Evaluating Indian Language Performance in LLMs

Tattle

Aatman Vaidya

The Growth of Large Language Models

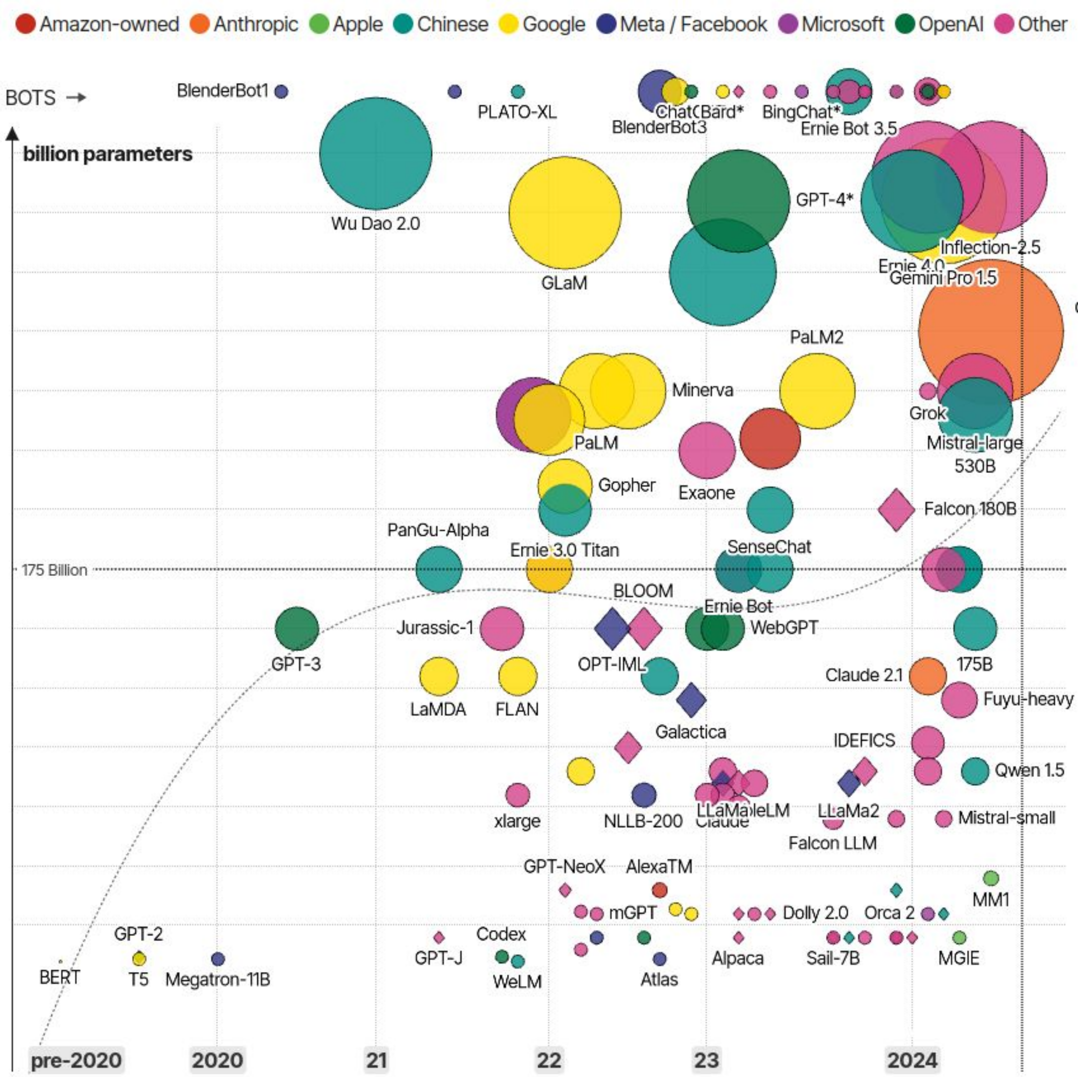
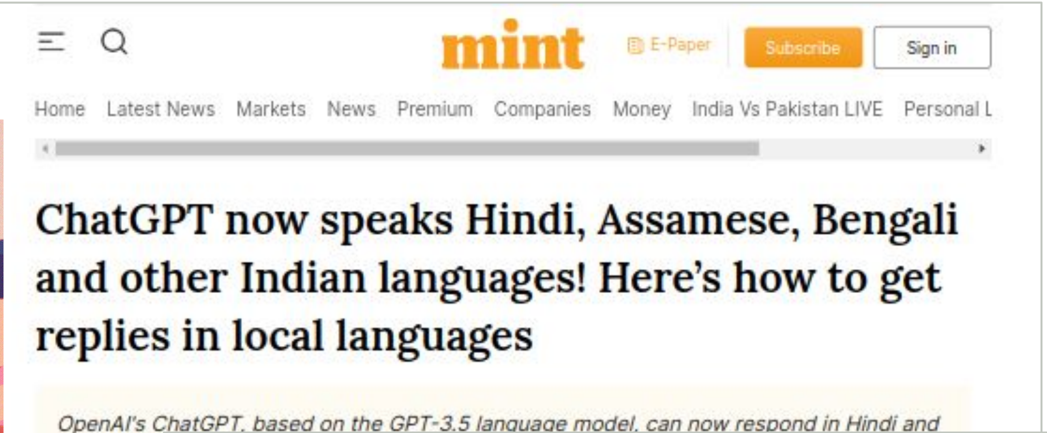


