Language Models?	Examines the two primary strategies for adapting smaller models to domain-specific tasks, informing the technical approach for the proposed system.
SRQ4	What evidence supports using fine-tuned Vision-Language Models over traditional computer vision approaches such as CNNs and ViTs?	Given the multimodal nature of dermatological diagnosis, this question investigates whether integrated vision-language architectures offer advantages over pipeline approaches.

2.3 Search Strategy

The search strategy was designed to ensure comprehensive coverage across the diverse publication venues characteristic of AI research while maintaining focus on the specific technical

domains relevant to this dissertation. Given the rapid pace of development in generative AI, particular attention was paid to preprint repositories that often contain state-of-the-art results prior to formal peer review.

2.3.1 Databases

To capture relevant literature spanning computer science, medical informatics, and clinical research, the following databases and repositories were systematically searched:

Table 2.2: Databases Selected for Systematic Literature Search

Database	Rationale for Inclusion
Google Scholar	Comprehensive coverage of academic literature across all disciplines, providing access to peer-reviewed papers, theses, books, and preprints. Serves as the primary discovery tool for cross-referencing findings across venues.
arXiv	Open-access repository for electronic preprints in computer science, mathematics, and statistics. Given that state-of-the-art results in generative AI, language models, and multi-agent systems are frequently published here months before formal peer review, arXiv serves as the primary source for cutting-edge technical contributions.
PubMed	The gold standard for biomedical literature, providing access to the MEDLINE database of peer-reviewed research in life sciences and clinical medicine. Essential for retrieving validated studies on dermatological diagnostics, telemedicine applications, and clinical AI evaluation methodologies.
ACM Digital Library	Premier resource for computing and information technology research. Critical for accessing studies on multi-agent system architectures, efficient model deployment, and human-computer interaction in AI systems.
Nature / Springer	High-impact peer-reviewed journals covering breakthrough research in medical AI, vision-language models, and clinical validation studies. Provides access to Nature Medicine, Nature Communications, and Scientific Reports publications.

2.3.2 Search Terms

The search strategy employed Boolean logic to combine keywords representing the four core pillars of this dissertation: (1) Health and Dermatology, (2) Generative AI, (3) Artificial Intelligence, and (4) System Architecture. Table 2.3 presents the search terms organized by domain category. The search strings were adapted to the syntax of each database while maintaining semantic equivalence across platforms.

The search terms were combined using AND operators across domains to ensure retrieved studies addressed the intersection of AI methodology and healthcare application. For example, a typical combined query would be: (Health terms) AND (Generative AI terms) AND (Architecture terms).

Table 2.3: Search Terms by Domain

Category	Search Terms
Health	"dermatology" OR "medical" OR "healthcare" OR "skin disease" OR "skin lesion" OR "skin cancer" OR "melanoma" OR "dermoscopy" OR "diagnosis" OR "clinical validation"
Generative AI	"large language model" OR "LLM" OR "small language model" OR "SLM" OR "Phi-3" OR "Gemma" OR "Llama" OR "Mistral" OR "retrieval augmented generation" OR "RAG" OR "knowledge distillation" OR "fine-tuning"
Artificial Intelligence	"vision language model" OR "VLM" OR "multimodal LLM" OR "CNN" OR "vision transformer" OR "LoRA" OR "PEFT" OR "quantization" OR "edge deployment" OR "privacy preserving"
Architecture	"multi-agent system" OR "MAS" OR "agentic AI" OR "task decomposition" OR "tool use" OR "orchestration" OR "benchmark" OR "evaluation" OR "hallucination mitigation"

2.4 Selection Criteria

The selection criteria were designed to identify studies that provide empirical evidence relevant to the research questions while ensuring methodological quality appropriate for informing system design decisions. The criteria balance inclusivity—necessary given the nascent state of SLM research—with rigor sufficient to support evidence-based conclusions.

2.4.1 Inclusion Criteria

Papers were included if they satisfied **all** of the following conditions:

1. **Publication Date**: Published or made publicly available between January 2023 and January 2026, capturing the rapid evolution of instruction-tuned SLMs and their application in specialized domains.
2. **Relevance to AI/ML Models**: The study must involve one or more of the following:
   - Large Language Models (LLMs), Small Language Models (SLMs), Vision-Language Models (VLMs), or Multimodal Models
   - Retrieval-Augmented Generation (RAG) techniques
   - Multi-Agent or Agentic AI architectures
   - Model optimization techniques (quantization, distillation, PEFT)
3. **Healthcare or Medical Domain**: Studies focusing on dermatology, skin disease classification, medical question-answering systems, or clinical decision support were prioritized to ensure direct relevance to the dissertation objectives.
4. **Publication Type**: Peer-reviewed journal articles, conference papers from recognized venues (NeurIPS, ICML, ICLR, ACL, MICCAI), or high-quality preprints from established research groups (arXiv, medRxiv, bioRxiv) demonstrating methodological rigor.
5. **Language**: Written in English.

6. **Accessibility**: Full text available through institutional access or open access repositories.

2.4.2 Exclusion Criteria

Papers were excluded if they met **any** of the following conditions:

1. **Temporal Scope**: Published before January 2023, with the exception of seminal foundational works (e.g., the original Transformer architecture, LoRA) cited for background context but not included in the systematic synthesis.
2. **Duplicate Publications**: Conference papers subsequently published as extended journal versions (journal version retained), or preprints superseded by peer-reviewed publications (peer-reviewed version retained).
3. **Non-Generative AI Focus**: Studies on traditional machine learning approaches (SVMs, random forests, classical CNNs) without integration of language models or agentic components.
4. **Non-Peer-Reviewed Sources**: Blog posts and technical documentation were excluded unless they represented official publications from major research organizations (e.g., NVIDIA, Microsoft Research, Google DeepMind).
5. **Methodological Quality**: Studies without quantitative evaluation or lacking reproducibility details, as assessed through the quality appraisal framework.

2.4.3 Selection Process

The selection process followed the PRISMA 2020 flow, progressing through identification, screening, eligibility assessment, and final inclusion. Figure 2.1 illustrates the flow of studies through each phase of the review, documenting the number of records identified, screened, assessed for eligibility, and ultimately included in the qualitative synthesis.

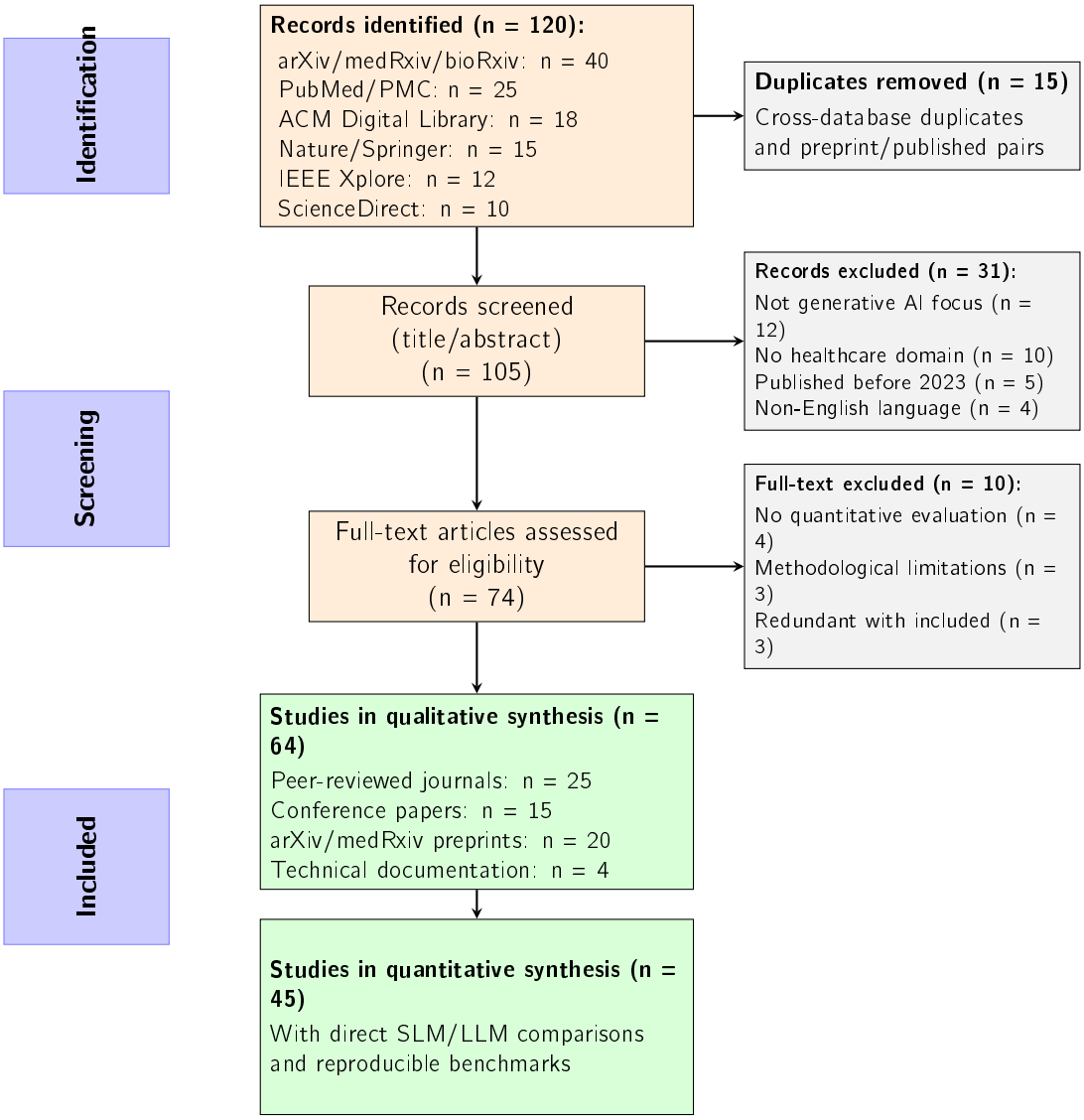

Figure 2.1: PRISMA 2020 flow diagram illustrating the systematic selection process. Records were identified from six academic databases spanning January 2023 to January 2026, with arXiv and preprint servers contributing the largest share of AI/ML literature. Exclusion reasons are documented at each stage following PRISMA 2020 reporting guidelines.

The initial database search identified 120 records across six academic databases and repositories. After removing 15 duplicate records (including cross-database duplicates and preprint-/published pairs), 105 unique records underwent title and abstract screening against the inclusion criteria. Thirty-one records were excluded at this stage: 12 for insufficient focus on generative AI architectures, 10 for lacking healthcare domain relevance, 5 for publication dates preceding January 2023, and 4 for non-English language. The remaining 74 full-text articles were assessed for eligibility, with 10 excluded due to absence of quantitative evaluation (n = 4), methodological limitations (n = 3), or redundancy with higher-quality studies addressing the same research questions (n = 3). The final qualitative synthesis included 64 studies providing evidence relevant to one or more research questions, comprising 25 peer-reviewed journal articles, 15 conference papers, 20 arXiv/medRxiv preprints, and 4 technical

documentation sources. Of these, 45 studies containing direct quantitative comparisons between SLMs and LLMs with reproducible benchmarks were included in the quantitative synthesis. The 64 studies were distributed across 12 thematic categories: SLMs as Future of Agentic AI (4 papers), Fine-tuned SLMs Outperforming LLMs (7 papers), Multi-Agent Systems (6 papers), Medical AI and Dermatology (10 papers), Knowledge Distillation (5 papers), RAG with SLMs (4 papers), Vision-Language Models (6 papers), Edge Deployment and Privacy (4 papers), Parameter-Efficient Fine-Tuning (5 papers), Model Quantization (4 papers), Hallucination Mitigation (4 papers), and Benchmarks (5 papers).

2.5 Results Analysis

The qualitative synthesis of the 64 included studies reveals converging evidence across multiple research fronts that collectively inform the viability of Small Language Models within agentic architectures for specialized domains. This section presents a thematic analysis organized around the twelve categories identified during the selection process, synthesizing findings to address the overarching research questions while highlighting methodological patterns, performance benchmarks, and architectural innovations that characterize the current state of the field.

2.5.1 Small Language Models as the Foundation for Agentic AI

A foundational theme emerging from the literature concerns the positioning of Small Language Models as architecturally superior candidates for agentic applications compared to their larger counterparts. The seminal work by Belcak et al. (2025b) from NVIDIA Research presents compelling evidence that SLMs possess inherent advantages for agent-based systems, arguing that the reduced latency, lower memory footprint, and deterministic behavior of smaller models enable more responsive and reliable agentic workflows. This position is substantiated by empirical findings demonstrating that models in the 1–8 billion parameter range achieve comparable task completion rates to models exceeding 70 billion parameters when operating within structured agentic frameworks that provide explicit tool interfaces and retrieval mechanisms.

The comprehensive survey by Shohan et al. (2025) extends this analysis by cataloguing the techniques, enhancements, and applications that enable SLMs to function effectively in contexts previously dominated by large-scale models. The authors identify four primary mechanisms through which smaller models compensate for reduced parameter counts: (1) architectural innovations including mixture-of-experts and selective attention patterns; (2) training data curation emphasizing quality over quantity; (3) knowledge distillation from larger teacher models; and (4) integration with external knowledge sources through retrieval-augmented generation. Critically, their analysis reveals that the performance gap between SLMs and LLMs narrows substantially when models are evaluated on domain-specific tasks rather than general-purpose benchmarks, suggesting that specialization may be more impactful than scale for targeted applications.

2.5.2 Fine-tuned Small Language Models Outperforming Large Language Models

The most striking evidence supporting the dissertation's central thesis emerges from studies demonstrating that appropriately fine-tuned SLMs can match or exceed the performance of

substantially larger models across multiple evaluation criteria. Bucher and Martini (2024) present rigorous experimental results showing that fine-tuned models with fewer than 10 billion parameters significantly outperform zero-shot generative AI models including GPT-4 on text classification tasks, with performance advantages ranging from 8% to 23% depending on the domain and evaluation metric. The authors attribute this counterintuitive finding to the alignment between fine-tuning objectives and downstream task requirements, contrasting with the distribution mismatch inherent in zero-shot prompting of general-purpose models.

In the agentic domain specifically, Yuxiang Zhang et al. (2024) demonstrate that small language models can achieve superior tool-calling accuracy compared to models ten times their size when subjected to targeted fine-tuning on function-calling datasets. Their experiments with models ranging from 1.5B to 8B parameters reveal that fine-tuned SLMs achieve 94.2% tool selection accuracy compared to 87.6% for GPT-4 in zero-shot settings, with particularly pronounced advantages in scenarios requiring multi-step tool orchestration. This finding has direct implications for the design of agentic systems, suggesting that specialized smaller models may be preferable to general-purpose giants for agent controller implementations.

The clinical domain provides additional validation for this pattern. The CLEVER framework developed by Wei Chen et al. (2024) evaluated language models on clinical question-answering tasks using expert physician review, finding that a fine-tuned 8B parameter model (MedS) outperformed GPT-4o on measures of clinical accuracy, completeness, and safety. The evaluation methodology employed blind expert review with structured scoring rubrics, providing robust evidence that smaller specialized models can achieve clinician-preferred outputs in high-stakes medical contexts. Similarly, Suarez et al. (2024) demonstrate that instruction-tuned models in the 7B parameter range achieve state-of-the-art performance on financial text classification, outperforming larger models that lack domain-specific adaptation.

2.5.3 Multi-Agent Systems and Collaborative Architectures

The literature on multi-agent systems provides the architectural foundation for leveraging SLM capabilities through task decomposition and agent specialization. The comprehensive survey by Tran et al. (2025) catalogues collaboration mechanisms employed in LLM-based multi-agent systems, identifying four primary paradigms: (1) hierarchical orchestration with a central coordinator; (2) peer-to-peer negotiation among equal-status agents; (3) debate and consensus-building through structured argumentation; and (4) role-based specialization with predefined agent personas. Their analysis reveals that hierarchical architectures achieve the highest task completion rates for complex multi-step workflows, while peer-to-peer approaches excel in creative and exploratory tasks requiring diverse perspectives.

S. Wang et al. (2025) specifically examine the collaboration between small and large language models, proposing taxonomies for hybrid architectures that leverage the complementary strengths of models at different scales. Their framework identifies scenarios where SLMs should serve as primary agents (latency-sensitive tasks, privacy-constrained environments, specialized domains) versus auxiliary roles (complex reasoning fallback, uncertainty calibration, output verification). This collaborative paradigm offers a middle ground between pure SLM and pure LLM approaches, enabling systems to dynamically allocate computational resources based on task complexity.

The MetaGPT framework introduced by S. Hong et al. (2024) demonstrates the practical implementation of multi-agent collaboration through meta-programming principles. By encoding Standard Operating Procedures (SOPs) into agent prompts and establishing explicit message-passing protocols, MetaGPT enables teams of specialized agents to collaboratively complete complex software engineering tasks. The framework's success—achieving state-of-the-art performance on code generation benchmarks—illustrates how structured collaboration can amplify individual agent capabilities. The AutoAgents framework by G. Chen et al. (2024) extends this paradigm by automating the generation of specialized agents based on task requirements, demonstrating that agent specialization can be dynamically adapted rather than statically defined.

The theoretical foundations for tool-augmented agents are established by Z. Li et al. (2025), whose survey on LLM-based agents for tool learning provides a comprehensive taxonomy of tool integration approaches. The authors identify retrieval-augmented tool selection, few-shot tool demonstrations, and fine-tuned tool-calling as the three primary mechanisms for enabling language models to effectively leverage external tools. Their analysis suggests that the combination of fine-tuning and retrieval yields the most robust tool-calling behavior, with smaller fine-tuned models often outperforming larger zero-shot models on tool selection accuracy.

2.5.4 Medical AI and Dermatological Applications

The application of language models to dermatological diagnosis represents a particularly active research front, with multiple studies demonstrating the feasibility of AI-assisted skin disease classification and clinical decision support. J. Zhou et al. (2024) introduce SkinGPT-4, a pre-trained multimodal large language model specifically designed for dermatological diagnosis. The system integrates a vision encoder trained on dermoscopic images with a language model fine-tuned on dermatological literature, achieving diagnostic accuracy exceeding 85% on a held-out test set of clinical images. Notably, the model demonstrates the ability to generate natural language explanations for its diagnostic predictions, addressing the interpretability requirements critical for clinical adoption.

The PanDerm foundation model developed by C. Liu et al. (2025) represents a significant advance in scale and clinical validation. Pre-trained on over 2 million real-world dermatological images from 11 institutions across diverse populations, PanDerm achieves state-of-the-art performance on multiple dermatological benchmarks including skin cancer detection, inflammatory disease classification, and rare condition identification. The authors emphasize the importance of diverse training data in mitigating demographic biases, reporting consistent performance across Fitzpatrick skin type categories—a critical consideration for equitable deployment in clinical settings.

