

Safe Medical AI for Low-Resource Languages

Evidence-Aware Health Information Assistants for Vietnamese
and Underserved Language Communities

NGO TRUNG KIEN

Manuscript version 0.9.0 — design-complete draft

Not clinically validated. Not a medical device.

All frameworks are design proposals awaiting qualified human review.

September 13, 2026

Contents

Abbreviations and Acronyms	xvii
Glossary	xix
I Foundations	1
1 Introduction — Why Low-Resource Medical AI Is Different	3
1.1 The Promise and Peril of Medical AI	4
1.2 Why Translation Is Not Enough	4
1.2.1 Linguistic Structure Creates Safety Risks	4
1.2.2 Regional Variation Is Medically Relevant	5
1.2.3 Informal Language Is the Norm	6
1.2.4 Traditional Medicine Is Culturally Embedded	6
1.3 The Evidence-Aware Health-Information Assistant Concept	6
1.4 Book Scope, Audience, and Limitations	7
1.4.1 Scope	7
1.4.2 Audience	7
1.4.3 Limitations	7
1.5 MedAI Position Statement	7
1.6 How to Read This Book	8
2 Language Variation and Clinical Safety	11
2.1 Introduction	12
2.2 Vietnamese Linguistic Structure	12
2.2.1 Tonal System	12
2.2.2 Syllable Structure	12
2.3 Regional Dialect Variation	14
2.4 Missing-Diacritic Ambiguity in Medical Contexts	14
2.5 Code-Switching Patterns in Health Queries	14

2.6	Informal and Colloquial Symptom Descriptions	16
2.7	Traditional Medicine Vocabulary	16
2.8	Health Literacy Variation	17
2.9	Language-Mediated Safety Failure Taxonomy	17
2.10	Evaluation Dimensions for Language Robustness	17
2.11	Limitations	18
2.12	Chapter Summary	18
3	Evidence, Citation, and Provenance	19
3.1	Introduction	20
3.2	The Citation Problem in Medical AI	20
3.2.1	Hallucinated Citations	20
3.2.2	Citation Laundering	20
3.2.3	Outdated Evidence	21
3.3	Evidence Types and Strength Levels	21
3.4	Provenance Chains from Source to Claim	21
3.5	Retrieval-Augmented Generation for Evidence Grounding	23
3.6	Citation Verification Architecture	24
3.7	Traditional-Use Evidence vs. Clinical Evidence	24
3.8	Source Policy and Verification Requirements	24
3.9	Limitations	25
3.10	Chapter Summary	25
II	Safety Architecture	27
4	Clinical Risk Classification	29
4.1	Introduction	30
4.2	Risk Level Definitions	30
4.3	Classification Decision Logic	31
4.3.1	Algorithm	31
4.3.2	Critical Design Rules	31
4.4	High-Risk Population Modifiers	31
4.5	Uncertainty Handling	33
4.6	Failure Modes and Their Medical Consequences	33
4.7	Limitations	33
4.8	Chapter Summary	34

5	Emergency Detection and Escalation	37
5.1	Introduction	38
5.2	The Multi-Layer Detection Architecture	38
5.2.1	Why Multiple Layers Are Necessary	38
5.2.2	Detection Pipeline	38
5.2.3	Critical Design Rules	39
5.3	Vietnamese Emergency Red-Flag Lexicon	39
5.3.1	Lexicon Design Principles	39
5.3.2	Lexicon Coverage by Emergency Category	40
5.3.3	Negation Handling	40
5.4	Emergency Response Templates	40
5.4.1	Template Design Principles	40
5.4.2	General Emergency Template	41
5.4.3	Template Variants	41
5.4.4	Prohibited Template Content	42
5.5	Emergency Bypass Path	42
5.6	Detection Evaluation Criteria	43
5.6.1	Metrics and Design Targets	43
5.6.2	Required Test Coverage	43
5.6.3	Failure Analysis Protocol	44
5.7	Over-Escalation Management	44
5.8	Special Emergency Scenarios	44
5.8.1	Third-Party Queries	44
5.8.2	Patient Dismissal	44
5.8.3	Multiple Symptoms with One Emergency	45
5.9	Limitations	45
5.10	Chapter Summary	45
6	Response Policy and Output Safety	47
6.1	Introduction	48
6.2	Response Structure by Level	48
6.2.1	R0 — General Health Education	48
6.2.2	R1 — Low-Risk Self-Care	49
6.2.3	R2 — Professional Consultation	49
6.2.4	R3 — Urgent Evaluation	49
6.2.5	R4 — Emergency	49
6.3	Prohibited Response Patterns	49
6.4	Output Safety Layer Architecture	50
6.5	Safe Fallback Design	50
6.6	Mental Health Crisis Handling	50

