Burmese-Microbiology-1K

Min Si Thu

July 2024

1 Introduction

Burmese is a language mainly and officially used in Myanmar. Being a low-resource language, the Burmese language needs a lot of quality language resources to train Large Language Models.

2 Purpose

Before this Burmese Clinical Microbiology 1K dataset, the open-source resources to train the Burmese Large Language Model in Medical fields were rare. Thus, the high-quality dataset needs to be curated to cover medical knowledge for the development of LLM in the Burmese language.

3 Motivation

I found an old notebook in my box. The book was from 2019. It contained written notes on microbiology when I was a third-year medical student. Because of the need for Burmese language resources in medical fields, I added more facts, and more notes and curated a dataset on microbiology in the Burmese language.

To our knowledge, this dataset is the first open-source, human-generated Burmese Language dataset on clinical microbiology specifically designed to enable large language models to exhibit the highly interactive capabilities of ChatGPT. Min Si Thu created the Burmese-Microbiology-1k dataset in July of 2024. These training records are natural and expressive.

Figure 1: Burmese Microbiology 1K Dataset Logo

4 How we implement

The dataset includes two columns – **Instruction, Output** in CSV format.

The dataset for microbiology in the Burmese language contains 1262 rows of instruction and output pairs in CSV format. The dataset mainly focuses on clinical microbiology foundational knowledge, abstracting basic facts on culture medium, microbes – bacteria, viruses, fungi, parasites, and diseases caused by these microbes.

4.0.1 Examples

1. ငှက်ဖျားရောဂါဆိုတာ ဘာလဲ?,ငှက်ဖျားရောဂါသည် Plasmodium ကပ်ပါးကောင်ကြောင့် ဖြစ်ပွားသော အသက်အန္တရာယ်ရှိနိုင်သည့် သွေးရောဂါတစ်မျိုးဖြစ်သည်။ ၎င်းသည် ငှက်ဖျားခြင်ကိုက်ခြင်းမှတဆင့် ကူးစက်ပျံ့နှံ့သည်။

2. Influenza virus အကြောင်း အကျဉ်းချုပ် ဖော်ပြပါ။,Influenza virus သည် တုပ်ကွေးရောဂါဖြစ်စေသော RNA ဗိုင်းရပ်စ် ဖြစ်သည်။ Orthomyxoviridae မိသားစုဝင် ဖြစ်ပြီး type A၊ B၊ C နှင့် D ဟူ၍ အမျိုးအစား လေးမျိုး ရှိသည်။

3. Clostridium tetani ဆိုတာ ဘာလဲ,Clostridium tetani သည် မေးခိုင်ရောဂါ ဖြစ်စေသော gram-positive၊ anaerobic bacteria တစ်မျိုး ဖြစ်သည်။ မြေဆီလွှာတွင် တွေ့ရလေ့ရှိသည်။

4. Onychomycosis ဆိုတာ ဘာလဲ?,Onychomycosis သည် လက်သည်း သို့မဟုတ် ခြေသည်း များ တွင် ဖြစ်ပွား သော မှို ကူး စက်မှု ဖြစ်သည်။ ၎င်း သည် လက်သည်း သို့မဟုတ် ခြေသည်းများကို ထူထဲစေပြီး အရောင်ပြောင်းလဲစေသည်။

5 Where to download the dataset

Github – https://github.com/MinSiThu/Burmese-Microbiology-1K

Zenodo – https://zenodo.org/records/1280363

Hugginface – https://huggingface.co/datasets/jojo-ai-mst/Burmese-Microbiology-1K

Kaggle – https://www.kaggle.com/datasets/minsithu/burmese-microbiology-1k

5.1 Applications

Burmese Microbiology 1K Dataset can be used in building various medical-related NLP applications.

1. The dataset can be used for pretraining or finetuning the dataset on Burmese Large Langauge Models.
2. The dataset is ready to use in building RAG-based Applications.

6 Acknowledgments

Special thanks to magickospace.org for supporting the curation process of the Burmese Microbiology 1K Dataset.

7 References for this dataset

1. https://openstax.org/details/books/microbiology – For medical facts

2. https://moh.nugmyanmar.org/my/ – For burmese words for disease names
3. https://myordbok.com/dictionary/english – English–Myanmar Translation Dictionary