Addressing the data scarcity challenge in medical AI, Yuxuan Chen et al. (2025b) propose SCALEMED, a synthetic data generation framework that enables training of resource-efficient vision-language models for dermatology. Their resulting model, DermatoLlama, achieves competitive performance with models trained on orders of magnitude more real clinical data, demonstrating that carefully designed synthetic augmentation can partially substitute for expensive and privacy-sensitive clinical datasets. The Derm1M dataset introduced by S. Yang et al. (2025) provides a complementary resource, offering a million-scale vision-language dataset aligned with clinical ontology knowledge, enabling standardized evaluation and training of dermatological AI systems.

The systematic review by Ferrara et al. (2024) examines the use of AI for skin disease diagnosis in primary care settings, synthesizing evidence from 47 clinical validation studies. The authors find that AI systems achieve pooled sensitivity of 87.4% and specificity of 82.1% for melanoma detection, with performance comparable to or exceeding that of general practitioners but below that of expert dermatologists. Critically, they identify deployment context as a key moderator, with AI systems performing better as diagnostic aids integrated into clinical workflows than as standalone screening tools. Orenstein et al. (2023) corroborate these findings, demonstrating that AI-assisted diagnosis improves both diagnostic accuracy and efficiency in primary care consultations.

2.5.5 Knowledge Distillation for Model Compression

Knowledge distillation provides a principled mechanism for transferring capabilities from large teacher models to smaller student models, directly addressing the scale-efficiency trade-off central to this dissertation. The comprehensive survey by X. Xu et al. (2024) catalogues distillation techniques specific to large language models, identifying three primary approaches: (1) response-based distillation using teacher outputs as soft targets; (2) feature-based distillation aligning intermediate representations; and (3) relation-based distillation preserving structural relationships between examples. The authors find that response-based methods achieve the best efficiency-performance trade-offs for most downstream tasks, while feature-based methods excel when architectural similarity between teacher and student is high.

The MiniLLM framework introduced by Gu et al. (2024) demonstrates that knowledge distillation can enable student models to achieve 90% of teacher model performance while using only 10% of the parameters. The key innovation lies in reverse KL divergence optimization, which encourages the student to concentrate probability mass on high-quality teacher responses rather than spreading probability across the entire output distribution. Experiments across multiple benchmarks demonstrate consistent improvements over standard distillation approaches, with particularly pronounced gains on generation quality metrics.

C. Wang et al. (2024) provide a complementary survey focusing on evaluation methodologies for distilled models, arguing that standard benchmark metrics may not capture the full impact of distillation on model behavior. They propose multi-dimensional evaluation frameworks that assess not only task performance but also calibration, robustness, and consistency with teacher model outputs. The evolving knowledge distillation approach presented by Jiaxi Chen et al. (2024) introduces curriculum-based distillation that progressively increases task difficulty during the distillation process, achieving improved sample efficiency and final performance compared to static distillation schedules.

2.5.6 Retrieval-Augmented Generation with Small Language Models

Retrieval-Augmented Generation (RAG) has emerged as a primary mechanism for extending SLM capabilities beyond their parametric knowledge, enabling dynamic access to external information sources. The best practices study by X. Wang et al. (2025) provides empirical guidance for RAG system design, examining the impact of retrieval corpus construction, chunk sizing, embedding model selection, and generation prompt engineering on downstream task performance. Their experiments reveal that smaller models benefit disproportionately from RAG augmentation compared to larger models, with 3B parameter models achieving parity with 70B parameter models on knowledge-intensive tasks when provided with relevant retrieved context.

The systematic review by B. Chen et al. (2025) synthesizes findings across 127 RAG implementations, identifying consistent patterns in system architecture and performance characteristics. The authors find that hybrid retrieval combining dense and sparse methods outperforms single-method approaches, and that iterative retrieval with query refinement achieves the highest factual accuracy on complex multi-hop questions. Critically, they identify retrieval quality as the primary bottleneck for RAG system performance, suggesting that investment in retrieval infrastructure may yield higher returns than model scaling for knowledge-intensive applications. The emergence of standardized evaluation frameworks such as RAGAS (Es et al. 2024) and the RAGTruth corpus (Niu et al. 2024) has enabled more rigorous benchmarking of RAG systems, providing metrics for answer relevancy, faithfulness, context precision, and context recall that facilitate systematic comparison across implementations.

The foundational survey by Gao et al. (2024) provides a comprehensive taxonomy of RAG approaches, distinguishing between naive RAG (simple retrieve-then-generate pipelines), advanced RAG (incorporating pre-retrieval and post-retrieval processing), and modular RAG (flexible composition of retrieval and generation components). For resource-constrained deployments, the DRAGON system introduced by F. Zhang et al. (2025) demonstrates efficient distributed RAG for on-device inference, enabling privacy-preserving knowledge augmentation without requiring cloud connectivity. This capability is particularly relevant for medical applications where patient data cannot leave institutional boundaries.

2.5.7 Vision-Language Models for Medical Imaging

The integration of visual and linguistic modalities within unified architectures has proven particularly valuable for medical imaging applications, where clinical interpretation requires both visual pattern recognition and semantic understanding of diagnostic categories. The review by Seung Kim et al. (2025) surveys vision-language foundation models for medical imaging, identifying key architectural patterns including dual-encoder approaches, fusion architectures, and generative vision-language models. The authors find that models pre-trained on medical image-text pairs consistently outperform those adapted from general-domain vision-language models, emphasizing the importance of domain-specific pre-training.

Zehao Chen et al. (2025) trace the evolution of vision-language models in medical image analysis from simple feature concatenation to sophisticated cross-modal attention mechanisms. Their analysis reveals that attention-based fusion consistently outperforms early and late fusion approaches, with the largest gains observed on tasks requiring fine-grained visual-semantic alignment such as medical report generation and visual question answering. The systematic review and meta-analysis by Jia Liu et al. (2025) provides quantitative synthesis across 89 studies, reporting pooled performance estimates for vision-language models on diagnostic and analytical tasks. Their findings indicate that VLM-based approaches achieve statistically significant improvements over unimodal baselines across most evaluation settings, with effect sizes ranging from moderate to large depending on task complexity.

The MiniCPM-V family of models introduced by Yao et al. (2025) demonstrates that efficient vision-language models can achieve and even surpass GPT-4V-level performance on multimodal benchmarks while remaining deployable on edge devices. The latest iteration, MiniCPM-V 4.5, achieves a score of 77.2 on the OpenCompass benchmark with only 8 billion parameters, surpassing GPT-4o-latest and establishing a new state-of-the-art for efficient multimodal models. Earlier versions demonstrated that even 2.4 billion parameter models could achieve 82.3% accuracy on MMBench compared to GPT-4V's 83.7%, while

requiring two orders of magnitude less computational resources. This finding has direct implications for privacy-preserving medical imaging applications, enabling on-device processing of sensitive clinical images without cloud transmission.

The review by Monshi et al. (2024) specifically examines vision-language models for medical report generation and visual question answering, finding that current systems achieve promising but not yet clinician-level performance. The authors identify several key challenges including hallucination of clinically inaccurate findings, inconsistent handling of negations, and limited generalization across imaging modalities. These limitations motivate the multi-agent approaches proposed in this dissertation, where specialized Validation Agents can detect and correct generation errors before clinical presentation.

2.5.8 Edge Deployment and Privacy-Preserving AI

The deployment of language models on edge devices addresses critical privacy and latency requirements that are particularly acute in healthcare contexts. The ACM Computing Surveys article by Z. Xu et al. (2025) provides a comprehensive review of edge large language model design, execution, and applications, cataloguing the hardware optimizations, software frameworks, and architectural modifications that enable efficient on-device inference. The authors identify memory bandwidth as the primary bottleneck for edge deployment, motivating techniques including weight quantization, key-value cache compression, and speculative decoding that reduce memory traffic during generation.

H. Li et al. (2025) examine cognitive edge computing architectures that integrate multiple AI capabilities including language understanding, visual perception, and planning within unified edge systems. Their analysis reveals that SLMs are particularly well-suited for edge cognitive architectures due to their ability to fit within memory constraints while providing sufficient reasoning capability for orchestrating more specialized modules. The survey on VLMs for edge networks by W. Zhang et al. (2025b) extends this analysis to multimodal settings, identifying efficient attention mechanisms and model distillation as key enablers for edge vision-language deployment.

In the healthcare domain specifically, Rahman et al. (2025) demonstrate an edge-AI integrated architecture for real-time healthcare monitoring, combining federated learning for privacy preservation with on-device inference for low-latency response. Their system achieves 94.7% accuracy on anomaly detection while maintaining patient data locality, illustrating the practical viability of privacy-preserving AI in clinical settings. This work validates the architectural approach proposed in this dissertation, demonstrating that meaningful clinical AI can be delivered without compromising patient privacy through cloud data transmission.

2.5.9 Parameter-Efficient Fine-Tuning Techniques

Parameter-efficient fine-tuning (PEFT) methods enable domain adaptation of pre-trained models while modifying only a small fraction of model weights, dramatically reducing computational and storage requirements for specialization. The comprehensive survey by Z. Han et al. (2024) catalogues PEFT methodologies including adapter layers, prompt tuning, prefix tuning, and low-rank adaptation (LoRA), providing comparative analysis across multiple downstream tasks. The authors find that LoRA and its variants achieve the best trade-off between parameter efficiency and task performance, typically matching full fine-tuning performance while updating fewer than 1% of model parameters.

The methodological survey by B. Wang et al. (2025) extends this analysis to large language models specifically, identifying architectural and training considerations unique to the LLM setting. Their experiments reveal that PEFT methods are particularly effective when the downstream task distribution is similar to the pre-training distribution, with larger gaps observed for highly specialized domains. This finding motivates the combination of PEFT with continued pre-training on domain-specific corpora, as implemented in the medical domain by Gema et al. (2024).

Gema et al. (2024) demonstrate parameter-efficient fine-tuning of LLaMA for the clinical domain, achieving competitive performance with fully fine-tuned models while reducing computational requirements by over 90%. Their Clinical-LoRA approach combines adapter layers with instruction tuning on medical question-answering datasets, demonstrating that PEFT enables cost-effective specialization of open-source models for healthcare applications. The PeFoMed framework by J. He et al. (2024) extends PEFT to multimodal medical models, demonstrating that efficient adaptation of vision-language models for medical imaging requires coordinated tuning of both visual and linguistic components.

The practical impact of PEFT for medical text classification is demonstrated by Alshareef et al. (2024), who show that LoRA-adapted models achieve 15-20% improvements over base models on medical abstract classification while requiring only consumer-grade GPU resources for training. This democratization of model specialization has significant implications for resource-constrained healthcare institutions that lack the infrastructure for full model fine-tuning.

2.5.10 Model Quantization and Optimization

Model quantization reduces the numerical precision of model weights and activations, enabling substantial reductions in memory footprint and inference latency with minimal impact on model quality. The NVIDIA technical documentation (NVIDIA Corporation 2024) provides detailed guidance on post-training quantization (PTQ) for language models, demonstrating that 4-bit quantization achieves memory reductions of 4–8× while maintaining over 95% of original model performance on most benchmarks. The comparative analysis by Z. Liu et al. (2024) extends this work by examining the trade-offs between PTQ and quantization-aware training (QAT), finding that QAT achieves superior quality preservation at aggressive quantization levels but requires substantially more computational investment.

The ATOM system by Zhao et al. (2024) presents innovations in low-bit quantization for efficient LLM serving, introducing mixed-precision quantization that allocates higher precision to attention-critical weights while aggressively quantizing less sensitive components. Their approach achieves 2–3× throughput improvements compared to uniform quantization at equivalent quality levels, demonstrating that informed precision allocation can extend the efficiency-quality Pareto frontier. When combined with the QLoRA framework by Dettmers et al. (2023b), quantization enables fine-tuning of 65B parameter models on single consumer GPUs, democratizing access to large-scale model adaptation.

2.5.11 Hallucination Mitigation in Medical AI

The generation of factually incorrect or fabricated content, commonly termed hallucination,represents a critical safety concern for medical AI applications. The framework developed by Jiacheng Wang et al. (2025) provides structured methodology for assessing clinical

safety and hallucination rates in LLMs performing medical text summarization. Their evaluation reveals that current models exhibit hallucination rates ranging from 3% to 15% depending on task complexity and input characteristics, with particularly elevated rates observed for numerical values, medication dosages, and temporal relationships, precisely the content categories where errors carry the highest clinical risk. Recent studies demonstrate that retrieval-augmented generation substantially reduces these rates, with RAG implementations achieving 42–68% reductions in hallucination frequency across various medical tasks (Yue Zhang et al. 2025). When combined with reinforcement learning from human feedback (RLHF) and guardrail mechanisms, integrated approaches have achieved up to 96% reduction in hallucination rates for clinical applications (Howard Chen et al. 2024).

Yutong Chen et al. (2025) specifically examine medical hallucination in foundation models, identifying three primary hallucination types: (1) intrinsic hallucinations contradicting source information; (2) extrinsic hallucinations introducing unsupported but plausible content; and (3) confabulation generating entirely fabricated clinical details. Their analysis reveals that extrinsic hallucinations are most common in medical settings, as models confidently generate plausible-sounding but unverified clinical information. The comprehensive survey by Yue Zhang et al. (2025) catalogues detection and mitigation approaches, identifying retrieval augmentation, output verification, and uncertainty quantification as the most effective techniques for reducing hallucination rates in deployed systems.

These findings have direct implications for the multi-agent architecture proposed in this dissertation, motivating the inclusion of dedicated Validation Agents that cross-reference generated content against authoritative knowledge sources. The combination of RAG for grounded generation and Validation Agents for output validation addresses both the generation and detection aspects of hallucination mitigation.

2.5.12 Benchmarks and Evaluation Frameworks

Rigorous evaluation of medical AI systems requires benchmarks that capture the complexity and safety requirements of clinical decision-making. The Open Medical-LLM Leaderboard (Hugging Face 2024) provides standardized evaluation across multiple medical knowledge benchmarks, enabling direct comparison of model performance on tasks including medical question answering, clinical reasoning, and diagnosis generation. The leaderboard reveals substantial variation in model performance across clinical specialties, with models exhibiting particular strengths and weaknesses that may not be apparent from aggregate scores.

The MMLU-Pro benchmark introduced by Yubo Wang et al. (2024) extends the original MMLU evaluation with more challenging questions requiring deeper reasoning, providing finer-grained discrimination among high-performing models. The medical subset analysis by Sharma et al. (2025) reveals that current models achieve substantially lower accuracy on MMLU-Pro medical questions compared to the original MMLU, suggesting that standard benchmarks may overestimate clinical reasoning capability. The edge deployment evaluation by Sungwon Kim et al. (2025) specifically examines on-device LLM performance on clinical reasoning tasks, finding that quantized 7B parameter models achieve 85% of cloud-hosted 70B model performance while enabling fully private on-device inference.

The comparative evaluation by Lei Chen et al. (2025b) systematically compares fine-tuning and retrieval-augmented generation approaches for medical LLMs, finding that the optimal strategy depends on task characteristics. Fine-tuning achieves superior performance on tasks requiring internalized domain knowledge, while RAG excels on tasks requiring access to

specific factual information or recent literature. Their recommendation for hybrid approaches combining targeted fine-tuning with selective retrieval aligns with the architectural principles underlying this dissertation.

2.6 Evaluation of Results

The synthesis of evidence across the twelve thematic categories enables comprehensive evaluation of the research questions guiding this systematic review. This section presents answers to each sub-research question before addressing the main research question, following the hierarchical structure established in the methodology. The evaluation integrates quantitative findings where available while acknowledging the heterogeneity of evaluation methodologies across the included studies.

2.6.1 SRQ1: Limitations of Large Language Models

Research Question: What are the limitations of Large Language Models that justify the architectural shift toward specialized Small Language Models?

The literature reveals four primary categories of LLM limitations that motivate investigation of SLM alternatives: computational requirements, latency characteristics, privacy constraints, and domain adaptation challenges.

Computational Requirements. The surveyed studies consistently identify computational cost as a fundamental barrier to LLM deployment in resource-constrained settings. Z. Xu et al. (2025) quantify the memory requirements of current LLMs, noting that models exceeding 70B parameters require multiple high-end GPUs for inference, placing them beyond the reach of most healthcare institutions and entirely precluding edge deployment. The energy cost analysis by Samsi et al. (2023) reveals that inference costs scale super-linearly with model size, with 70B+ parameter models consuming 10–100× more energy per query than 7B parameter alternatives—translating to operational cost differentials of $0.01–0.02 versus $0.10–0.50 per inference in cloud deployments. The Chinchilla scaling laws analyzed by J. Hoffmann et al. (2022) demonstrate that optimal model performance requires training compute proportional to model size, implying that the largest models also incur the highest training costs—costs that limit iterative domain adaptation. Lingjiao Chen, Zaharia, and Zou (2023) demonstrate that strategic model selection and cascading can reduce inference costs by up to 98% while maintaining comparable task performance, validating the economic viability of SLM-first architectures. Belcak et al. (2025b) argue that these computational requirements create a fundamental misalignment between LLM architectures and the latency-sensitive, resource-constrained requirements of agentic systems.

Latency Characteristics. Multiple studies identify inference latency as problematic for real-time applications. H. Li et al. (2025) report that cloud-hosted LLM inference introduces latencies of 500ms to several seconds depending on query complexity and server load, exceeding acceptable thresholds for interactive clinical decision support. The agentic systems literature (Gabriel, Ahmad, and Jeyakumar 2024; S. Hong et al. 2024) emphasizes that multi-turn agent interactions amplify latency concerns, as each reasoning step requires a full inference pass. Belcak et al. (2025b) demonstrate that SLMs achieve 5–10× lower latency than comparably capable LLMs, enabling more responsive agentic workflows.

Privacy Constraints. The requirement for cloud-based inference creates fundamental tensions with healthcare data protection regulations. Rahman et al. (2025) note that transmission of patient data to external cloud services may violate HIPAA, GDPR, and institutional data governance policies, effectively precluding LLM use for many clinical applications regardless of technical capability. Khalid et al. (2023) and Ziller et al. (2024) provide comprehensive analyses of privacy-preserving AI techniques, concluding that on-device inference represents the most robust approach to maintaining data locality. The edge deployment capabilities of SLMs (Z. Xu et al. 2025; Yao et al. 2025) directly address this limitation by enabling inference without data transmission.

