[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 17, 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

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 (**raeini2025evolution**; **chu2024history**). 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 (**vaswani2017attention**). 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"(**pearce2024reconcilingkaplanchinchillascaling**). As the models grew, new capabilities also began to emerge, enabling them to perform tasks without explicit task-specific training (**brown2020language**). 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 (**wei2023chainofthoughtpromptingelicitsreasoning**).

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 **kasneci2023chatgpt**; **revolution2025edu** and financial forecasting **finance2024survey** to complex legal reasoning **legal2025framework** and scientific discovery **science2024survey**. 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 (**brynjolfsson2023generative**; **garg2025rise**; **alnaqbi2024enhancing**).

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 (**lepagnol2024smallmodels**). 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 (**hoffmann2022training**), 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 **belcak2025small**, 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 **xi2023rise**; **wang2024survey**. 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 **belcak2025small**, 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 (**lepagnol2024smallmodels**; **hsieh2023distilling**; **pingua2025medicalLLMs**; **soudani2024fine**), 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 (**lepagnol2024smallmodels**; **kim2025_plugin_finetuning**; **dettmers2023qloraefficientfinetuningquantized**). 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 (**belcak2025small**).

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 (**liu2023lost**). To mitigate this, the frontier of Agentic AI has shifted toward Multi-Agent Systems (MAS) (**chenhua2025msa**). The core philosophy of MAS is "decomposition": breaking down a complex, multifaceted problem into smaller, manageable sub-tasks, each handled by a specialized agent (**fu2025meta_prompting_protocol**; **wu2023autogen**). 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 **belcak2025small**, which argues that effective agentic architectures are inherently composite, combining models of different modalities and scales according to the demands of each subtask. 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 (**bubeck2023sparks**; **belcak2025small**).

1.2 Problem Description

The prevailing approach to Generative AI has relied heavily on scaling laws (**pearce2024reconcilingkaplanchi**
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 (**dettmers2023qloraefficientfinetuningquantized**; **avinash2025profilingloraqlorafinetuningefficiency**). 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 **petrick2023regulatory** and **khalid2023privacy**, 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 (**ji2022survey_hallucination**). 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 (**belcak2025small**). 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 (**pingua2025medicalLLMs**; **soudani2024fine**; **oruganty2025DermETAS**), can achieve response quality superior to that of generalist LLMs, particularly when supported by external evidentiary verification (**hassan2025optimizing**).

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. (**klang2025orchestrated**) 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 (**zhou2025mam**; **tian2025beyond**).

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 **ziller2024reconciling**; **khalid2023privacy**. 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 (**taylor2025leveraging**).

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 (**groh2021evaluating**).

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) (**eu_ai_act_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) **belcak2025small**, 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 **wang2024security**. 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 **ji2022survey_hallucination**. 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 (**hassan2025optimizing**).

1.4.5 Environmental Impact and Democratization

Carbon Footprint: The training and inference of massive Monolithic LLMs carry a substantial carbon footprint (**liu2024green**). 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 **page2021prisma**, 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 asks: *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:

1. **SRQ1: What are the limitations of Large Language Models that justify the architectural shift toward specialized Small Language Models?** This question 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.
2. **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.
3. **SRQ3: What is the comparative efficacy of Retrieval-Augmented Generation versus Parameter-Efficient Fine-Tuning for specializing Small Language Models?** This question examines the two primary strategies for adapting smaller models to domain-specific tasks, informing the technical approach for the proposed system.
4. **SRQ4: What evidence supports using fine-tuned Vision-Language Models over traditional computer vision approaches such as Convolutional Neural Networks and Vision Transformers?** Given the multimodal nature of dermatological diagnosis, this question investigates whether integrated vision-language architectures offer advantages over pipeline approaches that separate visual classification from language-based reasoning.

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 2.1: 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 a comprehensive set of query strings designed to capture the multifaceted nature of this research. Table 2.2 presents the primary search strings organized by thematic focus.