6.7	Poisoning Response Protocol	51
6.8	Traditional Medicine Response Rules	51
6.9	Limitations	51
6.10	Chapter Summary	52
7	Medical Safety Boundary	53
7.1	Introduction	54
7.2	The Boundary Defined	54
7.2.1	Absolutely Prohibited	54
7.2.2	Permitted	54
7.2.3	The Gray Zone	54
7.3	Framing Requirements for Medical Examples	55
7.4	High-Risk Population Safeguards	55
7.5	Uncertainty Communication	56
7.6	Compliance Verification Checklist	56
7.7	Legal and Regulatory Context	57
7.8	Limitations	57
7.9	Chapter Summary	57
III	Traditional Medicine and Pharmacology	59
8	Traditional Medicine Evidence Framework	61
8.1	Introduction	62
8.2	Vietnamese Traditional Medicine Context	62
8.3	Evidence Classification for Traditional Medicine	62
8.4	Medicinal Plant Record Schema	64
8.5	Product Quality Uncertainty	64
8.6	Regulatory Context	65
8.7	Respectful Communication Without Endorsement	65
8.8	Limitations	65
8.9	Chapter Summary	66
9	Medicinal Plants and Drug-Herb Interactions	67
9.1	Introduction	68
9.2	Medicinal Plant Records	68
9.3	Drug-Herb Interaction Records	68
9.4	Toxicity and Contraindication Documentation	69
9.5	Pregnancy and Pediatric Risks	69
9.6	Interaction Severity Classification	70

9.7	Response Rules for Traditional Medicine Queries	70
9.8	Emergency Exclusion Rule	70
9.9	Limitations	70
9.10	Chapter Summary	71
IV	Privacy, Security, and Governance	73
10	Privacy Architecture	75
10.1	Introduction	76
10.2	Health Data Classification	76
10.3	Data Minimization Principles	76
10.4	Data Flow Architecture	77
10.5	Model Provider Data Flows	77
10.6	Logging Restrictions	77
10.7	User Rights	78
10.8	Vulnerable Population Protections	78
10.9	Vietnamese Regulatory Context	78
10.10	Limitations	79
10.11	Chapter Summary	79
11	Security and Threat Model	81
11.1	Introduction	82
11.2	Threat Categories: STRIDE Analysis	82
11.3	Prompt Injection Defense	82
11.4	Retrieval Poisoning Defense	84
11.5	System Prompt Protection	85
11.6	Medical-Safety-Specific Threats	85
11.7	Attack Trees	85
11.8	Incident Response	86
11.9	Limitations	86
11.10	Chapter Summary	86
12	Runtime Governance and Continuous Safety	89
12.1	Introduction	90
12.2	Deployment Maturity Model	90
12.3	Runtime Safety Monitoring	90
12.4	Incident Response Procedures	92
12.5	Continuous Evaluation	92
12.6	Model Version Management	92

