MSc RESEARCH PROPOSAL · METHODOLOGY DEFENCE · UNIVERSITY OF MORATUWA

Cultural Alignment of Large Language Models for Sinhala

A Culturally Validated Benchmark, Failure Diagnosis, and Grounded Alignment Study

Janith Ramanayake

Supervisors: Dr. Nisansa de Silva · Dr. Surangika Ranathunga

OVERVIEW

Agenda

01

Introduction

- Language, culture, and LLMs
- What is cultural alignment?
- Why Sinhala, why now
- The research gap

02

Research Objectives

- Problem statement
- Six objectives
- Research questions
- Preregistered hypotheses

03

Proposed Methodology

- Six-phase research arc
- Ontology and benchmark
- Evaluation and diagnosis
- Alignment, generalization, timeline

04

Challenges

- Resources
- Expected contributions

Language as a Lens on Culture

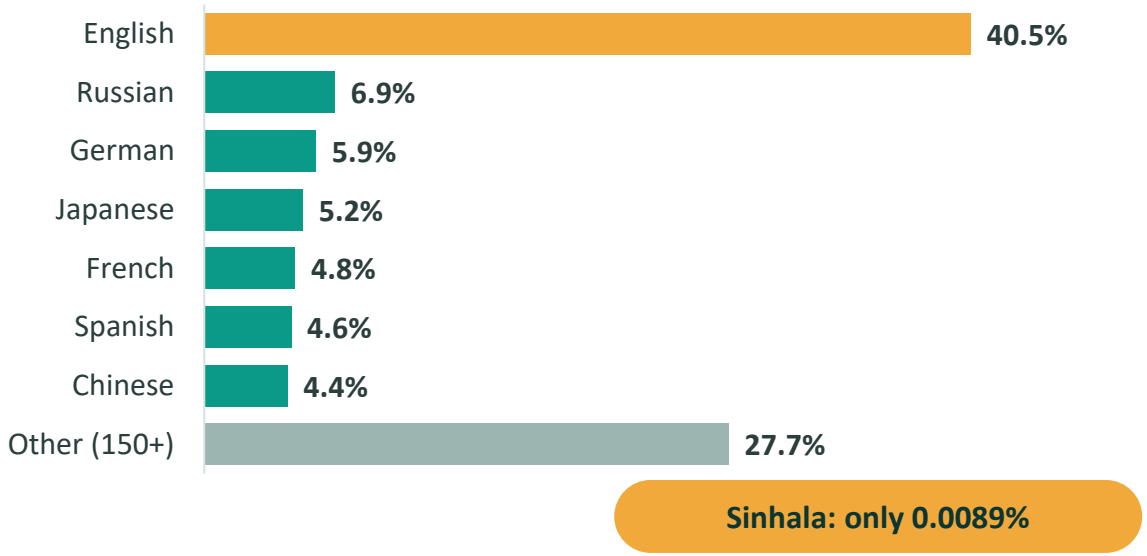
Language is a practical entry point

Culture has no rigid boundaries, but each language maps to a community whose shared usage lets us study culturally situated knowledge and norms [1].

Many LLMs inherit web imbalance

Foundation models such as GPT-3 [2] and Llama [3] train wholly or partly on Common Crawl-derived text, which is heavily skewed to English.

Share of Common Crawl by language [4]



Fluency transfers. Multilingual pretraining gives grammar and syntax across languages.

Knowledge doesn't. Cultural facts, values, and norms are learned only from culture-specific text.

[1] Kramsch (1998). Language and Culture. Oxford Univ. Press. [2] Brown et al. (2020). Language Models are Few-Shot Learners. NeurIPS 2020. [3] Touvron et al. (2023). Llama 2. arXiv:2307.09288. [4] Common Crawl Foundation (2026). CC-MAIN-2026-34 Crawl Statistics: Languages.

What Is Cultural Alignment?

An LLM's ability to **understand and appropriately respond** to culture-specific knowledge, values, norms, traditions, and context - not just to produce grammatically correct text.

Language Adaptation

- Vocabulary
- Grammar and syntax
- Fluency
- Script and pronunciation
- Translation accuracy

≠

Cultural Alignment

- Cultural knowledge
- Values and beliefs
- Social norms
- Traditions and practices
- Contextual appropriateness

Example: a grammatically flawless condolence message that opens with a festive greeting — perfect grammar, wrong register. Such missteps can make a model unusable or offensive in deployment.