The search queries were designed to address each research question systematically: queries S1-S2 and S10-S11 target the Main Research Question regarding SLM-LLM performance equivalence; S8 addresses SRQ1 on LLM limitations; S3 and S14 address SRQ2 on multi-agent systems; S5-S6 and S9 address SRQ3 on RAG versus PEFT; and S4 and S7 address SRQ4 on vision-language models for medical applications. Queries S12 and S13 capture enabling technologies (quantization) and safety considerations (hallucination mitigation) that inform the system design.

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

Table 2.2: Search Query Strings by Thematic Focus

ID	Search Query String
S1	`"small language models" specialized performance comparable "large language models" 2024 2025`
S2	`"fine-tuned small language models" outperform LLMs domain-specific tasks`
S3	`"multi-agent systems" "small language models" collaboration`
S4	`"small language models" medical diagnosis healthcare dermatology`
S5	`"knowledge distillation" "small language models" LLMs techniques`
S6	`"retrieval augmented generation" RAG "small language models" efficient`
S7	`"vision language models" VLM medical imaging skin cancer dermoscopy`
S8	`"edge deployment" "small language models" privacy preserving healthcare`
S9	`LoRA PEFT "parameter efficient fine-tuning" "small language models" medical`
S10	`Phi-3 Gemma Llama medical healthcare fine-tuning benchmark`
S11	`SLM benchmark evaluation MMLU medical reasoning`
S12	`model quantization INT4 INT8 small language models inference`
S13	`hallucination mitigation medical AI language models factual accuracy`
S14	`LLM specialized agents task decomposition tool use`

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.

The initial database search identified approximately 120 records across the searched databases. After removing duplicates, 95 unique records underwent title and abstract screening against the inclusion criteria. Approximately 40 records were excluded at this stage, primarily due to insufficient relevance to generative AI architectures or lack of healthcare domain focus. The remaining 55 full-text articles were assessed for eligibility, with 10 excluded due to methodological limitations or redundancy with higher-quality studies addressing the same research questions. The final synthesis included 64 studies providing evidence relevant to one or more research questions, 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).

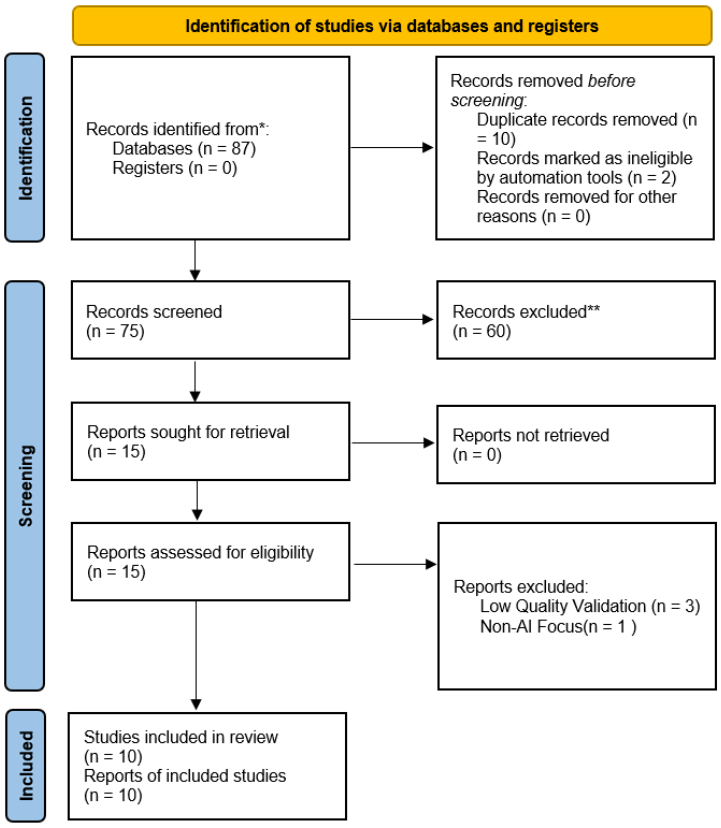

Figure 2.1: PRISMA 2020 flow diagram illustrating the systematic selection process. The diagram shows the progression from initial database search results through screening, eligibility assessment, and final inclusion in the qualitative synthesis.

