**State of Indic language is very poor on world stage**

India is home to a vast number of languages, with estimated 19,500 dialects and mother tongues. The Indian Constitution officially recognizes 22 languages spoken by 1.46 billion people. With a vast population and being the fastest growing and the fourth largest economy in the world, one would assume that India would be a powerhouse in AI, and home to a large pool of language models. However, a survey of Hugging Face website, that has largest collection of worldwide AI models, suggests otherwise.

To understand how the AI model research is going for Indic languages, we surveyed the HuggingFace web site. The HuggingFace model has the largest collection of models with total of 1.96 million models available for download. These models perform various tasks, that include speech, image and text processing. Over 450,000 models that support natural language processing that deal with understanding language.

A total of 221,062 support English languages (Figure 1), whereas 12,862 support Chinese Mandarin. There are French, Spanish and German languages have over 10k models. Indic languages have less than 7500 models. Hindi, the most common Indic language, spoken by over 600 million users has 5,957 models.

Figure 1: Total number of models that support each language

To put this in context, consider the number of speakers for each language (See Figure 2). English is the most commonly spoken worldwide with over 1.5 billion speakers. This is followed closely by Indic languages which collectively make up 1.45 billion speakers, the second highest. Mandarin Chinese is the third most spoken language in the world with 1410 speakers. Spanish and Arabic round the top five with 559 and 332 million speakers respectively.

Figure 2: Total speakers for top five languages worldwide

When correlating the number of speakers with the number of models a notable disparity emerges. While English has the highest number of speakers among the listed languages in the previous table, its lead in ‘supported models’ is disproportionately large, suggesting a much higher investment and development in AI models for English compared to other languages. Despite the collective large number of Indic speakers, Hindi's model count is comparatively low. Over 12,000 models support Chinese Mandarin – a figure significantly higher than Indic languages. The Chinese Mandarin is spoken by a smaller number of people compared to Indic languages, yet it features more models.

This disparity suggests that the development of AI models doesn't always directly scale with the number of speakers, potentially highlighting areas where more linguistic resources and development could be beneficial for less-represented languages.

This means that the number of models per speaker for India languages is significantly smaller. English language has over 140 models per speaker, and Chinese Mandarin has 10. Indic languages barely have 7 models per speaker (See Table 1 below).

**Lack of indigenous models in most downloaded Indic model list**

Majority of the models that support Indic languages have been developed in US. If you consider top five most downloaded models for Hindi for example, an indigenous model doesn’t even feature in top five. (See table). The most downloaded Indic model is from AI4Bharat initiative and the number of downloads are off by a magnitude in comparison!

|  |  |
| --- | --- |
| Model | Downloads |
| sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2 | 12.5 million |
| Xlm-roberta-base | 12.3 million |
| Llama-3.1-8B-Instruct | 9.8 million |
| Bert-base-multilingual-cased | 7.55 million |
| ….. |  |
| ai4bharat/indictrans2-en-indic-dist-200M | 109 thousand |

The AI4Bharat was one of the key initiatives from the Government of India and IIT Madras aimed at overcoming the lack of progress in the development of Indic language models, but the data suggests that it has failed to catch on.

**India must invest more in India models**

Investing in language models is paramount for a country's future given their transformative potential across numerous sectors. These advanced AI models can understand, contextualize and generate response to any question in a human understandable means of communication. A good set of models can revolutionize education by providing personalized learning experiences and bridging language barriers, thereby fostering a more skilled workforce. In the economic sphere, language models drive innovation in industries like customer service, content creation, and data analysis, enhancing productivity and creating new job opportunities. Furthermore, they play a crucial role in national security and diplomacy by enabling sophisticated intelligence gathering, secure communication, and efficient translation for international relations. By developing and integrating cutting-edge language models, a country can bolster its economic competitiveness, improve public services, and strengthen its geopolitical influence in an increasingly interconnected and data-driven world. India must do more to overcome this.

Part 2: The state of regional languages model is even worse!

Figure 3: Model support for top five Indic languages

|  |  |  |  |
| --- | --- | --- | --- |
|  | Models | Lang Speakers | Model/speaker |
| English | 221062 | 1528 | 144.6740838 |
| Chinese | 12862 | 1184 | 10.86317568 |
| French | 11714 | 312 | 37.54487179 |
| Spanish | 10668 | 558 | 19.11827957 |
| German | 10490 | 134 | 78.28358209 |
|  |  |  |  |
| Hindi | 5957 | 609 | 9.781609195 |
|  |  |  |  |
| Marathi | 1478 | 99 | 14.92929293 |
| Tamil | 1776 | 77 | 23.06493506 |
| Bengali | 2155 | 284 | 7.588028169 |
| Telugu | 1454 | 94 | 15.46808511 |

When you consider other widely spoken language in India, the gap widens further.