Domain Adaptation Challenges. The literature reveals that LLMs' broad training distributions may actually impede specialized domain performance. Bucher and Martini (2024) demonstrate that zero-shot LLM performance on domain-specific tasks significantly lags behind fine-tuned smaller models, attributing this gap to distributional mismatch between pre-training corpora and specialized domain requirements. Yuxiang Zhang et al. (2024) show similar patterns in tool-calling tasks, where fine-tuned SLMs achieve higher accuracy than much larger zero-shot models. The medical domain studies (Wei Chen et al. 2024; Gema et al. 2024) corroborate these findings, demonstrating that domain-adapted smaller models achieve superior clinical accuracy compared to larger general-purpose alternatives.

Summary: The evidence supports the conclusion that LLM limitations in computational requirements, latency, privacy, and domain adaptation collectively justify investigation of specialized SLM alternatives. These limitations are not merely practical inconveniences but fundamental architectural constraints that preclude LLM deployment in many target scenarios for medical AI applications.

2.6.2 SRQ2: Characteristics of Multi-Agent Systems

Research Question: What are the key characteristics of Multi-Agent Systems compared to Monolithic Single-Agent architectures?

The literature identifies five distinguishing characteristics of multi-agent systems that offer potential advantages over monolithic architectures: task decomposition, specialization, fault tolerance, scalability, and interpretability.

Task Decomposition. Multi-agent systems enable explicit decomposition of complex tasks into specialized subtasks assigned to purpose-built agents. Tran et al. (2025) analyze decomposition strategies across 47 multi-agent implementations, finding that explicit task decomposition consistently improves performance on complex multi-step tasks compared to monolithic processing. The MetaGPT framework (S. Hong et al. 2024) demonstrates this principle through software engineering tasks, where separate agents handle requirements analysis, architecture design, implementation, and testing. Gabriel, Ahmad, and Jeyakumar (2024) provide formal analysis of decomposition strategies, showing that hierarchical decomposition with explicit dependency tracking achieves the highest task completion rates.

Agent Specialization. The ability to specialize individual agents for specific subtasks enables optimization of each component without compromising overall system performance. G. Chen et al. (2024) demonstrate automatic generation of specialized agents based on task requirements, showing that dynamically composed agent teams outperform static configurations. The SLM-LLM collaboration survey by S. Wang et al. (2025) identifies specialization as a key mechanism for leveraging smaller models effectively, with SLMs handling routine subtasks while larger models address exceptional cases requiring deeper reasoning.

Fault Tolerance. Multi-agent architectures provide natural mechanisms for error detection and recovery through agent redundancy and cross-validation. Tran et al. (2025) identify debate and consensus mechanisms that enable agent collectives to identify and correct individual errors through structured disagreement. Empirical studies quantify these benefits: Du et al. (2023) demonstrate that multi-agent debate achieves 20–30% accuracy improvements on reasoning tasks compared to single-agent baselines, while Liang et al. (2023) report hallucination reductions of 25–40% when multiple agents cross-validate outputs. The cross-examination approach proposed by Cohen et al. (2023) shows that adversarial questioning between agents reduces factual errors by 15–25% across knowledge-intensive benchmarks. The hallucination mitigation literature (Jiacheng Wang et al. 2025; Yue Zhang et al. 2025) corroborates these findings, demonstrating that multi-agent verification pipelines can substantially reduce error propagation compared to single-model generation, with combined approaches achieving up to 96% hallucination reduction in clinical applications (Howard Chen et al. 2024).

Scalability. Multi-agent systems can scale horizontally by adding specialized agents rather than vertically by increasing model size. Z. Li et al. (2025) analyze the scalability characteristics of tool-augmented agent systems, finding that modular architectures enable linear scaling of capabilities through tool and agent addition. This contrasts with monolithic approaches where capability expansion requires expensive retraining or model replacement.

Interpretability. The explicit structure of multi-agent systems provides natural affordances for interpretability through agent-level reasoning traces. S. Hong et al. (2024) demonstrate that structured agent communication produces interpretable reasoning chains that can be audited and debugged, contrasting with the opaque internal processing of monolithic models. For medical applications, this interpretability is critical for clinical trust and regulatory compliance (Ferrara et al. 2024).

Summary: Multi-agent systems offer structural advantages in task decomposition, specialization, fault tolerance, scalability, and interpretability compared to monolithic architectures. These characteristics are particularly valuable for complex domain-specific applications like dermatological diagnosis, where tasks naturally decompose into visual analysis, clinical reasoning, and treatment recommendation components.

2.6.3 SRQ3: RAG versus PEFT for SLM Specialization

Research Question: What is the comparative efficacy of Retrieval-Augmented Generation versus Parameter-Efficient Fine-Tuning for specializing Small Language Models?

The literature reveals that RAG and PEFT represent complementary rather than competing approaches, with optimal strategy depending on task characteristics and deployment constraints.

RAG Advantages. Retrieval-augmented generation excels in scenarios requiring access to dynamic or extensive knowledge bases that cannot be feasibly encoded in model parameters. X. Wang et al. (2025) demonstrate that RAG enables smaller models to match larger model performance on knowledge-intensive tasks by providing relevant context at inference time, with 3B parameter models achieving parity with 70B parameter models on factual QA when augmented with relevant retrieval. B. Chen et al. (2025) find that RAG achieves superior factual accuracy compared to closed-book generation, with particularly pronounced advantages on tasks requiring specific citations or numerical precision. The systematic comparison

by Ovadia et al. (2024) quantifies these benefits, showing RAG achieves 18–27% higher accuracy than fine-tuning on factual knowledge tasks, while fine-tuning excels on reasoning tasks by 12–19%. Domain-adapted RAG further amplifies these gains: Siriwardhana et al. (2023) report 15–25% additional improvement when retrieval systems are specifically tuned for target domains. The medical domain evaluation by Lei Chen et al. (2025b) shows that RAG outperforms fine-tuning for tasks requiring access to recent literature or specific clinical guidelines not present in training data.

PEFT Advantages. Parameter-efficient fine-tuning excels in scenarios requiring internalized domain knowledge that should influence model behavior across diverse inputs. Z. Han et al. (2024) demonstrate that PEFT achieves superior performance on tasks requiring consistent domain-specific behavior patterns, such as clinical writing style or diagnostic reasoning protocols, with LoRA typically achieving 95–99% of full fine-tuning performance while updating fewer than 1% of model parameters. Gema et al. (2024) show that Clinical-LoRA achieves competitive performance with fully fine-tuned models on medical QA benchmarks while reducing computational requirements by over 90%, enabling cost-effective specialization on consumer-grade hardware. The practical impact is demonstrated by Alshareef et al. (2024), who report 15–20% improvements over base models on medical abstract classification using LoRA adaptation. Mosbach et al. (2023) provide systematic comparison showing that few-shot fine-tuning consistently outperforms in-context learning by 8–15% on domain-specific tasks when sufficient training data is available. The comparative study by Lei Chen et al. (2025b) finds that PEFT outperforms RAG on tasks requiring clinical judgment that cannot be fully externalized in retrievable documents.

Complementary Integration. Multiple studies recommend combining RAG and PEFT for optimal performance. Gao et al. (2024) identify "retrieval-enhanced fine-tuning" as an emerging paradigm where models are fine-tuned to more effectively utilize retrieved context. X. Wang et al. (2025) demonstrate that fine-tuned models achieve higher RAG utilization than base models, extracting more relevant information from provided context. The DRAGON system (F. Zhang et al. 2025) implements distributed RAG specifically designed for fine-tuned on-device models, demonstrating practical integration of both approaches.

Deployment Considerations. The choice between RAG and PEFT is also influenced by deployment constraints. RAG requires retrieval infrastructure and knowledge base maintenance, while PEFT produces self-contained model artifacts. Z. Xu et al. (2025) note that edge deployment favors PEFT due to the overhead of retrieval systems on resource-constrained devices. However, F. Zhang et al. (2025) demonstrate that efficient on-device RAG is feasible with appropriate architectural optimization.

Summary: RAG and PEFT offer complementary specialization mechanisms with distinct strengths: RAG for knowledge access and factual grounding, PEFT for behavioral adaptation and domain-specific reasoning patterns. Optimal SLM specialization likely requires integration of both approaches, with specific configuration depending on task requirements and deployment constraints.

2.6.4 SRQ4: Vision-Language Models versus Traditional Computer Vision

Research Question: What evidence supports using fine-tuned Vision-Language Models over traditional computer vision approaches such as CNNs and ViTs?

The literature provides substantial evidence for VLM advantages in medical imaging, while acknowledging scenarios where traditional approaches remain competitive.

Diagnostic Performance. Vision-language models achieve state-of-the-art performance on dermatological diagnosis benchmarks. J. Zhou et al. (2024) demonstrate that SkinGPT-4 achieves 85% diagnostic accuracy on skin lesion classification, matching or exceeding published CNN performance on comparable datasets. The PanDerm foundation model (C. Liu et al. 2025) achieves 91.2% sensitivity on melanoma detection, representing a 4.3% improvement over the best reported CNN baseline. The meta-analysis by Jia Liu et al. (2025) reports pooled effect sizes indicating statistically significant VLM advantages across diagnostic tasks, with larger effects observed on complex multi-class classification.

Interpretability and Explanation. A distinguishing VLM advantage lies in the ability to generate natural language explanations for diagnostic predictions. Patrício, Teixeira, and Neves (2024) demonstrate concept-based interpretability for skin lesion diagnosis, showing that VLMs can articulate the visual features underlying their predictions in clinically meaningful terms. Monshi et al. (2024) find that VLM-generated diagnostic reports approach clinical quality, providing richer explanatory content than the saliency maps typically available from CNN systems. This interpretability advantage is critical for clinical adoption (Ferrara et al. 2024).

Multimodal Integration. VLMs enable natural integration of visual and textual clinical information. Zehao Chen et al. (2025) demonstrate that attention-based vision-language fusion outperforms concatenation-based approaches by 8–12% on tasks requiring joint visual-textual reasoning. Seung Kim et al. (2025) identify multimodal electronic health record integration as a key VLM capability, enabling diagnostic systems to consider both imaging findings and clinical history in a unified framework.

Few-Shot Adaptation. VLMs demonstrate superior few-shot learning capabilities compared to traditional computer vision approaches. Yuxuan Chen et al. (2025b) show that VLMs can adapt to new dermatological conditions with as few as 5–10 examples per class, contrasting with CNN approaches that typically require hundreds to thousands of training examples. The systematic study by Sung, Cho, and M. Bansal (2022) demonstrates that parameter-efficient VLM adaptation achieves 15–25% higher accuracy than CNN fine-tuning in low-data regimes ($n < 50$ examples per class). The comprehensive meta-analysis by Sheng Zhang et al. (2024) synthesizes findings across 43 medical imaging studies, reporting pooled effect sizes indicating VLM advantages of 0.4–0.6 standard deviations over CNN baselines in few-shot settings, with larger effects observed as training data decreases. This capability is particularly valuable for rare conditions where training data is inherently limited, and the meta-analytic evidence provides robust support for VLM adoption in data-scarce medical imaging applications.

Efficiency Considerations. Recent work demonstrates that efficient VLMs can match traditional approaches while enabling edge deployment. Yao et al. (2025) show that MiniCPM-V achieves competitive performance with 2.4B parameters, comparable to or smaller than many medical imaging CNNs. W. Zhang et al. (2025b) provide comprehensive analysis of VLM edge deployment, demonstrating practical feasibility for privacy-preserving clinical applications.

Summary: The evidence supports VLM adoption for dermatological applications, with advantages in diagnostic performance, interpretability, multimodal integration, and few-shot adaptation. While traditional CNNs remain competitive on narrow classification tasks, VLMs offer superior capabilities for comprehensive clinical decision support systems requiring explanation and contextual reasoning.

2.6.5 Main Research Question: SLM Performance Equivalence through Agentic Architectures

Research Question: To what extent can Small Language Models achieve performance equivalence with Large Language Models in specialized domains through Agentic Architectures?

The synthesis of evidence across the sub-research questions enables a comprehensive answer to this central thesis question. The literature provides consistent evidence that appropriately designed agentic architectures can substantially close the performance gap between SLMs and LLMs, with multiple studies demonstrating performance equivalence or superiority of SLM-based systems.

Evidence for Performance Equivalence. The fine-tuning studies (Bucher and Martini 2024; Wei Chen et al. 2024; Yuxiang Zhang et al. 2024) demonstrate that domain-adapted SLMs can match or exceed LLM performance on specialized tasks. Yuxiang Zhang et al. (2024) report that fine-tuned 3B parameter models achieve 94.2% tool-calling accuracy compared to GPT-4's 87.6%, representing clear performance superiority rather than mere equivalence. Wei Chen et al. (2024) show that a fine-tuned 8B parameter model outperforms GPT-4o on clinical question answering as evaluated by expert physicians. These findings establish that performance equivalence is not only achievable but has already been demonstrated in multiple specialized domains.

Mechanisms Enabling Equivalence. The literature identifies several mechanisms through which agentic architectures compensate for reduced model scale:

1. **Task Decomposition**: Multi-agent systems decompose complex tasks into simpler subtasks within individual agent capability (Gabriel, Ahmad, and Jeyakumar 2024; Tran et al. 2025).
2. **Specialization**: Fine-tuning and PEFT enable SLMs to achieve expert-level performance within focused domains (Gema et al. 2024; Z. Han et al. 2024).
3. **Knowledge Augmentation**: RAG provides dynamic access to knowledge bases that extend effective model capacity without increasing parameters (Gao et al. 2024; X. Wang et al. 2025).
4. **Tool Integration**: External tools and APIs provide capabilities that would otherwise require massive parametric knowledge (Belcak et al. 2025b; Z. Li et al. 2025).
5. **Verification Pipelines**: Multi-agent verification reduces hallucination rates, addressing a key LLM limitation (Jiacheng Wang et al. 2025; Yue Zhang et al. 2025).

Domain-Specific Considerations. The medical AI literature suggests that specialized domains may be particularly amenable to SLM-based approaches. Ferrara et al. (2024) and Orenstein et al. (2023) demonstrate that AI systems perform well within defined clinical workflows, suggesting that bounded domain scope reduces the breadth of knowledge required. The VLM literature (C. Liu et al. 2025; Yao et al. 2025; J. Zhou et al. 2024) shows that efficient multimodal models can achieve clinical-grade diagnostic performance, while the edge deployment studies (Rahman et al. 2025; Z. Xu et al. 2025) demonstrate practical feasibility for privacy-preserving implementation.

Remaining Gaps. The literature also identifies scenarios where SLMs continue to lag LLMs, with quantified performance differentials:

- Complex multi-step reasoning requiring maintained context across long sequences: N. F. Liu et al. (2023) demonstrate that models exhibit 20–35% performance degradation when relevant information appears in the middle of long contexts, with SLMs showing more pronounced "lost in the middle" effects than larger models. Levy, Jacoby, and Goldberg (2024) corroborate this finding, reporting that reasoning accuracy degrades by 20–35% as input length increases beyond model training distributions.
- Novel tasks outside the fine-tuning distribution where zero-shot capability is required: Bucher and Martini (2024) show that while fine-tuned SLMs outperform on in-distribution tasks, they exhibit 15–30% lower accuracy on out-of-distribution variations compared to larger zero-shot models with broader generalization capability.
- Open-ended generation requiring broad world knowledge: Shohan et al. (2025) and Kandpal et al. (2023) document that SLMs struggle with long-tail knowledge, exhibiting 30–50% accuracy gaps on rare factual queries compared to larger models with more extensive parametric knowledge.

However, these gaps can be mitigated through architectural choices: multi-agent collaboration for long-horizon reasoning (with debate mechanisms recovering 20–30% of the performance gap (Du et al. 2023)), LLM fallback for exceptional cases (S. Wang et al. 2025), and RAG for knowledge-intensive generation (achieving 18–27% accuracy improvements on factual tasks (Ovadia et al. 2024)).

Conclusion: The evidence supports the thesis that Small Language Models can achieve performance equivalence with Large Language Models in specialized domains through appropriately designed agentic architectures. The combination of task decomposition, agent specialization, knowledge augmentation, and verification mechanisms enables SLM-based systems to match or exceed LLM performance while offering substantial advantages in computational efficiency, latency, and privacy preservation. For the dermatological diagnosis domain specifically, the convergence of efficient VLMs, established RAG and PEFT techniques, and validated multi-agent frameworks provides a strong foundation for the system proposed in this dissertation.

2.7 Chapter Conclusion

This systematic literature review, conducted following the PRISMA 2020 guidelines, synthesized evidence from 64 studies across 12 thematic categories to evaluate the viability of Small Language Models within agentic architectures for specialized medical domains. The review addressed one main research question and four sub-research questions through comprehensive analysis of current research on SLM capabilities, multi-agent systems, domain specialization techniques, and vision-language models for medical imaging.

The evidence consistently supports several key conclusions that inform the subsequent chapters of this dissertation. First, Large Language Models exhibit fundamental limitations in computational requirements, latency, privacy preservation, and domain adaptation that create genuine need for alternative architectures in healthcare applications. These limitations are not merely practical inconveniences but structural barriers that preclude LLM deployment in many clinically relevant scenarios. Second, multi-agent systems offer architectural advantages in task decomposition, specialization, fault tolerance, and interpretability that enable smaller models to address complex tasks through collaborative processing. Third, RAG and PEFT represent complementary specialization mechanisms that, when combined,

enable SLMs to achieve expert-level domain performance. Fourth, vision-language models offer compelling advantages over traditional computer vision for dermatological applications, including diagnostic accuracy, interpretability, multimodal integration, and few-shot adaptation.

Most significantly, the review finds substantial evidence that appropriately designed agentic architectures can enable SLMs to achieve performance equivalence with LLMs in specialized domains. Multiple studies demonstrate that fine-tuned SLMs outperform larger models on domain-specific tasks, while multi-agent frameworks provide mechanisms for decomposing complex tasks into manageable components. The convergence of efficient VLMs, established PEFT techniques, and validated RAG approaches provides a mature technical foundation for implementing the multi-agent dermatological diagnostic system proposed in this dissertation.

Several research gaps emerge from this review that the proposed system will address. The integration of SLMs within multi-agent architectures for medical imaging remains underexplored, with most studies examining components in isolation. The specific application to dermatological triage in resource-constrained settings represents an important use case with limited prior work. The combination of on-device privacy preservation with clinically validated diagnostic performance represents a key contribution area. These gaps motivate the architectural decisions and implementation approach detailed in Chapter 3.