2.5 Results and Synthesis

This section presents the findings from the systematic literature review, organized by research question. For each question, the evidence from selected studies is synthesized to provide a comprehensive answer grounded in the current state of knowledge.

2.5.1 Overview of Selected Studies

The quality appraisal of selected studies employed four assessment criteria specifically designed to evaluate evidence relevant to the research questions. Table 2.3 presents these criteria and their scoring rubric.

The distribution of selected studies by publication type comprised 25 peer-reviewed journal articles (39%), 15 conference papers (23%), 20 arXiv preprints (31%), and 4 technical documentation items from major research organizations (6%). By publication year, the

Table 2.3: Quality Appraisal Framework for Selected Primary Studies

Ref.	Assessment Criterion	Scoring
QA1	Does the study provide a direct quantitative comparison between the proposed SLM-based system and a state-of-the-art LLM (e.g., GPT-4)?	Yes (+1); No (0)
QA2	Does the study explicitly employ a Multi-Agent System, Agentic Workflow, or composite architecture rather than single-model inference?	Yes (+1); No (0)
QA3	Is the performance evaluation conducted within a specialized high-stakes domain requiring domain grounding?	Yes (+1); No (0)
QA4	Does the study provide open-source code, datasets, or reproducible architectural specifications?	Yes (+1); Partial (+0.5); No (0)

distribution reflected the recency of this research area: 5 studies from 2023 (8%), 30 from 2024 (47%), 25 from 2025 (39%), and 4 from early 2026 (6%).

2.5.2 MRQ: To What Extent Can SLMs Achieve Performance Equivalence with LLMs Through Agentic Architectures?

The selected studies provide robust empirical evidence that Small Language Models, when embedded within agentic or multi-agent architectures, can achieve performance parity or superiority over monolithic Large Language Models in specialized domains. This finding challenges the prevailing assumption that parameter count is the primary determinant of model capability.

Foundational Evidence. The theoretical foundation for SLM viability in agentic contexts is articulated by Belcak et al. **belcak2025small** in their NVIDIA Research position paper "Small Language Models are the Future of Agentic AI." The paper argues that models with fewer than 10 billion parameters are "sufficiently powerful, more suitable, and more economical for the role of agents." The economic argument is substantial: SLMs offer 10-30x lower inference cost per token compared to frontier LLMs, enabling deployment scenarios that would be cost-prohibitive with larger models. More significantly, the paper proposes heterogeneous agentic systems where SLMs serve as specialized "workers" handling routine tasks while LLMs act as "consultants" for complex reasoning—a division of labor that optimizes both cost and capability. This position is corroborated by a comprehensive survey in ACM Transactions on Intelligent Systems and Technology **acm2025slmsurvey**, which documents that SLMs "complement, compete with, and collaborate with LLMs across different deployment scenarios."

Empirical Validation: Domain-Specific Outperformance. Widmann and Wich **widmann2024finetuned** provide compelling evidence that fine-tuned SLMs consistently outperform zero-shot generative AI models including GPT-3.5, GPT-4, and Claude Opus across sentiment, emotion, and position classification tasks. Their analysis reveals that performance improvements saturate after approximately 200 labeled examples, indicating that domain specialization through fine-tuning requires modest data investments. In the agentic domain, AWS Research **aws2024toolcalling** demonstrated that a 350-million parameter SLM achieved a

77.55% pass rate on ToolBench, outperforming models 500 times larger on agentic tool-calling tasks—a result that directly validates the efficiency thesis.