The World Has Started. Sinhala Hasn't.

India DRISHTIKON	**Bangladesh** BLUCK	**Korea** CLIcK	**Indonesia** IndoCulture	**Ghana** AMAMMERE
Russia RuBIN	**Romania** RoCulturaMCQ	**Iran** PerCul	**Pakistan** PakBBQ	**Japan** JMMMU
Kazakhstan KazBench-KK	**Mongolia** MonCulture-Eval	**Italy** ITALIC	**Tamil (Sri Lanka)** ILAKKANAM	**Arabic · 22 countries** PalmX · multi-country only

Sinhala

No dedicated cultural-alignment benchmark yet

WHAT EXISTS FOR SINHALA

SinhalaMMLU
Curriculum knowledge

LKValues
Societal values

Global PIQA
Physical commonsense

None measures lived cultural practice, diagnoses how models fail, or tests whether alignment transfers.

Benchmarks shown: [5]–[19]. Sinhala resources: [20] SinhalaMMLU, [21] LKValues, [22] Global PIQA.

Problem Statement and Objectives

PROBLEM STATEMENT

- No existing resource measures whether LLMs understand Sinhala culture as a multi-dimensional, lived construct.
- Cultural benchmarks in other languages mostly test facts with single-string gold answers — they cannot represent context-dependent practice and risk encoding stereotypes.
- Nothing diagnoses how models fail culturally, or whether alignment transfers beyond what it was trained on.

O1 Develop a taxonomy of Sinhala cultural alignment

Define

O2 Create a culturally validated benchmark

Build

O3 Evaluate existing LLMs across the cultural dimensions

Evaluate

O4 Develop a failure taxonomy for cultural misalignment

Diagnose

O5 Investigate grounded alignment methods

Align

O6 Measure generalization to unseen cultural contexts

Generalize

Research Questions and Hypotheses

RESEARCH QUESTIONS

RQ1 — MEASUREMENT
How well do LLMs understand and appropriately respond to Sinhala cultural contexts, per dimension, versus a native human reference?

RQ2 — STRUCTURE
Are the six dimensions empirically distinguishable — do models that know the facts still fail on meaning, reasoning, and nuance?

RQ3 — MISALIGNMENT
What failure types characterize cultural misalignment, and how do they vary by model and dimension?

RQ4 — ALIGNMENT
Do grounded methods (prompting, RAG, instruction tuning) improve alignment without degrading Sinhala quality?

RQ5 — GENERALIZATION
Do improvements transfer to unseen cultural practices never present in alignment data?

DIRECTIONAL HYPOTHESES · PREREGISTERED ON OSF

H1 — All models fall below the human reference, and the gap grows from knowledge toward reasoning and nuance.

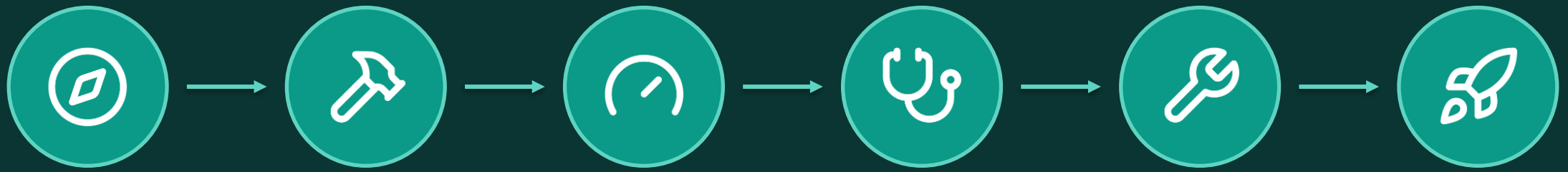
H2 — Models systematically overgeneralize and default to other-culture answers at measurable rates.

H3 — Grounded interventions improve alignment, with retrieval reducing hallucination most.

H4 — Alignment gains partially transfer to unseen practice clusters, with decay.

Six Phases, One Pipeline

Define → Build → Evaluate → Diagnose → Align → Generalize — each phase's output is the next phase's input.



