Privacy-Preserving Medical Triage: On-Device Deployment of Compact Language Models for New Zealand Healthcare

Manseung Choi
School of Computer Science
University of Auckland
Auckland, New Zealand
mcho868@aucklanduni.ac.nz

November 2025

Abstract

Privacy concerns and regulatory constraints limit cloud-based large language model adoption in New Zealand healthcare, whilst rural workforce shortages necessitate offline-capable clinical decision support. This research investigates on-device medical triage using compact language models within strict mobile constraints ($\leq$400 MB memory, sub-3-second latency).

A Retrieval-Augmented Generation (RAG) system was developed combining structured-agent-based chunking of New Zealand and international medical sources with quantised small language models (135M–360M parameters at 4-bit/8-bit precision). Four LoRA adapter configurations were trained to explore accuracy-safety trade-offs on 13,825 synthetically-generated triage cases. The system was evaluated across 96 configurations and validated on an iPhone 14 Pro.

Fine-tuned SmolLM2-135M-4-bit achieved 70.4% accuracy within 300 MB memory and 2.4-second latency, demonstrating technical feasibility with 30–50$\times$ fewer parameters than existing compact medical models. Structured chunking outperformed traditional methods by 7–10 percentage points (59.5% Pass@5). However, no configuration met clinical deployment criteria. Maximum emergency recall reached 70.3% versus 95% required, with severe class imbalance. Retrieval augmentation improved class balance but reduced accuracy by 13 percentage points and reasoning quality by 26 points.

The research proves on-device medical triage is technically achievable within mobile constraints, establishing foundations for equitable, offline-capable AI-assisted healthcare. However, synthetic data limitations, insufficient emergency detection, and accuracy-safety trade-offs prevent clinical deployment. Future work requires real-patient-data validation, clinical trials, regulatory approval and accuracy improvements before health-system integration.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Context and Problem Statement

New Zealand faces an escalating medical workforce crisis that threatens the stability, accessibility, and equity of its healthcare system. Severe understaffing has produced cases in which patients lack timely medical attention, particularly in rural hospitals affected by closures and increased dependence on telehealth [1, 2]. More than a third of adults are unable to access the required specialist care due to doctor shortages [3]. This shortage is due to structural factors, including the heavy reliance on International Medical Graduates (IMGs), which comprise more than 43% of the workforce, but only 25% remain after ten years [4]. Migration to Australia, where the remuneration and physician-to-population ratios are higher, further intensifies the problem [5]. The Te Whatu Ora 2023/24 health workforce plan estimates a shortage of 1,700 doctors, including general practitioners, which is well short of the ASMS estimates [6]. These shortages disproportionately impact Māori and Pacific peoples, who remain severely under-represented in the medical profession, exacerbating health inequities. In response to these shortfalls, government initiatives include on-the-job training pathways to rapidly advance graduates into general practice [7]. John and Jackson [8] illustrate how such shortages strain both general practices and Emergency Departments (ED). Severe ED crowding, often caused by "access block" when admitted patients cannot be transferred to the wards, is worsened by the high number of low-acuity presentations. These frequently arise from overwhelmed GPs referring non-urgent cases to EDs, exposing critical inefficiencies in patient flow and the absence of effective early-stage guidance.

1.2 Motivation and Significance

To mitigate these workforce pressures and improve healthcare efficiency, New Zealand has embraced telenursing as a cornerstone of its telehealth strategy. The nurse-led Healthline service [9] serves millions of New Zealanders as a primary triage point, demonstrating nationwide acceptance of remote nursing as a means to expand access and optimise scarce clinical resources [10]. Evidence confirms its impact. One regional study estimated that Healthline prevents 23,000 unnecessary ED visits annually in Te Manawa Taki by diverting 14.6% of potential presentations to more appropriate care [11]. Building on this foundation, advances in Artificial Intelligence (AI) present new opportunities to enhance the effectiveness of such services [12–16]. AI tools can automate administrative tasks, analyse patient data in real-time, and support clinical decision-making. A New Zealand survey by Whittaker et al. [13] found that 47% of GPs believed

AI could save 30 minutes to two hours daily per clinician, although some warned that verifying AI outputs could offset these gains. Despite these potential benefits, AI adoption in New Zealand remains constrained [14]. Te Whatu Ora's guidance prohibits unapproved large language models (LLMs) for clinical use when patient data are involved, citing privacy, bias, and transparency concerns. These restrictions have produced a domestic AI utilisation gap. Clinicians lack approved AI support, and patients have virtually no direct AI-based assistance. Closing this gap requires a privacy-preserving, locally-deployable technical approach consistent with New Zealand's strict safety standards.

1.3 Research Approach and Proposed Solution

The prevailing model for AI in telenursing is one of augmentation, where AI acts as an intelligent co-pilot, automating preliminary triage and providing decision-support [15,16]. The technological leap making this possible is the emergence of powerful Large Language Models (LLMs), which have demonstrated expert-level performance across specialised domains [17]. The recent development of efficient Small Language Models (SLMs), such as SmolVLM-256M, enables such intelligence to run directly on consumer devices [18]. However, the clinical application of on-device AI at this scale remains largely unexplored. On-device inference offers decisive advantages for a national telehealth service. It offers minimal latency, guaranteed privacy, and offline functionality that bridges the digital divide for rural populations [19]. However, SLMs possess limited internal knowledge and require specialised techniques to ensure clinical safety and accuracy. This research investigates whether an on-device SLM can serve as a reliable, privacy-preserving triage assistant within a strict 400 MB mobile memory budget. To this end, it evaluates two key techniques. The first is RAG [20] to supply verified medical knowledge. The second is Parameter-Efficient Fine-Tuning (PEFT) using Low-Rank Adaptation (LoRA) [21] to align the model's reasoning with established triage protocols. The central aim is to determine **whether these methods in combination can enable safe medical triage on-device**. The research systematically evaluates 96 system configurations, testing advanced RAG strategies and LoRA adapters with safety-oriented configurations and post-training selection to quantify trade-offs between accuracy, safety, and performance, culminating in a feasibility demonstration on an iPhone 14 Pro. The subsequent chapters present the previous work, experimental design, implementation, and results, followed by a discussion of findings and future directions. Rather than proposing a new model architecture, this study examines whether existing techniques, individually and in combination, can deliver safe and accurate triage within a 400 MB on-device memory budget.

Chapter 2

Previous Work

This literature review examines the development of medical artificial intelligence from large centralised models to smaller on-device systems, highlighting the technological, ethical, and practical factors driving this transition. It reviews the efficiency methods that enable SLMs, the role of RAG in addressing knowledge limitations, and the deployment barriers that influence clinical adoption. Together, these themes identify the research gap: the absence of a systematic evaluation of advanced retrieval methods and parameter-efficient fine-tuning for safe, privacy-preserving medical triage on mobile devices.

2.1 Evolution of Medical AI: From Large to Small Models

Artificial Intelligence (AI) in medicine is evolving from centralised, proprietary systems toward smaller, specialised, and more transparent models. Early breakthroughs with large language models (LLMs) proved that generative AI can reach expert level reasoning, yet these systems remain expensive, opaque, and difficult to deploy in safety and latency-sensitive environments. The field has therefore shifted toward SLMs: compact, efficient systems that trade scale for accessibility, interpretability, and cost-effectiveness. Google's Med-PaLM 2 [22], built upon PaLM 2 with targeted medical domain-specific fine-tuning, achieved 86.5% on MedQA, effectively passing the U.S. Medical Licensing Exam. While this demonstrated that AI can master vast medical corpora, the model's closed-source nature and high compute requirements limited reproducibility. This limitation inspired open-source efforts such as ChatDoctor [23], which fine-tuned LLaMA-7B on patient-doctor dialogues and introduced an external retrieval mechanism that occasionally outperformed larger systems like GPT-3.

More recently, the drive for efficiency has produced compact yet capable medical models. MedMobile [24], a parsimonious adaptation of phi-3-mini (3.8 B parameters selected for its enhanced reasoning capabilities), became the first mobile-sised system to pass MedQA with 75.7% accuracy. Other mid-scale systems such as BioMistral-7B (derived from Mistral-7B-Instruct v0.1 and pre-trained on PubMed Central), UltimateMedLLM-Llama3-8B, and the Meerkat series [25–27] have matched or exceeded much larger competitors. The Meerkat family, built upon Mistral-7B-v0.1 and Meta-Llama-3-8B-Instruct, distilled Chain-of-Thought (CoT) reasoning from medical textbooks to enhance clinical logic. Meerkat-7B reached 77.1% on MedQA and even surpassed human experts on NEJM Case Challenges [27].

While these medical SLMs represent significant progress, parallel developments in general-purpose small models have established the foundation for edge device deployment. Efficiency techniques such as knowledge distillation and architectural innovations have enabled remarkably capable models at ultra-small scales. DistilBERT [28] demonstrated that distillation can retain

~97% of BERT-base accuracy while reducing parameters by ~40% and CPU inference time by ~60%. MobileBERT [29] introduced a "deep but thin" bottleneck with layer-wise distillation: at ~25 M parameters (77% smaller than BERT-base), it achieved comparable GLUE performance and ran ~5.5× faster on a Pixel 4. More recent models push these boundaries further: SmolLM2 [30], trained on ~11 T tokens, is available in 135 M, 360 M, and 1.7 B variants, while Gemma3-270M [31] reports strong IFEval scores at only 270 M parameters, with INT4 deployment on Pixel 9 Pro consuming ~0.75% battery across 25 conversations. However, medical models at this ultra-small scale (<1 B parameters) remain largely underexplored, despite their potential for widespread deployment on resource-constrained devices.

At the same time, even larger compact medical models continue to exhibit factual vulnerabilities. For example, Meerkat occasionally produced incorrect drug dosages [27], illustrating how reduced parameter budgets can endanger reliability in clinical settings. These limitations underscore the need for methods that enhance SLM factual reliability and safety without sacrificing efficiency advantages. The subsequent sections examine approaches that may enable such capability on resource-constrained devices.

2.2 Enhancing Small Language Model Capability

Specialised techniques are needed to increase the capability of models without increasing the resources required to run them. The challenge is fundamental because expert level reasoning must be delivered within the severe constraints of mobile and edge devices, where memory, computation, and power are strictly limited. The subsequent sections examine how compression, knowledge transfer, and efficient adaptation enable SLMs to approach the performance of their larger counterparts while remaining deployable on consumer hardware. These techniques address the reasoning and adaptation dimensions of clinical AI, though as we shall see, they leave unresolved the critical challenge of parametric knowledge limitations.

2.2.1 Model Compression Techniques

Specialised architectural designs and compression methods form the foundation for achieving high performance on resource-limited devices. At the architectural level, MobileLLM [32] explicitly incorporates a deep and thin design for sub-billion-scale models, maximising representational capacity within tight parameter budgets. Quantisation has proven particularly effective in this context. Research by Jin et al. [33] demonstrated that 4-bit quantisation preserves performance comparable to non-quantised versions, though precision below 3-bits introduces significant degradation. Post-training methods like GPTQ [34] can accurately compress large models to 3 or 4-bits, offering substantial improvements in accessibility and deployment viability.

Pruning provides a complementary compression pathway by eliminating less significant model components. SparseGPT [35] successfully removed up to 60% of parameters without meaningful accuracy loss. The SpQR method [36] refined this approach by selectively quantising weights according to their importance, maintaining high precision for critical parameters whilst aggressively compressing others. A large-scale evaluation by Kurtic et al. [37] confirmed that these one-shot compression techniques can dramatically reduce the cost of large language models without expensive retraining, establishing them as practical tools for clinical deployment.

2.2.2 Knowledge Transfer via Distillation

Whilst compression reduces model size, knowledge distillation enables smaller models to acquire sophisticated reasoning capabilities from larger, more powerful teacher models. This technique has proven pivotal for bootstrapping SLM performance in medical domains. The process typically involves fine-tuning smaller models on datasets enriched with Chain of Thought (CoT) reasoning paths generated by advanced systems like GPT-4. MedMobile [24], for instance, achieved expert level clinical capabilities at only 3.8 billion parameters by training on CoT responses distilled from GPT-4 alongside content from 18 medical textbooks. This approach allows compact models to internalise advanced problem solving processes without the prohibitive costs of training large models from scratch.

Knowledge transfer extends beyond reasoning patterns to architectural knowledge. MobileBERT [29] leverages transfer from a specially designed teacher model incorporating an inverted bottleneck BERT Large architecture during pre-training, achieving competitive results with significantly fewer parameters and faster inference. However, distillation requires careful calibration. Orr et al. [18] found that for compact models, excessive CoT data can actually harm performance, suggesting that minimal or sparse inclusion of CoT data proves more beneficial than extensive use. This finding underscores an important constraint, namely that knowledge distillation, whilst powerful for transferring reasoning capabilities, must be carefully tailored to the absorptive capacity of resource-constrained models.

2.2.3 Task-Specific Adaptation via Parameter-Efficient Fine-Tuning

Even with effective compression and distilled reasoning, adapting pre-trained models to specific clinical tasks requires fine-tuning. Traditional fine-tuning updates all model parameters, demanding substantial computational resources and memory that exceed the capabilities of consumer hardware. parameter-efficient fine-tuning (PEFT) methods solve this problem by updating only a small subset of parameters, making adaptation feasible even on modest devices.

Low Rank Adaptation (LoRA) [21] exemplifies this approach. Rather than modifying the entire weight matrix, LoRA freezes pre-trained weights and injects small trainable matrices into selected linear layers. Instead of updating a full weight matrix $W \in \mathbb{R}^{d\times k}$, LoRA learns a low rank update $\Delta W = BA$, where $A \in \mathbb{R}^{r\times k}, \quad B \in \mathbb{R}^{d\times r}, \quad r \ll \min(d, k)$.

At inference, ΔW is merged into W, typically without measurable latency overhead. Quantised LoRA (QLoRA) [38] extends this efficiency by applying LoRA to quantised base models. Notably, QLoRA reduced the memory required to fine-tune a 65 billion parameter model from over 780 GB to under 48 GB whilst achieving near-state-of-the-art chatbot performance, with Guanaco reaching 99.3% of the Vicuna benchmark score.

2.2.4 The Parametric Knowledge Bottleneck

Even with strong compression and knowledge transfer, compact models struggle to internalise the breadth of clinical facts needed for safe medical decision support [27]. Limited parameters and narrow context windows constrain factual coverage and long-form reasoning capacity [27]. The memory and computational requirements of transformer architectures scale quadratically with sequence length, typically limiting inputs to around 512 tokens and restricting applicability for tasks requiring larger context [29]. These constraints directly impact medical applications, where inaccurate information such as incorrect dosages stems not from reasoning flaws but from inherent limitations in smaller models' medical knowledge and memorisation capabilities [27].

Although parametric knowledge staleness is often discussed as a limitation, small models reduce this concern because they can be retrained and refreshed efficiently on consumer hardware [21, 30, 38]. The primary challenge is therefore not update speed, but the finite capacity of parametric memory and the need to verify information against authoritative medical sources. However, retrieval pipelines developed for server environments do not automatically translate to mobile settings [19, 29, 39]. On-device retrieval must operate within strict memory, latency, and battery constraints whilst ensuring accurate filtering to prevent incorrect or irrelevant information from reaching the model. These errors pose direct clinical risk, as language models that suggest incorrect treatments or fail to recognise critical symptoms could put users' safety, health, and life at risk [23, 26].

