

PRIVACY-PRESERVING CLINICAL AI

MobiCureVN

An offline, bilingual on-device AI companion for post-surgical colorectal recovery

Phung Thi Minh Anh s3986878 · Do Le Trang Hanh s3977994 · Nguyen Minh Quan s3979391 · Le Thien Son s3977955 · Trinh Phuong Thao s3979297 | Supervisors: Dr. Arthur Tang, Dr. Tom Huynh · Company: RMIT SSET

ABSTRACT

Patients recovering from colorectal cancer surgery in Vietnam go home with a healing wound or a stoma and very little continuous support: reliable guidance is written in English for clinicians, and the questions that matter most arrive at night. MobiCureVN puts a Vietnamese-speaking recovery assistant on the patient's own iPad. Every component runs on the device — a 4-bit quantised Qwen 3.5 4B model served through MLX Swift, a local retrieval-augmented generation index built from 39 curated clinical documents (1,238 chunks, hybrid vector + full-text search in SQLite), Apple's on-device Translation framework, and a vision-language model for wound-photo triage. An emergency detector runs on the patient's original Vietnamese wording *before* translation or generation, so red-flag symptoms trigger a direct “seek care now” redirect rather than a generated answer. No patient text, chat history, or wound photograph ever leaves the iPad — there is no server to send it to.

01 WHY IT MATTERS

Colorectal cancer is rising fast in Vietnam, and most recovery happens at home, in Vietnamese, without a nurse in the room. General-purpose chatbots answer in English first, invent details, and send private symptoms to servers abroad — unacceptable in post-operative cancer care.



02 OBJECTIVES

- 01 Answer recovery questions in Vietnamese, in the words a patient actually uses
- 02 Ground every answer in retrieved clinical sources and show the citation
- 03 Keep all inference, retrieval and translation on the device
- 04 Escalate urgent symptoms before the model is ever invoked
- 05 Measure retrieval and safety rather than assume them

03 WHAT WE BUILT

- ▶ An iPad app in Vietnamese: chat, wound photo log, medication reminders
- ▶ An offline ingestion pipeline producing a `vectorstore.db` shipped in the app bundle — 39 documents, 1,238 chunks
- ▶ An on-device hybrid retriever: vector KNN plus FTS5 keyword search, fused with Reciprocal Rank Fusion
- ▶ A three-stage safety layer: emergency detector, input guardrail, output guardrail

04 WHERE WE ARE

Requirement	Target	Status
Data locality	100% on-device	Met
Bilingual VI / EN	Full interaction	Met
Retrieval accuracy	doc-hit@5 ≥ 0.75	0.770
Safety coverage	Full suite passing	337
Response latency	Under 5 s end to end	In eval
Multimodal input	Text, voice, image	Building

05 ON THE DEVICE

Component	What ships
Language model	Qwen 3.5 4B, 4-bit quantised, MLX Swift
Embedder	BGE-small, CoreML build, 384-dim
Clinical index	39 documents · 1,238 chunks · sqlite-vec + FTS5
Translation	Apple Translation framework, on-device
Footprint	≈ 2.5 GB installed · Metal cache capped at 512 MB
Network	None in the response path; one-time asset download

06 HOW ONE QUESTION IS ANSWERED

A patient asks, in Vietnamese: “Vết mổ của tôi chảy dịch vàng, có sao không?” Everything below happens on the iPad, with the network switched off.

- 01 **Understand the language**
The on-device model detects the language and repairs typos and VI/EN code-switching without changing the question.
- 02 **Check for an emergency — first, in the patient's own words**
Red-flag wording short-circuits the pipeline and returns a “seek care now” redirect. The model is never invoked on an emergency turn.
- 03 **Translate inward, then screen the request**
Apple Translation converts VI → EN locally; the input guardrail rejects jailbreaks, out-of-scope questions and self-harm content.
- 04 **Retrieve the evidence**
A CoreML build of BGE-small embeds the query on-device; vector similarity (sqlite-vec) and keyword search (FTS5 BM25) run at once, fused with Reciprocal Rank Fusion. Top 5 passages become the grounding context.
- 05 **Generate locally**
Qwen 3.5 4B (4-bit, via MLX Swift on the Neural Engine and GPU) answers using only those passages, the patient profile, and facts remembered from this conversation.
- 06 **Screen the answer, translate back, show the receipts**
The output guardrail inspects the complete response before the patient sees any of it — diagnostic phrasing stripped, citations checked, PII masked — then the model translates its own answer to Vietnamese and a validation pass confirms it. The reply arrives with the clinical sources it was built from.