12.7 Knowledge Base Governance	93
12.8 Human Oversight Requirements	93
12.9 Clinical Governance Integration	93
12.10 Limitations	94
12.11 Chapter Summary	94
V Evaluation	95
13 Benchmark Design	97
13.1 Introduction	98
13.2 Benchmark Schema	98
13.3 Case Categories and Coverage Requirements	98
13.4 Language Variants	99
13.5 Annotation Guide	100
13.6 Ground-Truth Policy	100
13.7 Subgroup Representation	100
13.8 Quality Gates for Benchmark Cases	100
13.9 Test Contamination Prevention	101
13.10 Benchmark Statistics	101
13.11 Limitations	101
13.12 Chapter Summary	103
14 Evaluation Harness	105
14.1 Introduction	106
14.2 Deterministic Validators	106
14.3 Model-Assisted Evaluation	106
14.4 Metrics Definition	106
14.5 Subgroup Analysis	107
14.6 Aggregate vs. Critical Metric Reporting	108
14.7 Reproduction Methodology	108
14.8 Failure Analysis Workflow	108
14.9 Limitations	109
14.10 Chapter Summary	109
15 Human Expert Review Protocol	111
15.1 Introduction	112
15.2 Reviewer Roles and Qualifications	112
15.3 Review Procedures by Domain	112
15.4 Author Response Process	113

15.5 Simulated vs. Real Review Labeling	114
15.6 Review Documentation Requirements	114
15.7 Escalation from Simulated to Real Review	114
15.8 Limitations	115
15.9 Chapter Summary	115
 VI Case Study and Synthesis	 117
 16 MedAI Case Study	 119
16.1 Introduction	120
16.2 MedAI Position and Evidence Status	120
16.3 Reference Architecture Walkthrough	120
16.4 What MedAI Demonstrates vs. What It Claims	122
16.5 Lessons for Vietnamese Medical AI Development	122
16.6 Gap Analysis: Design Proposal to Deployment	123
16.7 Limitations	123
16.8 Chapter Summary	124
 17 Synthesis and Open Research Agenda	 125
17.1 Integration of All Frameworks	126
17.2 Open Research Questions	126
17.2.1 Language and NLP	126
17.2.2 Safety Architecture	127
17.2.3 Evaluation	127
17.2.4 Traditional Medicine	127
17.2.5 Governance and Deployment	127
17.3 Deployment Readiness Assessment	127
17.4 Path from Benchmark to Clinical Validation	128
17.5 Call for the Vietnamese Medical AI Research Community	128
17.6 Ethical Considerations for Future Work	129
17.7 Final Readiness Classification	129
17.8 Limitations	130
17.9 Chapter Summary	130

Chapter 1

Introduction — Why Low-Resource Medical AI Is Different

Abstract

Medical AI systems are increasingly deployed in health-information contexts, yet speakers of low-resource languages face unique safety risks that translation alone cannot address. This chapter frames the problem: tonal ambiguity, dialect variation, code-switching, missing diacritics, traditional medicine vocabulary, and the absence of language-specific safety benchmarks create failure modes invisible in English-language systems. Using Vietnamese as the primary case study, we establish the book’s central thesis, scope, audience, and limitations, and clarify the position of MedAI as a design proposal requiring validation.

Keywords: Medical AI safety Low-resource languages Vietnamese NLP Health information systems Language equity AI safety architecture

What this chapter gives you

- Why a medical AI system that is safe in English is not automatically safe in Vietnamese — and why translation does not fix this.
- Four concrete ways language itself becomes a safety risk: tone marks, regional vocabulary, informal typing, and traditional medicine.
- What an “evidence-aware health-information assistant” is allowed to do, and what it is forbidden to do.
- Where to start reading, depending on whether you come from medicine, linguistics, engineering, or policy.

In plain terms

Health advice is dangerous to get wrong, and most safety testing for AI has been done in English. This chapter argues that the gap is not a translation problem but a safety problem: Vietnamese writing loses tone marks when people type quickly, the same word means different things in different regions, and herbal remedies are part of everyday care. Each of those changes what a wrong answer costs. No specialist background is assumed — every Vietnamese example is glossed in English.

