MULTI-AGENT SYSTEM FOR ACCESSIBLE GOVERNMENT DATA RETRIEVAL IN SRI LANKA'S MULTILINGUAL LANDSCAPE

25-26J-093

Project Proposal Report

Charunya Thathsaranie Dissanayake

B.Sc. (Hons) Degree in Information Technology Specializing in Data Science

Department of Computer Science

Sri Lanka Institute of Information Technology
Sri Lanka

August 2025

MULTI-AGENT SYSTEM FOR ACCESSIBLE GOVERNMENT DATA RETRIEVAL IN SRI LANKA'S MULTILINGUAL LANDSCAPE

25-26J-093

Project Proposal Report

B.Sc. (Hons) Degree in Information Technology Specializing in Data Science

Department of Computer Science

Sri Lanka Institute of Information Technology
Sri Lanka

August 2025

ABSTRACT

Sri Lanka aspires to a truly inclusive digital government, yet the lack of a balanced, up-to-date trilingual public corpus—especially one that integrates Sinhala, Tamil, English, and their Romanized variants—remains a core limitation for AI-powered information retrieval. This project tackles that gap by creating, validating, and documenting a robust corpus in the health domain that supports (a) government domain annotation, (b) Singlish/Tamilish mapping, (c) bias and coverage analysis, and (d) sustainable, community-guided updates. Leveraging cutting-edge NLP methods and human-in-the-loop oversight, this resource is designed to support equitable citizen access, trustworthy model training, and scalable service rollouts across Sri Lankan digital governance.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

TABLE OF ABBREVIATIONS

Abbrev.	**Description**
AI	Artificial Intelligence
NLP	Natural Language Processing
SLM	Small Language Model
LID	Language Identification

Abbrev.	Description
API	Application Programming Interface
EGDI	E-Government Development Index

DECLARATION

I declare that this is my own work and this proposal does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any other university or Institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Group Member Name	Student ID	Signature
D.M.C.T.Dissanayake	IT22919700	

The above candidates are carrying out research for the undergraduate Dissertation under my supervision.

……. ……..

Signature of the supervisor:

(Dr. Lakmini Abeywardhana) Date:

1. INTRODUCTION

Sri Lanka, a nation of over 22 million, stands at the center of South Asia's linguistic and digital crossroads. Its constitutional commitment to trilingualism—Sinhala, Tamil, and English—has been both an opportunity and a challenge for digital public service. Though government policies have long mandated the availability of public documents and web portals in all three languages, the technology and underlying data resources have struggled to keep pace with actual citizen demand, especially as internet penetration and mobile-first usage have exploded in recent years [1], [2].

More than half of the population now accesses information primarily through smartphones, with social media platforms and messaging apps becoming de facto channels for news, government updates, and civic service queries [1]. In this landscape, language is not merely a medium of instruction or publication; it is the enabling bridge—or conversely, the invisible barrier—to national inclusion, economic opportunity, and transparent democratic engagement [2].

Current estimates suggest that Sri Lanka's digital literacy is rising, yet gaps remain, particularly among seniors, rural communities, and those less fluent in English [1]. This digital divide has linguistic dimensions: not only do many government sites continue to default to English for navigation or advanced queries, but the most natural and widely used forms of user input diverge dramatically from textbook Sinhala and Tamil [2]. Romanized scripts ("Singlish" and "Tamilish")—once considered internet slang or peripheral—have entered the mainstream, are used by millions for texting, online forms, and even e-government chatbot interfaces [3].

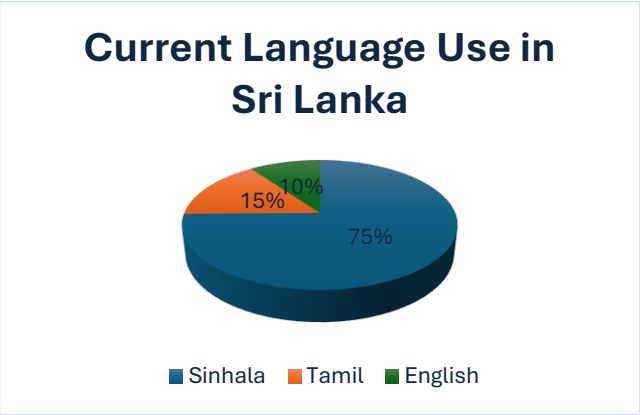

Figure 1.1: Planned language coverage for the health corpus, derived from population distribution [1], [2] and adjusted to capture Romanized digital usage [3], [7].

