Indic-TunedLens: Interpreting Multilingual Models in Indian Languages

Mihir Panchal[1], Deeksha Varshney[2], Mamta[3], Asif Ekbal[4]

[1] Dwarkadas J Sanghvi College of Engineering, [2] Indian Institute of Technology Jodhpur, [3] Kings's College London, [4] Indian Institute of Technology Patna

Introduction

- Multilingual large language models (mLLMs) are increasingly deployed in linguistically diverse regions like India, yet most interpretability tools remain tailored to English.
- **We propose Indic-TunedLens, a multilingual interpretability framework that:**

Aligns intermediate hidden states with final vocabulary distributions

Learns language-aware affine transformations

Enables faithful decoding of intermediate layers for 10 Indian languages

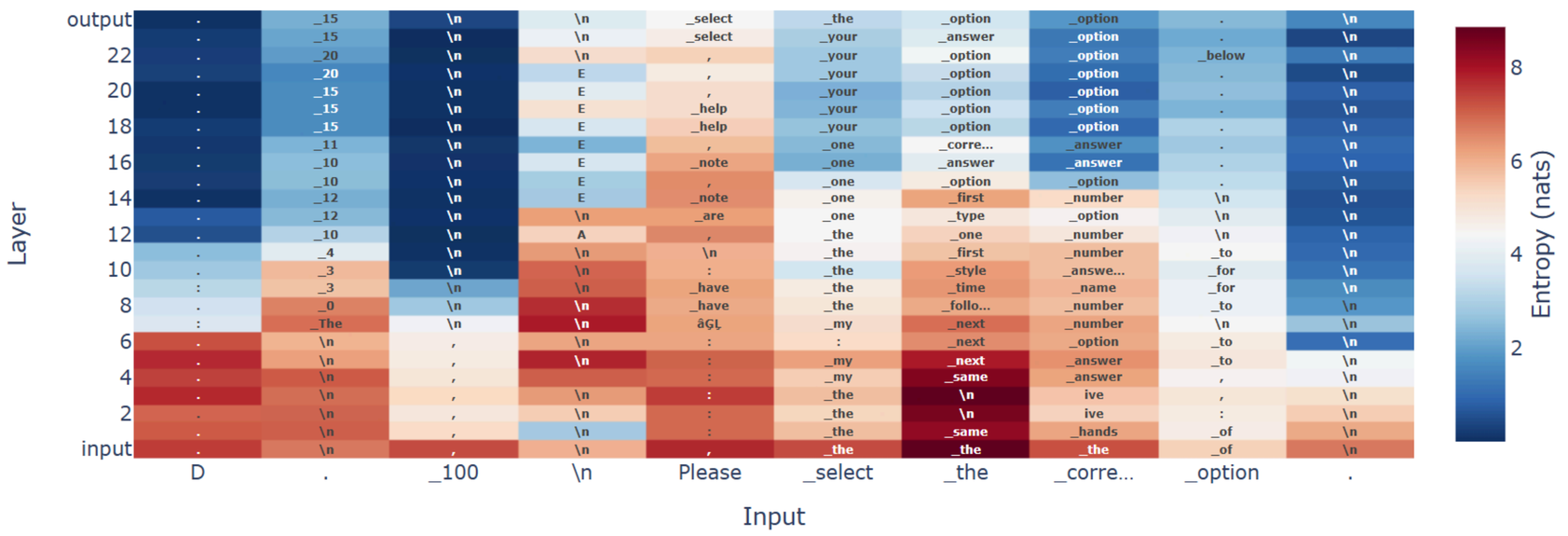

Figure 1: The high and irregular entropy across layers (standard Tuned Lens) suggests unstable intermediate representations and weak alignment for Indian languages, with predictions biased toward English tokens.