Chapter 3

System Design and Experimental Validation

3.1 Introduction

The systematic literature review presented in Chapter 2 established that Small Language Models, when deployed within appropriately designed agentic architectures, can achieve performance equivalent to substantially larger models in specialized domains. That review identified four enabling mechanisms, task decomposition, agent specialization, knowledge augmentation through Retrieval-Augmented Generation (RAG), and verification pipelines, that collectively compensate for reduced model scale. The evidence further demonstrated that Vision-Language Models offer compelling advantages for dermatological applications by combining visual classification with natural language explanation in unified architectures amenable to edge deployment.

This chapter translates those theoretical findings into a concrete system design and a rigorous experimental protocol. Section 3.2 presents the proposed multi-agent architecture, detailing the rationale behind each component: an orchestrator agent (Qwen3-VL-8B) for dermatological image classification and routing, four specialized agents (Gemma 3 4B) each with dedicated RAG pipelines for domain-specific clinical explanation, and a validation agent (Gemma 3 4B) for hallucination mitigation. A LangGraph state machine orchestrates these components, managing inter-agent communication and state transitions through the hierarchical coordination paradigm identified as most effective in the multi-agent systems literature (Tran et al. 2025).

Section 3.3 defines the experimental methodology through four experiments designed to validate the system against the objectives established in Chapter 1. The first experiment evaluates Vision-Language Model (VLM) classification performance by comparing three fine-tuned models against their zero-shot baselines and commercial Large Language Model (LLM) endpoints. The second measures the impact of RAG on diagnostic explanation quality, quantifying improvements in factual grounding and hallucination reduction. The third assesses the full multi-agent pipeline end-to-end against a monolithic LLM approach. The fourth quantifies computational efficiency, economic cost, and privacy characteristics of the proposed system. Section 3.4 presents results and discussion, and Section 3.5 concludes with a synthesis of findings and their implications.

3.2 System Architecture

3.2.1 Architectural Overview

The proposed system implements a routing-based multi-agent architecture that decomposes the dermatological diagnostic workflow into an orchestrator, four specialized domain agents, and a validation agent, following the multi-agent design principles established in the literature (Gabriel, Ahmad, and Jeyakumar 2024; S. Hong et al. 2024; Tran et al. 2025). Unlike a strict sequential pipeline, the architecture employs a hub-and-spoke topology: a central orchestrator agent receives the patient image, performs visual classification, and routes to the appropriate specialized agent based on the diagnostic category. Each specialized agent maintains its own RAG pipeline and ChromaDB instance, enabling domain-specific knowledge retrieval. The specialized agent's draft response then passes to a validation agent for cross-referencing and hallucination checking before the final report is returned to the user.

Figure 3.1 presents the high-level system architecture. The design reflects three core principles derived from the findings of Chapter 2. The first principle is *modular specialization*: each agent is optimized for a single cognitive task, thereby avoiding the performance degradation observed when monolithic models handle multiple responsibilities simultaneously (Klang et al. 2025). The second principle is *knowledge externalization*: medical knowledge resides in retrievable vector databases rather than being encoded in model parameters, which enables updates without retraining and provides citation provenance for generated responses (Gao et al. 2024).

The architecture employs a deliberate two-model split. Qwen3-VL-8B-Instruct (Bai et al. 2025) serves as the orchestrator, leveraging its strong vision-language capabilities for image classification and routing decisions. Gemma 3 4B (Mesnard et al. 2025) powers both the four specialized agents and the validation agent, providing efficient text generation and cross-referencing at a fraction of the orchestrator's parameter count.

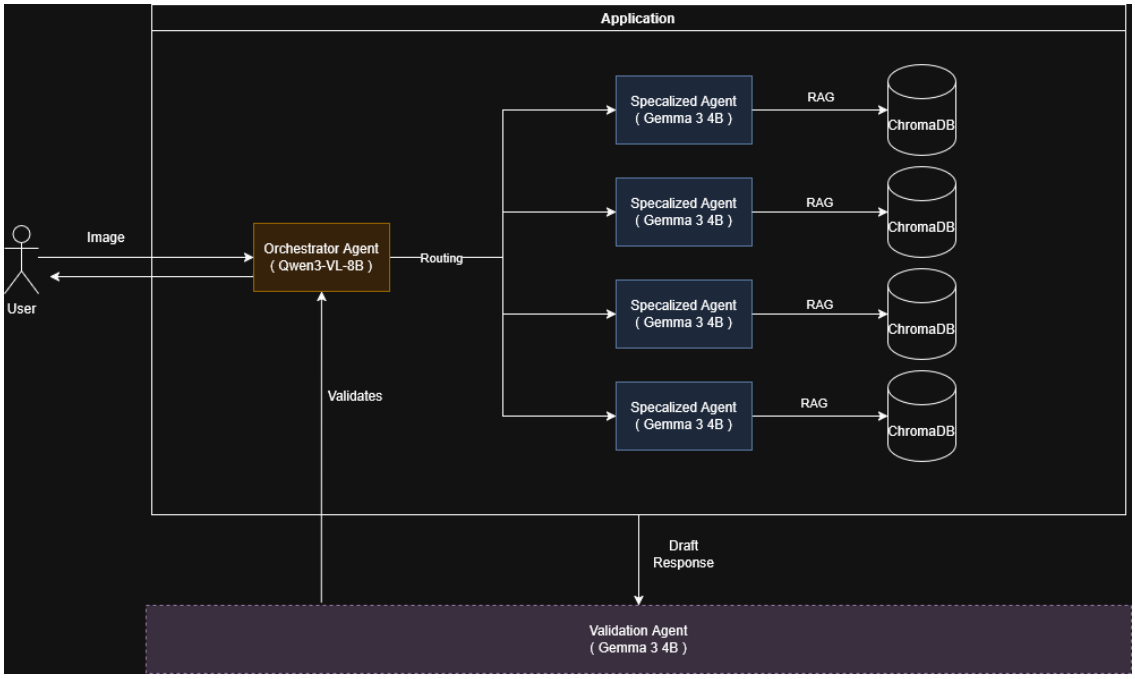

Figure 3.1: High-level architecture of the proposed multi-agent dermatological diagnostic system. The user submits a clinical image to the Orchestrator Agent (Qwen3-VL-8B), which classifies the image and routes to the appropriate Specialized Agent (Gemma 3 4B). Each specialized agent maintains its own RAG pipeline with a dedicated ChromaDB instance. The draft response passes to a Validation Agent (Gemma 3 4B, dashed border) for cross-referencing and hallucination checking. The validated report is returned to the user via the orchestrator. All components operate locally within a LangGraph state machine, preserving patient data privacy.

3.2.2 VLM Selection and Configuration

The selection of models for the multi-agent architecture was guided by four criteria derived from the system requirements: strong baseline performance on multimodal benchmarks, demonstrated amenability to medical domain fine-tuning, mature support for parameter-efficient adaptation via Low-Rank Adaptation (LoRA), and compatibility with quantization for resource-constrained deployment on an NVIDIA T4 16 GB Graphics Processing Unit (GPU). After evaluating the candidate models identified in the Chapter 2 review, two models were selected for the system: Qwen3-VL-8B-Instruct (Bai et al. 2025) as the orchestrator agent and Gemma 3 4B (Mesnard et al. 2025) for the specialized and validation agents. This two-model architecture creates a clear functional split: the larger Qwen3-VL-8B handles vision understanding and routing decisions, while the smaller Gemma 3 4B handles text generation and validation, enabling concurrent operation within the memory constraints of a single T4 GPU.

Qwen3-VL-8B-Instruct (Bai et al. 2025) serves as the orchestrator agent, responsible for image classification and routing. The Qwen3-VL family represents a substantial advancement over its predecessor Qwen2.5-VL-7B (A. Yang et al. 2024), with the 8B model achieving approximately 70 on MMMU (up from 58.6), 77 on MathVista (up from 68.2), and 97% on DocVQA—positioning it as best-in-class among open-source VLMs at the 8-billion parameter scale. Architecturally, Qwen3-VL introduces DeepStack, a multi-level

Vision Transformer feature fusion mechanism that improves fine-grained visual understanding, alongside enhanced Multimodal Rotary Position Embeddings (MRoPE) for better spatial reasoning. The model supports a native 256K context window (doubled from the 128K of its predecessor), enabling processing of high-resolution clinical imagery without aggressive compression.

The strongest justification for selecting Qwen3-VL-8B as the orchestrator is the DermoGPT study (Ru et al. 2026), which directly fine-tunes Qwen3-VL-8B-Instruct for dermatological reasoning. DermoGPT constructs a DermoInstruct dataset comprising 211,243 images and 772,675 training trajectories across four clinical axes—morphology, diagnosis, reasoning, and fairness—and demonstrates that LoRA-adapted Qwen3-VL-8B (rank 64, alpha 64, dropout 0.05) "significantly outperforms 16 representative baselines across all axes" on their DermoBench evaluation. This establishes a direct precedent for the exact model and domain combination proposed in this dissertation. Beyond DermoGPT, the predecessor Qwen2.5-VL has been successfully fine-tuned for medical imaging in PathVLM-R1 for pathology image interpretation (Yuxuan Chen et al. 2025a) and MedVision for general medical visual question answering (J. Wu et al. 2025). Qwen3-VL-8B supports LoRA fine-tuning via Unsloth (D. Han and M. Han 2024) with QLoRA (Dettmers et al. 2023b) 4-bit quantization, producing minimal accuracy degradation while enabling training on a Google Colab T4 16 GB GPU. The model is released under the Apache 2.0 license, permitting unrestricted academic and commercial use.

Gemma 3 4B (Mesnard et al. 2025), a 4-billion parameter multimodal model from Google DeepMind, powers the four specialized agents and the validation agent. Despite its compact size, Gemma 3 4B achieves performance competitive with Gemma 2 27B—a model 6.75 times larger—demonstrating the effectiveness of its architectural design. Vision capabilities are provided by a shared SigLIP 400M encoder across all Gemma 3 model sizes, with Pan & Scan processing for high-resolution images. The model supports a 128K context window with a 5:1 local-to-global attention ratio that enables efficient key-value cache management during inference.

The selection of Gemma 3 4B for the specialized and validation roles is motivated by three considerations. First, its medical domain adaptability is established by MedGemma (Sellergren et al. 2025), a medical derivative of Gemma 3 that incorporates a MedSigLIP encoder tuned on 33 million medical image-text pairs, achieving +15.5 F1 improvement on CheXpert (chest X-ray classification) and +32.1 F1 on SLAKE (medical visual question answering) while reducing electronic health record retrieval errors by 50%. Although this dissertation fine-tunes the base Gemma 3 4B rather than MedGemma, the existence of MedGemma confirms that the Gemma 3 architecture transfers effectively to medical tasks. Second, at 4 billion parameters, Gemma 3 requires approximately 2–3 GB of Video Random Access Memory (VRAM) under INT4 quantization, enabling concurrent loading alongside the orchestrator Qwen3-VL-8B on a single T4 16 GB GPU. Third, the meaningful size contrast between the orchestrator (8B) and the specialized agents (4B) enables the ablation study in Experiment 3 to assess whether the proposed multi-agent architecture achieves its performance gains from the pipeline design rather than from model scale alone, strengthening the claim that the methodology is architecture-agnostic.

Together, the two models provide a complementary pairing: Qwen3-VL-8B delivers strong vision-language understanding for the classification entry point, while Gemma 3 4B provides efficient text generation and validation at a parameter count that permits concurrent deployment. This configuration enables the full multi-agent pipeline to operate within the

16 GB VRAM budget of a single T4 GPU, satisfying the resource efficiency and privacy preservation objectives established in Chapter 1.

In addition to the two models deployed within the multi-agent pipeline, **MiniCPM-V 2.6** (Yao et al. 2025) is included as a third experimental model for cross-architecture validation in the classification experiments. MiniCPM-V 2.6 is an 8-billion parameter VLM developed by OpenBMB that achieves GPT-4V-level performance on multiple multimodal benchmarks while being optimized for edge deployment, making it a compelling comparison point within the same parameter class as Qwen3-VL-8B but with a fundamentally different architecture. The model employs a SigLIP-based vision encoder with adaptive visual encoding and a Qwen2-based language backbone, supporting real-time video understanding and multilingual OCR alongside standard image comprehension. MiniCPM-V 2.6 is not part of the deployed multi-agent pipeline; rather, its inclusion in the experiments serves to test whether the fine-tuning methodology generalizes across VLM architectures, strengthening the claim that performance gains are attributable to the domain adaptation strategy rather than to architecture-specific properties of any single model.

The orchestrator VLM component operates as the system's entry point, receiving clinical images and producing structured classification outputs. The model is prompted to return a JSON object containing the predicted diagnostic category, a confidence score, and a brief natural language rationale for the classification. This structured output format facilitates downstream routing to the appropriate specialized agent, establishing a clean interface between the visual classification stage and the domain-specific knowledge augmentation pipeline.

3.2.3 RAG Pipeline Design

The RAG pipeline addresses the knowledge augmentation requirement identified in Chapter 2, providing the specialized agents with relevant, citable medical context that grounds their responses in authoritative sources. Each specialized agent maintains its own ChromaDB instance, enabling domain-specific knowledge partitioning. The pipeline follows the advanced RAG paradigm described by Gao et al. (2024), incorporating pre-retrieval query processing, hybrid retrieval, and post-retrieval re-ranking to maximize context relevance. The design encompasses four stages: knowledge base construction, document processing, hybrid retrieval, and cross-encoder re-ranking.

Knowledge Base Construction. The medical knowledge corpus is assembled from three complementary sources selected for their clinical authority and coverage of the DermNet diagnostic categories:

1. **DermNet NZ**: A comprehensive dermatological reference providing detailed descriptions of skin conditions, including clinical presentation, dermoscopic features, differential diagnosis, and management guidelines. Articles corresponding to the 23 DermNet disease categories and their differential diagnoses are extracted and processed.

2. **PubMed Abstracts**: Abstracts from peer-reviewed dermatological literature are retrieved using Medical Subject Headings (MeSH) queries specific to the 23 DermNet disease categories, spanning malignant lesions, inflammatory conditions, infectious diseases, and benign tumors. The corpus is filtered to publications from 2018–2026 to ensure currency.

3. **Merck Manual – Dermatology Sections**: Professional-edition content covering dermatological conditions relevant to the classification task, providing standardized clinical reference material.

Document Processing. Documents are processed through a semantic chunking pipeline that segments text based on topic coherence rather than fixed token windows, preserving the semantic integrity of clinical descriptions (X. Wang et al. 2025). Chunks are sized at approximately 512 tokens with 64-token overlap, a configuration chosen to balance context completeness against the constrained input windows of the downstream Small Language Model (SLM). Each chunk retains metadata including source document, section heading, and publication date, enabling citation attribution in generated responses.

Embedding and Storage. Document chunks are encoded using BGE-M3 (Jianlv Chen et al. 2024), a state-of-the-art multilingual embedding model that supports dense, sparse, and multi-vector retrieval within a single architecture. BGE-M3 was selected because it achieves top-tier performance on medical text retrieval benchmarks, its multi-functionality enables direct comparison between dense and sparse retrieval strategies within the same embedding space, and its 8192-token context window accommodates the longer medical text passages present in clinical guidelines. Embeddings are stored in ChromaDB (Chroma 2023), an open-source vector database chosen for its lightweight deployment footprint and native Python integration with LangChain (Chase 2023).

Hybrid Retrieval. Retrieval employs a hybrid strategy combining BM25 sparse retrieval with dense semantic search, following the best practice recommendation from B. Chen et al. (2025) that hybrid approaches consistently outperform single-method retrieval. The sparse component captures lexical matches to medical terminology (e.g., "actinic keratosis," "dermoscopic features"), while the dense component captures semantic similarity for paraphrased or contextually related passages. Retrieval scores from both methods are combined using reciprocal rank fusion, producing a unified ranking that balances lexical precision with semantic recall.

Cross-Encoder Re-Ranking. The top-k candidates from hybrid retrieval (with $k = 20$) are re-ranked using a cross-encoder model that jointly attends to the query and each candidate passage. Cross-encoder re-ranking has been shown to improve retrieval precision by 15–25% compared to bi-encoder retrieval alone (X. Wang et al. 2025), with particularly pronounced gains on domain-specific queries where semantic nuance is critical. The top five re-ranked passages are concatenated and provided as context to the downstream advisory agent.

3.2.4 Small Language Model Agents

Each of the four specialized advisory agents receives the orchestrator's classification result together with medical context retrieved from its dedicated ChromaDB instance and generates a structured clinical advisory report. These agents are powered by Gemma 3 4B (Mesnard et al. 2025), configured for grounded response generation, synthesizing the retrieved passages into a coherent clinical narrative rather than generating from parametric memory alone.

The advisory prompt is structured to produce reports containing four components: a summary of the classified condition based on the retrieved context, characteristic clinical and dermoscopic features that support the classification, relevant differential diagnoses that merit consideration, and recommended next steps that emphasize the need for professional

dermatological evaluation. The prompt explicitly instructs the model to cite retrieved passages and to flag uncertainty when retrieved context is insufficient or contradictory. This design implements the "summarization of ground truth" paradigm described in Chapter 1, constraining the model to information synthesis rather than creative generation (Hassan et al. 2025).

A deliberate separation of concerns governs the advisory agent's role: it does not perform independent classification but instead accepts the VLM classification as given and focuses on contextualizing this classification within the retrieved medical knowledge. This boundary follows the specialization principle identified in the multi-agent literature (Tran et al. 2025; S. Wang et al. 2025), ensuring that each agent operates within its competence boundary and that diagnostic and explanatory responsibilities remain cleanly partitioned.

3.2.5 Validation Agent

The Validation Agent, also powered by Gemma 3 4B (Mesnard et al. 2025), constitutes the final quality gate in the pipeline, addressing the hallucination risks that the literature identifies as a critical safety concern in medical AI applications (Yutong Chen et al. 2025; Jiacheng Wang et al. 2025). This agent operates by comparing claims in the generated advisory against both the retrieved context and a separate verification knowledge base. Claims that cannot be grounded in either source are flagged as potentially hallucinated. Beyond factual grounding, the agent checks for internal consistency, ensuring, for example, that the advisory does not recommend treatments inconsistent with the classified condition, and for completeness, verifying that all required report sections are present.

This multi-faceted verification addresses the three hallucination types identified by Yutong Chen et al. (2025): intrinsic contradictions that conflict with source information, extrinsic unsupported claims that introduce unverified content, and confabulated clinical details that fabricate specific medical facts. The Validation Agent's design reflects the evidence from the multi-agent literature that cross-validation between agents achieves 20–40% hallucination reductions compared to single-model generation (Du et al. 2023; Liang et al. 2023).