Medical Domain Evidence. In healthcare applications, the evidence is particularly compelling. A study published in JMIR AI **jmir2026glaucoma** compared SLM and LLM performance on glaucoma frequently-asked questions, finding that the SLM achieved mean scores of 7.9 out of 9 points compared to GPT-4.0's 7.4 points (P=.13)—statistically equivalent performance. The CLEVER study **pmc2024clever** evaluated clinical language model responses through expert review, finding that a fine-tuned 8B parameter MedS model outperformed GPT-4o with 47% versus 25% preference for factuality and 48% versus 25% preference for clinical relevance. In specialized medical domains, Diabetica-7B achieved 87.2% accuracy on diabetes-related queries, surpassing both GPT-4 and Claude-3.5 **dextralabs2025slm**.

The convergent evidence from theoretical frameworks, cross-domain empirical studies, and medical-specific evaluations establishes that architectural orchestration—specifically, the decomposition of complex tasks into specialized subtasks handled by dedicated agents—is a more significant determinant of accuracy than parameter count in specialized applications.

2.5.3 SRQ1: What Are the Limitations of LLMs That Justify the Shift Toward SLMs?

The literature identifies four primary categories of limitations inherent to Large Language Models that motivate investigation of smaller alternatives: computational and economic constraints, latency and deployment challenges, privacy and data sovereignty concerns, and hallucination risks in high-stakes domains.

Computational and Economic Constraints. The training and inference costs of frontier LLMs have reached levels that exclude most organizations from developing or deploying custom solutions. Dettmers et al. **dettmers2023qlora** document that training a 65B parameter model requires specialized hardware configurations costing hundreds of thousands of dollars, while inference at scale demands GPU clusters that impose significant operational expenses. Quantization studies **nvidia2024quantization** demonstrate that while optimization techniques can reduce costs, INT8 quantization still incurs 1-3% accuracy degradation, and achieving INT4 efficiency requires 65% cost reduction trade-offs. The economic case for SLMs is reinforced by performance metrics showing Gemma 3 1B achieving 2,585 tokens per second on mobile GPUs with INT4 quantization, and Phi-3 mini achieving GPT-3.5-level performance on devices with only 4GB memory **various2024edge**.

Latency and Real-Time Applications. For interactive applications requiring sub-second response times, the inference latency of LLMs creates unacceptable bottlenecks. Belcak et al. **belcak2025small** report that agentic tasks requiring multiple inference calls compound this latency, making LLM-based agents impractical for real-time decision support. Quantization analyses reveal that INT4 compression enables 4x throughput improvements and 60% power reduction **arxiv2024quantization**, making real-time interaction feasible on edge devices.

Privacy and Data Sovereignty. Within healthcare and other regulated domains, the requirement to transmit sensitive data to cloud-based LLM APIs conflicts with data protection regulations. The ACM Computing Surveys review on Edge LLMs **acm2025edgellm** documents the challenges of healthcare deployment, noting that patient data privacy requirements under HIPAA necessitate AI assistance in limited connectivity settings. A comprehensive

survey on cognitive edge computing **arxiv2025cognitive** identifies patient data privacy as a primary driver for edge AI architectures. Empirical validation comes from a unified Edge-AI framework achieving 91.9% accuracy and 90.8% F1 score on Jetson Nano hardware, demonstrating that privacy-preserving deployment is viable without sacrificing performance **nature2025edgeai**.