1.1 The Promise and Peril of Medical AI

Artificial intelligence systems, particularly large language models (LLMs) [Thirunavukarasu et al. \[2023\]](#), have demonstrated remarkable capability in generating human-like text, answering questions, and providing information across diverse domains. Health information is among the most consequential of these domains: a helpful answer can guide a person toward appropriate care, while a harmful answer can delay emergency treatment, provide false reassurance, or suggest unsafe remedies [Jiang et al. \[2017\]](#).

The promise is clear. In communities where access to healthcare professionals is uneven—rural areas, underserved urban populations, regions with few specialists—an AI system that provides understandable, evidence-aware health information could help people make better decisions about when and how to seek care [Chen et al. \[2019\]](#). For speakers of languages with limited digital health resources, such a system could bridge an information gap that translation alone cannot close.

The peril is equally clear. Medical AI systems can hallucinate citations [Huang et al. \[2023\]](#), provide false reassurance for serious symptoms, miss emergency red flags, suggest unsafe dosages, and present traditional remedies as proven treatments. These failures are not merely inconvenient—they can cause serious harm or death. And in low-resource languages, these failures may be harder to detect, harder to evaluate, and harder to fix [Joshi et al. \[2020\]](#).

This book investigates how to build health-information AI systems that are useful, understandable, and safe for speakers of low-resource languages, using Vietnamese as the primary case study.

1.2 Why Translation Is Not Enough

A common assumption in medical AI development is that a system working well in English can be made to work in other languages through translation. This assumption is wrong, and understanding why it is wrong is central to this book.

1.2.1 Linguistic Structure Creates Safety Risks

Vietnamese is a tonal language with six tones distinguished by diacritical marks. The word *ma* can mean ghost (*ma*), mother (*mã*), horse (*mã*), tomb (*mã*), rice seedling (*mạ*), or but

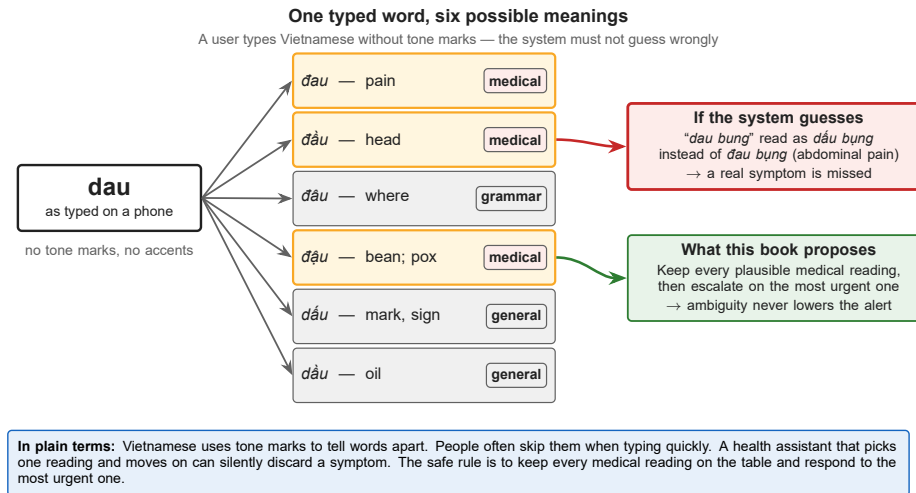


Figure 1.1: One typed word, six possible meanings. When tone marks are omitted — as they routinely are on a phone keyboard — the letters *dau* correspond to several unrelated Vietnamese words, three of which are medically relevant. A system that silently selects one reading can discard a reported symptom. The safety response is to retain every plausible medical reading and act on the most urgent.