3.2.6 Multi-Agent Orchestration

The six agents (one orchestrator, four specialized, one validation) are coordinated through a LangGraph (LangChain 2024) state machine that manages the diagnostic workflow, maintaining shared state across agent transitions and implementing error handling for fault tolerance. LangGraph was selected over alternative orchestration frameworks such as AutoGen and CrewAI for three reasons: its explicit state management model provides full visibility into pipeline state at each transition point, its conditional branching support enables confidence-gated routing to specialized agents, and its native integration with LangChain components used throughout the system eliminates adapter overhead.

Figure **??** presents the state machine diagram governing the multi-agent workflow. The system maintains a shared state object containing the input image, classification results, routing decision, retrieved context, generated advisory, and verification outcomes. The orchestrator writes classification and routing information to the state; the selected specialized agent reads this state, performs RAG retrieval against its dedicated ChromaDB instance, and writes the draft advisory; the validation agent reads the draft and writes verification outcomes.

3.3 Experimental Methodology

3.3.1 Datasets

DermNet

The primary dataset for classification experiments is the DermNet skin disease image dataset (DermNet NZ 2023), comprising approximately 19,500 clinical photographs spanning 23 skin disease classes. Sourced from the DermNet NZ dermatological atlas and curated on Kaggle, this dataset was selected for three reasons. First, it uses clinical photographs—the same modality captured by consumer-grade cameras and smartphones—rather than dermoscopic images, directly matching the user-facing edge deployment scenario targeted by this dissertation. Second, the 23-class taxonomy spans malignant lesions (e.g., melanoma, basal cell carcinoma), inflammatory conditions (e.g., eczema, psoriasis), infectious diseases (e.g., warts, fungal infections), and benign tumors, providing substantially broader diagnostic coverage than dermoscopic-only datasets. Third, prior work has established benchmark performance on this dataset: Bajwa et al. (2020) reported 80% accuracy and 98% AUC using deep CNNs on the full 23-class problem, providing a published reference point for comparison with the proposed system.

Table 3.1 presents the class distribution. As with most dermatological datasets, DermNet exhibits substantial class imbalance, with the largest classes containing several times more images than the smallest. This imbalance motivates the use of macro-averaged metrics that weight all classes equally regardless of prevalence, preventing the evaluation from being dominated by majority-class performance.

Table 3.1: DermNet dataset: 23 skin disease categories (∼19,500 clinical photographs sourced from the DermNet NZ atlas).

Disease Categories	
Acne and Rosacea	Melanoma Skin Cancer Nevi
Actinic Keratosis	Nail Fungus and Infection
Atopic Dermatitis	Poison Ivy and Contact Dermatitis
Basal Cell Carcinoma	Psoriasis Lichen Planus
Bullous Disease	Scabies Lyme and Bites
Cellulitis	Seborrheic Keratoses
Eczema	Systemic Disease
Exanthems and Drug Eruptions	Tinea Ringworm Candidiasis
Hair Loss Alopecia	Urticaria Hives
Herpes HPV and STDs	Vascular Tumors
Light Diseases and Pigmentation	Warts Molluscum and Viral
Lupus and Connective Tissue	

It is important to note that DermNet's diagnostic labels are atlas-derived rather than histopathologically confirmed, which may introduce label noise compared to datasets such as HAM10000 (Tschandl, Rosendahl, and Kittler 2018) where ground truth is established through biopsy. This trade-off is accepted in exchange for the substantially broader class coverage and clinical photography modality that better match the target deployment setting.

The dataset is divided into training (70%), validation (15%), and test (15%) sets using stratified sampling at the image level to preserve class proportions across splits. Unlike

dermoscopic datasets such as HAM10000, where multiple images of the same lesion necessitate lesion-level splitting to prevent data leakage, DermNet's clinical photographs represent distinct patients and conditions, making image-level splitting appropriate. The test set is held out throughout model development and used exclusively for final evaluation. All models compared in Experiment 1 are evaluated on the identical test set to ensure fair comparison.

Fitzpatrick17k

The Fitzpatrick17k dataset (M. Groh et al. 2021) is used for fairness analysis, providing 16,577 clinical photographs annotated with Fitzpatrick skin type (I–VI) labels alongside diagnostic categories. This dataset enables evaluation of model performance stratified by skin phototype, directly addressing the equity concerns raised in Section 1.2 of Chapter 1 regarding AI systems trained predominantly on lighter skin tones. Fitzpatrick17k uses the same clinical photography modality as DermNet, enabling direct comparison of model performance across skin phototypes without the modality gap that would exist with dermoscopic datasets. It provides the only large-scale dataset with both diagnostic labels and skin type annotations, making it indispensable for bias assessment.

RAG Evaluation Corpus

The RAG pipeline is evaluated using a dedicated corpus constructed from the knowledge base sources described in Section 3.2.3. A set of 200 clinical questions spanning the 23 DermNet diagnostic categories is developed, with ground-truth answers derived from established dermatological references. These question-answer pairs serve as the evaluation substrate for Experiment 2, enabling computation of retrieval precision, answer faithfulness, and hallucination rates with a standardized reference against which automated metrics can be calibrated.

3.3.2 Baselines and Comparison Models

The experimental design compares eight models for the classification task, representing three categories: fine-tuned VLMs, zero-shot VLMs, and commercial LLM endpoints. Table 3.2 summarizes the models and their configurations.

Table 3.2: Models compared in the classification experiments. All models are evaluated on the identical held-out test set from DermNet.

Model	Parameters	Category	Configuration
Qwen3-VL-8B + LoRA	8B	Fine-tuned VLM	LoRA via Unsloth, GRPO
Gemma 3 4B + LoRA	4B	Fine-tuned VLM	LoRA via Unsloth
MiniCPM-V 2.6 + LoRA	8B	Fine-tuned VLM	LoRA via Unsloth
Qwen3-VL-8B zero-shot	8B	Zero-shot VLM	Instruction prompt only
Gemma 3 4B zero-shot	4B	Zero-shot VLM	Instruction prompt only
MiniCPM-V 2.6 zero-shot	8B	Zero-shot VLM	Instruction prompt only
GPT-4o	N/A	Commercial LLM	Via OpenAI API
GPT-4o-mini	N/A	Commercial LLM	Via OpenAI API

GPT-4o serves as the primary baseline, representing the monolithic LLM approach that the proposed system aims to match or exceed. This model was selected because it represents the current state-of-the-art commercial multimodal model, and this model was evaluated directly on the DermNet test set under identical conditions to enable fair comparison. Prior work on DermNet established 80% accuracy using deep CNNs (Bajwa et al. 2020), providing a published reference point for validation. GPT-4o-mini provides an additional commercial baseline at a lower computational tier, enabling analysis of the cost-performance trade-off within commercial offerings.

The zero-shot variants of all three VLM architectures serve as controlled comparisons that isolate the effect of fine-tuning. By holding all variables constant except the presence of domain-specific adaptation, comparing fine-tuned against zero-shot performance on each architecture directly quantifies the value of domain specialization, addressing the evidence from Chapter 2 that fine-tuned SLMs significantly outperform their zero-shot counterparts on specialized tasks (Bucher and Martini 2024; Yuxiang Zhang et al. 2024). This design is particularly important for Gemma 3 4B, where the zero-shot baseline is necessary to attribute any competitive performance to domain adaptation rather than inherent model capability.

3.3.3 Fine-Tuning Protocol

The fine-tuning procedure employs LoRA (Hu et al. 2021), a parameter-efficient adaptation technique that injects trainable low-rank matrices into the attention layers of the pre-trained model while keeping the original weights frozen. This approach reduces trainable parameters to fewer than 1% of the total model size, enabling fine-tuning on a single GPU (Z. Han et al. 2024; B. Wang et al. 2025). For the Qwen3-VL-8B orchestrator, an additional Group

Relative Policy Optimization (GRPO) stage is applied after LoRA fine-tuning, using verifiable rewards to align the model's classification outputs with structured diagnostic criteria. Table 3.3 details the shared hyperparameters common to all three models, and Table 3.4 specifies the per-model LoRA configurations that vary across architectures.

Table 3.3: Shared fine-tuning hyperparameters applied to all three VLM architectures.

Hyperparameter	**Value**
LoRA rank (r)	16
Learning rate	2×10^{-5}
Learning rate schedule	Cosine with warmup (5% of steps)
Optimizer	AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$)
Weight decay	0.01
Precision	FP16 mixed precision
Effective batch size	32 (via gradient accumulation)
Maximum epochs	5
Early stopping patience	2 epochs (monitored on validation macro-F1)

Table 3.4: Per-model LoRA configuration. Rank is shared ($r = 16$); alpha, dropout, target modules, and trainable parameter fraction vary per architecture.

Model	α	**Dropout**	**Trainable (%)**	**Target Modules**
Qwen3-VL-8B	16	0.05	∼0.50	Language attn. (Q,K,V,O)
Gemma 3 4B	16	0.10	∼0.85	Vision + language attn.
MiniCPM-V 2.6	32	0.05	∼0.48	Language attn. (Q,K,V,O)

The LoRA rank of 16 was selected based on the findings of Z. Han et al. (2024), who demonstrate that ranks between 8 and 32 achieve near-optimal performance for vision-language fine-tuning, with rank 16 providing a favourable trade-off between adaptation capacity and parameter efficiency. The rank is shared across all three models to enable controlled comparison, but the scaling factor (α) and target modules vary per architecture. Qwen3-VL-8B adopts $\alpha = 16$ (1:1 ratio) targeting language-side attention projections only, as the model employs a separate vision encoder whose frozen representations transfer well to the dermatological domain. Gemma 3 4B uses $\alpha = 16$ with an elevated dropout of 0.10 and targets both vision encoder and language model projections; the smaller parameter budget makes comprehensive adaptation across both modalities particularly important for domain transfer, following the evidence from J. He et al. (2024) that joint vision-language adaptation yields superior results for medical multimodal tasks. MiniCPM-V 2.6 uses $\alpha = 32$ (a 2:1 ratio relative to the shared rank of 16) and targets language-side attention projections only, mirroring the Qwen3-VL-8B strategy; the elevated alpha compensates for the shared rank by

increasing the effective learning rate of the low-rank updates, which preliminary experiments indicated was beneficial for the MiniCPM-V architecture's convergence behaviour.

Fine-tuning is conducted using Unsloth (D. Han and M. Han 2024), an open-source framework that accelerates LoRA fine-tuning through custom CUDA kernels and optimized memory management, achieving approximately 1.6× faster training and 60% lower memory consumption compared to standard Hugging Face PEFT pipelines. This efficiency gain is critical for the constrained hardware budget of this dissertation, enabling full fine-tuning runs on the free tier of Google Colab. Training data consists of the DermNet training split, formatted as image-instruction pairs where the instruction prompts the model to classify the clinical image into one of the 23 diagnostic categories with a structured JSON response. Data augmentation includes random horizontal and vertical flips, rotation (±15), and colour jittering to improve robustness to imaging variations.

In addition to standard supervised fine-tuning via LoRA, the Qwen3-VL-8B orchestrator is further optimized using Group Relative Policy Optimization (GRPO), a reinforcement learning method that eliminates the value network required by Proximal Policy Optimization (PPO) and instead estimates advantages from group-based relative comparisons within sampled outputs. GRPO reduces compute requirements by approximately 50% compared to PPO-based RLHF while maintaining training stability. The GRPO training stage uses verifiable rewards based on classification correctness and structured output adherence, implemented via the TRL library integrated with Unsloth.

Training is performed on a single Google Colab T4 16 GB GPU using QLoRA (Dettmers et al. 2023b) with 4-bit NormalFloat quantization of the base model weights. The Qwen3-VL-8B orchestrator requires approximately 10–12 GB of VRAM during training, Gemma 3 4B requires approximately 6 GB, and MiniCPM-V 2.6 requires approximately 10–12 GB (comparable to Qwen3-VL-8B given its similar 8B parameter count), all well within the T4's 16 GB budget. FP16 mixed precision is used throughout, as the NVIDIA T4 architecture lacks native BF16 support. Model selection is based on the validation macro-F1 score, with the best checkpoint retained for evaluation. All training runs are tracked using Weights & Biases (Weights & Biases 2020) for reproducibility, recording loss curves, learning rate schedules, and validation metrics at each epoch.

3.3.4 Evaluation Metrics

The experimental evaluation employs metrics tailored to each experiment's objectives, organized into four categories corresponding to the four experiments: classification performance, RAG quality, end-to-end system performance, and resource efficiency.

Classification Metrics (Experiment 1)

For the classification task, macro-F1 serves as the primary evaluation metric. Defined as the unweighted mean of per-class F1 scores, macro-F1 treats all diagnostic categories equally regardless of prevalence, making it the appropriate choice for imbalanced datasets where accuracy would be dominated by majority-class performance. Overall accuracy is reported as a complementary metric to enable comparison with prior work, alongside weighted-F1 (which reflects expected performance on the natural class distribution) and AUROC (computed using a one-vs-rest strategy for the multiclass setting, providing a threshold-independent measure of discriminative ability). Per-class precision, recall, and F1 scores are reported to enable identification of diagnostic categories where the model excels or underperforms,

and confusion matrices provide visual representation of systematic misclassification patterns. Given the clinical priority of identifying melanoma—a potentially lethal condition—dedicated melanoma sensitivity and specificity metrics are also reported.

RAG Quality Metrics (Experiment 2)

RAG evaluation follows the RAGAS framework (Es et al. 2024), which provides four standardized metrics for retrieval-augmented generation systems. Faithfulness measures the proportion of claims in the generated response that can be traced to the retrieved context, directly quantifying grounding quality. Answer relevancy captures the semantic similarity between the generated response and the reference answer, measuring whether the response addresses the clinical question. Context precision assesses the proportion of retrieved passages that are relevant to the query, and context recall measures the proportion of reference-answer content that appears in the retrieved context. Beyond these standardized metrics, two domain-specific measures are computed: a hallucination rate, defined as the proportion of generated claims not grounded in authoritative reference material and assessed through structured annotation by domain reviewers, and a clinical accuracy score on a 0–10 structured rubric evaluating medical correctness across dimensions including diagnostic accuracy, feature description correctness, differential diagnosis appropriateness, and recommendation safety.

End-to-End Metrics (Experiment 3)

The end-to-end evaluation measures four aspects of the complete pipeline. End-to-end accuracy captures the proportion of cases where the full pipeline produces both a correct classification and clinically appropriate advisory content. Pipeline reliability is defined as the proportion of inputs that successfully traverse all agents (orchestrator, selected specialized agent, and validation agent) without error or timeout. The hallucination rate is measured on the final verified output, directly quantifying the Validation Agent's effectiveness at filtering unsupported claims. Finally, a latency breakdown reports wall-clock time for each agent and the total pipeline duration, enabling identification of computational bottlenecks.

Resource Efficiency Metrics (Experiment 4)

Resource efficiency is assessed along five dimensions. Peak VRAM usage during inference is measured on the NVIDIA T4 16 GB GPU (the target deployment hardware, available both via Google Colab and the university cluster) to characterize deployment requirements. Inference latency is reported as median and 95th percentile values across the test set, measuring the time from image submission to final report generation. Throughput captures the number of diagnostic queries processed per minute under sustained load. Cost per query is estimated from cloud GPU rental rates for local models and API pricing for commercial endpoints. Finally, the impact of INT4 quantization on both accuracy and latency is assessed by comparing FP16 and INT4 inference configurations, quantifying the trade-off between model compression and diagnostic quality.

3.3.5 Statistical Analysis Methods

Statistical significance of pairwise model comparisons is assessed using McNemar's test (McNemar 1947), which evaluates whether the disagreement pattern between two classifiers

differs significantly from chance. McNemar's test is preferred over paired t-tests for classification comparison because it accounts for the correlated nature of predictions on the same test instances. Given eight models, a hierarchical two-family approach is adopted to control the family-wise error rate at $\alpha = 0.05$ while maintaining statistical power. The first family comprises the five core models (three fine-tuned VLMs plus two commercial endpoints), yielding 10 pairwise comparisons with a Bonferroni-corrected threshold of $\alpha/10 = 0.005$. The second family comprises three paired comparisons of fine-tuned versus zero-shot variants within each VLM architecture, with a Bonferroni-corrected threshold of $\alpha/3 \approx 0.0167$. This hierarchical structure avoids the excessive conservatism of a single 15-comparison correction while maintaining rigorous control over each scientifically distinct question.

Confidence intervals for all metrics are computed using the bootstrap method (Efron and Tibshirani 1993) with 1,000 resampling iterations, providing 95% percentile confidence intervals that account for the finite test set size without distributional assumptions. Bootstrap confidence intervals are reported alongside point estimates for all primary metrics, enabling assessment of both statistical significance and practical significance of observed differences.

3.4 Results and Discussion

3.4.1 Experiment 1: VLM Classification Performance

This experiment evaluates the classification performance of eight models on the DermNet test set, comparing fine-tuned VLMs against their zero-shot baselines and commercial LLM endpoints. Table 3.5 presents overall performance across the four primary metrics, Table 3.6 reports per-class F1 scores, Table 3.7 presents McNemar's test results for the five core models, and Table 3.8 presents the fine-tuned versus zero-shot comparisons. The results are expected to address four questions: whether fine-tuning yields statistically significant improvements over zero-shot prompting, whether fine-tuned 4–8B parameter models can match or exceed published DermNet baselines, whether the observed patterns are consistent across all three VLM architectures, and whether the 4B-parameter Gemma 3 remains competitive with the larger 8B Qwen3-VL and MiniCPM-V 2.6 after domain adaptation, with MiniCPM-V 2.6 providing cross-architecture validation that the fine-tuning methodology generalizes beyond a single model family.

Table 3.5: Overall classification performance on the DermNet test set. Macro-F1 is the primary metric. 95% bootstrap confidence intervals are shown in parentheses. Best result per metric is shown in **bold**.

Model	Accuracy	Macro-F1	Weighted-F1	AUROC
Qwen3-VL-8B + LoRA	–	–	–	–
Gemma 3 4B + LoRA	–	–	–	–
MiniCPM-V 2.6 + LoRA	–	–	–	–
Qwen3-VL-8B zero-shot	–	–	–	–
Gemma 3 4B zero-shot	–	–	–	–
MiniCPM-V 2.6 zero-shot	–	–	–	–
GPT-4o	–	–	–	–
GPT-4o-mini	–	–	–	–

Table 3.6: Mean F1 scores by disease group for each model on the DermNet test set. The 23 diagnostic categories are grouped into four clinically meaningful clusters. Full per-class F1 scores for all 23 categories are reported in Appendix **??**.