2.3 RAG for SLMs

RAG emerged as a natural solution to these parametric limitations by supplementing a model's internal memory with external, authoritative knowledge sources. RAG combines parametric memory (knowledge encoded in model weights) with non-parametric memory (external knowledge sources accessed via retrieval). This integration addresses several core limitations of traditional language models, including their inability to revise or expand memory post-training, provide transparent provenance, or reliably avoid hallucinations [20]. For compact medical models, RAG offers a pathway to access comprehensive clinical knowledge without requiring exhaustive memorisation within limited parameter budgets.

2.3.1 RAG Fundamentals

The standard RAG pipeline consists of two components. A retriever selects relevant documents from a corpus, and a generator produces responses conditioned on both the query and retrieved context. Formally,

$$\boxed{\text{RAG}(x) = \text{Generator}(x, d)} \tag{2.1}$$

$$\boxed{d = \text{Retriever}(x, C)} \tag{2.2}$$

where x is the input query, C is the document corpus, d represents retrieved documents, and the generator produces response y.

In practice, the retriever may rely on sparse methods such as BM25 [40], or dense methods such as embedding-based search with FAISS [41], where models like the Dense Passage Retriever (DPR) [42] generate semantic embeddings for queries and documents. Many systems adopt hybrid approaches that combine both sparse and dense signals to maximise recall. The generator, typically a pre-trained language model, is fine-tuned or adapted to condition on retrieved evidence. Together, this pipeline enables models to ground their responses in external knowledge rather than relying solely on parametric memory.

2.3.2 RAG Architectural Paradigms

Building on this foundation, the landscape of advanced RAG can be understood through three primary architectural philosophies, each focusing on a different part of the pipeline [43]. Retriever-based systems delegate architectural responsibility primarily to the retriever, positing that the fidelity and relevance of the retrieved context are paramount. In contrast, Generator-Based systems concentrate innovation on the language model's decoding process, assuming the retrieved context is sufficient and focusing on enhancing output quality through mechanisms

like self-verification. Finally, Hybrid systems tightly couple the retriever and generator, treating them as co-adaptive reasoning agents that learn through iterative feedback.

When deploying these architectures on SLMs, their compatibility with resource constraints becomes the deciding factor [19]. Both Generator-based and Hybrid systems place a significant computational burden on the language model itself, requiring it to handle complex tasks like iterative reasoning or intensive self correction processes that are ill-suited for a compact, efficient model. Retriever-based systems, however, align more naturally with the needs of small models. This approach front-loads the complexity into the retrieval stage, focusing on preparing a clean, concise, and highly relevant context before it is passed to the generator. By simplifying the task for the small model and allowing it to focus on its core strength of synthesis, this architecture offers the most viable and efficient path for on-device RAG.

2.3.3 Advanced Retriever-Based Techniques

Recent innovations focus on three key strategies for optimising retriever-based systems. Query Enhancement addresses the problem of ambiguous or overly simplistic user queries. Instead of using a query verbatim, these techniques transform it to improve retrieval accuracy. RAG Fusion [44], for example, uses a language model to generate multiple alternative queries from different perspectives, then retrieves documents for all query variants and uses a Reciprocal Rank Fusion (RRF) [45] algorithm to surface the most consistently relevant sources, leading to a much richer context for the generator.

Retriever Adaptation improves the retriever's ability to understand the nuances of a specific domain without costly manual annotation. Techniques like SimRAG [46] achieve this through self-training. The system generates synthetic but realistic question-answer pairs, which are then used to fine-tune the retriever, teaching it to better identify relevant passages for the types of questions it is likely to encounter.

Context Granularity Optimisation refines retrieved context to be as concise as possible. Standard retrieval often returns entire document chunks, which may contain irrelevant or distracting sentences. FILCO [47], for instance, is a post-retrieval filtering mechanism that identifies and removes irrelevant sentences or spans from retrieved passages. This ensures the final context provided to the small model is dense with useful information, respecting its limited context window and improving the quality of the generated output.

Anthropic's Contextual Retrieval [48] enhances RAG by addressing the loss of context in traditional document chunking. The method uses a large language model to analyse an entire document and generate a concise, explanatory context (50 to 100 tokens) that is prepended to each chunk before it is embedded and indexed for both dense and sparse retrieval. This preprocessing technique has been shown to dramatically improve retrieval accuracy, reducing the top 20 chunk retrieval failure rate by 35% with Contextual Embeddings, 49% when combined with Contextual BM25, and up to 67% when also including a reranking step.

Key implementation factors include the choice of chunk boundaries, such as fixed-length chunking where chunks are created at predetermined sizes, sentence or paragraph based boundaries for more semantic coherence, or agent-based chunking, which leverages a language model to help make intelligent chunk decisions [49]. The performance and efficiency of the embedding model is also critical, since the embedding model performs both embedding creation and retrieval.

2.3.4 Limitations of Conventional RAG for Small Models

Despite these sophisticated techniques, results have been mixed when applying RAG to resource-constrained models. MedMobile [24] reported that BM25-based RAG decreased accuracy by 12.6%, plausibly because verbose contexts exceeded the model's limited attention window. In resource-constrained settings, long or loosely relevant passages can dilute attention and introduce noise, offsetting any gains from added knowledge. This observation reveals a critical research gap. Context efficient, retriever-centric RAG tailored to small model limitations remains under-evaluated for clinical tasks.

However, technical capability alone does not guarantee clinical utility. The viability of compact medical AI systems depends equally on their ability to navigate the regulatory, ethical, and operational constraints that govern real-world healthcare deployment.

2.4 Deployment Barriers and Real-World Constraints

Translating compact medical AI from research prototype to clinical practice requires satisfying multiple interrelated constraints. These barriers not only justify the need for this research but also define the specific requirements that shape the experimental design and evaluation criteria.

2.4.1 Regulation and Safety Validation

Te Whatu Ora advises caution with unapproved LLMs due to limited local safety evidence [14]. International medical AI systems [22–25] similarly acknowledge the need for extensive validation before deployment. Rapid innovation outpaces regulatory development, creating uncertainty around certification pathways and approval timelines. For this research, regulatory concerns necessitate transparent, verifiable knowledge sources. RAG-based systems that cite authoritative medical guidelines offer clearer audit trails than purely parametric models, potentially accelerating regulatory approval by enabling clinicians to verify the provenance of AI recommendations.

2.4.2 Privacy and Security

Patient data protection and HIPAA equivalent compliance are non negotiable requirements [13]. On-device inference directly addresses these constraints by eliminating data transmission and keeping all computation local. This privacy imperative motivates the strict 400 MB memory budget that shapes this research. Unlike cloud-based systems that can access unlimited server resources, on-device deployment requires proving that advanced RAG and parameter-efficient fine-tuning can deliver clinical capability within consumer hardware constraints whilst maintaining complete data privacy.

2.4.3 Bias and Health Equity

Models trained on unrepresentative datasets risk replicating or amplifying health disparities [50]. In New Zealand, under-representation of Māori patients in training data can produce inequitable triage outcomes [13]. Whilst transparent reasoning mechanisms may help clinicians identify bias [27], systematic auditing and diverse training data remain essential. RAG offers a potential pathway to address bias by enabling dynamic updates to knowledge sources as health equity research evolves, without requiring complete model retraining. However, this benefit only

materialises if retrieval systems can effectively surface relevant equity considerations within the limited context windows of compact models.

2.4.4 Data Quality and Currency

Maintaining accurate, up-to-date medical knowledge is resource intensive. Approaches such as biomedical corpus pre-training [25] or synthetic patient simulations [51] partially address data scarcity, yet knowledge drift persists as medical guidelines evolve. This challenge directly motivates the RAG component of this research. By decoupling knowledge from model parameters, RAG enables updates to medical information without costly retraining cycles. However, this advantage disappears if retrieved contexts exceed the constrained context windows of SLMs [24], explaining why advanced retriever-centric techniques that produce concise, high fidelity contexts are central to this investigation.

2.4.5 Computational Cost and Hardware Constraints

Large models impose prohibitive infrastructure demands. Compact systems such as MedMobile and BioMistral [24, 25] demonstrate that expert level reasoning can operate on consumer hardware, yet deployment validation at scale is limited. For mobile deployment specifically, a crucial industry guideline suggests that applications should not exceed 10% of a device's total Dynamic Random Access Memory (DRAM) to ensure stable performance [32]. For phones with 4 to 8 GB RAM, this yields around 400 to 800 MB for the entire app, not just the model. This constraint directly determines the choice of base model, the size of retrievable knowledge bases, and the feasibility of combining RAG with fine-tuning. The 400 MB budget is not merely a technical specification but rather a fundamental requirement that shapes every design decision in this research.

2.4.6 Liability and Accountability

Erroneous AI recommendations create legal uncertainty [50]. Current safeguards predominantly rely on human oversight [15, 23, 51], which ensures safety but limits automation potential. Clear frameworks for assigning accountability in hybrid human AI decision making remain underdeveloped. This liability concern informs the safety-focused evaluation strategy in this research, particularly the emphasis on measuring emergency detection recall and false negative rates. A system that misses critical symptoms poses greater legal and clinical risk than one that over refers, shaping the relative importance of different performance metrics.

These barriers collectively define the success criteria for this research. The system must operate within 400 MB memory whilst achieving factual accuracy through RAG, maintain complete data privacy through on-device processing, enable knowledge updates without full retraining, provide transparent reasoning for regulatory compliance, and demonstrate safe triage performance that minimises clinical risk. For New Zealand specifically, meeting these requirements would enable deployment of a privacy-preserving triage assistant that addresses workforce shortages whilst ensuring equitable access through offline functionality in rural areas.

2.5 Research Gap and Contribution

The literature establishes that compact medical models possess strong reasoning capabilities through compression, distillation, and parameter-efficient fine-tuning, yet struggle with

factual knowledge limitations that compromise clinical safety [27]. Two technical solutions have emerged independently. RAG enables external knowledge access without parameter updates, whilst advanced retriever-centric techniques promise to overcome the context overflow problems that caused conventional RAG to reduce MedMobile's accuracy by 12.6% [24]. parameter-efficient fine-tuning, particularly LoRA [21], enables behavioural alignment with clinical protocols without prohibitive computational costs.

Despite the existence of these techniques, a critical gap persists. No prior work has systematically evaluated advanced RAG and parameter-efficient fine-tuning, individually or combined, as solutions for on-device medical SLMs under strict resource constraints. The literature demonstrates each technique's potential in isolation but provides no empirical evidence of their interaction effects, comparative effectiveness, or viability when deployed within mobile memory budgets.

This research addresses the following question: Can advanced retriever-centric RAG, parameter-efficient fine-tuning via LoRA, or their combination effectively bridge the knowledge accuracy gap in a medical small language model operating within a 400 MB memory budget whilst maintaining safe triage performance aligned with clinical protocols?

By systematically evaluating these approaches both individually and in combination, this research establishes empirically grounded guidance for deploying reliable medical AI systems within consumer mobile device constraints. For New Zealand specifically, this addresses healthcare workforce shortages through accessible, privacy-preserving telehealth technology that operates effectively under real-world resource limitations. The following chapter details the experimental design used to address this research question.

Chapter 3

Experimental Design

This chapter outlines the methodological framework adopted to evaluate the feasibility of on-device medical triage under strict mobile hardware constraints. It details the experimental design, evaluation strategy, and success criteria for systematically comparing RAG and parameter-efficient fine-tuning approaches within the 400 MB deployment budget established in Section 2.4.5.

The system architecture comprises two main components designed to enable systematic evaluation through controlled variation. The retrieval component processes patient queries to surface relevant medical knowledge from a curated knowledge base. Multiple chunking strategies and source prioritisation configurations are evaluated to identify approaches that maximise information density within small language model context windows. The generation component transforms queries and retrieved context into structured triage decisions, producing classifications, next step recommendations, and clinical reasoning. Three compact language models at two quantisation levels are compared, each fine-tuned using four adapter configurations. This modular design enables evaluation of retrieval and generation both independently and in combination.

This yields 96 unique system configurations (6 model variants × 4 adapters × 4 retrieval conditions), enabling systematic evaluation of how retrieval quality and model capacity interact to affect clinical safety and accuracy under resource constraints. Within the deployment constraint of 400 MB memory and sub-3-second inference latency, the primary objective is to determine how best to balance clinical safety, particularly the minimisation of false negatives in emergency cases, with computational efficiency. Success is measured through safety-critical metrics prioritising emergency detection recall alongside secondary metrics encompassing overall accuracy, inference latency, and memory utilisation.

The following sections detail the design rationale for the retrieval component, generation component, dataset preparation, fine-tuning methodology, and evaluation framework.

3.1 Retrieval Component Design

Following the retriever-centred approach identified in Section 2.3.3, the architecture front-loads complexity into the retrieval stage. This design enables systematic evaluation of how retrieval quality affects clinical safety and accuracy under resource constraints.

3.1.1 Dataset Choices

The medical knowledge base integrates content from three complementary sources. Healthify NZ provides patient-centred New Zealand-specific resources reflecting local healthcare pathways and cultural context. NHS UK offers evidence-based clinical protocols developed through systematic review processes. Mayo Clinic contributes specialist tertiary care insights. This multi-source design creates deliberate tension between local applicability and international medical consensus, enabling systematic evaluation of how source prioritisation affects triage accuracy. All content undergoes standardised cleaning to remove non-substantive elements whilst preserving medical information hierarchy, ensuring retrieval surfaces clinically actionable content rather than navigational text or formatting artefacts.

3.1.2 Chunking Strategy Evaluation

Chunking determines whether the system surfaces precise information or verbose passages that overwhelm small language models. Yet no consensus exists for medical content with its unique requirements around symptom lists, dosage precision, and treatment sequences. This research systematically compares six chunking paradigms ranging from simple deterministic methods to sophisticated LLM-guided approaches.

Fixed-length chunking employs character-based segmentation with varying sizes and overlap. This provides deterministic, tokeniser-agnostic baseline performance that ensures cross model compatibility but risks fragmenting semantically coherent medical concepts.

Sentence-based chunking preserves syntactic boundaries through natural language processing whilst targeting specific token thresholds with variable overlap. This maintains grammatical coherence crucial for medical instructions but potentially groups unrelated sentences to meet length requirements.

Paragraph-based chunking uses boundary detection with minimum length thresholds to preserve topical coherence, leveraging the convention that paragraph breaks indicate transitions between distinct medical concepts. However, paragraph boundaries may reflect design choices rather than semantic organisation.

Agent-based chunking addresses these limitations by employing Llama-3.3-70B to perform intelligent semantic segmentation. Using specialised medical prompts, the model identifies self-contained chunks whilst preserving critical information such as symptoms, dosages, and emergency warnings. This content-aware approach recognises domain specific boundaries that rule-based methods might fragment.

Contextual retrieval chunking implements Anthropic's methodology [48], augmenting base chunks with LLM-generated contextual summaries describing each chunk's role within the broader document. This addresses the fundamental limitation where isolated passages lose essential context. For medical content, this enrichment is valuable where understanding symptom relationships or treatment sequences requires awareness of the broader clinical picture.

