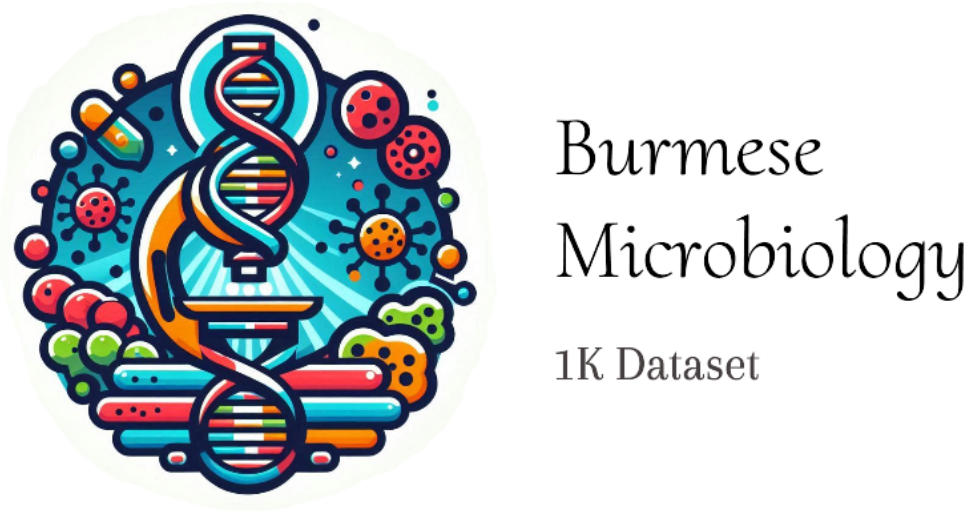
Burmese-Microbiology-1K

Min Si Thu, min@globalmagicko.com

Microbiology 1K QA pairs in Burmese Language



Purpose

Before this Burmese Microbiology 1K dataset, the open-source resources to train Burmese Large Language Model on Medical fields are very rare. Thus why, the high quality dataset needs to be curated for development of LLM in Burmese Language, covering medical knowledge.

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 on medical fields, I added more facts, more notes and curate a dataset on microbiology in burmese language.

About

The dataset for microbiology in burmese language contains **1131 rows of Instruction, Output pairs in csv format**. The dataset mainly focus on mircobiology foundational knowledge, abstracting basic facts on culture medium, microbes - bacteria, virus, fungi and parasite, and diseases caused by these microbes.

Examples

- ငှက်ဖျားရောဂါဆိုတာ ဘာလဲ?,ငှက်ဖျားရောဂါသည် Plasmodium ကပ်ပါးကောင်ကြောင့် ဖြစ်ပွားသော အသက်အန္တရာယ်ရှိနိုင်သည့် သွေးရောဂါတစ်မျိုးဖြစ်သည်။ ၎င်းသည် ငှက်ဖျားခြင်ကိုက်ခြင်းမှတဆင့် ကူးစက်ပျံ့နှံ့သည်။
- Influenza virus အကြောင်း အကျဉ်းချုပ် ဖော်ပြပါ။,Influenza virus သည် တုပ်ကွေးရောဂါ ဖြစ်စေသော RNA ဗိုင်းရပ်စ် ဖြစ်သည်။ Orthomyxoviridae မိသားစုဝင် ဖြစ်ပြီး type A၊ B၊ C နှင့် D ဟူ၍ အမျိုးအစား လေးမျိုး ရှိသည်။
- Clostridium tetani ဆိုတာ ဘာလဲ,Clostridium tetani သည် မေးခိုင်ရောဂါ ဖြစ်စေသော gram-positive၊ anaerobic bacteria တစ်မျိုး ဖြစ်သည်။ မြေဆီလွှာတွင် တွေ့ရလေ့ရှိသည်။

- Onychomycosis ဆိုတာ ဘာလဲ?,Onychomycosis သည် လက်သည်း သို့မဟုတ် ခြေသည်းများတွင် ဖြစ်ပွားသော မှို ကူးစက်မှုဖြစ်သည်။ ၎င်းသည် လက်သည်း သို့မဟုတ် ခြေသည်းများကို ထူထဲစေပြီး အရောင်ပြောင်းလဲစေသည်။

Where to download the dataset

- https://github.com/MinSiThu/Burmese-Microbiology-1K/blob/main/data/Microbiology.csv
-
-
-

Applications

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

- The dataset is added to ChatGPT services for custom QA bot applications.
    - Link to be published sooner
- The dataset can be used for pretraining or finetuning the dataset on Burmese Large Langauge Models.

Ackowledgements

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

References for this datasets

- https://openstax.org/details/books/microbiology - For medical facts
- https://moh.nugmyanmar.org/my/ - For burmese words for disease names
- https://myordbok.com/dictionary/english - English-Myanmar Translation Dictionary

License - **CC BY SA 4.0**

How to cite the dataset