PHASE 1

Define

Cultural ontology and construct

O1

PHASE 2

Build

SinCulBench: ≈3,000 validated items

O2

PHASE 3

Evaluate

10–12 LLMs vs a native human reference

O3

PHASE 4

Diagnose

Seven-class failure taxonomy

O4

PHASE 5

Align

Prompting · retrieval · instruction tuning

O5

PHASE 6

Generalize

Seen vs unseen practice clusters

O6

STAGE A · MONTHS 1–6 · PUBLISHABLE ON ITS OWN

STAGE B · MONTHS 7–9

Phase 1: Building the Cultural Ontology

Phase 1 defines what counts as Sri Lankan culture before any question is written. Its output is a map of 12 domains and 150–200 practices, the frame every later question is sampled from.

Why start with an ontology?

Free question-writing gives random, uneven coverage. Map the culture first, then sample from the map — nothing is left to chance.

EVERY PRACTICE IS SCOPE-TAGGED

- National ↔ community-specific
- Urban ↔ rural
- Traditional ↔ contemporary

Variation is built in, not discovered later.

- Festivals
- Rituals
- Religious Practice
- Food
- Life-Cycle
- Family & Kinship
- Games & Music
- Livelihood
- Health & Astrology
- Everyday Norms
- Significant Places
- Material Culture

Ministry of Cultural Affairs catalogue
The authoritative source list

→

Classification rules into 12 domains
Inclusion and exclusion rules per domain

→

Scope tags added
Variation recorded up front

→

150–200 scoped practices
The sampling frame for authors

Phase 2: Not One Task, but Six

Authors take each topic from the Phase 1 map and write questions for it using these six tasks. Any number of questions can be written per topic; each task tests a different kind of cultural understanding.

	TASK	FORMAT	WHAT IT TESTS	EXAMPLE QUESTION → ANSWER
T1	**Knowledge**	4-option MCQ	Does the model know the basic fact?	"What is Ankeliya?" → **pick 1 of 4 options**
T2	**Meaning & Explanation**	4-option MCQ	Does it understand why the practice exists?	"What does oil anointing at Avurudu symbolise?" → **pick the meaning, 1 of 4**
T3	**Practice Understanding**	Open response	Can it describe how the practice actually happens?	"Describe, in order, what happens during Ankeliya." → **written answer, scored on key points**
T4	**Scenario Reasoning**	Open response	Can it reason through a real situation?	"A guest at a thovil uses flash. What should they know?" → **written answer, scored on key points**
T5	**Appropriateness**	Binary + justification	Can it judge appropriate from inappropriate?	"A visitor enters a shrine room wearing shoes." → **Appropriate / Inappropriate, and why**
T6	**Variation & Misconception**	Binary + correction	Can it spot a false generalisation?	"This ritual is identical in every community." → **Accurate / Inaccurate, plus a correction**

Phase 2: Building the Benchmark



Topic

Picked by the author from the classified practice inventory

Source content

Ministry of Cultural Affairs catalogue plus academic, ethnographic, and institutional sources. LLMs may assist drafting, never ground truth.

Question generation

Questions written for each task (T1–T6), each with a ground-truth record and controlled distractors

Quality control

Non-author review, double-validation, and automated checks

SinCulBench

Validated items with development, public, and private splits, plus a datasheet

4–6
authors to recruit

~60
questions per author, per week

≈3,000
questions in SinCulBench

Development + public test | **~50% private held-out · never exposed to alignment**

Phase 3: Evaluating Models Against a Human Baseline

Phase 3 runs 10 – 12 models on the benchmark under one fixed protocol, set before any run, and compares every score with a native human baseline, so no result can change how models are prompted or scored.

Models · 10–12

Frontier proprietary

GPT · Claude · Gemini

Open-weight

Qwen · Llama · Gemma, several scales

Sinhala-adapted

SinLlama

Exact versions recorded; scale compared within a family.

Prompting conditions

Primary

Sinhala zero-shot, temperature 0

Secondary

Sinhala 5-shot, native exemplars

Control

English instructions

MCQs scored over 3 option permutations.

Scoring

Closed-form

Accuracy, averaged over permutations

Open-ended

Key-point rubric + two trained native raters

Refusals and unparseable answers count as incorrect and are reported separately.

