



Large Language Model Development for Sinhala Language

H.M.V.R. Herath¹, G.M.R.I. Godaliyadda¹, M.P.B. Ekanayake¹, R.G. Ragel¹, Bandara Herath²
¹University of Peradeniya. ²Cornell University

Abstract- The developed language model is expected to empower various applications in Sinhala text analysis, information retrieval, machine translation, and content generation, thereby fostering digital inclusivity and linguistic diversity in the era of artificial intelligence.

GPT4

write a 100 word letter to your teacher explaining why you were unable to attend school due to fever



Dear Teacher,

I hope this message finds you well. I am writing to explain my absence from school. I was unable to attend due to a high fever. I experienced severe headaches, body aches, and fatigue, which made it difficult to get out of bed. Following my doctor's advice, I rested at home to recover properly. I regret missing valuable class time and any assignments or activities during my absence. I am now feeling better and look forward to returning to school. I will catch up on any missed work and seek your guidance on anything I need to complete.

Sincerely,
[Your Name]

ගැටළු සාරාංශය

- පවතින විශාල භාෂා ආකෘතිවලට සිංහල භාෂාවේ සංකීර්ණත්වය එලදායී ලෙස ග්‍රහණය කරගත නොහැක.
- පුහුණු දත්ත ප්‍රමාණය මෙන්ම පුහුණු දත්තවල ගුණාත්මක භාවය ද මෙම අඩුපාඩුවට හේතු වේ.
- අති නවීන ගැඹුරු ඉගෙනුම් ශිල්පීය ක්‍රම සිංහල භාෂාවට අනුගත වීම අභියෝගාත්මක කාර්යයකි.
- මෙයට හේතුව එහි රූප විද්‍යාත්මක සංකීර්ණත්වය, එකතු කිරීමේ ස්වභාවය සහ පුළුල් ලෙස අධ්‍යයනය කරන ලද භාෂාවලට සාපේක්ෂව ඩිජිටල් සම්පත් සීමිත වීමයි.

GPT4

write a 100 word letter in Sinhalese to your teacher explaining why you were unable to attend school due to fever



ආචාර්යවරයාණනි,

ආයුබෝවන්! ගරුකම් සලකමින්, මට පෙර දින පාසැලට පැමිණීමට නොහැකි වූයේ උණ රෝගයට ගොදුරු වීම නිසාය. මට ගරීර උණුසුම් ඉහළ යාම, හිස කැක්කුම්, සහ නතරෙන් වේදනාවක් දැනිණි. වෛද්‍යවරයාගේ උපදෙස් පරිදි නිවසේදී විවේක ගත කළේය. මෙම අවස්ථාවෙන් හට ගත පීඩාවට සමාව ඉල්ලන අතර, ඉදිරි දිනවලදී පාසැල් කටයුතු වලට සෘජුව සහභාගී වීමට බලාපොරොත්තු වෙමි.

ඔබගේ අවගාච්චිය සහ සහයෝගය මැනවින් ඉල්ලා සිටිමි.

ඔබගේ ශිෂ්‍යයා,
[ඔබේ නම]



Path to the solution

- Use open source large language models.
- Collect as much digital data as possible that represents the breath and depth of the Sinhala Language.
- Train large language models with collected data taking into consideration morphological complexity and agglutinative nature of the language

Contact details

Name : Vijitha Herath
Tel. No.: +94 716732688
Email : vijitha@eng.pdn.ac.lk

Multidisciplinary AI Research Centre (MARC)
University Research Council
University of Peradeniya
Peradeniya, 20400, Sri Lanka