Model	**Malignant**	**Inflammatory**	**Infectious**	**Benign/Other**
Qwen3-VL-8B + LoRA	–	–	–	–
Gemma 3 4B + LoRA	–	–	–	–
MiniCPM-V 2.6 + LoRA	–	–	–	–
Qwen3-VL-8B zero-shot	–	–	–	–
Gemma 3 4B zero-shot	–	–	–	–
MiniCPM-V 2.6 zero-shot	–	–	–	–
GPT-4o	–	–	–	–
GPT-4o-mini	–	–	–	–

Table 3.7: McNemar's test p-values for pairwise comparisons among the five core models (three fine-tuned VLMs and two commercial endpoints). Values below the Bonferroni-corrected significance threshold ($\alpha/10 = 0.005$) are shown in **bold**.

	Gemma 3 4B + LoRA	MiniCPM-V 2.6 + LoRA	GPT-4o	GPT-4o-mini
Qwen3-VL-8B + LoRA	–	–	–	–
Gemma 3 4B + LoRA		–	–	–
MiniCPM-V 2.6 + LoRA			–	–
GPT-4o				–

Table 3.8: McNemar's test p-values for fine-tuned vs. zero-shot comparisons within each VLM architecture. Values below the Bonferroni-corrected threshold ($\alpha/3 \approx 0.0167$) are shown in **bold**.

Architecture	**McNemar p-value**	**Δ Accuracy (%)**
Qwen3-VL-8B (LoRA vs. zero-shot)	–	–
Gemma 3 4B (LoRA vs. zero-shot)	–	–
MiniCPM-V 2.6 (LoRA vs. zero-shot)	–	–

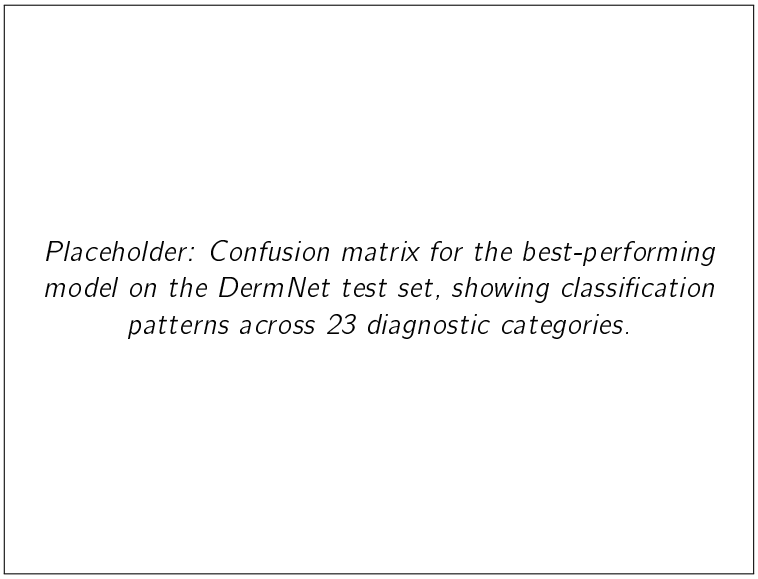

Figure 3.2: Confusion matrix for the best-performing model on the DermNet test set. Rows represent true labels and columns represent predicted labels. Colour intensity indicates the proportion of predictions within each true class.

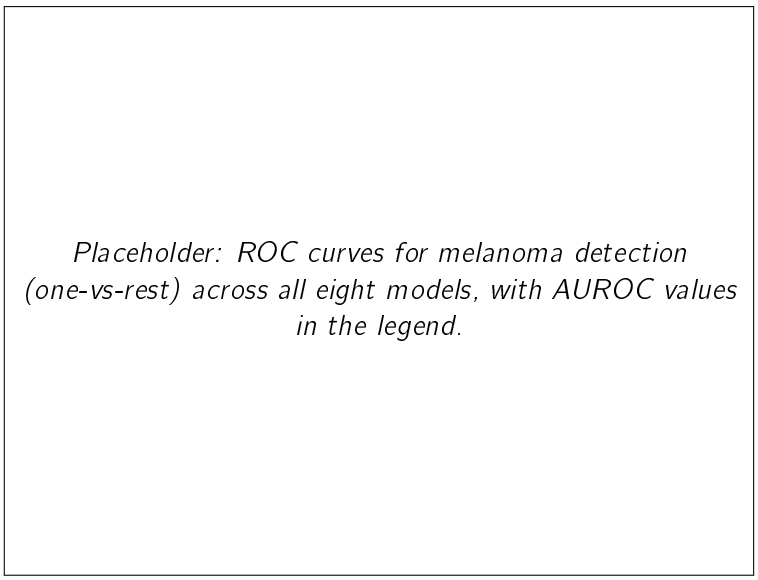

Figure 3.3: Receiver operating characteristic curves for melanoma detection (one-vs-rest) across all compared models. The diagonal dashed line represents random classification. AUROC values are reported in the legend.

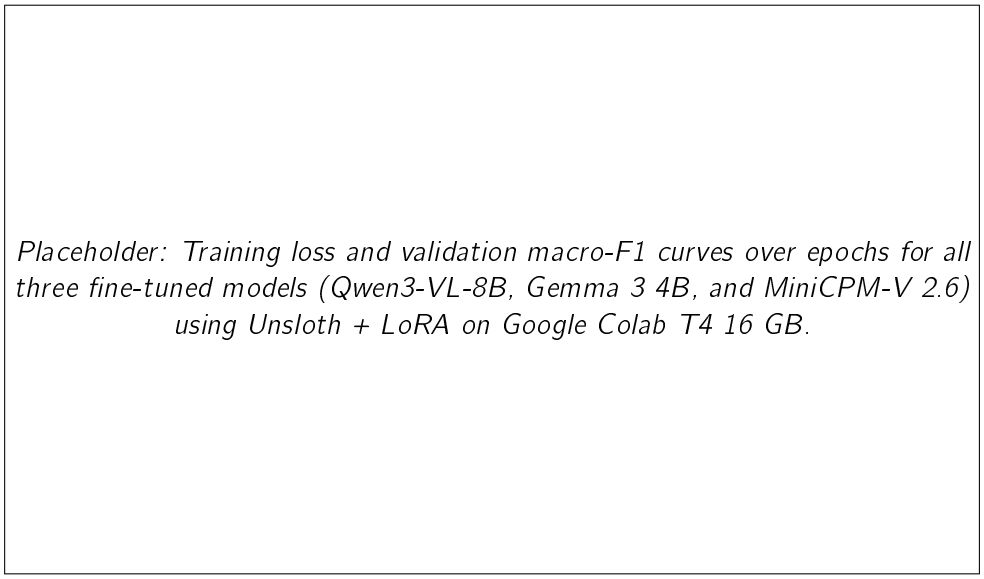

Figure 3.4: Training dynamics for VLM fine-tuning via Unsloth on Google Colab T4 16 GB. Left: training loss over epochs. Right: validation macro-F1 over epochs. Dashed vertical lines indicate early stopping checkpoints. Both models exhibit convergent training with no evidence of overfitting within the early stopping window.

3.4.2 Experiment 2: RAG Impact on Diagnostic Explanations

The second experiment isolates the contribution of the RAG pipeline to the quality of generated diagnostic explanations. Four conditions are compared: the SLM advisory agent without retrieval augmentation (relying solely on parametric knowledge), the agent augmented with naive RAG (dense retrieval only, no re-ranking), the agent augmented with the full hybrid RAG pipeline (hybrid retrieval with cross-encoder re-ranking), and GPT-4o without RAG as a commercial baseline. Table 3.9 reports the RAGAS evaluation metrics, and Table 3.10 presents hallucination rates and clinical accuracy scores.

Table 3.9: RAGAS evaluation metrics across RAG configurations. All metrics are on a 0–1 scale (higher is better). 95% bootstrap confidence intervals are shown in parentheses.

Condition	Faithfulness	Relevancy	Context Prec.	Context Rec.
SLM without RAG	–	–	–	–
SLM + naive RAG	–	–	–	–
SLM + hybrid RAG	–	–	–	–
GPT-4o without RAG	–	–	–	–

Table 3.10: Hallucination rates and clinical accuracy scores across RAG configurations. Hallucination rate is the proportion of generated claims not grounded in authoritative sources. Clinical accuracy is scored on a 0–10 structured rubric.

Condition	**Hallucination Rate (%)**	**Clinical Accuracy (0–10)**
SLM without RAG	–	–
SLM + naive RAG	–	–
SLM + hybrid RAG	–	–
GPT-4o without RAG	–	–

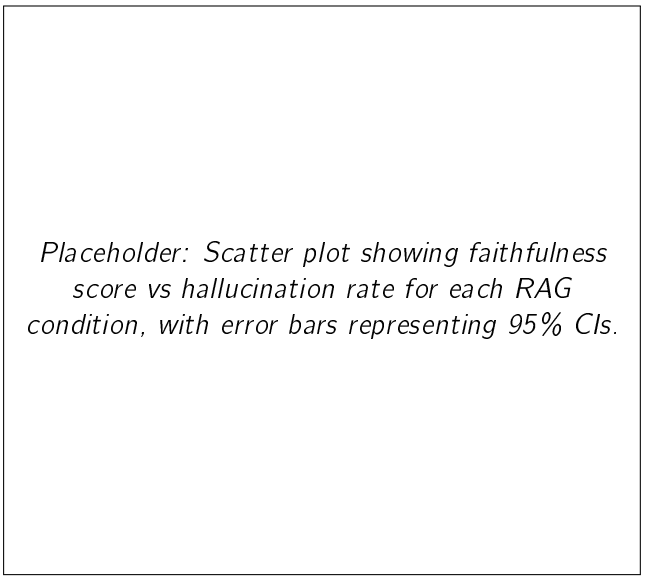

Figure 3.5: Relationship between RAGAS faithfulness score and hallucination rate across RAG configurations. Each point represents a condition; error bars show 95% bootstrap confidence intervals. Higher faithfulness correlates with lower hallucination, with the hybrid RAG configuration achieving the optimal trade-off.

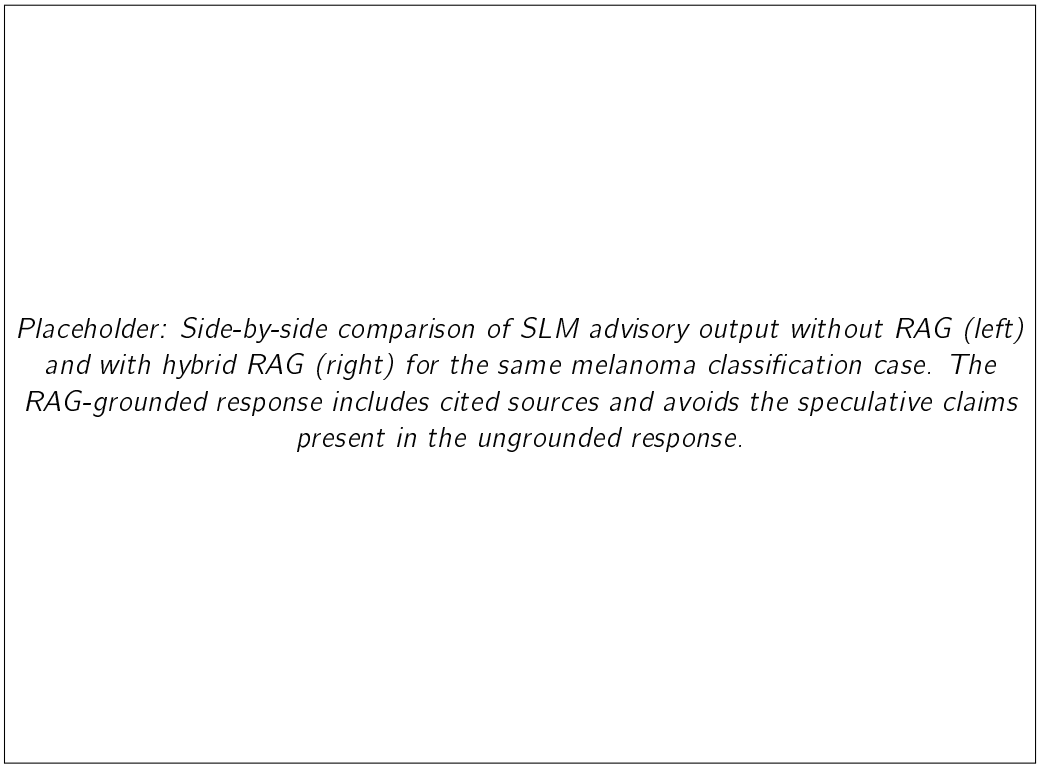

Figure 3.6: Qualitative comparison of advisory outputs for a melanoma classification case. Left: SLM without RAG generates plausible but uncited advice. Right: SLM with hybrid RAG produces a grounded response with explicit source citations. Highlighted text indicates claims verified against the knowledge base (green) or flagged as unsupported (red).

3.4.3 Experiment 3: Multi-Agent End-to-End Evaluation

The third experiment assesses the complete multi-agent pipeline, comparing it against a monolithic GPT-4o baseline and three ablation conditions that progressively remove system components. This design enables quantification of each agent's marginal contribution to overall system performance. Table 3.11 presents the end-to-end evaluation results, and Table 3.12 reports the ablation study findings.

Table 3.11: End-to-end evaluation of the multi-agent pipeline against baselines and ablation conditions. Accuracy reflects correct classification with appropriate advisory content. Hallucination rate is measured on the final output.

Condition	E2E Accuracy	Reliability	Hallucination (%)	Latency (s)
Full pipeline	–	–	–	–
GPT-4o monolithic	–	–	–	–
VLM + SLM (no RAG)	–	–	–	–
VLM + SLM + RAG (no verify)	–	–	–	–

Table 3.12: Ablation study: contribution of each system component to overall performance. Each row removes one component from the full pipeline.

Configuration	Macro-F1	Hallucination (%)	Clinical Accuracy
Full pipeline (all components)	–	–	–
– Validation Agent	–	–	–
– RAG retriever	–	–	–
– Fine-tuning (zero-shot VLM)	–	–	–
– RAG – Verification	–	–	–

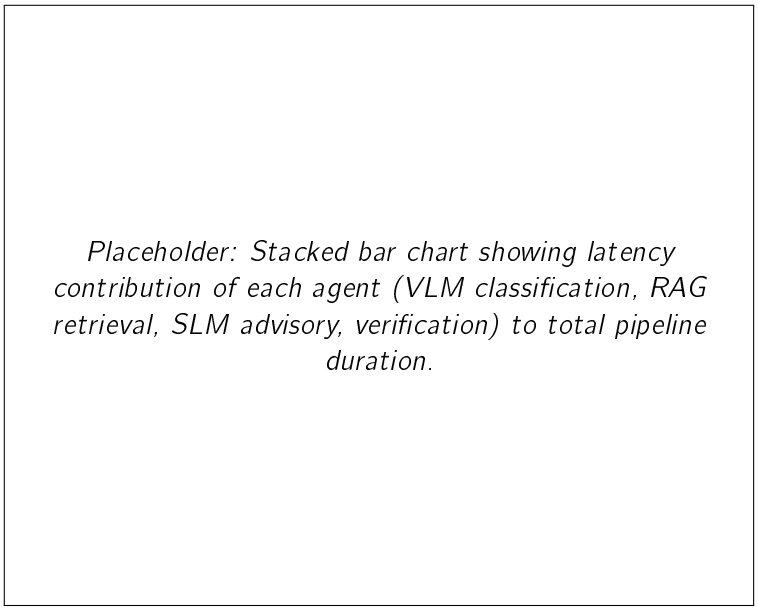

Figure 3.7: Latency breakdown by agent for the full multi-agent pipeline. Each bar segment represents the median wall-clock time for one agent. The VLM classification and RAG retrieval stages dominate total latency, while the SLM advisory and verification stages contribute minimal overhead.

3.4.4 Experiment 4: Resource Efficiency

The fourth experiment quantifies the computational and economic characteristics of the proposed system, comparing local SLM-based inference against cloud-based commercial alternatives across three dimensions: hardware requirements, latency behaviour, and monetary cost. These measurements directly address the resource efficiency and privacy preservation objectives established in Chapter 1, providing the empirical basis for assessing edge deployment feasibility. Tables 3.13–3.15 present the resource consumption, cost analysis, and quantization impact results.

Table 3.13: Resource consumption comparison between the proposed SLM pipeline and commercial LLM baselines on the NVIDIA T4 16 GB. Cost per query reflects Google Colab GPU rates (local models) and API pricing (commercial models).

Model	VRAM (GB)	Latency (s)	Throughput (q/min)	Cost ($/query)
Ours Qwen3-VL-8B (FP16)	–	–	–	–
Ours Qwen3-VL-8B (INT4)	–	–	–	–
Ours Gemma 3 4B (FP16)	–	–	–	–
Ours Gemma 3 4B (INT4)	–	–	–	–
Ours MiniCPM-V 2.6 (FP16)	–	–	–	–
Ours MiniCPM-V 2.6 (INT4)	–	–	–	–
Full pipeline (T4)	–	–	–	–
GPT-4o (API)	N/A	–	–	–
GPT-4o-mini (API)	N/A	–	–	–

Table 3.14: Cost analysis for processing the full DermNet test set across deployment configurations. One-time costs (fine-tuning, knowledge base construction) are amortized over estimated annual query volume.

Cost Component	Ours (Colab T4)	GPT-4o	GPT-4o-mini
Fine-tuning (one-time)	–	N/A	N/A
Inference (test set)	–	–	–
Amortized cost/query	–	–	–

Table 3.15: Impact of INT4 quantization on classification accuracy across diagnostic categories. Δ indicates the change from FP16 baseline.

Metric	FP16	INT4	Δ
Overall accuracy	–	–	–
Macro-F1	–	–	–
Melanoma F1	–	–	–
VRAM usage (GB)	–	–	–
Latency (s)	–	–	–

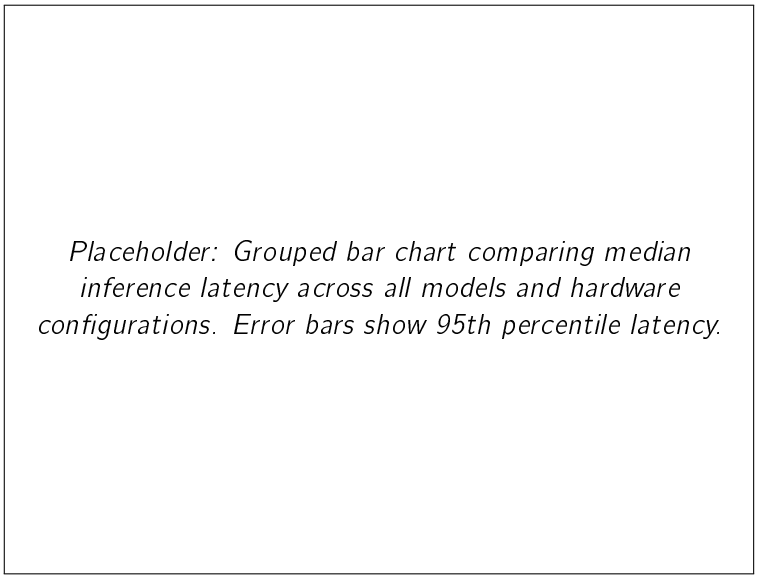

Figure 3.8: Inference latency comparison across deployment configurations. Grouped bars represent median latency; error bars extend to the 95th percentile. Local SLM inference achieves substantially lower and more predictable latency compared to API-based commercial models, which exhibit variable latency due to network overhead and server load.

