A Compact Hindi Language Model Based on Gemma-270M

Utpal Anand (3rd Year BS-MS)
Supervisor: Bedartha Goswami

Machine Learning For Weather and Climate Lab, Data Science Department, IISER Pune

Abstract

Gemma-270M is a lightweight open model from Google, yet its multilingual pre-training provides limited depth for Hindi. As a result, downstream performance for Hindi tasks—especially those involving morphology, script handling, and domain-specific text—remains suboptimal.

We adapt the Gemma-270M architecture into a compact Hindi language model by retraining it on a curated, high-quality Hindi corpus and optimizing tokenization for Devanagari. This specialization yields noticeably improved fluency, comprehension, and factual consistency compared to the base multilingual model, while preserving the efficiency benefits of a 270M-parameter design.

Our work demonstrates that even small models benefit significantly from Hindi-focused training, enabling practical deployment in low-resource, on-device, and real-time applications.

Model Architecture

Our Hindi language model is based on a compact variant of the Gemma-270M architecture, redesigned for efficiency and Hindi-centric performance. The model employs a lightweight Transformer stack with only **18 layers**, significantly reducing computational cost while retaining strong representational capacity for linguistic structure.

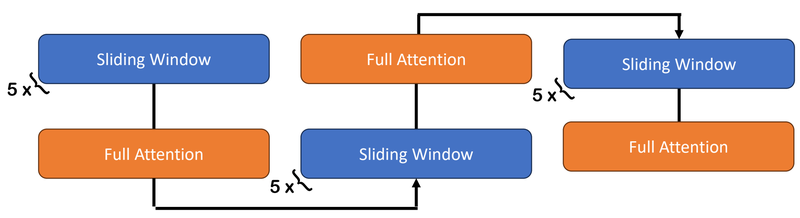
To enable scalable sequence processing on limited hardware, we integrate **sliding-window attention**, which restricts attention computation to a fixed contextual window. This greatly lowers memory usage while preserving local contextual understanding essential for Hindi morphology and sentence structure.

We further adopt **Rotary Positional Embeddings (RoPE)**, which provide stable positional encoding over long sequences and improve generalization across varying text lengths. Combined with optimized Devanagari tokenization and Hindi-focused pre-training, this architecture supports efficient and accurate modeling of Hindi text in small-parameter regimes.

Methodology

Model Development Pipeline

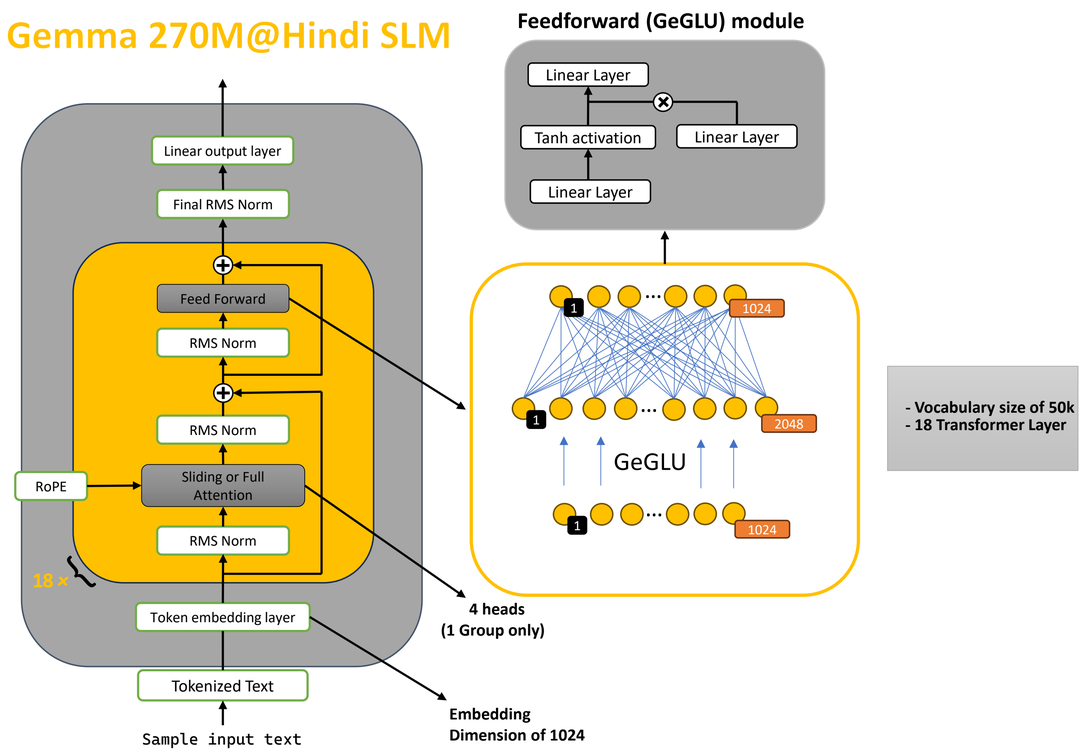
We adapt the Gemma-270M architecture into a compact Hindi language model using a Hindi-optimized training pipeline. The workflow includes corpus curation, tokenizer adaptation for Devanagari, model specialization, and evaluation on Hindi benchmarks.