(*mà*) depending on the tone mark [Nguyen et al. \[2020\]](#). In medical contexts, missing or incorrect diacritics create genuine ambiguity:

- *dau* without diacritics could mean *đau* (pain), *đầu* (head), *đâu* (where), or *đậu* (bean/pox).
- *mat* could mean *mắt* (eye), *mất* (lost), *mật* (honey/bile), or *mặt* (face).
- *ruou* could mean *rượu* (alcohol) or *rười* (half).

A system that processes Vietnamese health queries must handle these ambiguities safely. An English-language system, even a multilingual one, may not have been evaluated against these specific failure modes.

1.2.2 Regional Variation Is Medically Relevant

Vietnamese has three major dialect regions—Northern (Hà Nội), Central (Huế/Đà Nẵng), and Southern (Sài Gòn)—with vocabulary differences that affect medical comprehension:

- “Vomit” is *nôn* in the North but *ói* in the South.
- “Dizzy” is *chóng mặt* in standard Vietnamese but *xây xẩm* in Southern colloquial speech.

- “Sick/ill” is *ốm* in the North, *đau* in the Central region, and *bệnh* in the South—and *đau* also means “pain,” creating ambiguity.

A keyword-based emergency detection system trained on Northern vocabulary may miss Southern descriptions of emergency symptoms. This is not a usability problem—it is a safety problem.

Safety rule

A red-flag lexicon validated on one dialect region has not been validated for the others. Coverage must be demonstrated separately for Northern, Central, and Southern vocabulary, because a term that is absent from the lexicon is indistinguishable, at runtime, from a symptom the user never reported.

1.2.3 Informal Language Is the Norm

Real health queries are not written in formal, clinical Vietnamese. They are typed quickly on phones, often without diacritics, using abbreviations, slang, and code-switching:

- *e bị sốt 3 hôm r* (I’ve had fever for 3 days)—abbreviated, no diacritics.
- *đau bụng + khó thở, cần đi BV ko?* (abdominal pain + dyspnea, need hospital?)
- *em bị acid reflux, ợ nóng* (code-switched with English medical term).
- *Tào Tháo đuổi* (humorous slang for diarrhea, literally “Cao Cao chases”).

A system evaluated only on formal, well-formed Vietnamese text will overestimate its safety in real-world use.

1.2.4 Traditional Medicine Is Culturally Embedded

Vietnamese culture has a rich tradition of traditional medicine (*thuốc nam, thuốc bắc, đông y*) [World Health Organization \[2014\]](#), [Do \[2006\]](#). Many users will ask about herbal remedies, medicinal plants, and traditional treatments. These queries require respectful acknowledgment of cultural practices, clear distinction between traditional use and clinical evidence, disclosure of known toxicity and interactions, and warnings about product quality variability [Izzo and Ernst \[2009\]](#), [Ernst \[2004\]](#).

1.3 The Evidence-Aware Health-Information Assistant Concept

This book proposes the concept of an *evidence-aware health-information assistant*: an AI system that provides understandable, culturally appropriate health information while maintaining strict safety boundaries. Such a system:

- Does **not** diagnose, prescribe, or provide dosage information.

- **Does** provide evidence-graded health information with source attribution.
- **Does** detect emergencies and escalate immediately.
- **Does** handle traditional medicine queries with cultural respect and safety disclosure.
- **Does** communicate uncertainty honestly.
- **Does** protect user privacy and resist adversarial attacks.

1.4 Book Scope, Audience, and Limitations

1.4.1 Scope

This book covers the safety architecture, evaluation methodology, and governance model for health-information AI in low-resource languages. Vietnamese is the primary case study. The frameworks are designed to be adaptable to other low-resource languages.

1.4.2 Audience

The primary audience comprises medical-AI researchers, AI-safety researchers, clinical NLP researchers, and healthcare software engineers. The secondary audience includes clinicians participating in AI evaluation, public-health researchers, and graduate students. No Vietnamese language knowledge is required; all examples include English glosses.

1.4.3 Limitations

- All frameworks are **design proposals** requiring validation.
- No implementation exists; MedAI is an architectural specification.
- No empirical validation has been conducted.
- All reviews in this book are simulated unless otherwise stated.
- Vietnamese is one case study; generalization requires further research.