Studies of mobile search logs and IVR query submissions show that these forms are the default for a significant portion of citizens under 40, especially when using transliteration keyboard tools or when under time pressure [3]. The result? Most state-of-the-art government NLP systems remain out of reach for everyday people, especially outside Colombo, because these Romanized utterances are either not recognized or are processed with very low fidelity [3], [4].

Compounding the challenge is the fragmentation of data sources and policy implementation. Ministries, departments, and local authorities each operate their own websites, sometimes with incomplete trilingual coverage and radically different standards for metadata, document formatting, and data openness [2], [5]. The government's ambitious right-to-information framework and participation in the Open Government Partnership have created new mandates for data transparency, yet the lack of a unified, high-quality annotated corpus hinders large-scale automation, digital assistant deployment, evidence-based policy evaluation, and language tech entrepreneurship [6].

The international experience in low- and middle-income countries is instructive. Where trilingual, regularly maintained, and domain-tagged corpora exist, both the efficiency of digital service provision and citizen engagement measures have measurably improved [7], [8]. Moreover, the rapid progress of AI—particularly in natural language processing for

low-resource settings—underscores that access to high-quality, balanced, and annotated corpora is the essential foundation for further innovation, from smart helpdesks and multi-lingual IVRs to bias-aware sentiment analysis and AI-powered public service auditing [9]. A robust, living corpus for Sri Lanka, then, must do more than mirror what exists in the English-speaking West. It should account for the country's unique patterns of code-mixing, informal usage, demographic subgroups, and evolving policy landscape. It must also be structured for updatability, transparency, and accessible integration with both government and civil society technologies. Backed by leading research from Dr. Lakmini Abeywardhana and the SLIIT computing research cluster, this proposal puts forward a national corpus pipeline built for flexibility, inclusivity, and sustainability [4].

2. LITERATURE REVIEW

2.1 The Evolution of Multilingual Public Corpora

As globalization intersects with national and regional policy, the demand for high-quality multilingual datasets has surged. In Sri Lanka, early government digitization attempts relied heavily on manual translation teams and outsourced linguistic support, often producing static, English-first web portals [1], [2]. Modern corpus initiatives have instead favored scalable, semi-automated data pipelines that blend manual curation with supervised and unsupervised NLP-driven annotation. International efforts such as the **FLoRes**, **SinMin**, and **Swa-Bhasha** projects have established templates for generating and updating parallel text in Sinhala, Tamil, and English [3], [10]. These benchmarks illuminate best practices such as cross-lingual sentence alignment, layered metadata, and code-mixed sample collection, but also highlight ongoing deficiencies—particularly in government domain annotation and coverage of Romanized input [3].

In the Government of India's **Digital India** program, structured multilingual corpus development, open sourcing, and frequent dataset refresh cycles are directly tied to both the emergence of public AI solutions and improvements in citizen digital satisfaction indices [11]. Comparative studies show that countries which deploy annotated corpora as foundational "infrastructure" (akin to roads or electricity) rapidly outpace those that treat language resources as afterthoughts, especially in low- and middle-income contexts. For Sri Lanka, where linguistic, ethnic, and digital divides often co-exist, a corpus designed for continuous expansion and rigorous human-in-the-loop validation holds the promise of reducing geographic, age, and social sector disparities in access to information and e-services [7], [8].

2.2 Recent Advances in Corpus Augmentation and Data Annotation

Leading edge work by Dr. Abeywardhane has showcased how saliency-based token selection and domain-agnostic augmentation can substantially boost NLP performance on low-resource language tasks—especially classification and intent prediction [4]. Their experiments with **DistilBERT** and dynamic SHAP-based value thresholding provide a

scalable framework for automatically discerning key tokens, ensuring that noise and data paucity do not overwhelm real signal even in very small or specialized datasets. These techniques are now influencing corpus design standards regionally, moving the field beyond brute-force crawling and basic labeling to real-time, contextually relevant updates and persistent community review [4].