Structured chunking represents a novel contribution, curating content directly into JSON format with fields for Condition, Symptoms, Triage Level, and Next Steps using Llama-3.3-70B. This schema-driven approach imposes standardised structure mirroring the triage task itself, making each chunk self-contained and immediately actionable.

3.1.3 Retrieval Architecture and Index Design

The retrieval system employs hybrid search combining dense semantic matching with sparse lexical matching, addressing the complementary limitations of each approach. As discussed in

Section 2.3.1, pure semantic methods may miss domain-critical keywords whilst pure lexical methods struggle with paraphrased queries [42].

Dense retrieval is designed for efficient cosine similarity search using vector indexing, capturing conceptual similarity between patient descriptions and medical documentation. Sparse retrieval employs keyword frequency scoring like BM25 to ensure exact medical terminology receives appropriate weight. Outputs are combined using Reciprocal Rank Fusion [45] with weighting that favours semantic similarity whilst ensuring critical medical terms are not overlooked.

Rather than merging all content into a unified index, the system design maintains separate indices for each knowledge source (Healthify, NHS, Mayo Clinic). This architectural choice enables fine-grained control over source prioritisation during retrieval, directly addressing the tension between local relevance and international evidence. This design treats source selection as an experimental variable, tested through two configurations a New Zealand-prioritised approach (6:2:2 ratio favouring Healthify) and a balanced international approach (4:3:3 ratio).

To maximise information density within the small language model's limited context window, the system retrieves only the top k ranked documents after Reciprocal Rank Fusion, where k is varied as an experimental parameter. This constraint ensures retrieved context remains concise, aligned with the finding that verbose contexts degrade small model performance [24]. The optimal k value represents a critical trade-off between information completeness and attention dilution.

3.2 Generation Component Design

The generation component transforms retrieved medical knowledge into structured triage decisions. Unlike the retrieval stage where complexity can be front-loaded, the generation component faces the challenge of delivering accurate clinical reasoning within a compact parameter budget. This section details the model selection rationale, synthetic dataset development, and Lora fine-tuning methodology.

3.2.1 Quantisation Strategy

Both 4-bit and 8-bit quantisation are evaluated to assess the accuracy-efficiency trade-off in medical contexts. Whilst research demonstrates that 4-bit quantisation preserves performance in general domains [33], medical triage introduces unique challenges around precise numerical reasoning and rare but critical vocabulary. The evaluation therefore treats quantisation as an experimental variable rather than assuming optimal precision a priori. Quantisation-Aware Training principles [38] ensure models maintain quality under compression, though the interaction between quantisation level, model size, and medical task performance remains an empirical question.

3.2.2 Model Selection

Candidate language models are selected based on parameter efficiency to remain within the 400 MB deployment budget and availability of open weights for task-specific adaptation. The selected models represent different points on the efficiency capability frontier, enabling systematic evaluation of how model capacity affects triage performance under resource constraints.

SmolLM2-135M and SmolLM2-360M [30] establish lower-bound baselines for reasoning capacity under strict memory budgets. These models are notable for their extensive training [30], directly addressing the data efficiency challenge identified for compact models [18].

Gemma3-270M [31] tests whether marginally higher parameter counts yield meaningful accuracy gains in medical triage whilst remaining within deployment constraints. This model sits between the two SmolLM2 variants, creating a three-point comparison that reveals the relationship between model size and medical performance.

The embedding model, all MiniLM-L6-v2, was chosen as it provides strong retrieval accuracy with minimal memory usage and low computational demand, fitting comfortably within the device constraint. Together, these language models and quantisation levels form six candidate configurations, three language models multiplied by two precision levels, as summarised in Table 3.1.

Language Model	**Quantisation**
SmolLM2 135M	4-bit (INT4)
SmolLM2 360M	4-bit (INT4)
Gemma3 270M	4-bit (INT4)
SmolLM2 135M	8-bit (INT8)
SmolLM2 360M	8-bit (INT8)
Gemma3 270M	8-bit (INT8)

Table 3.1: Evaluated language model configurations using the all MiniLM-L6-v2 embedding model for retrieval.

3.2.3 Synthetic Dataset Development

Adapting general-purpose language models to medical triage requires training data that captures the task structure, clinical reasoning patterns, and safety-critical decision boundaries. However, no public New Zealand-specific medical triage dataset exists with the required granularity. This gap necessitates synthetic data-generation, following the successful pattern demonstrated by MedMobile [24].

The synthetic dataset development proceeds through several validation stages. First, medical conditions are extracted from the curated knowledge base and validated through LLM filtering to exclude non-conditions such as medications or procedures. This filtering ensures the dataset focuses on actual clinical presentations.

Second, each validated condition is mapped to symptoms, triage levels (ED, GP, HOME), and evidence-based next steps drawn directly from the knowledge base. This grounding in source documents ensures that the synthetic dialogues reflect actual medical guidance rather than model-generated assumptions. When conditions appear across multiple sources with inconsistent triage advice, a most frequent aggregation mechanism determines the final label. In the event of a tie, the system applies a safety-oriented hierarchical rule prioritising clinical severity (ED > GP > HOME), ensuring that ambiguous cases default to the more conservative classification. This approach preserves representational fidelity while reducing the risk of unsafe under-triage.

Associated symptoms and next steps are consolidated across sources. Seven synthetic dialogues are then generated per condition, creating conversational variety. Each dialogue

follows a standardised schema with fields for symptom, patient query, triage classification, next steps, and clinical reasoning. This structure enables the model to learn both the decision outcome and the reasoning process.

The clinical reasoning component deserves particular attention given the conflicting evidence about Chain of Thought effectiveness in compact models [27]. The dataset therefore employs intentionally sparse Chain of Thought annotations, constrained to a maximum of two sentences per case [18]. This design captures essential clinical logic without overwhelming the student model's limited capacity.

3.2.4 Dataset Partitioning and Distribution

The complete synthetic dataset is partitioned into training, validation, and test subsets at a 5:1:1 ratio. This allocation provides sufficient training data for adaptation whilst reserving adequate held-out samples for robust validation and final testing. Stratification preserves the clinical distribution across all splits, ensuring each subset reflects real-world triage patterns derived from Healthline operational data. This ensures the model encounters representative case mixes during training and evaluation.

3.2.5 PEFT

With base models selected and training data prepared, the generation component requires task-specific adaptation. Traditional fine-tuning updates all model parameters, demanding substantial resources. parameter-efficient fine-tuning through Quantised Low Rank Adaptation (QLoRA) [38] addresses this constraint by updating only a small fraction of parameters.

QLoRA enables adaptation of already quantised base models, stacking memory efficiency with parameter efficiency. The low rank decomposition freezes pre-trained weights and injects small trainable matrices. These adapters subsequently merge into frozen weights at inference without latency overhead [21], critical for maintaining the sub-3-second response time.

Medical triage is a safety-critical task where the consequences of misclassification are asymmetric. A cost framework was defined to guide evaluation rather than training. The framework assigns higher penalties to false negatives involving emergency department (ED) cases and lower penalties to over-triage errors.

Four adapter configurations were developed to explore the trade-off between overall accuracy and emergency recall. Each configuration varied learning rate, batch size, and LoRA rank to examine different regions of the performance space. Conservative settings employed lower learning rates to favour higher ED recall, whereas balanced and higher capacity settings aimed to preserve general accuracy.

Although class weights and penalty matrices were defined, they were not incorporated into the loss computation. The fine-tuning procedure used a standard cross entropy objective. The experimental design includes consistent defaults for reproducibility, such as a fixed random seed and a maximum sequence length. Evaluation checkpoints were recorded at regular intervals to monitor convergence and prevent overfitting.

3.3 Evaluation Framework and Success Criteria

The evaluation framework systematically tests 96 unique configurations. This grid comprises the six quantised model variants, four adapter sets, and four retrieval conditions (no RAG

plus the three top retrieval strategies). Base models without LoRA adapters are excluded as preliminary testing shows they cannot consistently produce the structured outputs required for triage.

3.3.1 Safety-Critical Evaluation and Deployment Criteria

Evaluation employs safety-centred metrics that prioritise emergency detection over general accuracy, reflecting the asymmetric risk inherent to medical triage. The framework defines strict quantitative thresholds for both performance and resource efficiency.

Emergency Recall measures the proportion of true emergency department (ED) cases correctly identified. A minimum of 95% is required for deployment consideration. **False Negative Rate** quantifies the percentage of ED cases incorrectly classified as general practice (GP) or home care. A maximum of 5% is accepted. **F_2 Score** provides a harmonic mean weighted toward recall (with beta equal to 2) over precision, emphasising safety sensitivity. A target of 90% is set for the emergency class.

Configurations failing any of these safety criteria are deemed unsafe and excluded from further evaluation regardless of their overall accuracy. Secondary performance metrics include total triage accuracy across all classes, per class precision, recall, and F1 scores. For retrieval analysis, Pass@K is used to assess retrieval quality, and interpretability is evaluated by having Llama-3.3-70B independently review clinical reasoning coherence.

A configuration qualifies as deployment ready only if all of the following conditions are met:

- Overall triage accuracy $\geq 85\%$
- Emergency recall $\geq 95\%$
- False negative rate $\leq 5\%$
- Inference latency ≤ 3 seconds per case
- Total memory footprint ≤ 400 MB

These thresholds collectively balance clinical safety, user experience, and mobile feasibility, ensuring that any model advanced to deployment is both technically efficient and clinically trustworthy.

3.4 Ethical Considerations

This research constitutes an experimental feasibility study aimed at validating the technical potential of on-device medical triage systems. No real patient data or live deployment was conducted. All textual data originated from publicly available medical sources, and synthetic datasets were generated exclusively for controlled experimentation. The findings therefore represent proof-of-concept outcomes rather than clinically deployable evidence. Ethical limitations relating to unverified data, model bias, and absence of clinical validation are discussed further in Limitations.

Chapter 4

Implementation

The implementation proceeds through six sequential phases as illustrated in Figure 4.1, from knowledge base construction through mobile deployment validation. The implementation emphasises reproducibility, modularity, and hardware efficiency, integrating Python-based data processing, FAISS retrieval, MLX-based model training, and iOS deployment through SwiftUI. Together, these components form an end-to-end system that validates the technical feasibility of privacy-preserving medical AI operating entirely on mobile hardware. The following sections detail each implementation phase. **The complete source code for this project, including data processing scripts, MLX model training, and the iOS application, is publicly available at: https://github.com/mcho868/hft.**

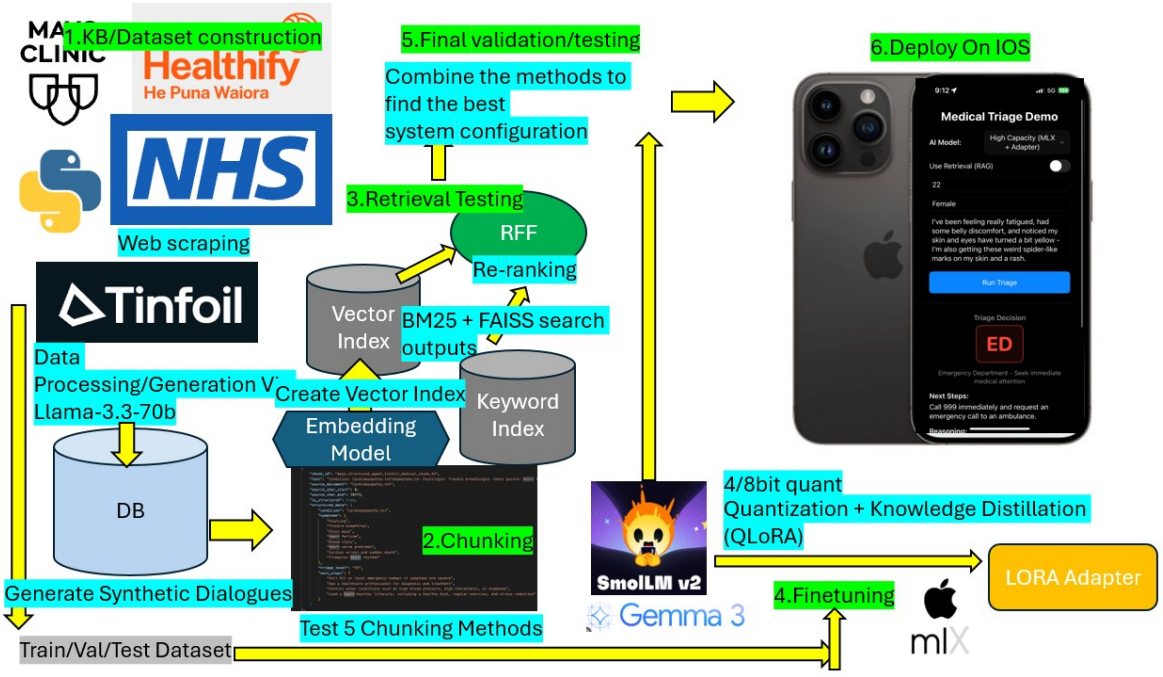

Figure 4.1: Complete implementation workflow showing the six technical phases.

4.1 Knowledge Base and Dataset Construction

4.1.1 Knowledge Base Collection and Cleaning

The medical knowledge base was constructed from three sources. These were Healthify NZ, NHS UK, and Mayo Clinic. Data scraping was implemented in Python using the `requests` and `BeautifulSoup` libraries, with randomised delays of 1 to 3 seconds and a custom `User-Agent` header to ensure responsible web access. Discovery crawled the Healthify and Mayo A-Z indexes and the NHS conditions list.

Parsers were source specific. NHS content was extracted from the `<main id="maincontent">` tag, Healthify prioritised accordion-style sections with fallback to standard headers and paragraphs, and Mayo Clinic pages combined "Symptoms & Causes" with "Diagnosis & Treatment" into a unified document.

All scrapers supported resume functionality through a JSON-based URL ledger to prevent duplication. A cleaning pipeline removed non-substantive HTML elements while preserving medical hierarchy using markers for section headers (e.g., `<h2>`, `<h3>`). Each document was stored as JSON with fields {*source, url, title, cleaned_text*}, ensuring provenance and traceability for retrieval experiments.

4.1.2 Synthetic Medical Dataset Generation

Dataset curation proceeded through several stages to ensure clinical validity. First, condition names were extracted from scraped filenames and filtered in batches of 100 by Llama-3.3-70B, producing 2,036 validated conditions. Non-condition entities (e.g., medications, procedures, general topics) were excluded through rule-based filtering.

Second, a structured data-generation process mapped each condition to symptoms, triage levels (ED, GP, HOME), and next steps drawn directly from the knowledge base. When triage advice conflicted across sources, a most frequent aggregation mechanism was applied as described earlier in Section 3.2.3. Associated symptoms and next steps were subsequently consolidated and deduplicated to preserve clinical completeness while maintaining safety consistency across the dataset.

Using this validated foundation, synthetic triage dialogues were generated using enhanced prompts grounded in verified symptoms, decisions, and next steps. Seven dialogues were produced per condition, yielding approximately 13,825 samples. Each entry followed a standardised JSON schema with fields {*patient_query, clarifying_question, patient_response, triage, next_steps, reasoning*}, preserving traceability to source content.

Finally, the dataset was split 5:1:1 into training (9,875), validation (1,975), and test (1,975) subsets, stratified to maintain the class distribution described in Section 3.4.