Human reference: benchmark authors (excluding their own items) and volunteers answer ~300 stratified items under the same instructions — a reference, not a ceiling.

Phase 3: Developing the OCAS Scoring Model

The six tasks test culture at different depths and are scored differently, so no single task gives an overall picture. We develop OCAS to combine accuracy and agreement-adjusted rubric scores into one standardised measure for comparing models.

Accuracy-based

Tasks with predefined answers: T1, T2 and the T5 and T6 labels. Score = % answered correctly.

Agreement-based

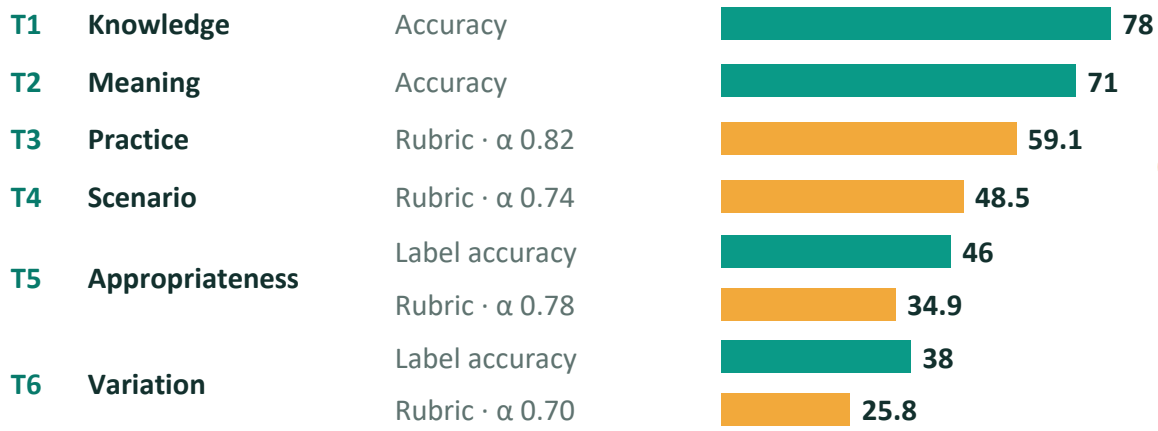
Rubric tasks: T3, T4 and the T5 and T6 justifications. Score = raters' mean, adjusted by their agreement (Krippendorff's α).

Combine

OCAS = weighted mean of the six task scores.

Worked example: Model X

Illustrative — no results



OCAS ≈ 50.6

Phase 4: Seven Ways Models Fail Culturally

Phase 4 moves from how much models fail to how they fail. We sort their failures into seven cultural-misalignment classes, plus operational failures, so results show what goes wrong and not just who scores lower.

F1 Cultural hallucination

Invents a practice, detail, or etymology

F2 Over-generalization

Absolute claim where practice varies

F3 Westernization

Generic or other-culture default answer

F4 Cultural omission

Ignores the cultural dimension at issue

F5 Context collapse

One context's practice treated as universal

F6 Cultural conflation

Merges practices of different communities

F7 Incorrect reasoning

Knows the practice, misreads its meaning

F8 Operational

Refusals and unparseable outputs, reported separately

Model outputs → **Failure detection** → **Three independent coders** → **Agreement reported (α)**

After classification: we count each failure type per model and dimension, draw the taxonomy map as a main results figure, examine the largest classes in detail, and use them to guide the Phase 5 interventions.

Phase 5: Three Grounded Interventions

Phase 5 tests whether grounding models in Sri Lankan cultural knowledge improves their alignment. Three interventions of increasing intensity run under the same protocol as the baseline, so the comparison stays fair.

Expected ordering M1 < M2 < M3 — a preregistered hypothesis [23]

M1

Cultural prompting

System-level instructions ground answers in Sri Lankan context and encourage careful handling of variation and uncertainty.

M2

Cultural RAG

Retrieve validated passages from SinCulKB through a bilingual index; settings tuned on the development set only.

M3

Instruction tuning

LoRA on 2–3 open-weight families with ~2,500–3,000 native-filtered examples, including nuance-positive ones.