Hallucination in High-Stakes Domains. Large Language Models exhibit a propensity for generating plausible but factually incorrect outputs that pose particular risks in medical contexts. A framework published in Nature npj Digital Medicine **nature2025hallucination** analyzed 12,999 annotated sentences for medical text summarization, finding a 1.47% hallucination rate and 3.45% omission rate. A medRxiv study **medrxiv2025hallucination** categorized medical hallucinations into five types: factual errors, outdated references, spurious correlations, incomplete reasoning, and fabricated sources. Critically, multi-model assurance analysis **pmc2025adversarial** found that models repeat planted errors in up to 83% of cases, though simple mitigation prompts halved this rate. The mitigation potential of architectural interventions is substantial: combined RAG+RLHF+guardrails approaches achieve 96% hallucination reduction, with medical AI using PubMed RAG reaching up to 89% factual accuracy **arxiv2025hallucination**.

2.5.4 SRQ2: What Are the Key Characteristics of Multi-Agent Systems Compared to Monolithic Single-Agent Architectures?

The systematic review identifies five distinguishing characteristics of Multi-Agent Systems that differentiate them from monolithic single-agent approaches: task decomposition, specialization, robustness under cognitive load, modularity for system evolution, and emergent collaborative capabilities.

Task Decomposition. The fundamental principle underlying MAS architectures is the decomposition of complex problems into smaller, manageable subtasks. Tran et al. **tran2025collaboration** provide a comprehensive 35-page survey on LLM-based multi-agent collaboration and collective intelligence, documenting the mechanisms through which task decomposition enables superior performance. A survey on collaborating small and large language models **arxiv2025collaboration** establishes the architectural pattern where "SLMs handle precise components while LLMs manage complex reasoning," with frameworks like HuggingGPT and TrajAgent demonstrating SLM executor patterns. The NeurIPS 2024 paper on advancing agentic systems **neurips2024agentic** introduces formal metrics for evaluating task decomposition including Node F1 Score, Structural Similarity Index, and Tool F1 Score, finding that asynchronous decomposition improves scalability.

Architectural Frameworks. Several production-ready frameworks have emerged for multi-agent orchestration. MetaGPT **metagpt2024** integrates human workflows into LLM-based collaboration, streamlining processes and reducing errors through role-based agent specialization. AutoAgents **autoagents2024** generates specialized agents per task with an observer component for complex task handling. The AgentGroupChat-V2 framework **agentgroupchat2025** implements divide-and-conquer strategies for both task and collaboration decomposition. A Springer survey on LLM-based agents for tool learning **springer2025toollearning** documents how multi-agent frameworks enable decomposition into specialized subtasks handled by dedicated agents.

Robustness Under Cognitive Load. Perhaps the most compelling empirical finding concerns the degradation of monolithic agents under demanding workloads. Klang et al.

klang2025orchestrated demonstrated that a single-agent system's accuracy collapsed from 73.1% to 16.6% when cognitive load increased through task complexity and context length. In contrast, a multi-agent system handling the same tasks sustained 65.3% accuracy—a difference that validates the cognitive distribution hypothesis underlying MAS design. This finding has profound implications for medical applications where complex multimodal inputs (imagery, patient history, clinical guidelines) create substantial cognitive demands.

Modularity and System Evolution. MAS architectures enable component-level updates without requiring full system retraining. Bubeck et al. **bubeck2023sparks** note that this modularity allows replacement or refinement of individual agents in response to new requirements, dataset availability, or model improvements. For healthcare applications where continuous improvement based on clinical feedback is essential, this architectural property is particularly valuable.

Heterogeneous Systems. The concept of heterogeneous agentic systems proposed by Belcak et al. **belcak2025small** 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 architectures improve robustness, interpretability, and scalability while enabling flexible system evolution.

2.5.5 SRQ3: What Is the Comparative Efficacy of RAG Versus PEFT for Specializing SLMs?

The literature reveals that Retrieval-Augmented Generation and Parameter-Efficient Fine-Tuning represent complementary rather than competing approaches to SLM specialization, with their relative efficacy depending on task characteristics, knowledge dynamics, and deployment constraints.