Sumanathilaka and colleagues have demonstrated the efficiency and accuracy of neural and rule-based transliteration engines, with **Swa-Bhasha** providing tools that are already being deployed in conversational government bots and multilingual text-based helplines [3]. The granularity of their error correction testing, combined with field trials in code-mixed settings, means that Romanized corpus coverage can finally keep pace with the realities of Sri Lankan user behavior [3]. In tandem, the **Open Government Partnership's** new Global Report emphasizes that the next-generation public corpora must extend beyond token-level translation to support annotation for domain, source, confidence, and even bias metrics [6].

2.3 Code-Mixing, Demographics, and Annotation Bias

The challenge of demographic and regional bias—long recognized in corpus studies—has moved front and center for Sri Lankan researchers and policymakers. Not only are older, rural citizens less likely to benefit from digital policy shifts if corpora remain urban-centered and formalist, but groups such as Muslims, estate Tamils, and recent returnees often use mixed, hybrid language patterns interpolated with English, Arabic, or other heritage scripts [7], [8]. Existing state and independent datasets generally under-sample these populations, undermining both the precision of intent-based services (e.g., welfare, disaster alerts) and the generalizability of AI models trained on them [8].

As the **UN** and **Verité Research** note in recent digital governance reports, periodic corpus audits, bias mapping, and proactive stakeholder engagement (i.e., feedback, correction, and dispute mechanisms open to all user groups) now mark the international gold standard [1], [2]. These requirements are reflected in the corpus creation strategy presented here—recruitment of diverse annotators, multi-layer validation, and open dashboards for bias and source tracking are core components, not afterthoughts.

2.4 SLIIT's Leading Role and the Road Ahead

With its substantial contributions to corpus construction, community review, and open-source language resources within and beyond Sri Lanka, the Department of Computing at SLIIT—guided by Dr. Lakmini Abeywardhana—has become an acknowledged regional leader [4]. Their cycle of academic research, government partnership, and digital tool roll-out has kept pace with international advances in low-resource AI, ensuring that corpus development remains linked to real policy and technology deployment milestones [9].

Despite this, the gap between Sri Lankan promise and on-the-ground delivery persists. The present project therefore offers not only a technical framework but an institutional vision: a living corpus pipeline where government, academia, and the citizenry work in tandem to continually narrow the divide between official aspiration and practical inclusion.

3. RESEARCH GAP AND PROBLEM STATEMENT

Though Sri Lanka possesses a trilingual policy framework and increasing digital infrastructure, the gulf between linguistic policy and citizen experience is substantial. A closer inspection of government data offerings reveals a landscape where web portals regularly present incomplete translations, inconsistent document formatting, and poorly maintained metadata for Sinhala, Tamil, and English [1], [2], [5]. Often, interactive features and search capabilities work only in English or textbook Sinhala/Tamil, leaving the majority of real-world citizen queries—especially those made in Romanized or blended forms—either misunderstood or unanswered [3].

The limitations become especially salient at the interface between citizens and automated systems, such as government chatbots or IVR platforms. When analyzing transcripts from these services, it is clear that the language variety used by citizens diverges from both official corpora and standardized government communications. Urban youth may employ a hybrid of Sinhala-English ("Singlish"), rural Tamil speakers often inject dialectal nuances, and even government staff respond with ad-hoc abbreviations [3], [8]. Recent

government surveys show that over 30% of digital helpdesk queries are either code-mixed or employ non-standard spelling, which most AI systems cannot comprehend [1].

Meanwhile, the lack of a shared, richly annotated corpus for training and benchmarking these applications not only affects service delivery but also stifles further research and innovation. Without access to representative, up-to-date datasets, start-ups, policy researchers, and corporate partners must assemble fragmented solutions from scratch—wasting resources and reinforcing urban–rural and majority–minority divides [7], [8]. The problem is compounded when corpora are static or one-off: they quickly become outdated in a fast-evolving digital environment, making periodic audit and continual updates essential for real-world system accuracy [6].

What further complicates matters is that much of the existing linguistic data was collected for translation-only purposes, neglecting intent recognition, domain specificity (e.g., legal, welfare, health), entity annotation, or error mapping [10]. Such limitations cripple the next generation of government chatbots, document retrieval agents, and digital intermediaries, which must operate with high precision and nuance [9]. Academic pilots—such as those led by Dr. Abeywardhana and SLIIT—have proven that targeted, bias-aware annotation (even on a moderate scale) can measurably improve model performance [4], but scaling this to national coverage remains a daunting, unsolved challenge.