HOW THE INSTRUCTION DATA IS CREATED

- LLM drafts from SinCulKB and dev/public items
- Native annotators filter and edit every one
- Nuance-positive examples included
- Deduplicated; no private-split content

SinCulKB (retrieval corpus for M2, grounding source for M3) — built as a by-product of benchmark construction: source passages are curated per practice (Sinhala + English), each with its source, scope tags and an annotator-validated summary. ~400–600 passages; contains no benchmark item text.

[23] Xu et al. (2025). Self-Pluralising Culture Alignment (CultureSPA). NAACL 2025.

Phase 6: Competence or Memorization?

Phase 6 asks whether gains from alignment are real cultural competence or memorization of the benchmark. We compare the gain each method gives on practices it saw with its gain on practices held out from all training and retrieval.

Seen clusters

Festivals & Rituals · Family, Kinship & Hierarchy · Food & Livelihood. M1–M3 may train and tune here.

Δ

Gain = aligned score − baseline score, on each side

Unseen — held out

Everyday Norms & Places: 2–3 ontology sub-branches, ~15–20 practices. Excluded from instruction data, the retrieval index, and few-shot demonstrations.

Unseen gain ≈ 0 → possible memorization: the method only helped on what it saw

Unseen gain positive but smaller than seen gain → partial transfer: evidence of generalization (H4)

Beyond the Core Plan: Three Ideas We Could Try

Three extensions could sharpen the study. None is needed for the core plan: each would be an exploratory, preregistered addition if time allows, or future work.

Search agent

THE IDEA

A model that chooses what to search, rewrites queries in Sinhala and English, and follows up when results are weak.

WHERE IT FITS

Phase 5, extending M2 (retrieval)

Reasoning

THE IDEA

Compare a model with thinking on and off, and study where its reasoning goes wrong.

WHERE IT FITS

Phase 3 as a condition; Phase 4 as trace analysis

Multiple agents

THE IDEA

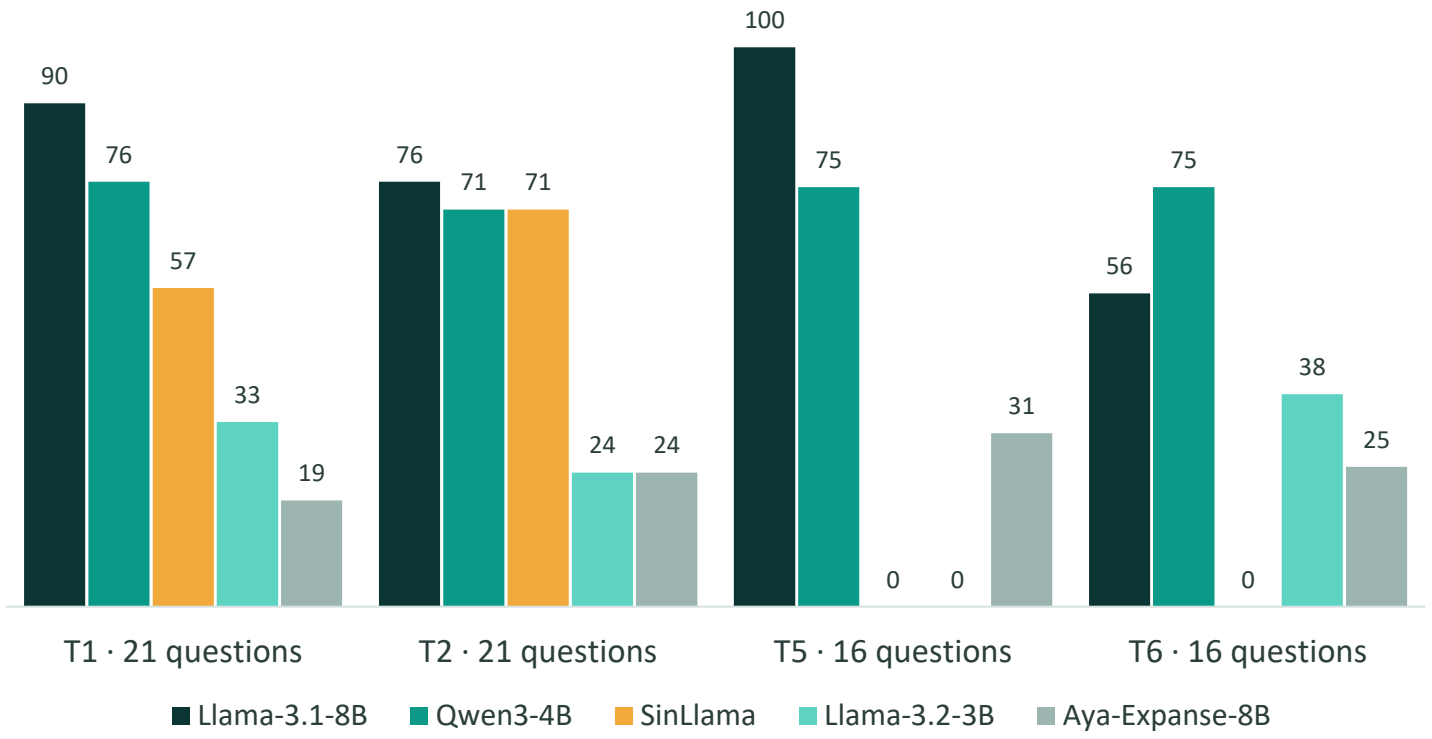
A proposer, a retrieval agent and a critic, each aimed at failure classes such as overgeneralization.

