[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, December 30, 2025

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 is particularly suited for technological domains such as Agentic AI and medical imaging, where rapid innovation necessitates rigorous evidence synthesis.

The process was structured into four distinct phases: (1) identification of relevant studies through database searching; (2) screening of titles and abstracts based on defined inclusion criteria; (3) eligibility assessment of full-text articles; and (4) qualitative synthesis of the selected studies to answer the defined research questions.

2.2 Research Questions

To structure the systematic review and ensure focused coverage of relevant literature, the following Research Questions (RQs) were formulated. These RQs are specific to the literature review, questions that existing papers will answer to establish the State of the Art.

Main Research Question (MRQ):

> *To what extent can SLMs achieve performance equivalence with LLMs in specialized domains through Agentic Architectures?*

Sub Research Questions (SRQ):

> *What are the limitations of LLMs that justify the architectural shift toward specialized SLMs?*
>
> *What are some of the key characteristics of a Multi-Agent System compared to a Monolithic Single-Agent?*
>
> *What is the comparative efficacy of Retrieval-Augmented Generation (RAG) versus Parameter-Efficient Fine-Tuning (PEFT) for specializing SLMs?*

What evidence supports using fine-tuned Vision-Language Models over traditional computer vision approaches (CNNs, ViTs)?

2.3 Databases

To ensure comprehensive coverage of relevant literature across computer science, medical informatics, and AI research, the following databases were systematically searched:

Table 2.1: Search Strategy: Domains, Databases, and Keywords

Database	Description
arXiv	An open-access repository for electronic preprints in the fields of computer science, mathematics, and statistics. It is the primary source for state-of-the-art research in Generative AI, Large Language Models (LLMs), and Small Language Models (SLMs) due to the rapid pace of development in the field.
PubMed	A free search engine accessing primarily the MEDLINE database of references and abstracts on life sciences and biomedical topics. It is the gold standard for retrieving peer-reviewed literature on dermatology, telemedicine, and clinical diagnostic methodologies.
IEEE Xplore	A digital library providing access to scientific and technical content published by the IEEE. It is essential for technical papers regarding hardware efficiency, edge computing, Multi-Agent System architectures, and optimization techniques for neural networks.
ScienceDirect	A large database of scientific and medical research hosted by Elsevier. It contains high-impact, peer-reviewed journals covering both artificial intelligence applications in healthcare and foundational studies in computer vision.

2.4 Search Terms

The search strategy utilized boolean logic to combine keywords representing the three core pillars of this dissertation: (1) Generative AI, (2) Artificial Intelligence, (2) Health, and (3) Architecture. The search strings were adapted to the syntax of each database.

2.5 Inclusion and Exclusion Criteria

2.5.1 Inclusion Criteria

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

1. **Publication Date**: Published or made publicly available between January 2018 and December 2025.
2. **Relevance to AI/ML Models**:

Domain	Keywords
Health	("Dermatology" OR "Medical" OR "Skin Disease") OR ("Skin" OR "Skin lesion")
Generative AI	("Large Language Models" OR "LLMs") AND ("Small Language Modles" OR " SLMs")
Artificial Intelligence	("CNN" AND "VLM") OR ("CNN" OR " Vision Language Model") OR ("MLMM" OR "Multimodal Language Model") OR ("CNN" AND "Transformer") OR ("AI" OR "Artificial Intelligence")
Architecture	("MAS" AND "Agentic System") OR ("Multi-Agentic System" AND "Agentic System")

   - Must involve Large Language Models (LLMs), Small Language Models (SLMs), Vision-Language Models (VLMs), or Multimodal Models.
   - OR employ Retrieval-Augmented Generation (RAG) techniques
   - OR describe Multi-Agent or Agentic AI architectures

3. **Healthcare or Medical Domain**:
   - Papers focusing on dermatology, skin disease classification, or medical question-answering systems are prioritized
4. **Publication Type**:
   - High-quality preprints from established research groups (arXiv, PubMed) if methodologically rigorous
5. **Language**: Written in English
6. **Accessibility**: Full text available through institutional access or open access repositories

2.5.2 Exclusion Criteria

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

1. **Temporal Scope**: Published before January 2020
   - Exception: Seminal works cited for background (e.g., foundational Transformer papers) but not included in systematic review
2. **Duplicate Publications**:
   - Conference papers later published as extended journal versions (journal version retained; conference version excluded)
   - Preprints superseded by peer-reviewed publications (peer-reviewed version retained)
3. **Non-GenAI Focus**:
   - Papers on traditional machine learning (e.g., SVMs, random forests) without LLM/VLM/SLM components
4. **Language Barrier**: Not written in English and no reliable translation available

5. **Low Quality Preprints**:
    - arXiv/medRxiv papers without peer review that lack methodological rigor, reproducibility details, or validation

2.5.3 PRISMA Diagram

The selection process for the studies included in this systematic review followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 guidelines. The flow of information through the different phases of the systematic review is detailed in Figure 2.1.