In summary, Sri Lanka's challenge is multidimensional: to create a living corpus that is representative across languages, scripts, geography, generation, and service area—a dataset that is continually updated, community-validated, and openly accessible, forming the backbone of Digital Sri Lanka's equitable future.

4. OBJECTIVES

Main Objective

To establish a sustainable pipeline for building, validating, and updating a large-scale, trilingual, bias-analyzed, domain-annotated **health corpus** for Sri Lankan e-government, integrating Sinhala, Tamil, and English (including Romanized forms and code-mixed language), and ensuring its adoption by AI/NLP researchers, health-sector ICT teams, and public service partners.

Sub-objectives

1. Comprehensive Health Data Collection:

Scrape structured and unstructured data from key national health sites, including:

- Colombo South Teaching Hospital (clinic schedules, circulars in image/PDF formats),
- National Hospital Kandy (annual health bulletins, procurement notices),
- National Institute of Mental Health (statistics, maps, clinic times),
- Epidemiology Unit (disease reports, circulars, PDFs),
- Health Promotion Bureau, National Dengue Control Unit,
- Ministry of Indigenous Medicine, and others.

Supplement this with "in the wild" health-related data: anonymized chatbot logs, SMS queries (with privacy), and hotline transcripts from public helpdesks.

2. Hybrid Annotation and Community Validation:

Define domain-specific ontologies (e.g., disease, epidemiology, hospital services) and combine expert curation (medical translators, health informaticians) with supervised community annotation across universities, NGOs, and civil society.

3. Question-Answer Data Curation and Annotation:

Curate and annotate the health corpus with high-quality, structured question-answer pairs relevant to both government and citizen health queries, mapping each pair to specific user intents and health topics while ensuring comprehensive trilingual and

code-mixed coverage to support effective benchmarking and training of conversational AI systems for the Sri Lankan health sector.

4. Romanized Variant Normalization and Enrichment:

Integrate the latest Swa-Bhasha transliteration models [3] to capture Romanized Sinhala/Tamil queries (e.g., “dengue clinic near me”) and expand to underrepresented digital health text sources such as TikTok comments, memes, and WhatsApp forwards.

5. Bias Monitoring and Dynamic Updating:

Audit each health corpus release for demographic and regional skew (e.g., rural clinics, upcountry Tamils, women’s health needs) and run corrective annotation sprints and publish open bias audit dashboards for transparency.

6. Accessible Distribution and Capacity Building:

Provide API access and user dashboards showing dataset freshness, coverage across health domains, and error reports and develop training modules for health-sector IT staff and students, ensuring knowledge transfer and long-term adoption.