4.2 Retrieval Architecture Implementation

4.2.1 Chunking Strategy Implementation

Five chunking families and one structured variant were implemented. Fixed-length chunking employed 13 character-based configurations from `c256` to `c1024` with 0 19.5% overlap (e.g., `c256_o0`, `c512_o0`, `c768_o200`). Sentence-based chunking targeted 384 1024 tokens with 0 3 overlap sentences, and paragraph-based chunking enforced minimum lengths of 25, 50, or 100 words.

Agent-based chunking used Qwen-3-4B (local) and Llama-3.3-70B (via the Tinfoil API) with a medical prompt to produce self-contained, clinically coherent segments. Structured chunking employed Llama-3.3-70B to output JSON-formatted chunks with fields {*Condition, Symptoms, Triage, NextSteps*}.

For contextual retrieval chunking, baseline `c512_o100` and `t1024_o2` chunks were augmented with 50-100 token summaries describing each chunk's role within its parent document.

4.2.2 Vector Database and Hybrid Retrieval

All chunks were embedded using `all-MiniLM-L6-v2` (384 dimensions) with L2 normalisation. The FAISS `IndexFlatIP` index was employed, making inner-product search equivalent to cosine similarity on normalised vectors [41].

Separate indices were maintained for each source (Healthify, NHS, Mayo). At query time, semantic FAISS results were combined with BM25 lexical matches using RRF (= 70:30). Per-source top-k quotas were then applied. Two retrieval configurations were tested, **6:2:2** (NZ-prioritised) and **4:3:3** (balanced), ensuring control over the mix of local and international medical content.

4.3 Retrieval and Chunking Evaluation Framework

Retrieval performance was evaluated across 144 configurations (36 chunking methods $\times$ 2 source mixes $\times$ 2 retrieval modes) using 200 representative clinical queries. To optimise throughput, the evaluation runner cached model instances and processed queries in batches of 32 64, resulting in an average transient memory overhead of approximately 450 MB per configuration.

Retrieval effectiveness was measured using Pass@K for $K \in \{1, 2, 3, 4, 5, 10, 20\}$.

4.4 Model Fine-Tuning

Base generators included `SmolLM2-135M`, `SmolLM2-360M`, and `Gemma3-270M`. Each model was converted from Hugging Face format to MLX and quantised to INT4 and INT8, yielding six base variants. fine-tuning used the QLoRA method described in Section 3.5.

Four adapter configurations (Table 4.1) were trained to represent different optimisation priorities along the accuracy to emergency department (ED) recall trade-off curve. The configuration parameters were proposed through AI-assisted design using an LLM to suggest initial values for rank, learning rate, and training steps, which were then implemented and trained manually in the MLX environment. All training employed the standard cross entropy loss objective with the AdamW optimiser, a weight decay of 0.01, and a warmup ratio of 0.1.

Configuration	Rank	Learning Rate	Training Steps	Target ED Recall
`ultra_safe`	4	5×10^{-6}	1500	$\geq 98\%$
`balanced_safe`	8	1×10^{-5}	1200	$\geq 96\%$
`performance_safe`	12	2×10^{-5}	1000	$\geq 95\%$
`high_capacity_safe`	16	8×10^{-6}	1400	$\geq 98\%$

Table 4.1: Summary of the four AI-assisted adapter configurations evaluated for safety performance trade-offs.

Across all configurations, fixed batch sizes ranging from two to six samples were used. Training, validation, and test losses were monitored continuously, with evaluations every 200 steps and logs every 50 steps. All experiments used seed 42, a sequence length of 2048 tokens, and masked prompt loss to ensure that optimisation occurred only on model-generated completions.

All training was executed on a Macbook Pro with a M4pro chip (12 core CPU 16 core GPU) with 24GB of unified memory.

4.5 Model Testing Framework

The five best validation-stage configurations were advanced to full testing on the 1,975-case held-out dataset. The inference engine loaded the corresponding 4-bit base model and LoRA adapter once per configuration and cached them for reuse. Each case generated a context-aware prompt, optionally including RAG results, and inference was performed with MLX.

Structured outputs were parsed automatically, with malformed or undecidable results marked as `UNKNOWN`. Metrics included overall accuracy, F_2-score, latency per case, and a 4×4 confusion matrix covering ED, GP, HOME, and UNKNOWN categories. Qualitative review of next-step recommendations and reasoning coherence followed Section 3.6.

4.6 Mobile Deployment Framework

For on-device inference, the SmolLM2-135M model was packaged with two LoRA adapters (rank 16, α 8.0; rank 12, α 12.0) stored as `safetensors` and applied to the final 16 of 30 transformer layers at runtime.

The iOS application used SwiftUI (MVVM) integrated with MLX-Swift for inference, supporting dynamic adapter loading. The interface collected age, gender, and symptoms to construct culturally appropriate prompts (e.g., "Kia ora, I'm a [age]-year-old [gender]. [symptoms]"). Outputs were post-processed via regular expressions to extract $\{$*triage, next_steps, reasoning*$\}$ alongside raw text.

Deployment was validated on an iPhone 14 Pro (6 GB RAM) and operated within the 400 MB constraint. This complete pipeline, from data collection through mobile deployment demonstrates the end-to-end feasibility of privacy-preserving, on-device medical AI.

Chapter 5

Results

This chapter reports empirical findings from the retrieval, modelling, and deployment experiments. This chapter first quantify retrieval quality and context depth choices, then compare model and quantisation variants with adapters and RAG conditions, before presenting full test set outcomes and mobile runtime measurements. Results are organised to trace the path from component performance to end-to-end feasibility under the 400 MB constraint. All latency measurements reported in this chapter are based on inference performed on an Apple M4 Pro chip, unless otherwise specified.

5.1 Retrieval and Chunking Strategy Evaluation

Performance was measured using Pass@K metrics for $K \in \{1, 2, 3, 4, 5, 10, 20\}$, average retrieval latency, and peak memory utilisation across 144 configurations and 200 clinical queries.

5.1.1 Optimal Context Window Selection

Marginal efficiency analysis identified Pass@5 as the optimal retrieval depth (Table 5.1). The transition from $K = 4$ to $K = 5$ yielded the highest efficiency gain of 8.82% per additional chunk, substantially exceeding all other transitions. Beyond $K = 10$, no further accuracy improvement was observed (Pass@10 = Pass@20 = 58.61%).

Table 5.1: Marginal efficiency analysis for optimal K selection across all 144 configurations.

Transition	From (%)	To (%)	Total Gain (%)	Gain/chunk (%)
$K = 1 \rightarrow 2$	20.52	24.95	4.43	4.43
$K = 2 \rightarrow 3$	24.95	27.45	2.50	2.50
$K = 3 \rightarrow 4$	27.45	29.41	1.95	1.95
$K = 4 \rightarrow 5$	29.41	38.22	**8.82**	**8.82**
$K = 5 \rightarrow 10$	38.22	58.61	20.39	4.08
$K = 10 \rightarrow 20$	58.61	58.61	0.00	0.00

5.1.2 Chunking Strategy Performance

Performance aggregated by chunking type is presented in Table 5.2 and Figure 5.1. Structured agent-based chunking achieved the highest Pass@5 (45.4%) and Pass@10 (69.9%) across all families.

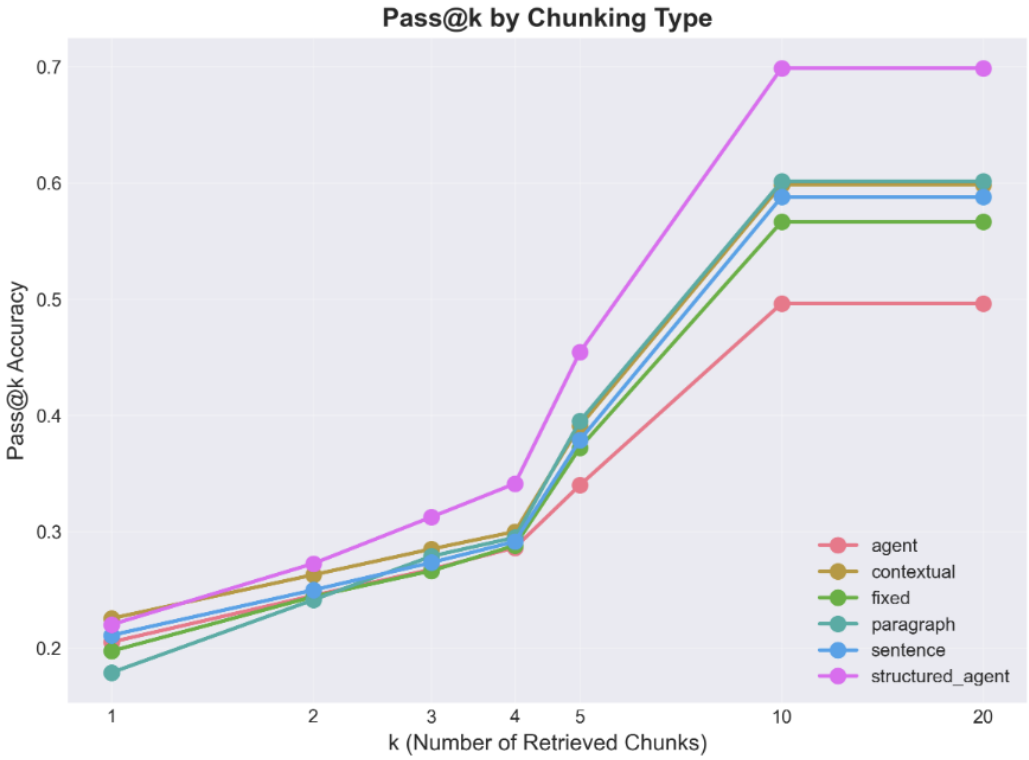

Figure 5.1: Retrieval accuracy by chunking strategy across Pass@K thresholds.

Table 5.2: Retrieval performance by chunking family.

Chunking Type	**P@5**	**P@10**	**P@20**	**Latency (s)**	*n*
Structured Agent	**0.454**	**0.699**	**0.699**	0.017	8
Paragraph	0.395	0.601	0.601	0.029	12
Contextual	0.391	0.598	0.598	0.369	16
Sentence	0.379	0.588	0.588	0.371	52
Fixed-Length	0.372	0.567	0.567	0.416	52
Agent	0.340	0.496	0.496	0.204	4

Within the top 10 individual configurations ranked by Pass@5, structured agent methods occupied the first and second positions (Table 5.3), with the leading configuration reaching 59.5% Pass@5. Among structured agent configurations, Llama-3.3-70B-generated chunks outperformed Qwen-3-4B-generated chunks by 10.0 pp at Pass@5 (59.5% vs 49.5%), despite comparable latency (0.034s vs 0.029s).

5.1.3 Retrieval Method Comparison

Performance aggregated across all chunking and bias configurations is presented in Table 5.4 and Figure 5.2. Contextual_rag achieved 2.1 pp higher Pass@5 than pure_rag (39.3% vs 37.2%), but pure_rag delivered 166× faster average retrieval (0.004s vs 0.664s). For the top-performing structured agent configuration specifically, contextual_rag reached 59.5% Pass@5 versus 59.0% for pure_rag (0.5 pp difference), while maintaining a 17× latency gap (0.034s vs 0.002s).

Table 5.3: Top 10 retrieval configurations ranked by Pass@5 performance.

Chunking Method	Retrieval	Bias	P@5	P@10	Latency (s)
structured_agent_tinfoil	contextual_rag	diverse	**0.595**	**0.770**	0.034
structured_agent_tinfoil	pure_rag	diverse	0.590	0.730	**0.002**
contextual_sentence_c1024_o2_tinfoil	contextual_rag	diverse	0.525	0.635	0.497
contextual_fixed_c512_o100	contextual_rag	diverse	0.520	0.650	0.992
sentence_t1024_o2	contextual_rag	diverse	0.520	0.645	0.422
fixed_c1024_o150	contextual_rag	diverse	0.495	0.625	0.434
structured_agent_qwen3_4b	contextual_rag	diverse	0.495	0.645	0.029
contextual_sentence_c1024_o2	contextual_rag	diverse	0.485	0.625	0.524
fixed_c384_o50	contextual_rag	diverse	0.485	0.615	0.990
sentence_t1024_o3	contextual_rag	diverse	0.485	0.615	0.469

Table 5.4: Retrieval performance by method, aggregated across 72 configurations per type.

Retrieval Type	P@5	P@10	P@20	Latency (s)
contextual_rag	**0.393**	**0.599**	**0.599**	0.664
pure_rag	0.372	0.574	0.574	**0.004**

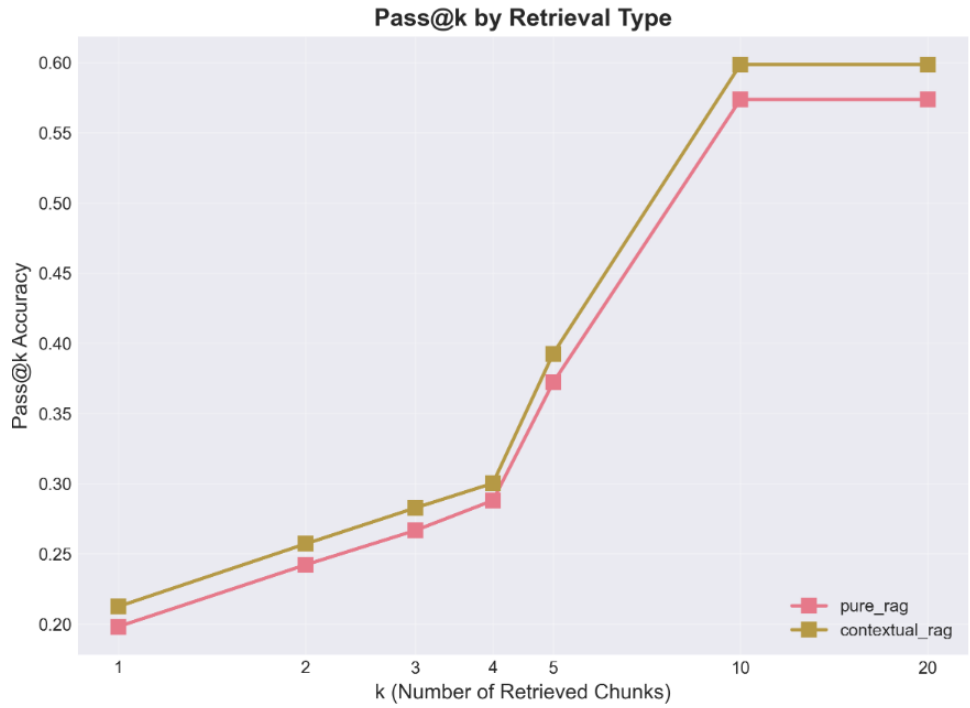

Figure 5.2: Pass@K curves comparing contextual_rag (hybrid semantic + lexical) and pure_rag (semantic only) retrieval methods.

5.1.4 Source Bias Configuration

Results aggregated across 72 configurations per bias setting are shown in Table 5.5 and Figure 5.3. The diverse bias configuration substantially outperformed healthify_focused at Pass@5 (45.3% vs 31.1%, a 14.2 pp difference). This pattern held across all six chunking families (Table 5.6). For the best-performing structured agent-based chunking, diverse bias achieved 54.0% Pass@5 compared to 36.9% for healthify_focused (17.1 pp difference). Performance at Pass@10 converged partially (59.3% vs 58.0%).