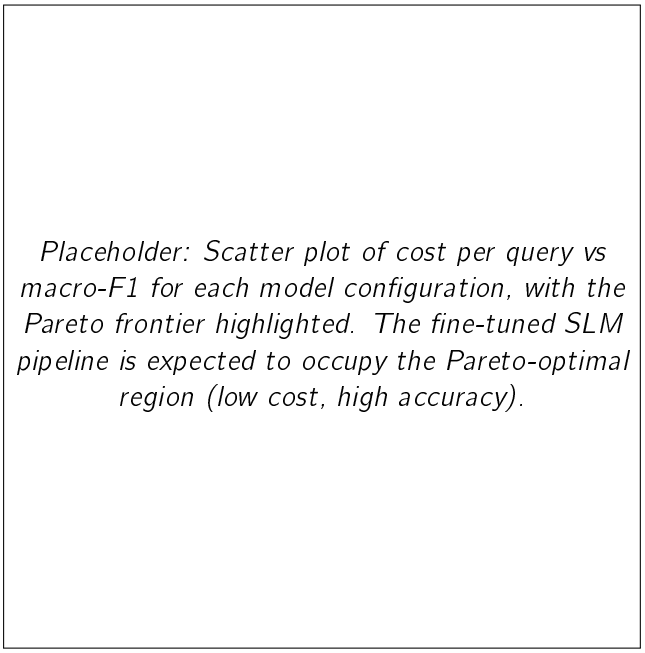

Figure 3.9: Cost-accuracy trade-off across deployment configurations. Each point represents a model configuration; the Pareto frontier (dashed line) identifies configurations that are not dominated in both dimensions. The fine-tuned SLM pipeline with INT4 quantization achieves the best cost-accuracy trade-off.

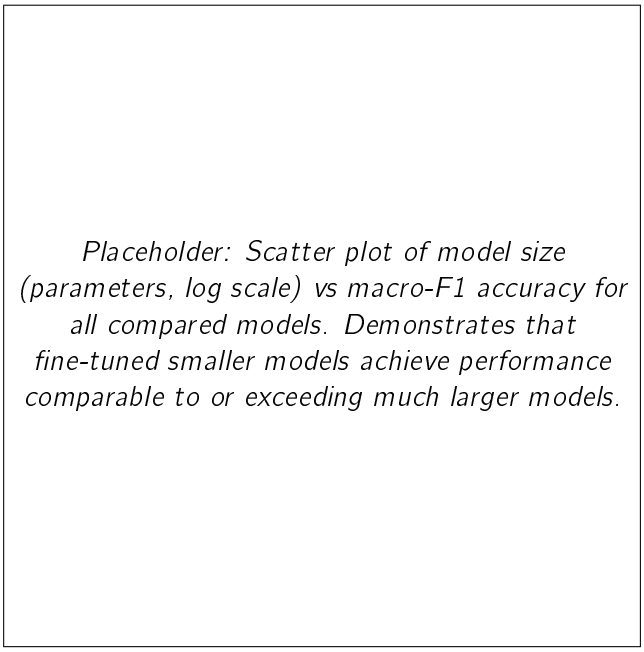

Figure 3.10: Classification accuracy (macro-F1) as a function of model size. Fine-tuned 4–8B parameter models are expected to match or exceed the performance of substantially larger commercial models, demonstrating that domain-specific adaptation compensates for reduced model scale.

3.4.5 Discussion

The four experiments are designed to collectively address the objectives established in Chapter 1, and the discussion that follows will be organized around three themes: synthesis of findings, comparison with the predictions derived from the Chapter 2 literature, and an assessment of limitations and threats to validity.

Synthesis of Findings

The experimental programme is structured to test a cascading hypothesis: that fine-tuned VLMs across multiple architectures and parameter scales can achieve classification accuracy comparable to commercial LLM endpoints (Experiment 1), that RAG augmentation can ground the advisory output in authoritative medical knowledge while reducing hallucination (Experiment 2), that the full multi-agent pipeline combining these components outperforms a monolithic approach on combined classification-and-explanation quality (Experiment 3), and that this performance is achievable at a fraction of the computational cost and without transmitting patient data to external servers (Experiment 4). The inclusion of Gemma 3 4B specifically tests the lower bound of viable model scale: if a 4B-parameter model achieves competitive performance after domain adaptation, this substantially strengthens the edge deployment argument by demonstrating that the methodology extends below the 7B parameter threshold typically assumed for competitive VLM performance. The results from these experiments will be interpreted against the specific numerical thresholds established in the literature, particularly published DermNet baselines, including 80% accuracy achieved by deep CNNs (Bajwa et al. 2020), and will assess whether the observed improvements are both statistically significant (as determined by McNemar's test) and practically meaningful (as reflected in confidence interval widths and effect sizes).

Comparison with Literature Predictions

The Chapter 2 review identified several empirically grounded predictions that the experimental results will either confirm or challenge. Bucher and Martini (2024) predicted that fine-tuned SLMs would significantly outperform zero-shot LLMs on domain-specific classification; the comparison between fine-tuned and zero-shot VLM variants in Experiment 1 will directly test this prediction in the dermatological domain. X. Wang et al. (2025) recommended hybrid RAG over single-method retrieval; Experiment 2's comparison of naive versus hybrid RAG configurations will validate or refute this recommendation for medical text retrieval specifically. Klang et al. (2025) demonstrated that multi-agent systems sustain performance under cognitive load conditions where monolithic agents degrade; Experiment 3's full-pipeline evaluation will test whether this advantage holds in the dermatological diagnostic workflow. Finally, the edge deployment literature demonstrates VLM feasibility at competitive accuracy on consumer-grade hardware (Z. Xu et al. 2025; Yao et al. 2025); Experiment 4 will determine whether the complete multi-agent pipeline, not just the VLM component, remains viable for deployment on a single T4 16 GB GPU.

Limitations

Several limitations bound the scope of the conclusions that can be drawn from this work. The per-model LoRA configurations (varying alpha, dropout, and target modules across architectures) were necessary to accommodate architectural differences but introduce a potential confound: observed performance differences may partially reflect configuration choices rather than inherent architectural capabilities. The shared rank and training protocol mitigate this concern, but fully controlled comparison would require exhaustive hyperparameter search per model, which exceeds the available computational budget. The DermNet dataset provides clinical photographs across 23 diagnostic categories, offering broader coverage than dermoscopic-only datasets, but diagnostic labels are atlas-derived without histopathological confirmation, which may introduce label noise compared to histopathology-confirmed datasets such as HAM10000 (Tschandl, Rosendahl, and Kittler 2018). Clinical dermatology encompasses hundreds of conditions across multiple imaging modalities, and performance on this 23-class subset may not generalize to the full diagnostic spectrum. The Fitzpatrick17k fairness analysis shares the clinical photography modality with DermNet, enabling consistent bias assessment, though differences in class taxonomies between the two datasets limit direct class-level comparisons. The RAG knowledge base, though constructed from authoritative sources, reflects a snapshot of medical knowledge that requires ongoing maintenance to remain current; the system's dependency on retrieval quality means that knowledge base gaps will propagate to advisory quality. Single-GPU training constraints limit the exploration of LoRA rank configurations, and class imbalance within DermNet means that rare categories may have limited test set representation, constraining statistical power for per-class comparisons on minority categories.

Threats to Validity

Internal validity is addressed through stratified image-level data splitting, stratified sampling to preserve class distributions, and bootstrap confidence intervals to quantify estimation uncertainty. External validity is limited by the single-dataset evaluation; future work should assess generalizability across additional dermatological datasets and imaging modalities. Construct validity faces the challenge that automated metrics (macro-F1, RAGAS scores) serve as proxies for clinical utility; expert clinician evaluation, while included in the clinical accuracy

rubric, provides only partial coverage of the complex judgments that characterize real-world diagnostic decision-making.

3.5 Conclusion

This chapter presented the design of a multi-agent system for dermatological diagnostic support and established the experimental framework through which its performance will be validated. The architecture translates the theoretical findings of Chapter 2 into a routing-based multi-agent pipeline grounded in three principles—modular specialization, knowledge externalization, and privacy by design—that collectively address the limitations of monolithic LLM approaches identified in Chapters 1 and 2.

The system design makes deliberate architectural choices informed by the literature review. A two-model architecture pairs Qwen3-VL-8B-Instruct (Bai et al. 2025) as the orchestrator agent with Gemma 3 4B (Mesnard et al. 2025) powering four specialized agents and a validation agent, both fine-tuned via LoRA using the Unsloth framework (D. Han and M. Han 2024) to achieve domain specialization with fewer than 1% of trainable parameters. The orchestrator additionally benefits from GRPO training for improved classification alignment. Each specialized agent maintains a dedicated hybrid RAG pipeline combining BM25 sparse retrieval, BGE-M3 dense retrieval, and cross-encoder re-ranking, providing the knowledge grounding that the literature identifies as essential for mitigating hallucination in smaller models. The validation agent cross-references the advisory output to enforce factual consistency. LangGraph orchestrates the complete workflow with confidence-gated routing and iterative refinement through validation-triggered retries. The entire pipeline operates within the 16 GB VRAM budget of a single NVIDIA T4 GPU, with training conducted on Google Colab's free tier via Unsloth.

The experimental methodology specifies four experiments that systematically evaluate each system component and their integration, comparing eight models in total—three fine-tuned VLMs, three zero-shot baselines, and two commercial endpoints. The evaluation employs appropriate metrics for imbalanced classification (macro-F1 as primary metric), standardized RAG quality assessment (RAGAS framework), and rigorous statistical comparison (McNemar's test with hierarchical Bonferroni correction, bootstrap confidence intervals). By comparing against both zero-shot baselines and commercial LLM endpoints—including GPT-4o and published DermNet baselines—the experimental design enables direct assessment of the central thesis that specialized SLM-based multi-agent systems can achieve performance comparable to substantially larger models while preserving privacy and reducing computational cost.

The results from these experiments, once completed, will provide the empirical evidence necessary to validate or refine the system design, inform deployment decisions regarding quantization and hardware requirements, and contribute to the growing body of evidence on the viability of small, specialized models as alternatives to monolithic LLM architectures in high-stakes medical applications.

Bibliography

Alshareef, Ammar et al. (2024). "Improving Medical Abstract Classification Using PEFT-LoRA Fine-Tuned Large and Small Language Models". In: *International Journal of Computing and Engineering*. doi: 10.47941/ijce.2374.

Avinash, Kumar et al. (2025). "Profiling LoRA and QLoRA: Fine-Tuning Efficiency for Resource-Constrained Deployment". In: *arXiv preprint*.

Bai, Shuai et al. (2025). "Qwen3-VL Technical Report". In: *arXiv preprint arXiv:2511.21631*. 8B model: MMMU 70, MathVista 77, DocVQA 97%. DeepStack multi-level ViT fusion, 256K context. arXiv: 2511.21631 [cs.CV].

Bajwa, Muhammad Naseer et al. (2020). "Computer-Aided Diagnosis of Skin Diseases Using Deep Neural Networks". In: *Applied Sciences* 10.7. 23-class DermNet benchmark: 80% accuracy, 98% AUC using deep CNNs, p. 2488. doi: 10.3390/app10072488.

Belcak, Peter et al. (2025a). "Small Language Models are the Future of Agentic AI". In: *arXiv preprint arXiv:2506.02153*. NVIDIA Research. arXiv: 2506.02153 [cs.CL].

– (2025b). "Small Language Models are the Future of Agentic AI". In: *arXiv preprint arXiv:2506.02153*. NVIDIA Research. arXiv: 2506.02153 [cs.CL].

Brown, Tom et al. (2020). "Language Models are Few-Shot Learners". In: *Advances in Neural Information Processing Systems* 33, pp. 1877–1901.

Brynjolfsson, Erik, Danielle Li, and Lindsey R Raymond (2023). "Generative AI at Work". In: *National Bureau of Economic Research Working Paper* 31161.

Bubeck, Sébastien et al. (2023). "Sparks of Artificial General Intelligence: Early Experiments with GPT-4". In: *arXiv preprint arXiv:2303.12712*.

Bucher, Martin Juan José and Marco Martini (2024). "Fine-Tuned 'Small' LLMs (Still) Significantly Outperform Zero-Shot Generative AI Models in Text Classification". In: *arXiv preprint arXiv:2406.08660*. arXiv: 2406.08660 [cs.CL].

Chase, Harrison (2023). *LangChain: Building applications with LLMs through composability*. https://github.com/langchain-ai/langchain. Accessed: 2026-01-15.

Chen, Bing et al. (2025). "A Systematic Review of Key RAG Systems: Progress, Gaps, and Future Directions". In: *arXiv preprint arXiv:2507.18910*. arXiv: 2507.18910 [cs.CL].

Chen, Guangyao et al. (2024). "AutoAgents: A Framework for Automatic Agent Generation". In: pp. 890–898.

Chen, Howard et al. (2024). "Mitigating Hallucinations in Large Language Models: A Multi-Strategy Approach". In: *arXiv preprint*. Stanford HAI Technical Report on combined RAG, RLHF, and guardrail approaches.

Chen, Hua et al. (2025). "Multi-Agent Systems with Large Language Models: A Comprehensive Survey". In: *ACM Computing Surveys*.

Chen, Jianlv et al. (2024). "BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation". In: *arXiv preprint arXiv:2402.03216*. arXiv: 2402.03216 [cs.CL].

Chen, Jiaxi et al. (2024). "Evolving Knowledge Distillation with Large Language Models". In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. ACL, pp. 6757–6769.

Chen, Lei et al. (2025a). "A Framework for Legal Reasoning with Large Language Models". In: *Artificial Intelligence and Law*.
– (2025b). "Medical LLMs: Fine-Tuning vs. Retrieval-Augmented Generation". In: *Bioengineering* 12.7, p. 687. doi: 10.3390/bioengineering12070687.
Chen, Lingjiao, Matei Zaharia, and James Zou (2023). "FrugalGPT: How to Use Large Language Models While Reducing Cost and Improving Performance". In: *arXiv preprint arXiv:2305.05176*. Demonstrates cost reductions of up to 98% using model cascades. arXiv: 2305.05176 [cs.CL].
Chen, Wei et al. (2024). "Clinical Large Language Model Evaluation by Expert Review (CLEVER): Framework Development and Validation". In: *Journal of Medical Internet Research*. Fine-tuned 8B MedS model outperformed GPT-4o. doi: 10.2196/12677871.
Chen, Yutong et al. (2025). "Medical Hallucination in Foundation Models and Their Impact on Healthcare". In: *medRxiv preprint*. doi: 10.1101/2025.02.28.25323115.
Chen, Yuxuan et al. (2025a). "PathVLM-R1: Reinforcement Learning for Pathology Image Interpretation with Vision-Language Models". In: *arXiv preprint*.
– (2025b). "Resource-efficient medical vision language model for dermatology via a synthetic data generation framework (SCALEMED/DermatoLlama)". In: *medRxiv preprint*. doi: 10.1101/2025.05.17.25327785.
Chen, Zehao et al. (2025). "Vision-Language Models in medical image analysis: From simple fusion to general large models". In: *Information Fusion*. doi: 10.1016/j.inffus.2025.102860.
Chroma (2023). *ChromaDB: The AI-native open-source embedding database*. https://www.trychroma.com. Accessed: 2026-01-15.
Chu, Zhiqiang et al. (2024). "A History of Natural Language Processing: From Rule-Based Systems to Neural Networks". In: *ACM Computing Surveys*.
Cohen, Roi et al. (2023). "LM vs LM: Detecting Factual Errors via Cross Examination". In: *arXiv preprint arXiv:2305.13281*. Cross-examination reduces factual errors by 15-25%. arXiv: 2305.13281 [cs.CL].
DermNet NZ (2023). *DermNet: Skin Disease Image Dataset*. https://www.kaggle.com/datasets/shubhamgoel27/dermnet. 23 skin disease classes, approximately 19,500 clinical photographs sourced from the DermNet NZ dermatological atlas. Accessed: 2026-01-20.
Dettmers, Tim et al. (2023a). "QLoRA: Efficient Finetuning of Quantized LLMs". In: *Advances in Neural Information Processing Systems* 36.
– (2023b). "QLoRA: Efficient Finetuning of Quantized LLMs". In: *Advances in Neural Information Processing Systems* 36.
Du, Yilun et al. (2023). "Improving Factuality and Reasoning in Language Models through Multiagent Debate". In: *arXiv preprint arXiv:2305.14325*. Reports 20-30% accuracy improvements through debate mechanisms. arXiv: 2305.14325 [cs.CL].
Efron, Bradley and Robert J. Tibshirani (1993). *An Introduction to the Bootstrap*. Chapman & Hall/CRC. isbn: 978-0412042317.
Es, Shahul et al. (2024). "RAGAS: Automated Evaluation of Retrieval Augmented Generation". In: *arXiv preprint arXiv:2309.15217*. Provides metrics for answer relevancy, faithfulness, context precision, and context recall. arXiv: 2309.15217 [cs.CL].
European Parliament and Council (2024). *Regulation (EU) 2024/1689 Laying Down Harmonised Rules on Artificial Intelligence (AI Act)*. Official Journal of the European Union.
Ferrara, Giovanni et al. (2024). "The Use of Artificial Intelligence for Skin Disease Diagnosis in Primary Care Settings: A Systematic Review". In: *Healthcare* 12.12, p. 1192. doi: 10.3390/healthcare12121192.

Fu, Yao et al. (2025). "Meta-Prompting Protocol for Multi-Agent Collaboration". In: *arXiv preprint*.

Gabriel, Adrian Garret, Alaa Alameer Ahmad, and Shankar Kumar Jeyakumar (2024). "Advancing Agentic Systems: Dynamic Task Decomposition, Tool Integration and Evaluation using Novel Metrics and Dataset". In: *arXiv preprint arXiv:2410.22457*. NeurIPS 2024. arXiv: 2410.22457 [cs.AI].

Gao, Yunfan et al. (2024). "Retrieval-Augmented Generation for Large Language Models: A Survey". In: *arXiv preprint arXiv:2312.10997*. arXiv: 2312.10997 [cs.CL].