5. METHODOLOGY

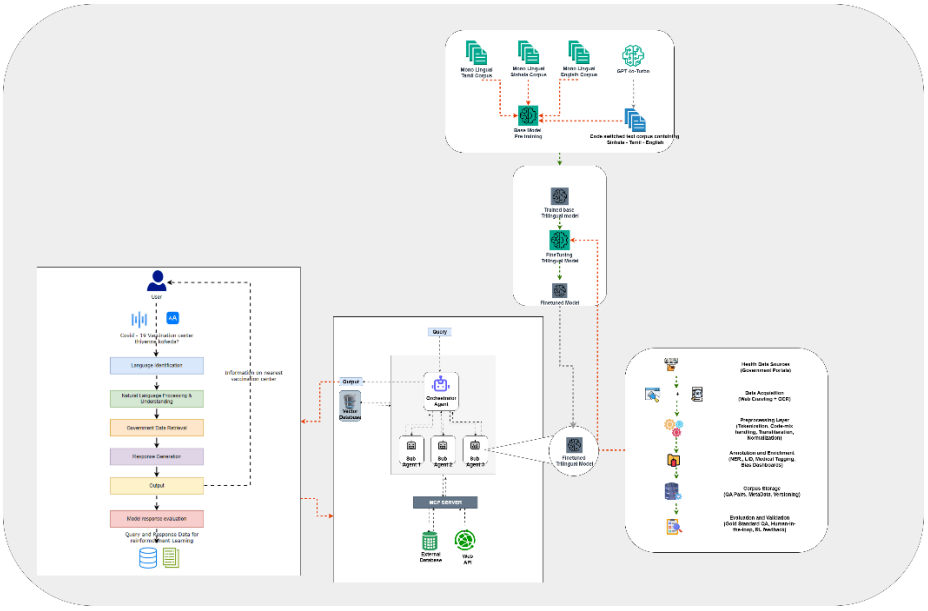

Figure 5.1: System overview Diagram of the project

To ensure inclusivity, the corpus design adopts a balanced allocation across Sinhala, Tamil, English, and Romanized forms. This planned coverage accounts for both official population distributions and the widespread use of Romanized inputs in digital health queries.

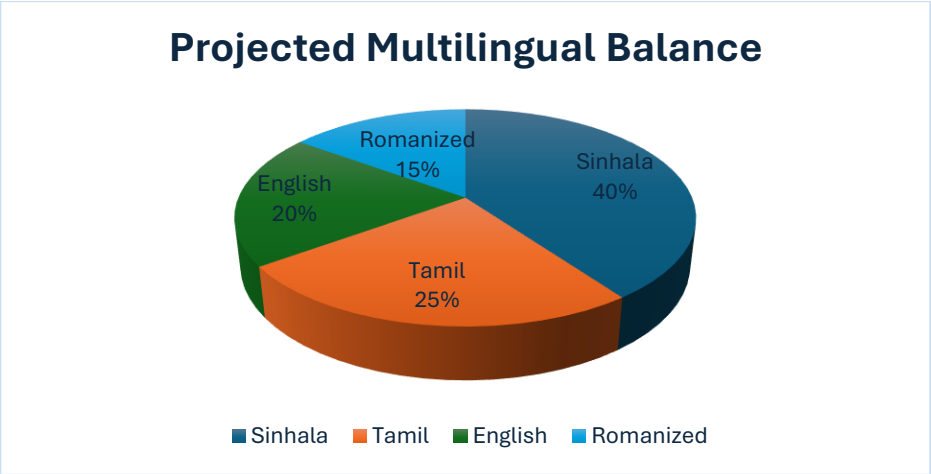

Figure 5.2: Planned language coverage for the health corpus, derived from population distribution [1], [2] and adjusted for widespread Romanized usage [3], [7].

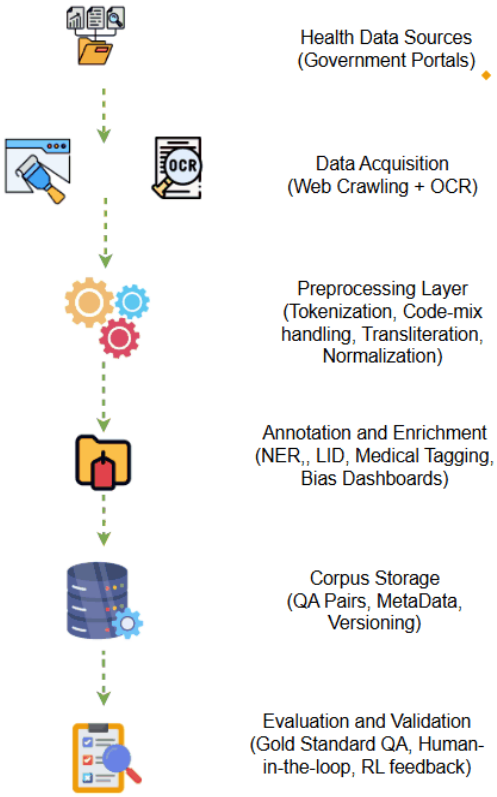

Figure 5.3: Health Corpus Data Pipeline — End-to-end workflow showing data sources (hospitals, health agencies, helpdesks), multilingual crawling and ingestion, expert and community annotation, iterative bias auditing, and final corpus outputs

5.1. Data Collection and Acquisition

- A comprehensive audit of health-sector government resources (websites, APIs, PDF repositories, hotline datasets) will be conducted to identify relevant data sources in Sinhala, Tamil, English, and Romanized/code-mixed formats.
- Advanced multi-script web crawlers and OCR tools will be deployed to automatically extract structured and unstructured health data from web portals as well as scanned documents, circulars, and images.
- Data collection cycles (nightly, weekly, monthly) will be synchronized with government data release calendars to capture timely updates and new releases.