Model Pipeline

Architecture Overview

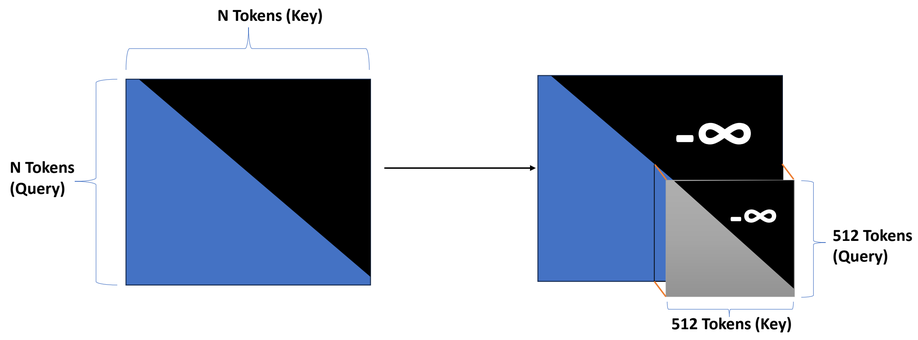
Our model uses a lightweight **18-layer Transformer**, optimized for efficiency while maintaining strong representational ability for Hindi grammar and morphological structure.



Model Architecture

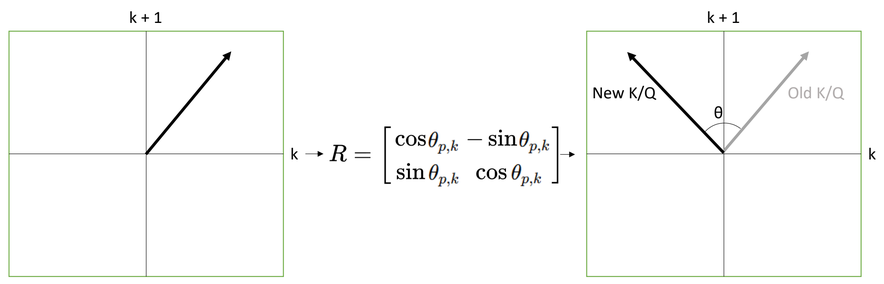
Key Mechanisms

Sliding-Window Attention To reduce memory and speed bottlenecks, attention is restricted to a local window, enabling long-sequence modeling with dramatically lower cost.



Sliding Window Attention Mechanism

Rotary Positional Embeddings (RoPE) RoPE provides smooth rotational positional encoding, making the model stable for longer contexts and improving sentence-level coherence in Hindi.



$$\text{For each pair } (k, k+1): \quad \theta_{p,k} = \frac{p}{\theta_{\text{base}}^{k/d_{\text{head}}}}$$

For sliding attention: $\theta_{\text{base}} = 10^4 = 10{,}000$ **For full attention:** $\theta_{\text{base}} = 10^6 = 1{,}000{,}000$

Rotary Positional Embedding

Data Collection & Processing

We use the **CC100 Hindi** dataset as our primary training corpus, containing approximately **1 billion Hindi tokens**. The text is cleaned, deduplicated, and normalized to ensure high-quality input.

A custom **Byte-Pair Encoding (BPE)** tokenizer with a **50k vocabulary size** is trained specifically on Hindi text to better capture Devanagari script patterns and subword structure.

These processed tokens serve as the input for model pretraining and evaluation.