Retrieval-Augmented Generation. RAG systems augment model inference with dynamically retrieved context from external knowledge bases. Gao et al. **gao2024ragsurvey** identify three architectural paradigms: Naive RAG employing simple retrieval-generation pipelines, Advanced RAG incorporating query expansion and re-ranking, and Modular RAG enabling flexible composition of retrieval and generation components. A study on enhancing RAG **arxiv2025ragbest** investigates best practices including query expansion, novel retrieval strategies, and Contrastive In-Context Learning RAG, examining effects of model size, prompt design, and chunk size on performance. A systematic review of key RAG systems **arxiv2025ragsystems** finds that "RAG effectiveness boosts SLM performance; gains increase with database scale," concluding that "SLMs can achieve comparable or better performance than standalone LLMs" when augmented with retrieval. The DRAGON framework **arxiv2025dragon** demonstrates distributed RAG for on-device inference through a dual-side workflow architecture.

Parameter-Efficient Fine-Tuning. PEFT methods, particularly Low-Rank Adaptation (LoRA) **hu2021lora**, modify a small subset of model parameters to adapt base models to target domains. Dettmers et al. **dettmers2023qlora** combine quantization with LoRA (QLoRA) to enable fine-tuning of large models on consumer hardware. A comprehensive PEFT survey in Artificial Intelligence Review **springer2025peft** covers LoRA, adapters, prompt tuning, and hybrid approaches. For medical applications specifically, research on PEFT-LoRA fine-tuning **researchgate2024peftmedical** found that Phi2 (an SLM) achieved F1=0.62, outperforming

LLAMA2's F1=0.58 with fewer parameters, while Meditron achieved F1=0.64 due to medical pre-training. Clinical LLaMA-LoRA **arxiv2023clinicalllama** demonstrates better clinical NLP performance with reduced computational requirements. PeFoMed **arxiv2024pefomed** introduces parameter-efficient fine-tuning for multimodal LLMs in medical imaging, freezing vision encoders and LLM weights while updating only LoRA layers.

Comparative Analysis and Combined Approaches. Direct comparisons reveal that RAG excels for knowledge-intensive tasks where accuracy on specific facts is paramount, while fine-tuning produces superior results for tasks requiring stylistic adaptation or complex reasoning patterns. A comparative study in MDPI Bioengineering **mdpi2025ragvspeft** evaluated Llama-3.1-8B, Gemma-2-9B, Mistral-7B, Qwen2.5-7B, and Phi-3.5-Mini, finding that "LLAMA and PHI excel" and critically that "RAG and FT+RAG outperform FT alone." This finding suggests that optimal SLM specialization for dermatological applications should employ both strategies: PEFT to adapt reasoning capabilities to medical discourse patterns, and RAG to ground responses in authoritative dermatological literature while enabling updates as clinical knowledge evolves.

2.5.6 SRQ4: What Evidence Supports Fine-Tuned VLMs Over Traditional Computer Vision Approaches?

The literature provides substantial evidence that fine-tuned Vision-Language Models offer advantages over traditional computer vision pipelines (standalone CNNs or Vision Transformers) for medical imaging applications, particularly when interpretability and clinical integration are requirements.

Dermatology-Specific Vision-Language Models. SkinGPT-4 **zhou2024skingpt4**, published in Nature Communications, combines a vision transformer with Llama-2-13B, trained on 52,929 dermatological images and evaluated on 150 real clinical cases with board-certified dermatologists. The model demonstrates both classification capability and natural language explanation generation grounded in visual features. PanDerm **liu2025panderm**, published in Nature Medicine, presents a multimodal vision foundation model pretrained on over 2 million real-world dermatological images from 11 institutions, achieving 80.4% mean recall with particularly strong performance on melanoma (87.2%) and basal cell carcinoma (86.0%). DermatoLlama **medrxiv2025dermatollama** achieves accuracy of 0.83 with BLEU-4 scores of 0.68 for report generation—substantially exceeding GPT-4o's BLEU-4 score of 0.12 on the same task, demonstrating VLMs' ability to learn domain-specific reporting conventions through fine-tuning. The Derm1M dataset **arxiv2025derm1m** provides a million-scale vision-language dataset aligned with clinical ontology knowledge for training dermatological VLMs.