WHERE IT FITS

Phase 5, built on the search agent

The Pipeline Already Runs

Pilot accuracy by task (%), five open-weight models



WHAT THE PILOT SHOWS

Closed tasks are answerable

Llama-3.1-8B and Qwen3-4B score 71–100% on T1, T2 and T5.

Open-ended tasks break down

Smaller models repeat the question, give very short answers, or loop into broken Sinhala.

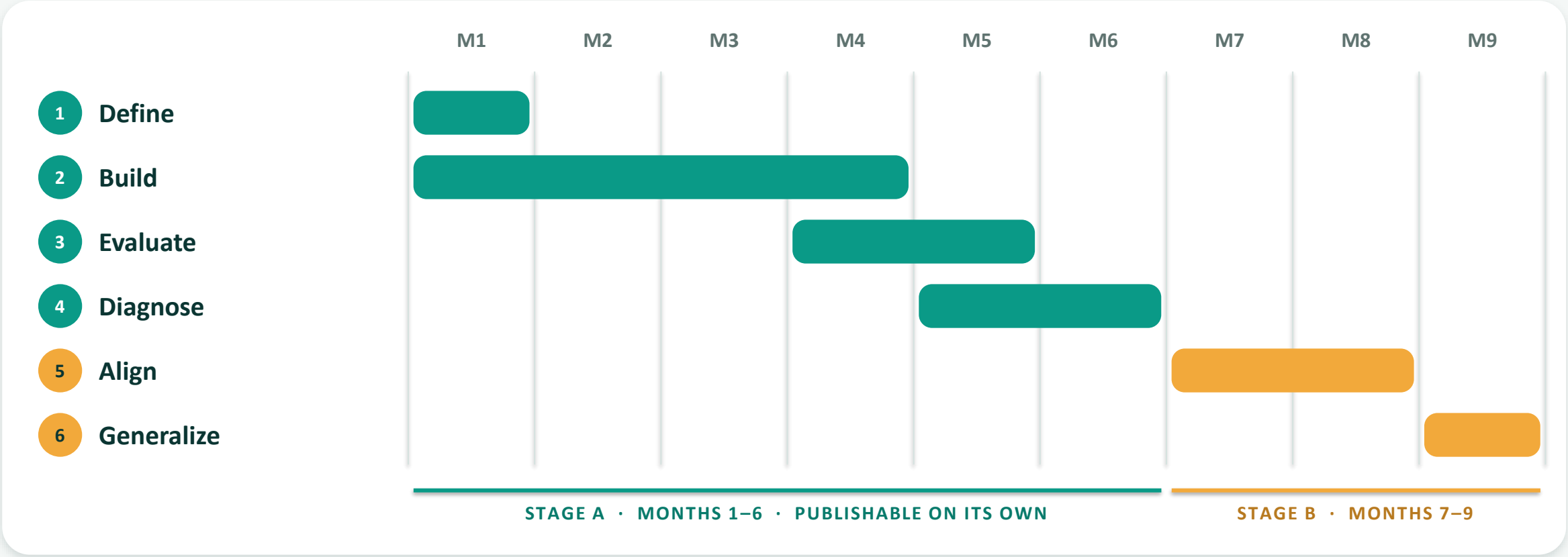
Possible causes, still to test

Weak Sinhala generation, limited instruction-following, little Sinhala data, model size, tokenization or decoding.