Table 5.5: Retrieval performance by source-bias configuration.

Bias Config	P@5	P@10	P@20	Latency (s)
diverse	**0.453**	**0.593**	**0.593**	0.333
healthify_focused	0.311	0.580	0.580	0.335

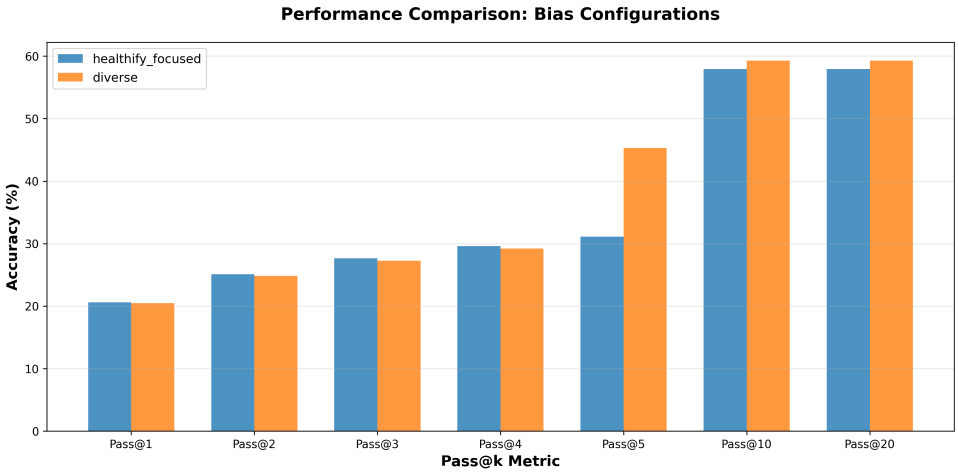

Figure 5.3: Pass@k performance comparing diverse (4:3:3) and healthify_focused (6:2:2) source-bias configurations across chunking families.

Table 5.6: Pass@5 performance by chunking family and source bias.

Chunking Type	Diverse	Healthify-Focused	Difference (pp)
Structured Agent	0.540	0.369	+17.1
Contextual	0.474	0.308	+16.6
Paragraph	0.458	0.333	+12.5
Sentence	0.451	0.306	+14.5
Fixed-Length	0.440	0.304	+13.6
Agent	0.380	0.300	+8.0

5.1.5 Selected Configurations for End-to-End Evaluation

Based on Pass@5 rankings, three retrieval configurations were selected for integration with fine-tuned language models (Table 5.7). Configuration 1 achieved the highest absolute Pass@5 (59.5%) and Pass@10 (77.0%). Configuration 2 delivered nearly identical Pass@5 (59.0%) with 17× faster retrieval. Configuration 3 represented the best-performing non-structured approach. All three configurations employed the diverse source-bias setting.

5.2 Model Evaluation Framework

Across 96 configurations evaluated on the validation set, mean accuracy was 22.9% ± 11.9%, F_2-score was 19.6% ± 13.7%, extraction success rate was 80.7% ± 22.7%, and inference time was 0.940 ± 0.259 seconds.

Table 5.7: Top three RAG configurations selected for downstream model evaluation.

Rank	Chunking Method	Retrieval Type	Config	P@5	P@10	Latency (s)
1	structured_agent_tinfoil_medical	contextual_rag	diverse	0.595	0.770	0.034
2	structured_agent_tinfoil_medical	pure_rag	diverse	0.590	0.730	0.002
3	contextual_sentence_c1024_o2_tinfoil	contextual_rag	diverse	0.525	0.635	0.497

5.2.1 Base Model and Quantisation Comparison

SmolLM2-135M at 4-bit quantisation achieved the highest mean accuracy (35.4% ± 18.5%) and maximum accuracy (68.0%), outperforming the larger 360M model by 9.1 pp despite having 62.5% fewer parameters (Table 5.8, Figure 5.4). Four-bit quantisation consistently outperformed 8-bit across all model families (26.5% vs 19.3% mean accuracy), with faster inference (1.010s vs 0.871s).

Table 5.8: Performance by model architecture and quantisation level ($n = 16$ configurations per variant).

Model	Quant	Mean Acc	Max Acc	Max F_2	Time (s)
SmolLM2-135M	4-bit	**0.354**	**0.680**	**0.667**	0.895
SmolLM2-360M	4-bit	0.263	0.465	0.459	1.076
SmolLM2-360M	8-bit	0.213	0.455	0.455	0.950
Gemma3-270M	8-bit	0.185	0.220	0.220	1.050
SmolLM2-135M	8-bit	0.180	0.355	0.385	0.612
Gemma3-270M	4-bit	0.177	0.215	0.146	1.058

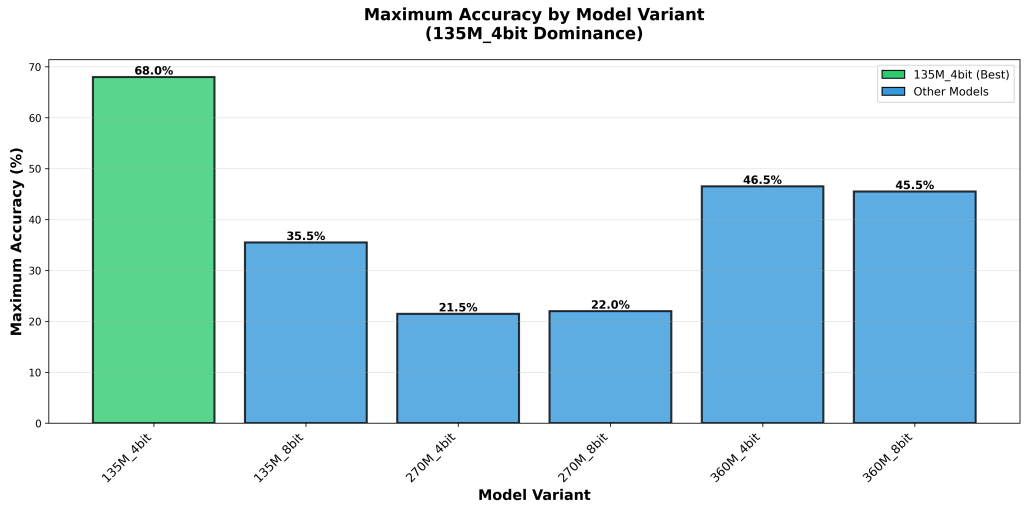

Figure 5.4: Maximum accuracy achieved by each model variant across all adapter and RAG configurations.

Gemma3-270M demonstrated superior format compliance (94.4% ± 8.9% extraction success, 11.2 UNKNOWN cases average) compared to SmolLM2 variants (73.8% ± 24.5% extraction success, 52.3 UNKNOWN cases average), but achieved lower triage accuracy (18.1% ± 1.9% vs 25.3% ± 14.0%).

5.2.2 Adapter Performance

Adapter performance varied by configuration (Table 5.9, Figure 5.5). The high_capacity_safe adapter achieved the highest mean accuracy (27.7% ± 14.4%) and maximum validation accuracy (68.0%), while ultra_safe achieved the lowest (17.1% ± 9.3%, maximum 31.5%) despite targeting the most conservative ED recall threshold (≥98%).

Table 5.9: Performance by adapter configuration (n = 24 configurations per adapter).

Adapter	Mean Acc	Max Acc	Max F_2	Time (s)	Target ED Recall
high_capacity_safe	**0.277**	**0.680**	**0.667**	0.946	98%
balanced_safe	0.253	0.595	0.591	0.940	96%
performance_safe	0.214	0.570	0.572	**0.908**	95%
ultra_safe	0.171	0.315	0.330	0.967	98%

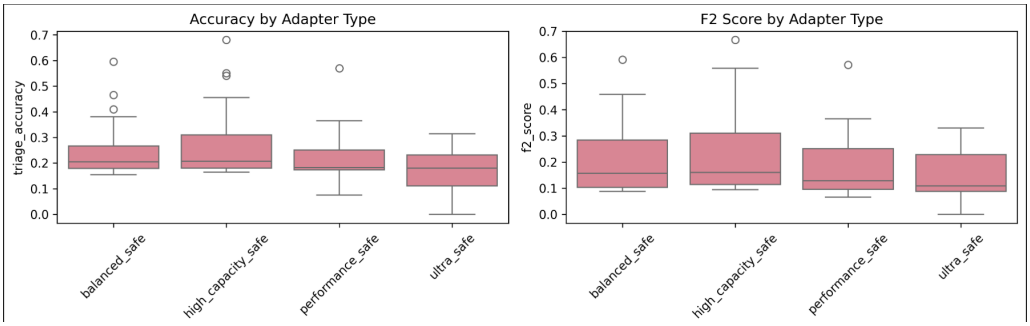

Figure 5.5: Mean accuracy and maximum F_2-score by adapter configuration.

5.2.3 RAG Impact

RAG integration reduced mean accuracy by 4.3 pp (26.1% without RAG vs 21.8% with RAG) but improved extraction success by 9.0 pp (73.9% vs 83.0%) and reduced UNKNOWN outputs by 18.0 cases on average (52.1 vs 34.1) (Table 5.10, Figure 5.6). For the top-performing high_capacity_safe adapter on SmolLM2-135M-4-bit, RAG reduced accuracy from 68.0% to 55.0% (contextual) and 54.0% (pure semantic) while maintaining high extraction reliability (97.0%).

Table 5.10: RAG impact on performance metrics (n = 48 configurations per condition).

Condition	Mean Accuracy	Extraction Success	Avg Unknown Cases
Without RAG	0.261	0.739	52.1
With RAG	0.218	0.830	34.1
Difference	-0.043	+0.090	-18.0

5.2.4 Top Configurations and Selection for Testing

The five highest-performing configurations by reliability score ($R = 0.7 \times \text{Acc} + 0.3 \times \text{Extract}$) all employed SmolLM2-135M-4-bit (Table 5.11). These configurations advanced to full test set evaluation (Section 4.3).

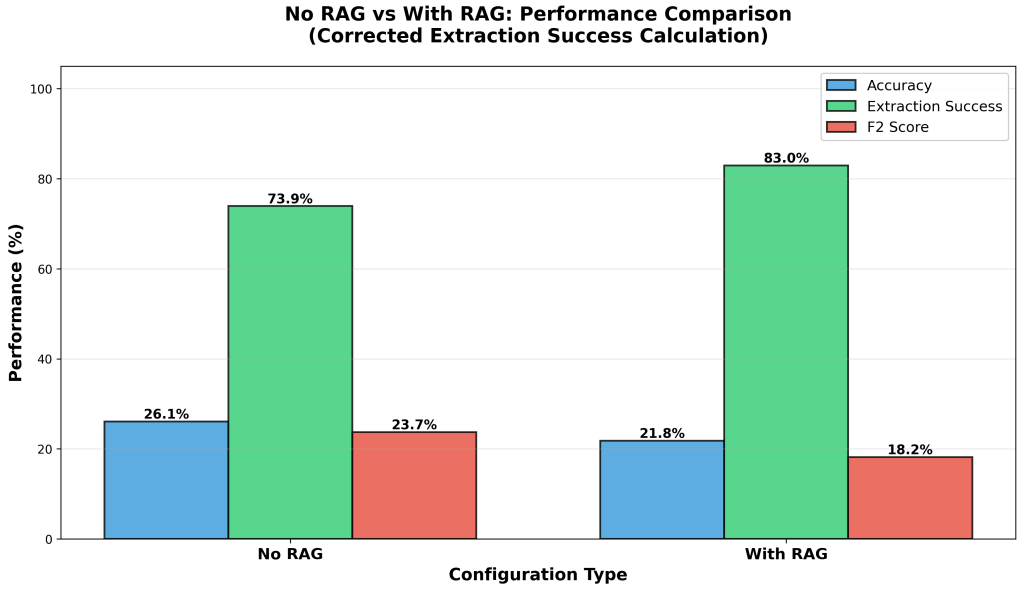

Figure 5.6: Effect of RAG on accuracy, extraction success, and UNKNOWN output rates.

Table 5.11: Top five configurations selected for full test set evaluation.

Rank	Configuration	Acc	F_2	Extract	Reliability
1	135M-4-bit + high_capacity_safe + NoRAG	0.680	0.667	0.995	0.774
2	135M-4-bit + high_capacity_safe + RAG_top1	0.550	0.558	0.970	0.676
3	135M-4-bit + performance_safe + NoRAG	0.570	0.572	0.920	0.675
4	135M-4-bit + high_capacity_safe + RAG_top2	0.540	0.547	0.970	0.669
5	135M-4-bit + balanced_safe + NoRAG	0.595	0.591	0.835	0.667

5.3 Model Testing Framework

The five top-performing validation configurations were evaluated on the full 1,975-case test set. Mean accuracy was 59.7% ± 6.6%, F_2-score was 59.4% ± 5.5%, and mean clinical reasoning quality (LLM-as-judge, 0-100 scale) was 44.9 ± 9.4 points.

5.3.1 Overall Test Performance

Performance metrics for all five configurations are presented in Table 5.12. The accuracy leader (high_capacity_safe, no RAG) achieved 70.4% accuracy with 99.2% extraction success and 0.497s inference time. The safety-optimised configuration (performance_safe, no RAG) achieved 70.3% ED recall but 53.1% overall accuracy. RAG-augmented configurations achieved intermediate performance (54.1-56.8% accuracy) with improved class balance.

Table 5.12: Test set performance for five selected configurations ($n = 1,975$ cases).

Configuration	Acc	F_2	Extract	ED Recall	Quality	Time (s)
high_capacity_safe + NoRAG	**0.704**	**0.684**	**0.992**	0.122	**0.581**	**0.497**
balanced_safe + NoRAG	0.640	0.629	0.867	0.034	0.521	0.382
performance_safe + NoRAG	0.531	0.536	0.908	**0.704**	0.417	0.554
high_capacity_safe + RAG_top1	0.568	0.573	0.979	0.535	0.320	0.966
high_capacity_safe + RAG_top2	0.541	0.548	0.981	0.513	0.317	0.966

5.3.2 Configuration-Specific Analysis

Accuracy Leader (high_capacity_safe + NoRAG): Achieved 70.4% accuracy (F_2 = 0.684) with near-perfect format compliance (15 UNKNOWN outputs). However, the confusion matrix (Figure 5.7) revealed strong GP dominance: GP recall was 89.6% (1,348/1,504), ED recall was 12.2% (43/353), and HOME recall was 0% (0/118). Of the 353 true ED cases, 308 (87.3%) were misclassified as GP, 43 (12.2%) were correctly identified, and 2 (0.6%) became UNKNOWN outputs.