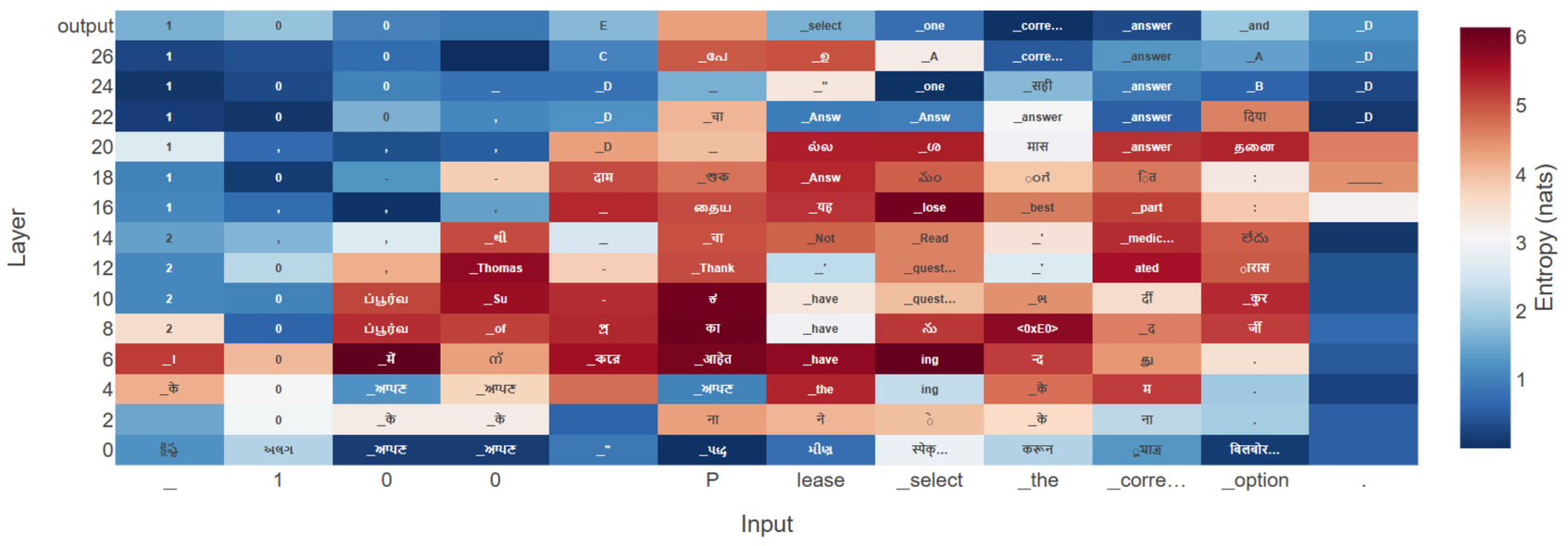

Figure 2: Entropy heatmap for the Indic-TunedLens, indicating improved semantic alignment, with intermediate predictions increasingly generating meaningful Hindi tokens.

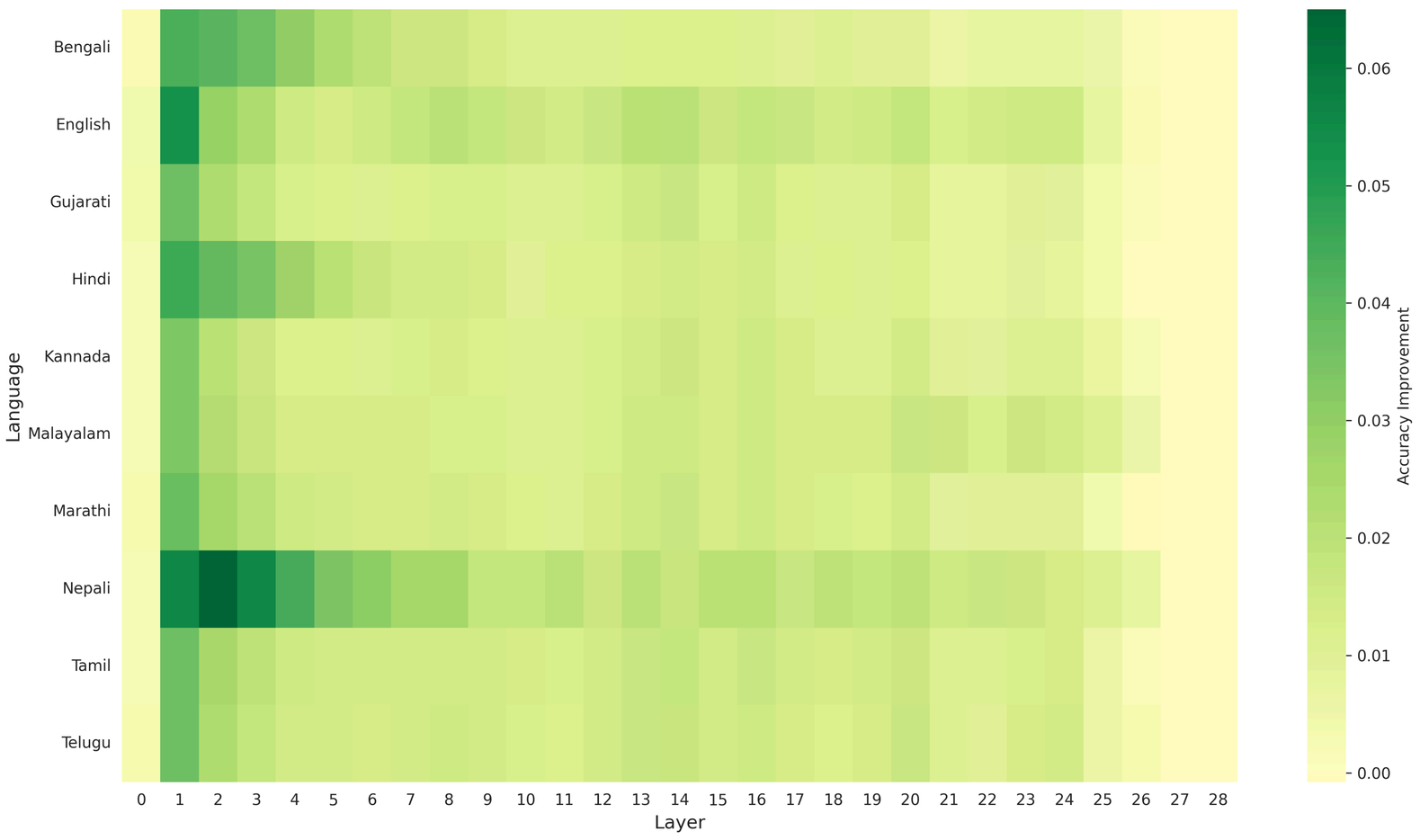

Figure 3: Layer-wise Improvement Patterns

Methodology

- Step 1: Take hidden state h_n
- Step 2: Apply affine transformation

$$\text{Indic-TunedLens}_n(h_n) = \text{LogitHead}(\tilde{h}_n)$$

$$\text{where } \tilde{h}_n = M_n h_n + b_n$$

- Step 3 : Pass transformed state through model's output head
- Step 4: Minimize KL divergence with final layer distribution

Key Findings

- Indic-TunedLens demonstrates superior accuracy over Logit Lens across all Indian languages, with particularly pronounced improvements in early and middle layers.
- Maximum improvements occur in early layers (1-8) with language specific patterns reflecting morphological complexity.
- Accuracy varies substantially across token positions, with specific positions showing language dependent spikes corresponding to answer tokens.

EACL 2026
RABAT • MOROCCO
Mars • March 24-29, 2026 • مارس