07 INSIDE ONE IPAD

PRESENTATION	SwiftUI chat, citation cards, wound log, medication tracker
APPLICATION	Chat orchestrator, emergency detector, input & output guardrails, session-fact memory
AI & RETRIEVAL	MLX Swift runtime, hybrid SQLite retriever, CoreML embedder, vision-language wound model, Apple Translation
DATA	Read-only clinical index in the app bundle; chat, profile and wound history in a local store

No account, no login, no telemetry, no crash reporter — and no server to send anything to.

08 WHAT WE LEARNED, WHAT IS NEXT

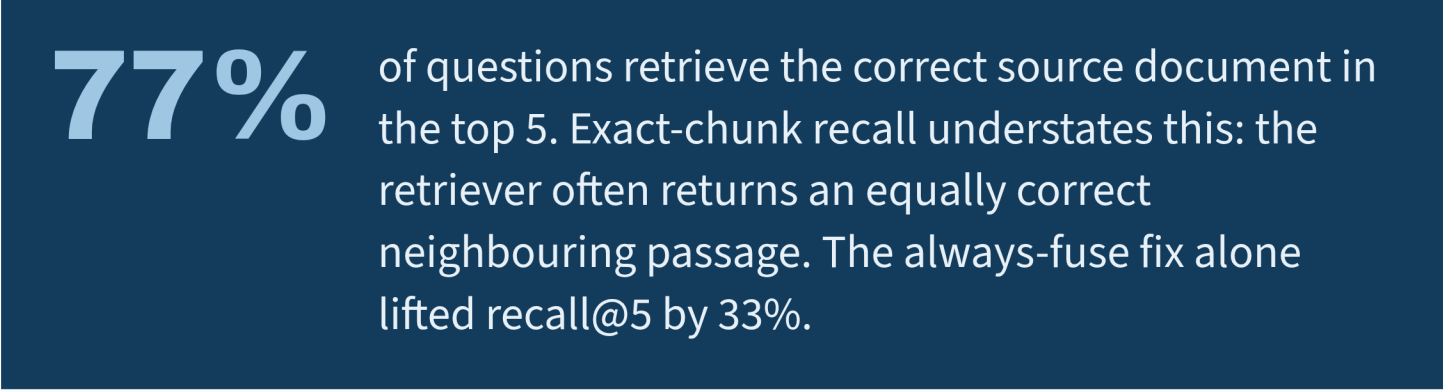
A small quantised model becomes genuinely useful once it is made to read from trusted documents instead of answering from memory. Privacy turned out to be architectural, not a setting: with no server, there is nowhere for patient data to go.

- Close the accuracy levers from our engineering audit: embedder pooling alignment, cross-encoder reranking, chunk-size limits
- Extend evaluation coverage to Vietnamese-language queries
- Finish voice input, benchmark latency, and test with patients