5.2. Preprocessing and Data Cleaning

- All incoming data will be subjected to robust deduplication to remove redundancies, followed by automated language detection and segmentation into standardized structures.
- Raw, intermediate, and cleaned datasets will be maintained in distinct collections to allow traceability, facilitate error correction, and support reprocessing when necessary.
- The pipeline will log all processing failures, unsupported encodings, and coverage gaps, with automated alerts for missing languages or corrupted inputs.

5.3. Ontology-Guided Annotation and QA Pair Curation

- Health-domain ontologies will be developed and employed to guide precise annotation of entities, intents, and thematic categories.
- A hybrid annotation workflow will be adopted, merging the expertise of medical translators, informaticians, and domain specialists with supervised community contributions from universities, NGOs, and civil society.
- The annotated corpus will include high-quality, structured question-answer pairs mapped to user intents and specific health topics, supporting effective benchmarking and the training of advanced conversational AI systems.

5.4. Model-Assisted, Human-in-the-Loop Refinement

- Annotation efficiency and quality will be enhanced using transformer-based NLP models (e.g., SinhalaBERT, TamilBERT) to pre-label data and suggest annotations.

- Reinforcement learning feedback loops and human-in-the-loop protocols will provide continuous improvement, especially for code-mixed and dialectal content.

5.5. Quality Assurance and Bias Auditing

- Interactive bias dashboards and scheduled manual audits will be implemented to monitor demographic, regional, and language coverage in the annotated data.
- Critical subsets, including curated QA pairs, will undergo periodic gold-standard manual reviews to ensure annotation validity, inclusivity, and benchmarking reliability.

5.6. Data Governance, Ethics, and Privacy

- All data collection and annotation will strictly adhere to privacy protocols and ethical guidelines, ensuring only publicly available or fully anonymized data are used.
- An independent advisory board or review group will provide oversight to maintain high ethical standards and support ongoing public trust in the project.

6. EXPECTED OUTCOMES AND IMPACT

This corpus initiative is set to deliver a transformation in how Sri Lankan citizens, policymakers, and technologists interact with government information. The immediate outcome will be a living, continually refreshed trilingual and Romanized language resource—comprised of millions of annotated entries—spanning the full breadth of government domains and public service channels.

6.1. Enhanced AI/NLP Model Performance

Language models and information retrieval tools fine-tuned on this corpus are projected to demonstrate marked improvements in intent recognition, sentiment accuracy, answer recall, and code-mix handling—especially in service to rural, minority, and mobile-first users. Baseline evaluations from pilot deployments anticipate accuracy gains of 20–30% over models trained on legacy datasets [9].

6.2. Inclusive Digital Governance

The resource will embolden government websites, chatbots, and multi-modal assistive interfaces to answer questions in Sinhala, Tamil, English, and blended forms, decreasing exclusion rates and digital frustration, especially among seniors and non-English speakers [1], [2].

6.3. Bias Detection and Regional Equity

With systematic annotation for region, demographic, and domain, the project will surface and remediate bias—making clear, through dashboards and open data publishing, who is included, who is not, and how targeted expansion is bridging divides. This positions Sri Lanka to set standards for fair, transparent corpus design—potentially influencing digital policy across South and Southeast Asia.

6.4. Research, Entrepreneurship, and Education

Universities, civic startups, and NGOs will access the corpus via APIs and documented download, using its benchmarks and tools to accelerate new products (from digital literacy

aids to AI-powered civic monitoring dashboards). By including student contributors, the project also builds a workforce of data-savvy linguists and policy analysts.

6.5. Innovation Pipeline

Routine audit, feedback, and update cycles—plus collaborative governance between university, government, and community—ensure this corpus will not become obsolete but grow with the country's technology and culture.

7. TIMELINE & WORK BREAKDOWN STRUCTURE

A successful national-scale corpus effort demands rigorous planning, iterative milestones, and multi-stakeholder coordination. The following table and Gantt chart narrative outline a realistic schedule, balancing quality control with speed.