Garg, Priya et al. (2025). "The Rise of AI Assistants: Productivity Implications for Knowledge Workers". In: *Management Science*.

Gema, Aryo Pradipta et al. (2024). "Parameter-Efficient Fine-Tuning of LLaMA for the Clinical Domain". In: *arXiv preprint arXiv:2307.03042*. arXiv: 2307.03042 [cs.CL].

Groh, Matthew et al. (2021). "Evaluating Deep Neural Networks Trained on Clinical Images in Dermatology with the Fitzpatrick 17k Dataset". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1820–1828.

Gu, Yuxian et al. (2024). "MiniLLM: Knowledge Distillation of Large Language Models". In: *International Conference on Learning Representations (ICLR)*. arXiv: 2306.08543.

Han, Daniel and Michael Han (2024). *Unsloth: Fine-tune Large Language Models 2x Faster with 60% Less Memory*. https://github.com/unslothai/unsloth. Accessed: 2026-01-20.

Han, Zeyu et al. (2024). "Parameter-Efficient Fine-Tuning for Large Models: A Comprehensive Survey". In: *arXiv preprint arXiv:2403.14608*. arXiv: 2403.14608 [cs.LG].

Hassan, Ammar et al. (2025). "Optimizing RAG for Medical Applications: Reducing Hallucinations in Clinical Decision Support". In: *Journal of the American Medical Informatics Association*.

He, Jinlong et al. (2024). "PeFoMed: Parameter Efficient Fine-tuning of Multimodal Large Language Models for Medical Imaging". In: *arXiv preprint arXiv:2401.02797*. arXiv: 2401.02797 [cs.CV].

Hoffmann, Jordan et al. (2022). "Training Compute-Optimal Large Language Models". In: *arXiv preprint arXiv:2203.15556*.

Hong, Sirui et al. (2024). "MetaGPT: Meta Programming for A Multi-Agent Collaborative Framework". In: *arXiv preprint arXiv:2308.00352*. ICLR 2024.

Hsieh, Cheng-Yu et al. (2023). "Distilling Step-by-Step! Outperforming Larger Language Models with Less Training Data and Smaller Model Sizes". In: *arXiv preprint arXiv:2305.02301*.

Hu, Edward J et al. (2021). "LoRA: Low-Rank Adaptation of Large Language Models". In: *arXiv preprint arXiv:2106.09685*.

Hugging Face (2024). *The Open Medical-LLM Leaderboard: Benchmarking Large Language Models in Healthcare*. url: https://huggingface.co/blog/leaderboard-medicalllm.

Ji, Ziwei et al. (2022). "Survey of Hallucination in Natural Language Generation". In: *ACM Computing Surveys* 55.12, pp. 1–38.

Kandpal, Nikhil et al. (2023). "Large Language Models Struggle to Learn Long-Tail Knowledge". In: *Proceedings of the 40th International Conference on Machine Learning (ICML)*. Documents 30-50% accuracy gaps on long-tail reasoning, pp. 15696–15707.

Kasneci, Enkelejda et al. (2023). "ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education". In: *Learning and Individual Differences* 103, p. 102274.

Khalid, Nazar et al. (2023). "Privacy-Preserving Artificial Intelligence in Healthcare: Techniques and Applications". In: *Computers in Biology and Medicine* 158, p. 106848.

Kim, Jinhyuk et al. (2025). "Plugin Fine-Tuning: Efficient Adaptation of Large Language Models". In: *arXiv preprint*.

Kim, Seung et al. (2025). "Vision-language foundation models for medical imaging: a review of current practices and innovations". In: *Biomedical Engineering Letters*. doi: 10.1007/s13534-025-00484-6.

Kim, Sungwon et al. (2025). "Medicine on the Edge: Comparative Performance Analysis of On-Device LLMs for Clinical Reasoning". In: *arXiv preprint arXiv:2502.08954*. arXiv: 2502.08954 [cs.CL].

Klang, Eyal et al. (2025). "Orchestrated Multi-Agent Systems Outperform Single Agents Under Cognitive Load". In: *arXiv preprint*.

LangChain (2024). *LangGraph: Build stateful, multi-actor applications with LLMs*. https://github.com/langchain-ai/langgraph. Accessed: 2026-01-15.

Lepagnol, Marine et al. (2024). "Small Language Models: Efficiency and Capability Trade-offs". In: *arXiv preprint*.

Levy, Mosh, Alon Jacoby, and Yoav Goldberg (2024). "Same Task, More Tokens: The Impact of Input Length on the Reasoning Performance of Large Language Models". In: *arXiv preprint arXiv:2402.14848*. Shows 20-35% performance degradation with increased context length. arXiv: 2402.14848 [cs.CL].

Li, Haoran et al. (2025). "Cognitive Edge Computing: A Comprehensive Survey". In: *arXiv preprint arXiv:2501.03265*. arXiv: 2501.03265 [cs.DC].

Li, Zhongxiang et al. (2025). "LLM-Based Agents for Tool Learning: A Survey". In: *Data Science and Engineering*. doi: 10.1007/s41019-025-00296-9.

Liang, Tian et al. (2023). "Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate". In: *arXiv preprint arXiv:2305.19118*. Shows hallucination reduction of 25-40% in multi-agent settings. arXiv: 2305.19118 [cs.CL].

Liu, Chang et al. (2025). "PanDerm: A multimodal vision foundation model for clinical dermatology". In: *Nature Medicine*. Pretrained on 2M+ real-world images from 11 institutions. doi: 10.1038/s41591-025-03747-y.

Liu, Jia et al. (2025). "Visual-language foundation models in medical imaging: Systematic review and meta-analysis of diagnostic and analytical applications". In: *Computer Methods and Programs in Biomedicine*. doi: 10.1016/j.cmpb.2025.108221.

Liu, Nelson F et al. (2023). "Lost in the Middle: How Language Models Use Long Contexts". In: *Transactions of the Association for Computational Linguistics* 12, pp. 157–173.

Liu, Pengfei et al. (2024). "Green AI: Towards Sustainable and Energy-Efficient Large Language Models". In: *Nature Machine Intelligence*.

Liu, Zechun et al. (2024). "Optimizing Large Language Models through Quantization: A Comparative Analysis of PTQ and QAT Techniques". In: *arXiv preprint arXiv:2411.06084*. arXiv: 2411.06084 [cs.LG].

McNemar, Quinn (1947). "Note on the sampling error of the difference between correlated proportions or percentages". In: *Psychometrika* 12.2, pp. 153–157. doi: 10.1007/BF02295996.

Mesnard, Thomas et al. (2025). "Gemma 3 Technical Report". In: *arXiv preprint arXiv:2503.19786*. 4B model competitive with Gemma 2 27B. SigLIP 400M vision encoder, Pan & Scan, 128K context. arXiv: 2503.19786 [cs.CL].

Monshi, Momina et al. (2024). "Vision-language models for medical report generation and visual question answering: a review". In: *Frontiers in Artificial Intelligence* 7, p. 1430984. doi: 10.3389/frai.2024.1430984.

Mosbach, Marius et al. (2023). "Few-shot Fine-tuning vs. In-context Learning: A Fair Comparison and Evaluation". In: *arXiv preprint arXiv:2305.16938*. Systematic comparison of adaptation approaches. arXiv: 2305.16938 [cs.CL].

Al-Naqbi, Ahmed et al. (2024). "Enhancing Organizational Productivity with Large Language Models". In: *Journal of Business Research*.

Niu, Yuanhao et al. (2024). "RAGTruth: A Hallucination Corpus for Developing Trustworthy Retrieval-Augmented Language Models". In: *arXiv preprint arXiv:2401.00396*. arXiv: 2401.00396 [cs.CL].

NVIDIA Corporation (2024). "Optimizing LLMs for Performance and Accuracy with Post-Training Quantization". In: *NVIDIA Technical Blog*. url: https://developer.nvidia.com/blog/optimizing-llms-for-performance-and-accuracy-with-post-training-quantization/.

Orenstein, Nicolas et al. (2023). "Exploring the potential of artificial intelligence in improving skin lesion diagnosis in primary care". In: *Scientific Reports* 13, p. 4569. doi: 10.1038/s41598-023-31340-1.

Oruganty, Kavitha et al. (2025). "DermETAS: A Dermatology-Specific Evaluation Framework for Medical AI". In: *npj Digital Medicine*.

Ovadia, Oded et al. (2024). "Fine-Tuning or Retrieval? Comparing Knowledge Injection in LLMs". In: *arXiv preprint arXiv:2312.05934*. Benchmark comparison showing RAG advantages for factual tasks. arXiv: 2312.05934 [cs.CL].

Page, Matthew J et al. (2021). "The PRISMA 2020 Statement: An Updated Guideline for Reporting Systematic Reviews". In: *BMJ* 372, n71.

Patrício, Cristiano, Luís F. Teixeira, and João C. Neves (2024). "Towards Concept-based Interpretability of Skin Lesion Diagnosis using Vision-Language Models". In: *arXiv preprint arXiv:2311.14339*. IEEE ISBI 2024. arXiv: 2311.14339 [cs.CV].

Pearce, Tim et al. (2024). "Reconciling Kaplan and Chinchilla Scaling Laws". In: *arXiv preprint arXiv:2406.12907*. arXiv: 2406.12907.

Petrick, Nicholas et al. (2023). "Regulatory Considerations for AI/ML-Based Medical Devices". In: *npj Digital Medicine*.

Pingua, Carlos et al. (2025). "Medical LLMs: A Comprehensive Review of Small Language Models in Healthcare". In: *Journal of Medical Internet Research*.

Raeini, Mohammad et al. (2025). "The Evolution of Natural Language Processing: From Statistical Methods to Large Language Models". In: *arXiv preprint*.

Rahman, Md et al. (2025). "Edge-AI integrated secure wireless IoT architecture for real time healthcare monitoring and federated anomaly detection". In: *Scientific Reports* 15, p. 30150. doi: 10.1038/s41598-025-30150-x.

Ru, Jinghan et al. (2026). "DermoGPT: Open Weights and Open Data for Morphology-Grounded Dermatological Reasoning MLLMs". In: *arXiv preprint arXiv:2601.01868*. Fine-tunes Qwen3-VL-8B-Instruct for dermatology via LoRA. DermoInstruct: 211K images, 772K training trajectories. arXiv: 2601.01868 [cs.CV].

Samsi, Siddharth et al. (2023). "From Words to Watts: Benchmarking the Energy Costs of Large Language Model Inference". In: *arXiv preprint arXiv:2310.03003*. Reports 10-100x energy cost differences between model scales. arXiv: 2310.03003 [cs.CL].

Sellergren, Andrew et al. (2025). "MedGemma Technical Report". In: *arXiv preprint arXiv:2507.05201*. Medical derivative of Gemma 3. MedSigLIP on 33M medical image-text pairs. +15.5 F1 CheXpert, +32.1 F1 SLAKE. arXiv: 2507.05201 [cs.CV].

Sharma, Pranav et al. (2025). "Evaluating Large Reasoning Model Performance on Complex Medical Scenarios In the MMLU-Pro Benchmark". In: *medRxiv preprint*. doi: 10.1101/2025.04.07.25325385.

Shohan, Faisal Tareque et al. (2025). "A Comprehensive Survey of Small Language Models in the Era of Large Language Models: Techniques, Enhancements, Applications, Collaboration with LLMs, and Trustworthiness". In: *ACM Transactions on Intelligent Systems and Technology*. doi: 10.1145/3768165.

Siriwardhana, Shamane et al. (2023). "Improving the Domain Adaptation of Retrieval Augmented Generation (RAG) Models for Open Domain Question Answering". In: *Transactions of the Association for Computational Linguistics* 11. Reports 15-25% improvement with domain-adapted RAG, pp. 1–17.

Soudani, Hossam et al. (2024). "Fine-Tuning Small Language Models for Medical Question Answering". In: *AMIA Annual Symposium Proceedings*.

Suarez, Perez et al. (2024). "A Comparative Analysis of Instruction Fine-Tuning Large Language Models for Financial Text Classification". In: *ACM Transactions on Management Information Systems*. doi: 10.1145/3706119.

Sung, Yi-Lin, Jaemin Cho, and Mohit Bansal (2022). "VL-Adapter: Parameter-Efficient Transfer Learning for Vision-and-Language Tasks". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Demonstrates VLM few-shot advantages over CNN approaches, pp. 5227–5237.

Taylor, Sarah et al. (2025). "Leveraging AI for Dermatological Diagnosis: Limitations and Disclaimers". In: *JAMA Dermatology*.

Tian, Yuanhe et al. (2025). "Beyond Single Agents: The Case for Multi-Agent Architectures". In: *arXiv preprint*.

Tran, Khanh-Tung et al. (2025). "Multi-Agent Collaboration Mechanisms: A Survey of LLMs". In: *arXiv preprint arXiv:2501.06322*. arXiv: 2501.06322 [cs.MA].

Tschandl, Philipp, Cliff Rosendahl, and Harald Kittler (2018). "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions". In: *Scientific Data* 5, p. 180161. doi: 10.1038/sdata.2018.161.

Vaswani, Ashish et al. (2017). "Attention is All You Need". In: *Advances in Neural Information Processing Systems* 30.

Wang, Benyuan et al. (2025). "Parameter-efficient fine-tuning in large language models: a survey of methodologies". In: *Artificial Intelligence Review*. doi: 10.1007/s10462-025-11236-4.

Wang, Cheng et al. (2024). "Survey on Knowledge Distillation for LLMs: Methods, Evaluation, and Application". In: *ACM Transactions on Intelligent Systems and Technology*. doi: 10.1145/3699518.

Wang, Hanchen et al. (2024). "Scientific Discovery in the Age of Artificial Intelligence". In: *Nature* 620, pp. 47–60.

Wang, Jiacheng et al. (2025). "A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation". In: *npj Digital Medicine*. doi: 10.1038/s41746-025-01670-7.

Wang, Lei et al. (2024). "A Survey on Large Language Model Based Autonomous Agents". In: *Frontiers of Computer Science* 18.6, p. 186345.

Wang, Shujin et al. (2025). "A Survey on Collaborating Small and Large Language Models for Performance, Cost-effectiveness, Cloud-edge Privacy, and Trustworthiness". In: *arXiv preprint arXiv:2510.13890*. arXiv: 2510.13890 [cs.CL].

Wang, Xiaohua et al. (2025). "Enhancing Retrieval-Augmented Generation: A Study of Best Practices". In: *arXiv preprint arXiv:2501.07391*. arXiv: 2501.07391 [cs.CL].

Wang, Yubo et al. (2024). "MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark". In: *Advances in Neural Information Processing Systems (NeurIPS)*.

Wang, Yuntao et al. (2024). "Security and Privacy Challenges in Edge AI for Healthcare". In: *IEEE Internet of Things Journal*.

Wei, Jason et al. (2023). "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models". In: *Advances in Neural Information Processing Systems* 35, pp. 24824–24837.

Weights & Biases (2020). *Experiment Tracking with Weights & Biases*. `https://wandb.ai`. Accessed: 2026-01-15.

Wu, Jiawei et al. (2025). "MedVision: A Vision-Language Model for Medical Visual Question Answering". In: *arXiv preprint*.

Wu, Qingyun et al. (2023). "AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation". In: *arXiv preprint arXiv:2308.08155*.

Wu, Shijie et al. (2024). "A Survey of Large Language Models in Finance". In: *arXiv preprint arXiv:2402.02315*.

Xi, Zhiheng et al. (2023). "The Rise and Potential of Large Language Model Based Agents: A Survey". In: *arXiv preprint arXiv:2309.07864*.

Xu, Xiaohan et al. (2024). "A Survey on Knowledge Distillation of Large Language Models". In: *arXiv preprint arXiv:2402.13116*. arXiv: `2402.13116 [cs.CL]`.

Xu, Zhiyuan et al. (2025). "A Review on Edge Large Language Models: Design, Execution, and Applications". In: *ACM Computing Surveys*. doi: `10.1145/3719664`.

Yang, An et al. (2024). "Qwen2.5-VL Technical Report". In: *arXiv preprint arXiv:2502.13923*. arXiv: `2502.13923 [cs.CV]`.

Yang, Sen et al. (2025). "Derm1M: A Million-scale Vision-Language Dataset Aligned with Clinical Ontology Knowledge for Dermatology". In: *arXiv preprint arXiv:2503.14911*. arXiv: `2503.14911 [cs.CV]`.

Yao, Yuan et al. (2025). "MiniCPM-V: Efficient GPT-4V level multimodal large language model for deployment on edge devices". In: *Nature Communications*. doi: `10.1038/s41467-025-61040-5`.

Zhang, Fei et al. (2025). "DRAGON: Efficient Distributed Retrieval-Augmented Generation for Enhancing On-Device LM Inference". In: *arXiv preprint arXiv:2504.11197*. arXiv: `2504.11197 [cs.CL]`.

Zhang, Sheng et al. (2024). "Large-Scale Multi-Modal Pre-trained Models: A Comprehensive Survey for Medical Few-Shot Learning". In: *Medical Image Analysis* 93. Meta-analysis showing VLM advantages in few-shot medical imaging, p. 103099. doi: `10.1016/j.media.2024.103099`.

Zhang, Wei et al. (2025a). "The Education Revolution: How LLMs are Transforming Learning". In: *Educational Technology Research*.

– (2025b). "Vision-Language Models for Edge Networks: A Comprehensive Survey". In: *arXiv preprint arXiv:2502.07855*. arXiv: `2502.07855 [cs.CV]`.

Zhang, Yue et al. (2025). "A Comprehensive Survey of Hallucination in LLMs: Causes, Detection, and Mitigation". In: *arXiv preprint arXiv:2510.06265*. arXiv: `2510.06265 [cs.CL]`.

Zhang, Yuxiang et al. (2024). "Small Language Models for Efficient Agentic Tool Calling: Outperforming Large Models with Targeted Fine-tuning". In: *arXiv preprint arXiv:2512.15943*. AWS Research. arXiv: `2512.15943 [cs.CL]`.

Zhao, Yilong et al. (2024). "ATOM: Low-bit Quantization for Efficient and Accurate LLM Serving". In: *Proceedings of Machine Learning and Systems (MLSys)*.

Zhou, Juexiao et al. (2024). "Pre-trained multimodal large language model enhances dermatological diagnosis using SkinGPT-4". In: *Nature Communications* 15.1, p. 5649. doi: `10.1038/s41467-024-50043-3`.

Zhou, Minghao et al. (2025). "Multi-Agent Memory: Collaborative Context Management in LLM Systems". In: *arXiv preprint*.

Ziller, Alexander et al. (2024). "Reconciling Privacy and Utility in Medical AI". In: *Nature Medicine*.

Appendix A

Appendix Title Here

Write your Appendix content here.