09 HOW WELL IT PERFORMS

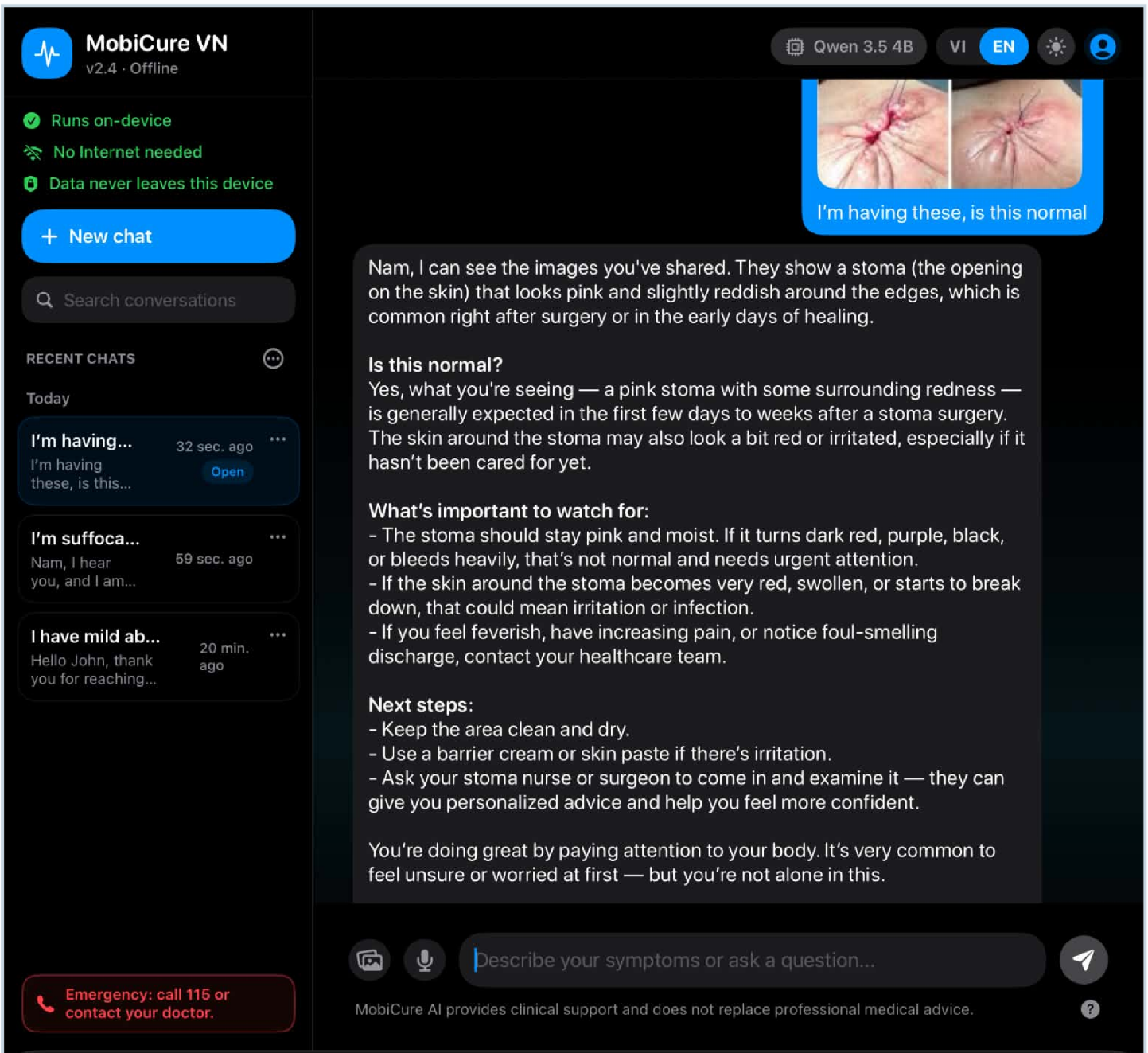
RETRIEVAL — 209-QUESTION GOLDEN SET

Configuration	recall@5	MRR@5	doc-hit@5
Original hybrid	0.187	0.097	0.689
Vector-only	0.239	0.170	0.756
Shipped hybrid + always-fuse	0.249	0.159	0.770