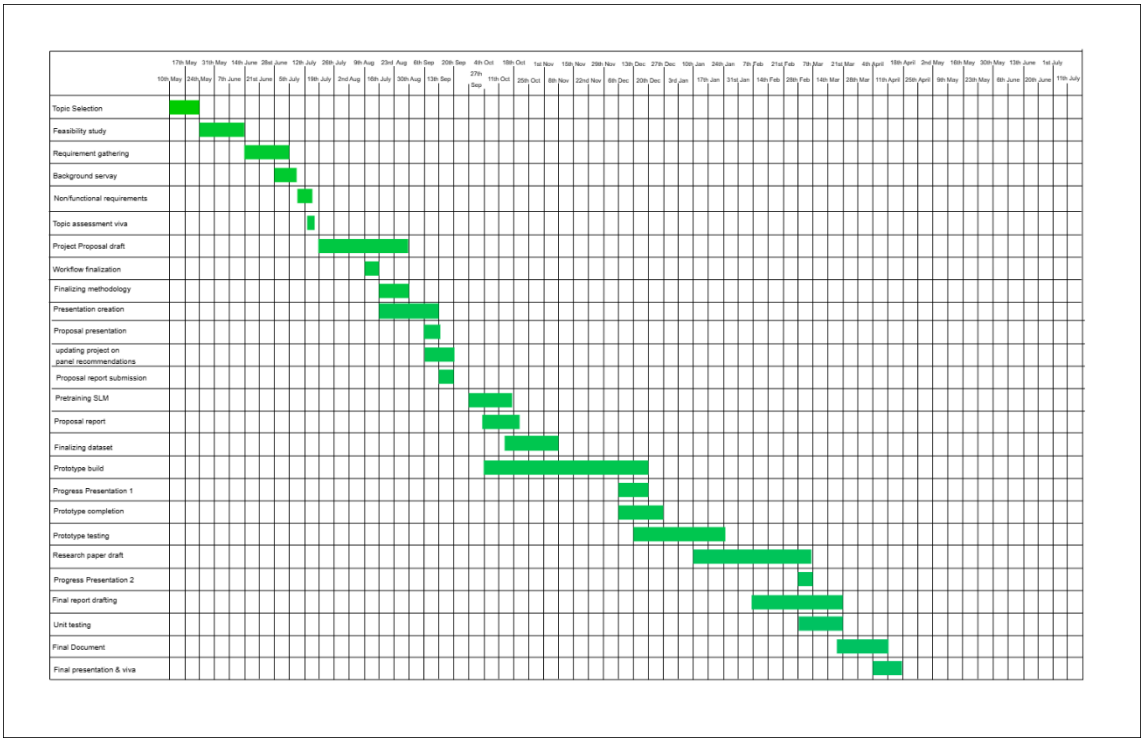

Table 7.1:Gantt Chart of the Project

8. RESOURCES, PERSONNEL, FACILITIES

Achieving the anticipated impact will require the deployment of the following specialized and collaborative resources:

1. Research Supervision and Direction:

   Dr. Lakmini Abeywardhana.

   Ms. Sanjeevi Chandrasiri (Co-supervisor).

2. Project Management:

   Scheduling, milestone tracking, and documentation handled by the student PI and SLIIT support staff using modern PM tooling (e.g., Jira, Trello).

3. Annotation & Validation Team:

   Core: 5–8 expert linguists, part-time.

   Community: rotating workshops of SLIIT students, NGO volunteers, rural annotators.

   QA: periodic focus group testers, pilot beneficiaries.

4. Technical Facilities:

   Cloud Compute Node (for data ingestion/preprocessing, annotation backend).

   MongoDB/AWS Enterprise Tier (storage, API, backups).

   Web-accessible annotation and feedback platforms.

5. Partnerships:

   MOUs with two or more ministries/departments for direct data access and pilot deployments.

   Collaboration with academic partners for external validation and benchmarking.

9. BUDGET AND JUSTIFICATION

A robust financial foundation ensures comprehensive data gathering, annotation quality, outreach, and sustainability.