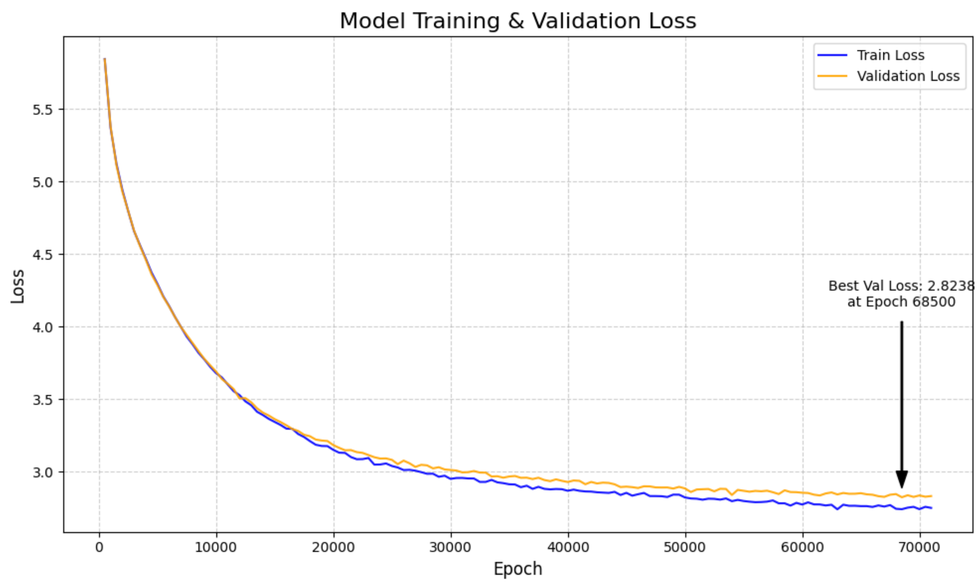
Results & Findings

Training Summary

Our model was trained for **110k iterations** using **mixed-precision training** on **14 × NVIDIA A30 GPUs**, taking approximately **14–15 hours**. We used a **3e-4 learning rate** with a warmup of **8000 steps**, **gradient accumulation = 2**, and a cosine learning rate scheduler.

- **World Size:** 14 GPUs
- **Batch per GPU:** 48
- **Effective Batch Size:** 672
- **Precision:** FP16 / BF16 Mixed Precision

Loss Curve



Training and Validation Loss Curve

Inference Results

The model produces coherent, context-aware Hindi text. Below are two prompt examples demonstrating smooth continuation and narrative consistency.

दिनेश कार्तिक भारतीय क्रिकेट खिलाड़ी हैं । उन्होंने साल 2007 में बांग्लादेश के खिलाफ अपना पहला टेस्ट मैच खेला था । उन्हें 2006 में वेस्टइंडीज के खिलाफ ईडन गार्डन्स में वनडे और टी - 20 की कप्तानी सौंपी गई थी । उनके नाम अब तक तीन शतक दर्ज है जिसमें एक शतक शामिल है । उन्होंने इस दौरान 11 अर्धशतक लगाए थे , वहीं टी20 में उनका सर्वश्रेष्ठ स्कोर 28 रन रहा था । कार्तिक ने आईपीएल 2017 में सनराइजर्स हैदराबाद टीम का नेतृत्व किया था । उनकी गेंदबाजी में भी कोई खास बदलाव नहीं आया लेकिन वह अपने पहले ही ओवर में तेज गेंदबाज ई शांत शर्मा को आउट कर सकते हैं । यह भी पढ़ें : IP L 2018 : कोहली ने एबी डि विलिय र्स से कहा - ' मुझे लगता है कि मैं अभी अच्छा प्रदर्शन करना चाहता हूं ' Pre v ious St ory भारत बनाम श्रीलंका : तीसरे वनडे में बने कुल दीप यादव के 6 हजारी N ext St ory भारत ने श्रीलंका को हरा दिया था वर्ल्ड कप 2019 में इंग्लैंड में होने वाले विश्व कप में जगह बनाने वाली टीम इंडिया का हिस्सा होंगे . इंडियन प्रीमियर लीग (आईपीएल) में खेलने पर चयन कर्ताओं ने इस

मौसम विभाग का अनुमान है कि अगले घंटे में दिल्ली में और अधिक बारिश होने की संभावना है । दिल्ली के कुछ इलाकों में तेज हवाएं चल रही हैं , जिसके चलते लोगों को काफी परेशानी हो सकती है । ऐसे में लोग घर से बाहर निकलने से पहले अपने साथ पानी लेकर निक लें । इससे आपको थोड़ी राहत मिलेगी । वहीं दूसरी ओर , बारिश के कारण यहां कई जगहों पर पेड़ गिरने की घटनाएं भी सामने आई हैं । मौसम विभाग ने चेतावनी जारी करते हुए कहा है कि अगर इस दौरान किसी तरह की अप्रिय घटना होती है तो इसके लिए प्रशासन जिम्मेदार होगा । हालांकि , अभी तक किसी प्रकार की लापरवाही या किसी व्यक्ति विशेष की अनदेखी नहीं हुई है । मौसम विभाग के अनुसार अगले 24 घंटों के दौरान उत्तराखंड , उत्तर प्रदेश , हरियाणा , चंडीगढ़ , पंजाब , राजस्थान , पश्चिमी यूपी और पूर्वी मध्य प्रदेश में आंधी - तूफान आने की आशंका जताई जा रही है । इस बीच राष्ट्रीय राजधानी दिल्ली सहित राज्य भर में धूल भरी आंधी चलने वाली है । इससे बचने के लिए लोग अपने घरों में ही रहें । इन दिनों में आंधी - तूफान के बाद अब तक किसी भी तरह की कोई खबर नहीं मिली है । जिससे लोगों को दिक् क

References

1. Gemma 3 Technical Report: https://arxiv.org/pdf/2503.19786
2. https://deepmind.google/models/gemma/
3. Attention Is All You Need: https://arxiv.org/pdf/1706.03762