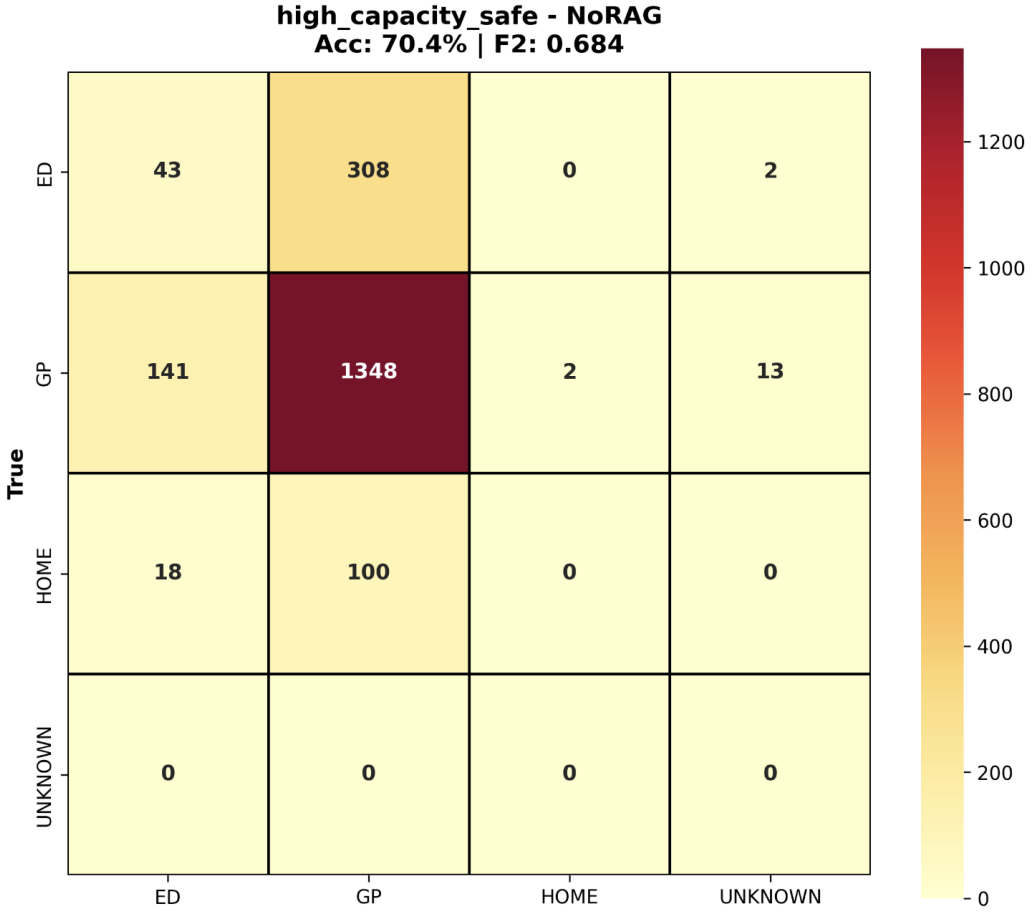

Figure 5.7: Confusion matrix for high_capacity_safe (no RAG) showing severe class imbalance.

Balanced Configuration (balanced_safe + NoRAG): Achieved 64.0% accuracy (F_2 = 0.629) but degraded ED recall further to 3.4% while producing the highest UNKNOWN output rate (262 cases, 13.3%). This configuration proved unreliable for safety-critical deployment.

Safety-optimised Configuration (performance_safe + NoRAG): Achieved 70.3% ED recall (248/353), substantially higher than other configurations, but overall accuracy dropped to 53.1% (F_2 = 0.536) with increased format failures (181 UNKNOWNs) and reduced GP recall (53.3%). HOME recall remained 0%, indicating persistent class blindness.

RAG-Augmented Configurations: Adding retrieval to high_capacity_safe rebalanced classification behaviour at the cost of accuracy and latency. With contextual RAG (top1), accuracy fell to 56.8% (F_2 = 0.573) while ED recall rose to 53.5% and HOME recall reached 2.5%. UNKNOWNs decreased to 41 cases, though latency nearly doubled (0.966s). Pure semantic RAG (top2) produced similar effects: accuracy 54.1%, ED recall 51.3%, HOME recall 5.1%, UNKNOWNs 37, and latency 0.966s.

5.3.3 Class Balance and Safety Trade-offs

Figure 5.8 illustrates the fundamental tension between overall accuracy and emergency detection. Configurations prioritising accuracy achieved 64-70% overall performance but 3-12% ED recall. The safety-optimised adapter achieved 70% ED recall but 53% accuracy. RAG-augmented configurations occupied intermediate positions (54-57% accuracy, 51-54% ED recall) while enabling partial HOME detection (2.5-5.1% recall) for the first time.

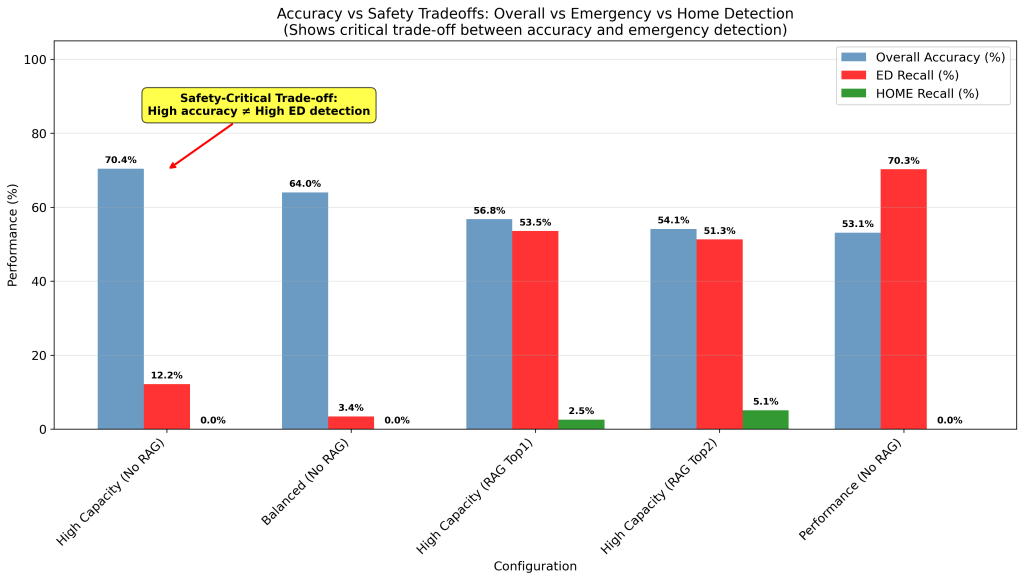

Figure 5.8: Accuracy vs safety trade-off across five test configurations.

5.3.4 LLM-as-Judge Quality Assessment

Clinical reasoning quality was evaluated using Llama-3.3-70B as judge (Table 5.13). Overall quality scores ranged from 31.7 to 58.4 (mean: 44.9 ± 9.4). The high_capacity_safe no-RAG configuration ranked highest across all qualitative dimensions: next-step quality (58.4), reasoning quality (57.5), and overall quality (58.1 ± 23.7).

Table 5.13: LLM-as-judge quality assessment on test set (0-100 scale).

Configuration	Reasoning Quality	Next Steps Quality	Overall Quality
high_capacity_safe + NoRAG	57.5	58.4	58.1
balanced_safe + NoRAG	52.1	54.0	52.1
performance_safe + NoRAG	44.8	44.8	41.7
high_capacity_safe + RAG_top1	32.0	38.1	32.0
high_capacity_safe + RAG_top2	31.7	36.9	31.7

A strong positive correlation (R^2 0.95) was observed between accuracy and quality scores for non-RAG configurations (Figure 5.9). RAG-enabled configurations exhibited a marked discrepancy: despite accuracies of 54-57%, their overall quality underperformed expectations by 26 percentage points, with reasoning quality falling to 31.7-32.0 compared with 44.8-57.5 for non-RAG variants (Figure 5.10).

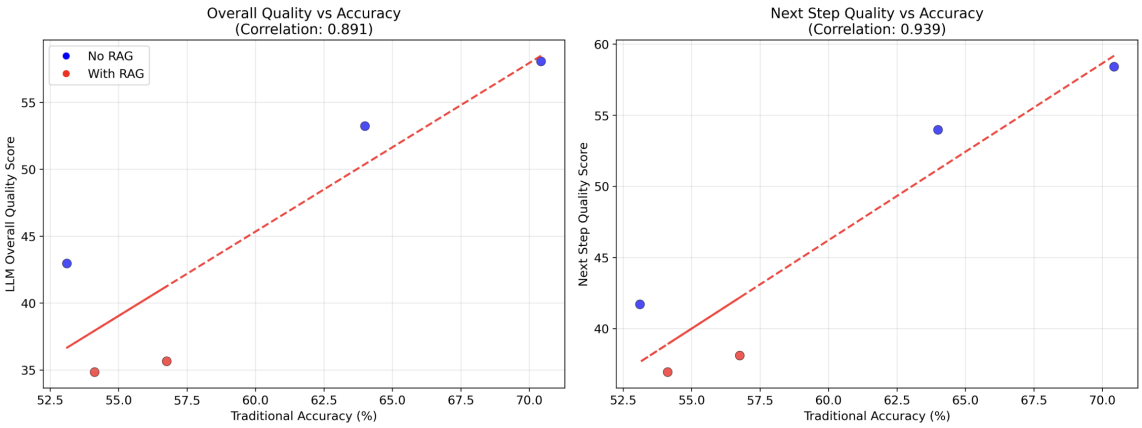

Figure 5.9: Correlation between triage accuracy and LLM-as-judge quality scores.

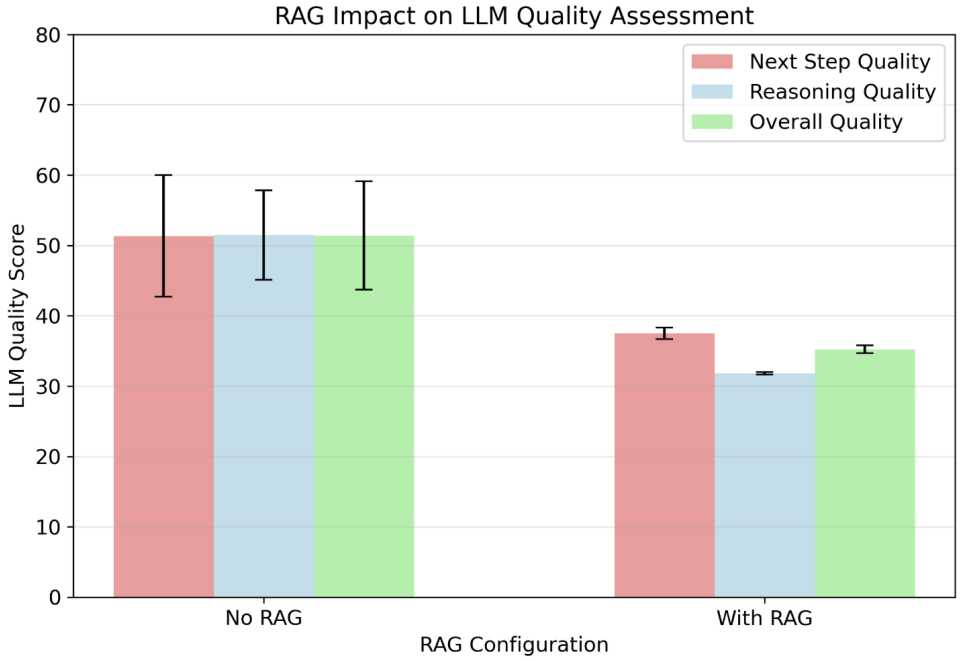

Figure 5.10: RAG impact on LLM-as-judge quality dimensions comparing no-RAG baseline (left) to RAG-augmented configurations (right).

5.3.5 Computational Performance Analysis

Figure 5.11 illustrates the computational cost-accuracy trade-off. No-RAG configurations achieved 0.382-0.554s inference time with 53-70% accuracy. RAG nearly doubled latency to 0.966s while reducing accuracy by 7-13 percentage points. The high_capacity_safe + NoRAG configuration achieved the optimal balance: 70.4 % accuracy at 0.497s per case, occupying the Pareto-optimal position in the latency-accuracy space.

5.3.6 Deployment Criteria Evaluation

None of the five tested configurations met all deployment criteria (Table 5.14). The primary failure mode was insufficient ED recall: the highest achieved was 70.3 %, below the 95 % threshold. All configurations met memory (≤400MB) requirements, but the safety-critical ED detection threshold remained unmet across all tested approaches.

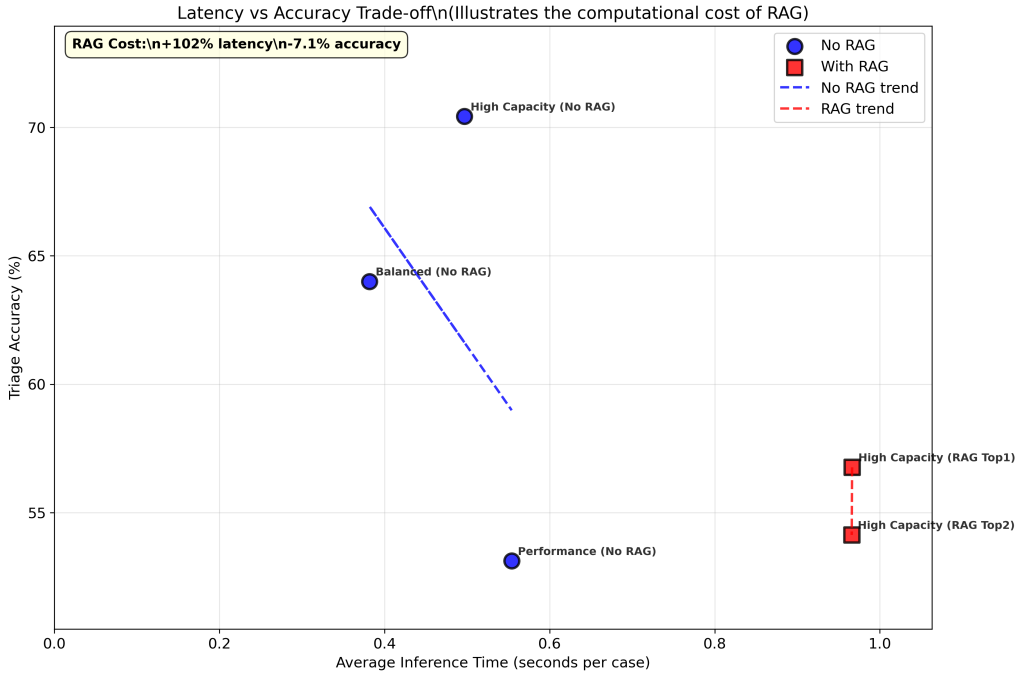

Figure 5.11: Latency vs accuracy trade-off across test configurations.

Table 5.14: Deployment criteria compliance for test configurations.

Configuration	**Acc≥85%**	**ED_R≥95%**	**FNR≤5%**	**Mem≤400MB**
high_capacity_safe + NoRAG	× (70.4%)	× (12.2%)	× (87.8%)	✓
balanced_safe + NoRAG	× (64.0%)	× (3.4%)	× (96.6%)	✓
performance_safe + NoRAG	× (53.1%)	× (70.3%)	× (29.7%)	✓
high_capacity_safe + RAG_top1	× (56.8%)	× (53.5%)	× (46.5%)	✓
high_capacity_safe + RAG_top2	× (54.1%)	× (51.3%)	× (48.7%)	✓

5.4 Mobile Deployment Validation

Although the final test set evaluation (Section 5.3.6) concluded that no configuration met the deployment-ready criteria for clinical safety, the top-performing model for *accuracy* (`SmolLM2-135M-4-bit` with the `high_capacity_safe` adapter) was selected to validate the *technical feasibility* of on-device execution. This test confirms whether the architecture operates within the strict 400 MB memory and sub-3-second latency constraints on real-world hardware. The performance was validated on an iPhone 14 Pro (6 GB RAM, A16 Bionic chip), as shown in Figure 5.12.