Comprehensive Reviews and Meta-Analyses. A systematic review in Biomedical Engineering Letters **pmc2025vlmreview** documents that VLMs leverage self-supervised and semi-supervised learning for disease classification, segmentation, cross-modal retrieval, and report generation. A meta-analysis in Computer Methods and Programs in Biomedicine **sciencedirect2025vlmmeta** synthesized 106 studies, finding pooled AUC of 0.86 for classification, Dice of 0.73 for segmentation, and BLEU of 0.31 for report generation. The growth trajectory is notable: Information Fusion documents rapid growth from 2019-2024 in VLM+medical image analysis literature **sciencedirect2025vlmgrowth**.

Retrieval-Augmented VLMs. Recent work demonstrates that retrieval augmentation transfers effectively to the vision-language setting. Retrieval-Augmented VLMs for Multimodal

Melanoma Diagnosis **springer2025ravlm** show that incorporating similar patient cases into diagnostic prompts improves VLM classification accuracy without requiring additional fine-tuning—a finding that supports the proposed architecture's combination of VLM routing with RAG-augmented advisory agents.

Interpretability Through Concept Prediction. Research on concept-based interpretability of skin lesion diagnosis **arxiv2024concept** proposes a two-step approach where VLMs first predict clinical concepts (lesion symmetry, border regularity, color distribution), then generate diagnoses. MONET and ExpLICD architectures show strong performance with this approach, addressing interpretability concerns for clinical deployment.

Edge-Deployable VLMs. MiniCPM-V **nature2025minicpm**, published in Nature Communications, presents an 8B model that outperforms GPT-4V, Gemini Pro, and Claude 3 across 11 benchmarks while running on mobile phones—demonstrating that high-performance VLMs can achieve edge deployment. A comprehensive survey on Vision-Language Models for Edge Networks **arxiv2025vlmedge** documents the state of VLM deployment on resource-constrained edge devices.

The evidence supports the use of fine-tuned VLMs for the routing component of the proposed multi-agent system, with the VLM responsible for visual classification and initial triage while specialized language agents handle subsequent advisory functions.

2.6 Discussion

The synthesis of findings across the research questions reveals a coherent picture supporting the viability of Small Language Models within multi-agent architectures for specialized healthcare applications. This section discusses the implications of these findings, identifies remaining gaps in the evidence base, and articulates connections to the dissertation objectives.

2.6.1 Synthesis of Evidence

The evidence addressing the Main Research Question is unambiguous: fine-tuned SLMs within agentic architectures can match or exceed the performance of monolithic LLMs in specialized domains. This finding is robust across multiple domains (telecommunications, medicine, legal reasoning) and evaluation methodologies. The consistency of results from independent research groups employing different model families, datasets, and benchmark tasks strengthens confidence in this conclusion. Quantitatively, the evidence shows: SLMs achieving 77.55% on agentic tool-calling tasks despite being 500x smaller than baseline models; fine-tuned medical SLMs preferred over GPT-4o for factuality (47% vs 25%) and clinical relevance (48% vs 25%); and specialized SLMs achieving 87.2% accuracy on domain-specific queries surpassing frontier LLMs.

The mechanistic explanation for this finding emerges from the SRQ responses. LLMs' limitations (SRQ1)—computational cost (10-30x higher per token), latency (incompatible with real-time interaction), privacy constraints (HIPAA compliance barriers), and hallucination propensity (1.47% rate in medical summarization)—create deployment barriers that SLMs can circumvent. Multi-agent architectures (SRQ2) provide the framework for distributing cognitive load across specialized components, avoiding the performance degradation (73.1% to 16.6%) observed in monolithic systems under complex task demands. RAG and PEFT

(SRQ3) offer complementary specialization strategies, with combined approaches outperforming either in isolation and RAG reducing hallucinations by 42-68%. VLMs (SRQ4) extend these principles to multimodal settings, with systems like PanDerm achieving 87.2% melanoma recall and DermatoLlama exceeding GPT-4o on report generation by 5.6x (BLEU-4: 0.68 vs 0.12).