Budget Line	LKR	Justification
Cloud Storage & Compute	30,000	Secure scalable storage, high-availability API/database hosting
Annotation Tools	10,000	Licenses for annotation software, Deployments, QA scripts, platform maintenance
Human Annotation	10,000	Honorarium for experts, community annotator stipends, error correction auditing
Contingency	5,000	Cost overruns, new data requirements, pilot expansion
Total	55,000	

10. COMMERCIALIZATION & SOCIETAL IMPACT

1. Transforming Access

This corpus will directly power the next generation of conversational agents and search tools for government portals—allowing seamless, accurate information flow for Sri Lankan citizens, regardless of their language, location, or digital literacy. For example, a single government IVR number could now natively answer in Romanized Tamil, SMS Singlish, or rural Sinhala dialect—improving trust, efficiency, and citizen outcomes.

2. Igniting Innovation

Universities, tech entrepreneurs, and NGOs gain an open, gold-standard dataset for research and start-up experiments—enabling the creation of new digital products (chatbots, voice assistants, sentiment engines) exportable not just in Sri Lanka, but as a template across the multilingual developing world.

3. Benchmarking & Bias Reduction

Public, ongoing bias dashboards set a precedent for other nations seeking fairer AI. Corpus releases, workshops, and pilots will become regional standards for inclusive, democratic e-governance.

4. Academic and Skills Development

Beyond government, the resource will accelerate the teaching of data annotation, evaluation, and responsible AI development in university curricula and technical training programs.

11. CONCLUSION

Sri Lanka's ambition to be a digital leader, bridging all communities, all domains, and all languages, demands infrastructural innovation of the highest order. By constructing a living, bias-aware, and exhaustively annotated trilingual corpus, this project creates the "backbone" not only for improved AI, search, and government service, but for a new era of participatory digital citizenship and evidence-based policy. With a foundation set in global best practice, open data, and inclusive design, the project's methods and outcomes will resonate for years, supporting research, business, and social change both within the nation and beyond.

12. REFERENCES

[1] United Nations, *E-Government Survey 2024: Inclusive Digital Governance*, Jun. 2024.

[2] Verité Research, *Rights Online and Language Policy in Government Portals*, 2025.

[3] D. Sumanathilaka, et al., "Swa-Bhasha Resource Hub: Romanized Sinhala to Sinhala Transliteration Systems and Data Resources," *arXiv preprint* arXiv:2507.09245, Jul. 2025.

[4] L. Abeywardhana, S. Rathnayake, and H. Ilangeshwaran, "Saliency-Based Token Swap: A Language-Agnostic Data Augmentation Method for Text Classification," in *Proc. IEEE 9th Int. Conf. on Information Technology Research (ICITR)*, 2024, pp. xx–yy.

[5] UN Department of Economic and Social Affairs, "Sri Lanka's E-Government Portal Data Quality Assessment," *EGDI Country Report*, 2024.

[6] Open Government Partnership, "OGP Global Report 2024: Multilingual Open Data Standards," Mar. 2024.

[7] S. Ranaweera, "A corpus-based study of pragmatic markers in spoken standard Sri Lankan English," Ph.D. dissertation, Univ. College London, 2021. [Online]. Available: https://discovery.ucl.ac.uk/10204129/13/Ranaweera_10204129_Thesis_sigs_removed.pdf

[8] S. Dissanayake, "Representation of Language in Sri Lankan English Newspapers: A Corpus-based Approach," *Sri Lanka Journal of Advanced Research Studies in Humanities and Social Sciences*, vol. 11, no. 1, pp. 59–82, Jan.–Jun. 2024.

[9] R. Dias and R. Jayakody, "Performance of Recent Large Language Models for a Low-Resourced Language," *arXiv preprint* arXiv:2407.21330, Jul. 2024.

[10] N. de Silva, "SinMin - Sinhala Corpus Project," University of Moratuwa, 2016. [Online]. Available: https://nisansads.staff.uom.lk/StudentTheses/Sinmin-Sinhala-Corpus-Project.pdf

[11] Ministry of Electronics and Information Technology (MeitY), *Digital India Programme Overview*, Government of India, 2023.

[12] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proc. Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2019, pp. 4171–4186.

[13] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," *arXiv preprint* arXiv:1910.01108, Oct. 2019.

[14] R. Sennrich, B. Haddow, and A. Birch, "Improving Neural Machine Translation Models with Monolingual Data," in *Proc. 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2016, pp. 86–96.