The SmolLM2-135M-4-bit model with high_capacity_safe LoRA adapter (rank 16, α 8.0) was deployed using MLX-Swift framework. Peak memory consumption during inference remained below 300 MB (284 MB measured), confirming suitability for resource-constrained mobile environments with 29% headroom within the deployment budget.

End-to-end inference time, measured from input prompt submission to structured triage output, averaged 2.4 seconds per case (2.377s measured in demo). This latency includes model initialisation, tokenisation, inference, output parsing, and UI rendering, meeting the sub-3-second responsiveness requirement.

The deployment demonstration (Figure 5.12) shows a complete triage workflow: patient symptom input with age/gender selection (left), real-time inference execution, and structured output displaying triage decision (ED/GP/HOME), next-step recommendations, and clinical rea-

soning (right). The system operates entirely on-device without network connectivity, preserving patient privacy while maintaining acceptable user experience.

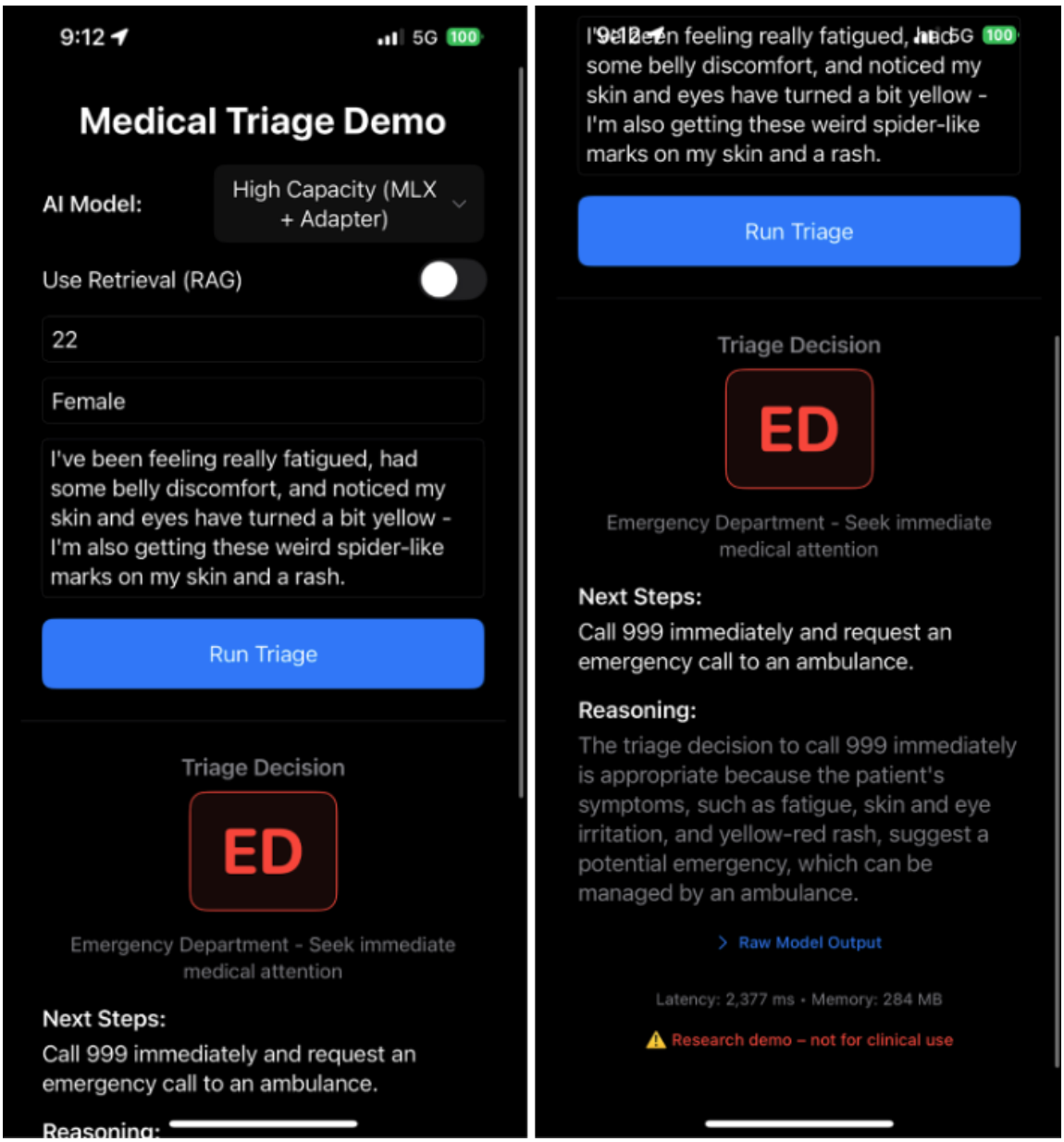

Figure 5.12: On-device deployment demonstration on iPhone 14 Pro.

These results validate that privacy-preserving, on-device medical triage is **technically feasible** on consumer mobile hardware, achieving the dual objectives of patient data protection and real-time responsiveness. However, as established in Section 5.3.6, the demonstrated model's poor clinical safety performance (12.2% ED Recall) confirms that while the **platform** is viable, the **model** itself does not meet the non-negotiable safety thresholds for real-world deployment.

Chapter 6

Discussion of Key Findings

This research demonstrates that privacy-preserving, on-device medical triage is technically feasible within current mobile hardware constraints. The SmolLM2–135M model at 4-bit quantisation achieved 70.4% triage accuracy while operating within a 300 MB peak memory footprint and 2.4 s average inference latency on an iPhone 14 Pro. This represents the first demonstration of medical triage inference under 400 MB, significantly smaller than existing compact medical models such as MedMobile and Meerkat–7B, both of which exceed on-device capacity. The results confirm that modern mobile processors can execute quantised transformer models with LoRA adapters efficiently enough for real-time clinical applications. This addresses Te Whatu Ora's concerns about the privacy risks of cloud-based AI and establishes a foundation for privacy-preserving medical assistants that could operate offline or in rural New Zealand settings.

Methodologically, the novel structured agent-based chunking method was a notable success. It achieved 59.5% Pass@5 retrieval accuracy and outperformed traditional chunking approaches by 7–10 percentage points. Using Llama–3.3–70B to generate structured JSON chunks (*Condition, Symptoms, Triage, NextSteps*) produced self-contained retrieval units that preserved medical completeness while reducing verbosity.

However, the introduction of RAG revealed a critical and unexpected phenomenon. Although RAG substantially improved class balance by raising ED recall from 12% to approximately 51–54% and enabling partial HOME detection, it also produced a 13 percentage point drop in overall accuracy and a 26 point reduction in reasoning coherence. This mirrors the 12.6% performance degradation reported in MedMobile [24] under BM25 retrieval, but the present results extend those findings. Whereas MedMobile relied on simpler retrieval, the current pipeline used structured, compact, agent-generated chunks that should have reduced retrieval noise. The fact that accuracy still collapsed suggests that the problem lies not in retrieval quality alone but in a deeper architectural constraint when external context is injected into ultra-compact models.

6.1 Interpreting the RAG Quality Paradox

Two explanatory hypotheses arise from these results. The first is a *Cognitive Overload Hypothesis*. A 135M parameter model must simultaneously parse patient symptoms, infer differential diagnoses, reason through triage pathways, and integrate several retrieved chunks of medical context. This multitask load exceeds the practical attention capacity of a model of this scale. Rather than improving reasoning, additional context segments the attention budget, dilutes relevant signal, and disrupts coherence. This interpretation is supported by the near-perfect

correlation between accuracy and reasoning quality ($R^2 \approx 0.95$) observed in non-RAG models and by the pronounced degradation in coherence once retrieval is introduced. This behaviour is consistent with scaling-law evidence showing that multi-step reasoning is an emergent ability that only appears in models with hundreds of billions of parameters, while smaller models frequently produce fluent but illogical intermediate steps [52]. Compact models exhibit steep performance degradation as task complexity increases, which mirrors the coherence collapse observed once retrieved context is added.

A second explanation is an *Architectural Mismatch Hypothesis*. Contemporary RAG pipelines have been principally designed and evaluated for multi-billion parameter LLMs with deep attention stacks and substantial redundancy. Such models can filter noisy retrieval content and maintain semantic stability across long contexts. Compact architectures do not possess these structural advantages. Even when retrieval content is compact and medically self-contained, the model lacks the representational depth required to disentangle relevant information from distractors. Under this interpretation, the performance collapse is not an artefact of retrieval accuracy but evidence that RAG, as currently designed, is not well suited for ultra-compact models. This aligns with findings on compact multimodal models, where architectures in the 135M–360M range struggle to maintain semantic stability over extended context lengths, while larger models preserve coherence [18]. Together, these observations indicate that the degradation observed here is not domain-specific, but reflects a general capacity limit in small transformers when external context exceeds their attentional filtering ability.

These findings contribute to the emerging literature on compact clinical models. Whereas existing systems such as MedMobile and Meerkat–7B [27] achieve reduced parameter counts but still assume server-side deployment, the present work demonstrates the feasibility of true on-device inference while exposing the limitations of applying large-model RAG architectures to small models. To the author's knowledge, this is the first evaluation of retrieval-augmented clinical reasoning in an on-device transformer below 200M parameters, extending compact-model limitations into a safety-critical medical setting. Together, these results identify a new research problem: retrieval mechanisms for compact transformers require architectural adaptation rather than direct inheritance from large-model paradigms.

Ultimately, no configuration in this study reached deployment-level safety targets (accuracy $\geq 85\%$, ED recall $\geq 95\%$, false-negative rate $\leq 5\%$). Models tuned for accuracy systematically under-triaged emergencies, while models tuned for safety experienced severe declines in overall accuracy. The trade-off observed reflects a fundamental interaction between model capacity, dataset limitations, and safety-critical task asymmetry. This represents a methodological limitation rather than an inherent failure and is explored further in Chapter 7.

6.2 Implications for Practice

6.2.1 Implications for Healthcare Administrators

For healthcare administrators such as Te Whatu Ora, the results demonstrate that the primary barrier to privacy-preserving triage AI is no longer computational capacity. Mobile hardware can already support 300 MB inference with acceptable latency, meaning that cloud-based processing is a strategic choice rather than a technical requirement.

The safety failures observed in this research arise from a combination of factors, including limitations in the synthetic dataset, the restricted capacity of a 135M parameter model, and the architectural challenges associated with applying RAG to ultra-compact transformers. Dataset

limitations played a substantial role, but they interacted with model scale and architecture rather than acting as the sole cause.

Improving safety therefore requires investment in high-quality, representative triage datasets with appropriate governance, anonymisation pipelines, and demographic stratification, particularly for Māori and Pacific populations. Such investment will yield significantly greater safety dividends than increased compute resources.

6.2.2 Implications for AI Developers

For AI developers, the findings highlight several conceptual lessons for building safety-critical small language models. First, the results show that standard accuracy metrics can be misleading when task classes carry asymmetric risk. The accuracy-optimised configuration achieved strong aggregate performance while systematically under-triaging emergency cases, illustrating the danger of relying on undifferentiated accuracy in clinical settings.

Second, the study demonstrates that methods developed for large language models cannot be assumed to transfer directly to compact architectures. The RAG Quality Paradox provides clear evidence that retrieval depth, context length, and chunk structure behave differently in small models and may degrade rather than enhance reasoning quality. Retrieval therefore needs to be treated as a safety-relevant design choice rather than a routine optimisation step.

Finally, the interaction between model capacity, dataset imbalance, and safety requirements indicates that compact clinical models require training and evaluation frameworks that foreground safety from the outset. This includes the need for metrics that reflect clinical risk, for evaluation procedures that capture under-triage explicitly, and for model design choices that recognise that high-level performance does not guarantee safe behaviour on rare but critical cases.

These implications are conceptual rather than technical. They complement the methodological improvements outlined in Chapter 8 and offer guidance for the broader development of small language models for clinical decision support. Building safe and reliable on-device systems will require coordinated investment in data governance, architecture design, evaluation frameworks, and regulatory oversight rather than reliance on standard accuracy optimisation alone.

Chapter 7

Limitations

7.1 Methodological and Dataset Limitations

The critical failures in safety performance stem primarily from limitations in the training dataset. This study relied entirely on synthetic data, with all 13,825 samples generated via Llama-3.3-70B without human validation. This approach enabled rapid dataset construction using verified medical sources but could not replicate the nuance, ambiguity, or cultural diversity of real triage conversations, which are difficult to obtain due to privacy, consent, and annotation constraints.

Two dataset properties contributed directly to the model's safety failures. The first was severe class imbalance (76% GP, 18% ED, 6% HOME), which hindered the model's ability to learn rare but critical emergency and home-care patterns. Even with hyperparameter tuning, the 135M model failed to generalise to HOME cases, achieving 0% recall. The second was systematic bias introduced by synthetic generation. LLM-generated patients articulate symptoms with unrealistic clarity, reasoning follows textbook logic, and averaged triage labels may obscure context-specific details. The model therefore likely overfitted to "idealised dialogue patterns" unrepresentative of real clinical language.

Although cost-sensitive penalties were incorporated conceptually into the training design, MLX's standard LoRA implementation did not support custom loss integration. Safety was therefore enforced through conservative hyperparameter selection and strict post-training filtering based on ED recall thresholds. As discussed in Chapter 8, enabling cost-sensitive loss within the training loop remains a priority for future development.

The evaluation framework also introduces methodological constraints. Qualitative reasoning quality was assessed solely using Llama-3.3-70B as an automated judge. This provided scalable, reproducible comparison across 96 configurations but lacks human clinical validation. Additionally, the test set was drawn from the same synthetic distribution as the training set, limiting generalisability and preventing robust claims about real-world clinical performance. The dataset also lacked demographic stratification (e.g., Māori and Pacific populations), precluding analysis of equity-related failure modes.

7.2 Mobile Deployment Limitations

The on-device evaluation demonstrated feasibility but was limited to a single high-end device (iPhone 14 Pro). Performance on older, lower-resource, or Android devices remains unverified. The measured 2.4-second inference latency reflects average single-turn performance and does not account for thermal throttling, sustained usage, or battery constraints, all of which may

degrade reliability in real-world scenarios.

The implementation also involved architectural compromises. LoRA adapters were applied only to the final 16 of 30 transformer layers to meet memory constraints, potentially limiting their influence on core reasoning behaviour. The user interface was a proof-of-concept and has not undergone usability or safety validation with actual telehealth users.

7.3 Multilingual and Cross-Cultural Limitations

This study evaluated model performance exclusively on English-language inputs. In New Zealand, where te reo Māori and Pacific languages are central to equitable healthcare delivery, the absence of multilingual evaluation is a significant limitation. Without linguistic diversity, the model cannot be assumed to perform safely or fairly for all population groups.

7.4 Model Calibration and Confidence Estimation

The study did not assess model calibration or confidence alignment. Although qualitative reasoning was evaluated, the correspondence between predictive confidence and correctness remains unknown. Miscalibration, particularly overconfident incorrect predictions poses substantial clinical risk. Incorporating calibration metrics, such as Expected Calibration Error (ECE), and uncertainty-aware scoring will be essential before clinical deployment.