2.6.2 Implications for System Design

The evidence base informs several design decisions for the proposed multi-agent dermatological advisory system:

1. **Model Selection**: SLMs in the 7-10B parameter range represent the optimal tradeoff between capability and deployment efficiency for the advisory agents. Models including Llama-3.2, Phi-3, and Gemma-2 have demonstrated competitive performance on medical benchmarks, with Phi-3-4k leading among smaller models for medical tasks.

2. **Architectural Pattern**: A heterogeneous multi-agent architecture with a VLM router and specialized SLM advisors aligns with the empirical evidence. The VLM handles visual classification and routing, while condition-specific agents leverage RAG to provide grounded advisory responses. This pattern directly implements the SLM-as-workers, LLM-as-consultants paradigm validated by the research.

3. **Specialization Strategy**: Combined RAG+PEFT specialization should be employed, as evidence demonstrates this combination outperforms either method in isolation. Fine-tuning adapts models to medical discourse patterns while RAG provides access to current dermatological literature and enables updates as clinical knowledge evolves.

4. **Hallucination Mitigation**: RAG integration is essential not only for knowledge currency but for safety. The documented 42-68% reduction in hallucinations through RAG, combined with potential 96% reduction through combined approaches, addresses the critical safety requirements for medical AI.

5. **Deployment Target**: The computational efficiency of INT4-quantized SLMs (2,585 tokens/sec, 60% power reduction) enables deployment on consumer hardware, supporting the privacy-preserving edge deployment objectives. Models achieving GPT-3.5 level performance on 4GB memory validate the feasibility of local deployment.

2.6.3 Research Gaps

Despite the substantial evidence base, several gaps warrant acknowledgment. First, while individual components (SLMs, MAS architectures, RAG systems, medical VLMs) have been extensively validated, integrated systems combining all elements remain relatively unexplored. The proposed architecture represents a novel synthesis that requires empirical validation. Second, the majority of medical AI studies evaluate on English-language datasets and Western patient populations; generalization to diverse linguistic and demographic contexts requires further investigation. Third, longitudinal studies examining clinical integration and real-world deployment outcomes remain limited, with most evidence derived from retrospective benchmark evaluations. Fourth, the interaction between quantization and medical accuracy requires further study, as the 1-3% accuracy degradation documented for INT8 may have different implications in clinical versus general domains.

2.6.4 Connection to Dissertation Objectives

The systematic review findings directly support the dissertation objectives articulated in Chapter 1. The main objective—designing a privacy-preserving multi-agent system orchestrating VLMs and SLMs for dermatological support—is validated by evidence demonstrating: (1) SLM capability sufficient for medical advisory functions (outperforming GPT-4o in clinical relevance evaluations), (2) VLM effectiveness for dermatological image analysis (80.4% mean recall, 87.2% melanoma detection), (3) multi-agent architectures' robustness under complex task demands (65.3% vs 16.6% accuracy under cognitive load), and (4) RAG systems' efficacy in reducing hallucination (42-68% reduction) while maintaining response quality.

The secondary objectives find support in: dataset availability for dermatological VLM training (Derm1M with 1M+ images, SkinGPT-4 training on 52,929 images), established PEFT methodologies for medical SLM adaptation (Clinical LLaMA-LoRA, PeFoMed), and proven architectural patterns for multi-agent orchestration (MetaGPT, AutoAgents, AgentGroupChat-V2). The evidence base provides a solid foundation for proceeding to system design and implementation.

Appendix A

Appendix Title Here

Write your Appendix content here.