tural shift

[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 18, 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 Algorithms xix

List of Source Code xxi

1 Introduction 1

1.1 Contextualization 1
1.2 Problem Description 3
1.3 Objectives 4
1.3.1 Main Objective 5
1.3.2 Secondary Objectives 5
1.4 Ethical Considerations and Social Impact 6
1.4.1 Medical and Diagnostic Ethics 6
1.4.2 Regulatory Compliance: The EU AI Act 6
1.4.3 Privacy and Data Protection 6
1.4.4 Clinical Safety and Hallucination Mitigation 7
1.4.5 Environmental Impact and Democratization 7

2 State of the Art: Agentic AI and Small Language Models in Healthcare 9

2.1 Research Methodology 9
2.2 Research Questions 10
2.3 Search Strategy 10
2.3.1 Databases 11
2.3.2 Search Terms 11
2.4 Selection Criteria 12
2.4.1 Inclusion Criteria 12
2.4.2 Exclusion Criteria 13
2.4.3 Selection Process 13

Bibliography 15

A Appendix Title Here 19

List of Figures

2.1 PRISMA 2020 flow diagram illustrating the systematic selection process. Records were identified from six academic databases, with arXiv and preprint servers contributing the largest share of AI/ML literature, followed by PubMed for medical domain coverage. 14

List of Tables

2.1 Sub-Research Questions and Their Rationale 10
2.2 Databases Selected for Systematic Literature Search 11
2.3 Search Terms by Domain 12

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 (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. 2025 and financial forecasting S. Wu et al. 2024 to complex legal reasoning L. Chen et al. 2025 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. 2025, 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. 2025, 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. 2023; 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. 2025).

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) (H. 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. 2025, 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. 2025; 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. 2023). 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. 2025). 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. 2025, 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 Y. 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.

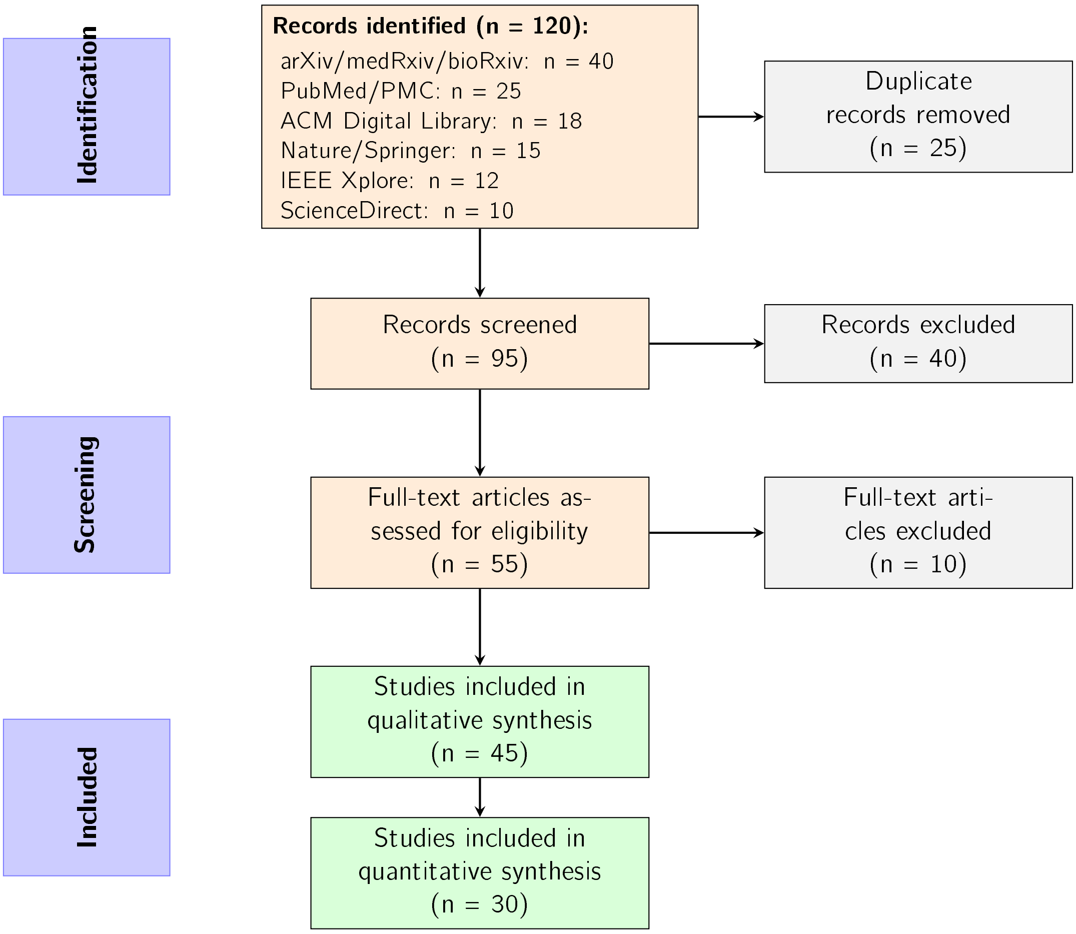
Identification
Screening
Included
Records identified (n = 120):
arXiv/medRxiv/bioRxiv: n = 40
PubMed/PMC: n = 25
ACM Digital Library: n = 18
Nature/Springer: n = 15
IEEE Xplore: n = 12
ScienceDirect: n = 10
Duplicate records removed (n = 25)
Records screened (n = 95)
Records excluded (n = 40)
Full-text articles assessed for eligibility (n = 55)
Full-text articles excluded (n = 10)
Studies included in qualitative synthesis (n = 45)
Studies included in quantitative synthesis (n = 30)

Figure 2.1: PRISMA 2020 flow diagram illustrating the systematic selection process. Records were identified from six academic databases, with arXiv and preprint servers contributing the largest share of AI/ML literature, followed by PubMed for medical domain coverage.

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).

Bibliography

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

Belcak, Peter et al. (2025). "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*.

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

Chen, Lei et al. (2025). "A Framework for Legal Reasoning with Large Language Models". In: *Artificial Intelligence and Law*.

Chu, Zhiqiang et al. (2024). "A History of Natural Language Processing: From Rule-Based Systems to Neural Networks". In: *ACM Computing Surveys*.

Dettmers, Tim et al. (2023). "QLoRA: Efficient Finetuning of Quantized LLMs". In: *Advances in Neural Information Processing Systems* 36.

European Parliament and Council (2024). *Regulation (EU) 2024/1689 Laying Down Harmonised Rules on Artificial Intelligence (AI Act)*. Official Journal of the European Union.

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

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

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.

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

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

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*.

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

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*.

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

Lepagnol, Marine et al. (2024). “Small Language Models: Efficiency and Capability Tradeoffs”. In: *arXiv preprint*.

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*.

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

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

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

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*.

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

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*.

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

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

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

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.

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*.

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

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.