Chapter 8

Future Works

8.1 Validation with Real-World Clinical Data

The most critical next step is validation using real anonymised patient data from New Zealand telehealth services and emergency departments. This would enable distribution shift analysis, demographic stratification, identification of equity impacts, and clinician-verified reasoning evaluations. However, such validation requires overcoming ethical, regulatory, and privacy barriers, including informed consent, Privacy Act compliance, data-sharing agreements, and institutional ethics approval.

8.2 Future Model and Architectural Directions

As mobile compute capacity increases, several architectural improvements become viable. Larger models within the strict memory budget, retrieval-augmented pre-training, and multimodal extensions incorporating vital signs or images (such as rashes or lesions) may meaningfully enhance clinical reasoning. These improvements must be coupled with retrieval mechanisms tailored specifically to compact transformers to avoid the failure modes observed in this study.

8.3 Safety-Critical Training and Inference Mechanisms

Future work should integrate safety considerations directly into training rather than relying solely on post-training filtering.

8.3.1 Cost-Sensitive Loss Functions

Embedding asymmetric risk directly into the optimisation objective is essential. Implementing a cost-sensitive loss function that encodes the penalty matrix, assigning greater cost for ED misclassification would shift the training focus toward high-recall emergency detection. This requires modifying MLX's training loop to support custom loss functions or adopting focal loss variants that implicitly upweight rare but critical cases.

8.3.2 Monte Carlo Dropout for Uncertainty Quantification

Enabling Monte Carlo dropout during inference would allow uncertainty-aware decision-making. Multiple stochastic forward passes can estimate epistemic uncertainty, flagging high-entropy predictions (e.g., entropy > 0.3) for automatic escalation to human triage operators. Implementing this requires model-level control over dropout states or a custom inference loop capable of batch stochastic sampling.

8.3.3 Ensemble Methods

Ensembling multiple fine-tuned adapters, each trained with different seeds or hyperparameters may improve both calibration and accuracy. Disagreement across ensemble members provides a robust uncertainty signal independent of dropout. However, this approach increases inference cost and must be optimised to remain feasible under mobile memory and latency constraints.

8.4 Retrieval Architecture Innovation

Addressing the RAG quality deficit requires retrieval architectures tailored to small models. Promising directions include retrieval-interleaved generation, fusion-in-decoder methods, retriever-generator co-training, and learned retrieval optimisation grounded in triage accuracy. Compressing or summarising retrieved context to reduce verbosity remains a priority, as compact transformers are highly sensitive to context length and noise.

8.5 Clinical and Regulatory Pathway

Clinical deployment requires alignment with regulatory standards, including Medsafe classification as a Class IIb medical device, randomised controlled trials, integration with existing Healthline infrastructure, and robust human-in-the-loop monitoring. Transparent failure reporting, bias audits, and ongoing model calibration will be essential components of a safe deployment strategy.

8.6 Concluding Remarks on Feasibility

Measured performance, 300 MB peak memory and 2.4 s latency on Apple's A16 Bionic demonstrates that on-device inference is already feasible on contemporary smartphones. With rapid advances in mobile NPUs, models of this scale will soon be deployable on mid-range devices. The remaining barriers are not computational but clinical: real-world validation, representative datasets, safety-focused training, and robust uncertainty estimation. As hardware continues to improve, the feasibility of on-device clinical AI will only increase.

Chapter 9

Conclusion

This research investigated whether a small, privacy preserving language model can deliver safe medical triage directly on consumer devices. The findings demonstrate that on-device triage is technically feasible within realistic mobile memory and latency constraints. A 135M parameter model operating within a 300 MB memory budget and achieving sub three second inference shows that local, offline triage support is already possible on contemporary smartphones. This represents a significant step toward equitable, offline capable AI support for rural and underserved communities across New Zealand.

The contributions of this work are threefold. First, a structured agent based chunking method achieved 59.5 percent Pass@5 retrieval accuracy, outperforming traditional chunking approaches while maintaining medical completeness. Second, LoRA fine tuning improved emergency sensitivity, reaching 70.3 percent ED recall under constrained conditions. Third, the complete end to end pipeline from knowledge base construction to mobile deployment demonstrates a reproducible framework for compact, privacy preserving medical AI.

These contributions must be interpreted alongside the limitations revealed by the results. No configuration met clinically acceptable safety thresholds, and the RAG quality paradox indicates an architectural mismatch between standard retrieval pipelines and ultra compact models. The reliance on synthetic and imbalanced data further restricted generalisation to rare critical cases. The system developed here is therefore a proof of concept rather than a deployable clinical tool.

The healthcare context of New Zealand underscores the relevance of this work. GP shortages, emergency department crowding, rural access barriers, and persistent inequities affecting Māori and Pacific peoples create conditions where timely care is often unavailable. Te Whatu Ora's restrictions on cloud based AI highlight legitimate concerns regarding privacy, fairness, and safety. This research shows that privacy preserving on-device AI can overcome these barriers while supporting data sovereignty, offline capability, and user trust.

Technical feasibility alone does not guarantee clinical readiness. Safe deployment will require high quality representative clinical data, robust calibration, clinician verified evaluation, and design approaches that prioritise equity for communities facing the greatest barriers to care. The findings provide an empirical foundation and highlight where future innovation must occur.

Looking forward, advances in mobile hardware, compact model research, and on-device AI governance suggest that clinically reliable on-device triage assistants are within reach. With rigorous validation, calibration aware training, and meaningful partnership with clinicians, Māori and Pacific health leadership, and Te Whatu Ora, the framework demonstrated here may evolve into a safe, culturally aligned, and locally governed triage support system for New Zealand.

Bibliography

[1] A. Murray and V. Chandrawidjaja, "Where have all the doctors gone?" 1News, Aug 2024, [Online]. Available: https://www.1news.co.nz/2024/08/01/where-have-all-the-doctors-gone/.

[2] Patient Voice Aotearoa, "More doctorless or closed hospitals in NZ?" Scoop, May 2025, [Online]. Available: https://www.scoop.co.nz/stories/GE2505/S00060/more-doctorless-or-closed-hospitals-in-nz.html.

[3] RNZ, "Specialist doctor shortage: More than a third of adults not getting healthcare they need," NZ Herald, May 2024, [Online]. Available: https://www.nzherald.co.nz/nz/specialist-doctor-shortage-more-than-a-third-of-adults-not-getting-healthcare-they-need/QSRXTN72ENDVTCTJFINZT5LZ6Y/.

[4] Medical Council of New Zealand, "The New Zealand medical workforce in 2024," Medical Council of New Zealand, Wellington, New Zealand, Tech. Rep., 2024, [Online]. Available: https://www.mcnz.org.nz/assets/Publications/Workforce-Survey/Workforce_Survey_Report_2024.pdf. [Online]. Available: https://www.mcnz.org.nz/assets/Publications/Workforce-Survey/Workforce_Survey_Report_2024.pdf

[5] Royal Australasian College of Physicians, "Workforce data shows physician shortages across Aotearoa New Zealand," Sep 2023, [Online]. Available: https://www.racp.edu.au/news-and-events/media-releases/workforce-data-shows-physician-shortages-across-aotearoa-new-zealand.

[6] Association of Salaried Medical Specialists, "Workforce: the make or break of the health reform," Association of Salaried Medical Specialists, Auckland, New Zealand, Tech. Rep., Sep 2023, [Online]. Available: https://asms.org.nz/wp-content/uploads/2023/09/Executive-summary_Final-sep-2023-update.pdf.

[7] R. Hill, "Doctor shortages: Plans for on the job training puts more pressure on GPs," RNZ, Mar 2025, [Online]. Available: https://www.rnz.co.nz/news/national/545134/doctor-shortages-plans-for-on-the-job-training-puts-more-pressure-on-gps.

[8] P. G. Jones and G. Jackson, "Emergency department crowding is not being caused by increased inappropriate presentations," *New Zealand Medical Journal*, Dec 2023.

[9] Te Whatu Ora – Health New Zealand, "Te Mauri o Rongo — The New Zealand Health Strategy," Wellington, New Zealand, 2024, [Online]. Available: http://health.govt.nz/strategies-initiatives/health-strategies/new-zealand-health-strategy.

[10] Whakarongorau Aotearoa, "Our Services," 2025, [Online]. Available: https://whakarongorau.nz/telehealth-services.

[11] ——, "Healthline reduces pressure on emergency departments," 2024, [Online]. Available: https://whakarongorau.nz/about/our-research.

[12] P. Sun, "How AI helps physicians improve telehealth patient care in real-time," Arizona Telemedicine Program, 2022, [Online]. Available: https://telemedicine.arizona.edu/blog/how-ai-helps-physicians-improve-telehealth-patient-care-real-time.

[13] R. Whittaker, R. Dobson, C. K. Jin, R. Style, P. Jayathissa, K. Hiini, K. Ross, K. Kawamura, P. Muir, and the Waitematā AI Governance Group, "An example of governance for AI in health services from Aotearoa New Zealand," *npj Digital Medicine*, vol. 6, no. 1, p. 164, 2023.

[14] New Zealand Medical Informatics Institute, "Te Whatu Ora advice about the use of artificial intelligence in healthcare," 2023, [Online]. Available: https://nzmii.co.nz/resources/fact-sheet/te-whatu-ora-advice-about-the-use-of-artificial-intelligence-in-healthcare.

[15] S. Y. Kim, D. H. Kim, M. J. Kim, H. J. Ko, and O. R. Jeong, "XAI-based clinical decision support systems: A systematic review," *Applied Sciences*, vol. 14, no. 15, p. 6638, 2024.

[16] A. Suksen and N. Nok, "Artificial intelligence-based symptom checkers for disease diagnosis: A systematic review," 2025.

[17] C. Hymel and H. Johnson, "Analysis of LLMs vs human experts in requirements engineering," 2025.

[18] A. Marafioti Orr *et al.*, "SmolVLM: Redefining small and efficient multimodal models," 2025.

[19] J. Xu, Z. Li, W. Chen, Q. Wang, and X. Xin, "On-device language models: A comprehensive review," 2024.

[20] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive NLP tasks," 2020.

[21] E. J. Hu *et al.*, "LoRA: Low-rank adaptation of large language models," 2021.

[22] K. Singhal *et al.*, "Towards Expert-Level Medical Question Answering with Large Language Models," 2023.

[23] Y. Li *et al.*, "ChatDoctor: A Medical Chat Model Fine-Tuned on a Large Language Model Meta-AI (LLaMA) Using Medical Domain Knowledge," 2023.

[24] K. Vishwanath, J. Stryker, A. Alaykin, D. A. Alber, and E. K. Oermann, "MedMobile: A mobile sized language model with expert-level clinical capabilities," 2024.

[25] E. Morin *et al.*, "BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains," 2024.

[26] J. Meribe and S. Zhang, "Ultimate MedLLM-Llama3-8B: Fine-tuning Llama 3 for Medical Question-Answering," Dept. Comput. Sci., Stanford Univ., Stanford, CA, USA, Stanford CS224N Custom Project Rep., 2024.

[27] H. Kim, H. Hwang, J. Lee *et al.*, "Small language models learn enhanced reasoning skills from medical textbooks," *npj Digital Medicine*, vol. 8, no. 1, p. 240, 2025.

[28] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smarter, faster, cheaper and lighter," 2019.

[29] Z. Sun, H. Yu, X. Song, R. Liu, Y. Yang, and D. Zhou, "MobileBERT: A compact task-agnostic BERT for resource-limited devices," 2020.

[30] L. Ben Allal *et al.*, "SmolLM2: When smol goes big – data-centric training of a small language model," 2025.

[31] Google, "Introducing Gemma 3: A new generation of open models," Google for Developers Blog, Sep 2024, [Online]. Available: https://developers.googleblog.com/en/introducing-gemma-3-270m/.

[32] Z. Liu *et al.*, "MobileLLM: Optimizing Sub-billion Parameter Language Models for On-Device Use Cases," in *Proceedings of the 41st International Conference on Machine Learning*, Vienna, Austria, 2024.

[33] R. Jin *et al.*, "A Comprehensive Evaluation of Quantization Strategies for Large Language Models," 2024.

[34] E. Frantar, S. Ashkboos, T. Hoefler, and D. Alistarh, "GPTQ: Accurate post-training quantization for generative pre-trained transformers," 2022.

[35] E. Frantar and D. Alistarh, "Massive language models can be accurately pruned in one-shot," 2023.

[36] T. Dettmers, R. Svirschevski, V. Egiazarian, D. Kuznedelev, E. Frantar, S. Ashkboos, A. Borzunov, T. Hoefler, and D. Alistarh, "SpQR: A sparse-quantized representation for near-lossless LLM weight compression," 2023.

[37] E. Kurtić, A. Marques, M. Kurtz, and D. Alistarh, "We ran over half a million evaluations on quantized LLMs—here's what we found," Red Hat Developer, Oct 2024, [Online]. Available: https://developers.redhat.com/articles/2024/10/17/we-ran-over-half-million-evaluations-quantized-llms.

[38] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLORA: Efficient Finetuning of Quantized LLMs," 2023.

[39] G. Hinton, O. Vinyals, and J. Dean, "Distilling the Knowledge in a Neural Network," 2015.

[40] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford, "Okapi at TREC-3," in *Proceedings of the Third Text REtrieval Conference (TREC-3)*, 1994.

[41] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with GPUs," 2017.

[42] V. Karpukhin, B. Oğuz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-t. Yih, "Dense Passage Retrieval for Open-Domain Question Answering," 2020.

[43] C. Sharma, "Retrieval-Augmented Generation: A Comprehensive Survey of Architectures, Enhancements, and Robustness Frontiers," 2025.

[44] Z. Rackauckas, "RAG-Fusion: A New Take on Retrieval-Augmented Generation," *International Journal of Natural Language Computing (IJNLC)*, vol. 13, no. 1, pp. 37–47, Feb 2024.

[45] G. V. Cormack, C. L. A. Clarke, and S. Büttcher, "Reciprocal Rank Fusion outperforms Condorcet and Individual Rank Learning Methods," in *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, Boston, MA, USA, Jul 2009, pp. 758–759.

[46] R. Xu *et al.*, "SimRAG: Self-Improving Retrieval-Augmented Generation for Adapting Large Language Models to Specialized Domains," 2025.

[47] Z. Wang, J. Araki, Z. Jiang, M. R. Parvez, and G. Neubig, "Learning to Filter Context for Retrieval-Augmented Generation," 2023.

[48] Anthropic, "Contextual Retrieval for RAG," May 2024, [Online]. Available: https://www.anthropic.com/news/contextual-retrieval.

[49] Mastering LLM, "11 Chunking Strategies for RAG — Simplified & Visualized," Medium, Nov 2024, [Online]. Available: https://masteringllm.medium.com/11-chunking-strategies-for-rag-simplified-visualized-df0dbec8e373.

[50] D. Amodei *et al.*, "Concrete Problems in AI Safety," 2016.

[51] J. Li *et al.*, "Agent Hospital: A Simulacrum of Hospital with Evolvable Medical Agents," 2025.

[52] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," in *Advances in Neural Information Processing Systems 36*, 2022.