1.5 MedAI Position Statement

Warning 1.1 (MedAI is a design proposal) *MedAI, the reference system discussed in this book, is a **design proposal**. No implementation exists. No performance metrics have been measured. No deployment has occurred. MedAI serves as an architectural specification and thought experiment, not as a validated system or product.*

1.6 How to Read This Book

The book is organized in six parts:

1. **Part I (Chapters 1–3):** Foundations—why low-resource medical AI is different, linguistic risks, and evidence handling.
2. **Part II (Chapters 4–7):** Safety architecture—risk classification, emergency detection, response policy, and safety boundary.
3. **Part III (Chapters 8–9):** Traditional medicine—evidence framework and medicinal plant safety.
4. **Part IV (Chapters 10–12):** Privacy, security, and governance.
5. **Part V (Chapters 13–15):** Evaluation—benchmark design, evaluation harness, and human review protocol.
6. **Part VI (Chapters 16–17):** Case study and synthesis.

Chapters can be read sequentially or selectively. Chapter 5 (Emergency Detection) is the most self-contained technical chapter. Chapter 17 (Synthesis) integrates all frameworks.

Research Questions

1. What is the measurable safety gap between English-language and Vietnamese-language medical AI systems?
2. Can a multi-layer safety architecture close this gap without prohibitive over-escalation?
3. What evaluation methodology is appropriate for medical AI safety in low-resource languages?
4. How can traditional medicine queries be handled safely without cultural dismissal?
5. What governance model enables responsible deployment of medical AI in underserved communities?

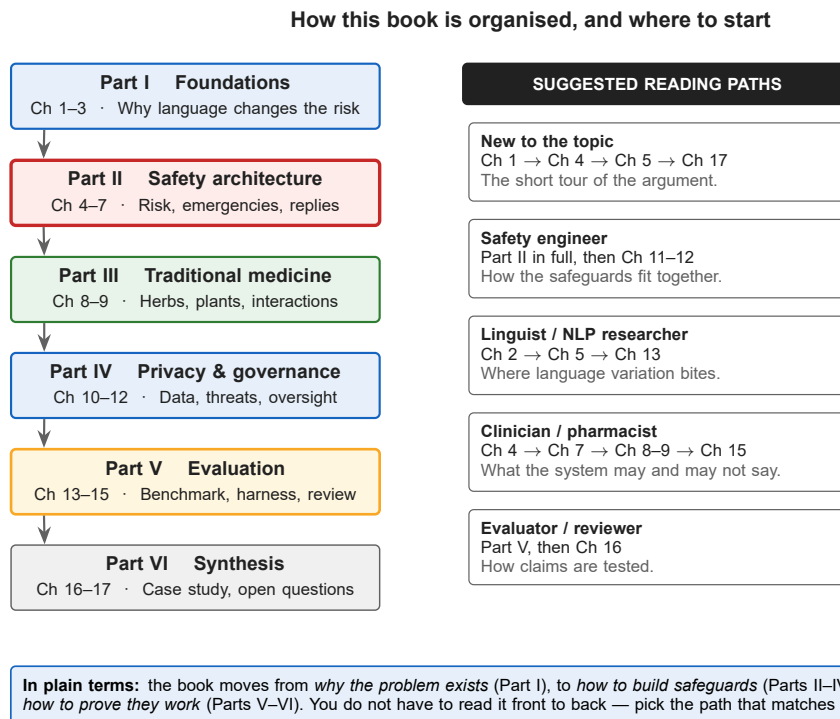


Figure 1.2: How this book is organised, and where to start. The six parts move from why the problem exists (Part I), through how to build safeguards (Parts II–IV), to how to demonstrate that they work (Parts V–VI). The right-hand column gives five reading paths for readers arriving from different disciplines; the book is not intended to be read only front to back.