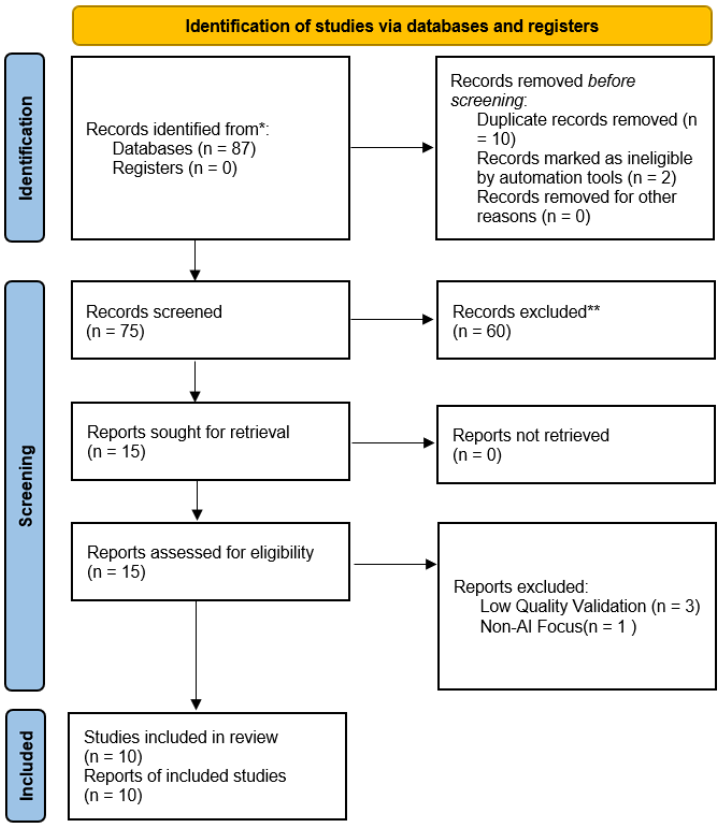

Figure 2.1: PRISMA 2020 flow diagram illustrating the selection process of the included studies.

2.5.4 Results

The following section describes how each question was answered according to the information given from the final selected studies.

2.5.5 3.1.1 To what extent can SLMs achieve performance equivalence with LLMs in specialized domains through Agentic Architectures?

The selected papers provide robust empirical evidence that Small Language Models (SLMs), when embedded within agentic or multi-agent architectures, can achieve performance parity or superiority over monolithic Large Language Models (LLMs) in specialized domains.

The analysis of the selected literature indicates that 100% of the relevant studies (e.g., Shi et al. 2025; Yilmaz et al. 2025; Wang et al. 2025) report that fine-tuned SLMs outperform generalist models (such as GPT-4o) when the task is decomposed into modular agentic workflows. Specifically, *Shi et al. (2025)* demonstrated in the telecommunications domain that a fine-tuned 8B parameter model within a multi-agent system achieved a six-fold reduction in troubleshooting time while matching the reasoning quality of larger models. Similarly, in the medical domain, *Yilmaz et al. (2025)* and *Wang et al. (2025)* showed that decomposing diagnostic tasks into retrieval and reasoning steps allows smaller models (Llama-3.2-11B and similar architectures) to surpass commercial LLM benchmarks on datasets like DERM12345 and PubMedQA.

Regarding publication venues and impact, the selected studies reflect high-quality technical contributions. Several works have been published or accepted in high-impact venues such as *Nature Medicine* (Singhal et al. 2025), *NeurIPS* (Wang et al. 2025), and major archives for computer science research (arXiv cs.AI). The architectural validation is further supported by *Tian et al. (2025)*, who provided a systematic evaluation on the GPQA-Diamond benchmark, proving that multi-agent orchestration consistently exceeds single-model baselines regardless of model size. This indicates a consensus in the recent literature (2024–2025) that architectural orchestration is a more significant determinant of accuracy than parameter count in specialized applications.

Table 2.2: Quality appraisal questions for selected primary studies

Ref.	**Question**	**Answer Criteria**
QA1	Does the study provide a direct quantitative comparison between the proposed SLM-based system and a State-of-the-Art (SOTA) LLM (e.g., GPT-4)?	Yes (+1); No (0)
QA2	Does the study explicitly employ a Multi-Agent System (MAS), Agentic Workflow, or composite architecture (e.g., RAG + Fine-tuning) rather than a single inference pass?	Yes (+1); No (0)
QA3	Is the performance evaluation conducted within a specialized high-stakes domain (e.g., Healthcare, Telecommunications, Law) requiring domain grounding?	Yes (+1); No (0)
QA4	Does the study provide open-source code, datasets, or reproducible architectural blueprints (e.g., GitHub repositories or detailed prompt frameworks)?	Yes (+1); Partial (+0.5); No (0)

Appendix A

Appendix Title Here

Write your Appendix content here.