What it means for the method

Fix decoding in advance and score degenerate output separately. SinLlama's 0% on T5/T6 is a 12–19% parse rate (F8/F9).

Delivery Timeline: Six Phases, Nine Months



No fixed deadline · target: *CL journal track via ACL Rolling Review

Validation gates are never compressed to hit a date

Resources

Models

- Frontier proprietary: GPT, Claude, Gemini
- Open-weight at several scales: Qwen, Llama, Gemma
- Sinhala-adapted: SinLlama
- Via HuggingFace and API access

Data

- SinCulBench — the benchmark itself
- SinCulKB — a by-product of benchmark validation
- Ministry of Cultural Affairs catalogue as the source frame

Compute and people

- Rented GPU compute for inference and LoRA
- 4–6 authors, being recruited
- Volunteer human-reference raters
- Crowdsourcing apps SCB and MTQE are live

Compute and model access are secured; the author team is the one resource still being recruited.

Five Expected Contributions

01

SinCulBench

First multi-dimensional cultural-alignment benchmark for Sinhala: ≈3,000 items, pluralistic ground truth, datasheet

***CL-venue paper**

Stage A findings: benchmark, evaluation, diagnosis

02

Methodology

A reusable construction and validation method for low-resource cultures

03

Measurement study

10–12 LLMs against a native human reference, with the dimension ladder

04

Failure taxonomy

Seven-class cultural-misalignment taxonomy, double-coded and agreement-reported

05

SinCulKB + alignment

Knowledge base and grounded alignment with a seen-vs-unseen generalization test

***CL-venue paper**

Stage A findings: benchmark, evaluation, diagnosis

Stage B findings: alignment, generalization

SOURCES

References

[1] Kramsch (1998). Language and Culture. Oxford Univ. Press.

[2] Brown et al. (2020). Language Models are Few-Shot Learners. NeurIPS 2020.

[3] Touvron et al. (2023). Llama 2. arXiv:2307.09288.

[4] Common Crawl Foundation (2026). CC-MAIN-2026-34 Crawl Statistics: Languages.

[5] Maji et al. (2025). DRISHTIKON. EMNLP 2025.

[6] Kabir et al. (2025). BLUCK. BLP Workshop 2025.

[7] Kim et al. (2024). CLIcK. LREC-COLING 2024.

[8] Koto et al. (2024). IndoCulture. TACL 2024.

[9] Acquaye et al. (2024). Susu Box or Piggy Bank (AMAMMERE). EMNLP 2024.

[10] Lazukova & Piontkovskaya (2026). RuBIN. LREC 2026.

[11] Iorga et al. (2026). RoCulturaMCQ. Electronics 2026.

[12] Moosavi Monazzah et al. (2025). PerCul. NAACL 2025.

[13] Hashmat et al. (2025). PakBBQ. EMNLP 2025.

[14] Onohara et al. (2025). JMMMU. NAACL 2025.

[15] Umbet et al. (2025). KazBench-KK. Field Matters Workshop 2025.

[16] Minggad et al. (2026). MonCulture-Eval. Findings of ACL 2026.

[17] Seveso et al. (2025). ITALIC. NAACL 2025.

[18] ILAKKANAM (2026). CHIPSAL 2026.

[19] Alwajih et al. (2025). PalmX 2025 Shared Task. ArabicNLP 2025.

[20] Pramodya et al. (2025). SinhalaMMLU. EMNLP 2025.

[21] Muthugala et al. (2026). LKValues. arXiv:2607.20410.

[22] Chang et al. (2025). Global PIQA. MRL Workshop, arXiv:2510.24081.

[23] Xu et al. (2025). Self-Pluralising Culture Alignment (CultureSPA). NAACL 2025.

Thank You

Presented by Janith Ramanayake

Supervised by Dr. Nisansa de Silva and Dr. Surangika Ranathunga

I welcome the panel's questions.