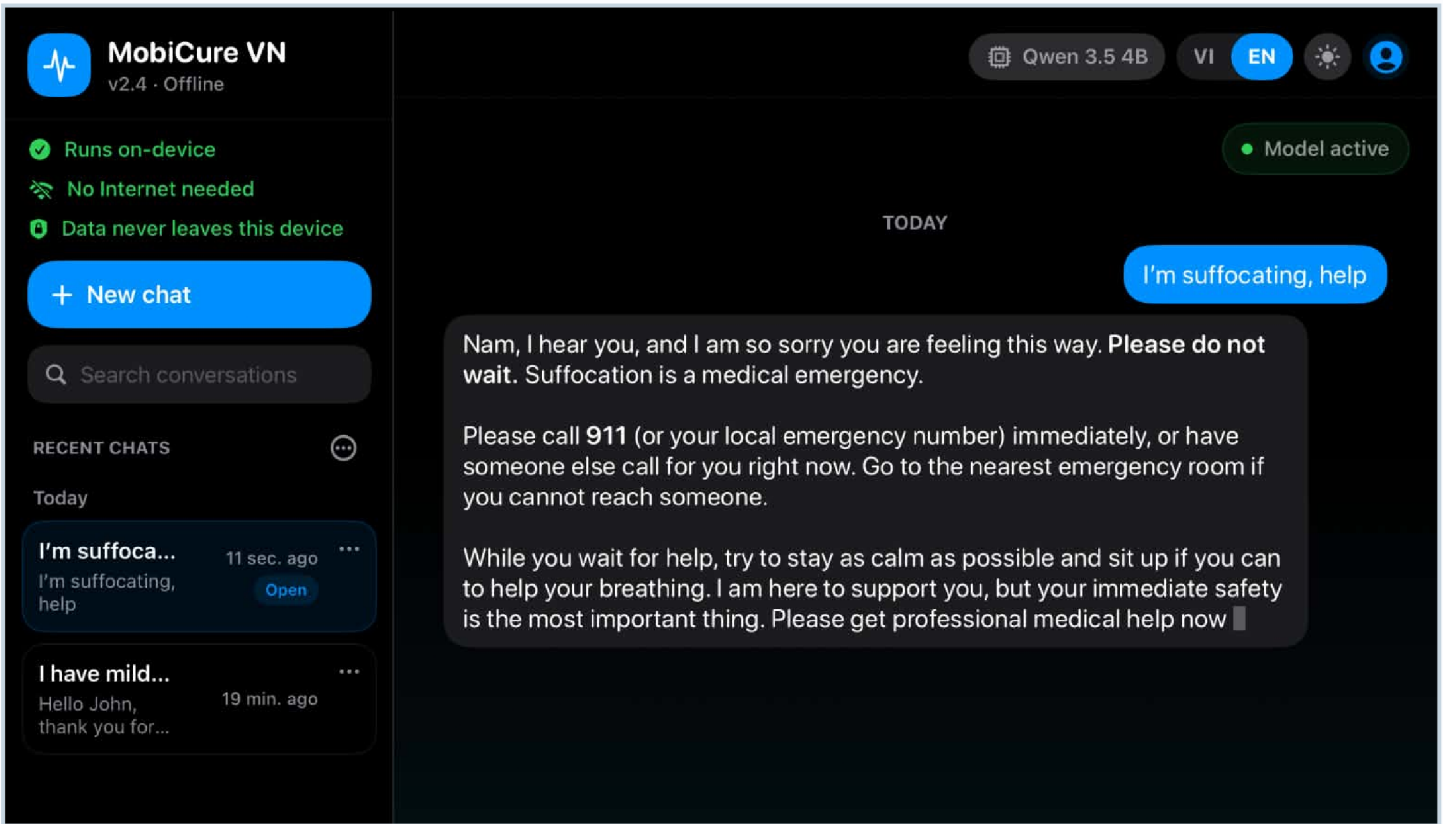


- ▶ **337 automated tests** across 21 suites — emergency detection (including obfuscated phrasings), guardrail bypass, output filtering, PII masking, citation integrity

10 THE APP



WOUND PHOTO TRIAGE — THE VISION-LANGUAGE MODEL READS THE IMAGE, THE ANSWER STAYS ON-DEVICE



EMERGENCY DETECTION — RED-FLAG WORDING SHORT-CIRCUITS TO A SEEK-CARE-NOW REDIRECT

ACKNOWLEDGMENTS

With thanks to Dr. Arthur Tang and Dr. Tom Huynh, and to RMIT SSET, for their guidance throughout this project. MobiCureVN is a student research prototype and does not replace advice from a qualified clinician.