Image source - https://informationisbeautiful.net/visualizations/the-rise-of-generative-ai-large-language-models-llms-like-chatgpt/

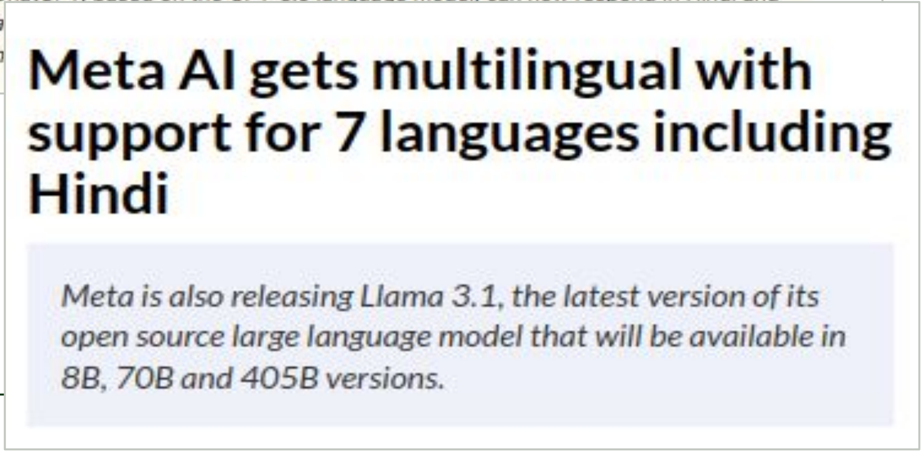
Multilingual Usage and Claims



ChatGPT now speaks Hindi, Assamese, Bengali and other Indian languages! Here's how to get replies in local languages

OpenAI's ChatGPT, based on the GPT-3.5 language model, can now respond in Hindi and other India

integration



Meta AI gets multilingual with support for 7 languages including Hindi

Meta is also releasing Llama 3.1, the latest version of its open source large language model that will be available in 8B, 70B and 405B versions.

Questions

- What are the capabilities of LLMs to understand and generate text in Indian Languages?

Questions

- What are the capabilities of LLMs to understand and generate text in Indian Languages?

- How are LLMs that support Indian languages developed? What are the key characteristics (license, access, ownership etc)?

Questions

- What are the capabilities of LLMs to understand and generate text in Indian Languages?
- How are LLMs that support Indian languages developed? What are the key characteristics (license, access, ownership etc)?
- How many resources are available for Indian languages, does this correlate with the number of speakers? How do resource gaps impact LLM performance?

Methodology

We conduct analysis using existing frameworks proposed by Ecosystem Graphs (Bommasani et al 2023)[1] and by KJ et al. 2024

LLMs

Evaluation

[1] - https://crfm.stanford.edu/ecosystem-graphs/

Methodology

We conduct analysis using existing frameworks proposed by Ecosystem Graphs (Bommasani et al 2023)[1] and by KJ et al. 2024

LLMs

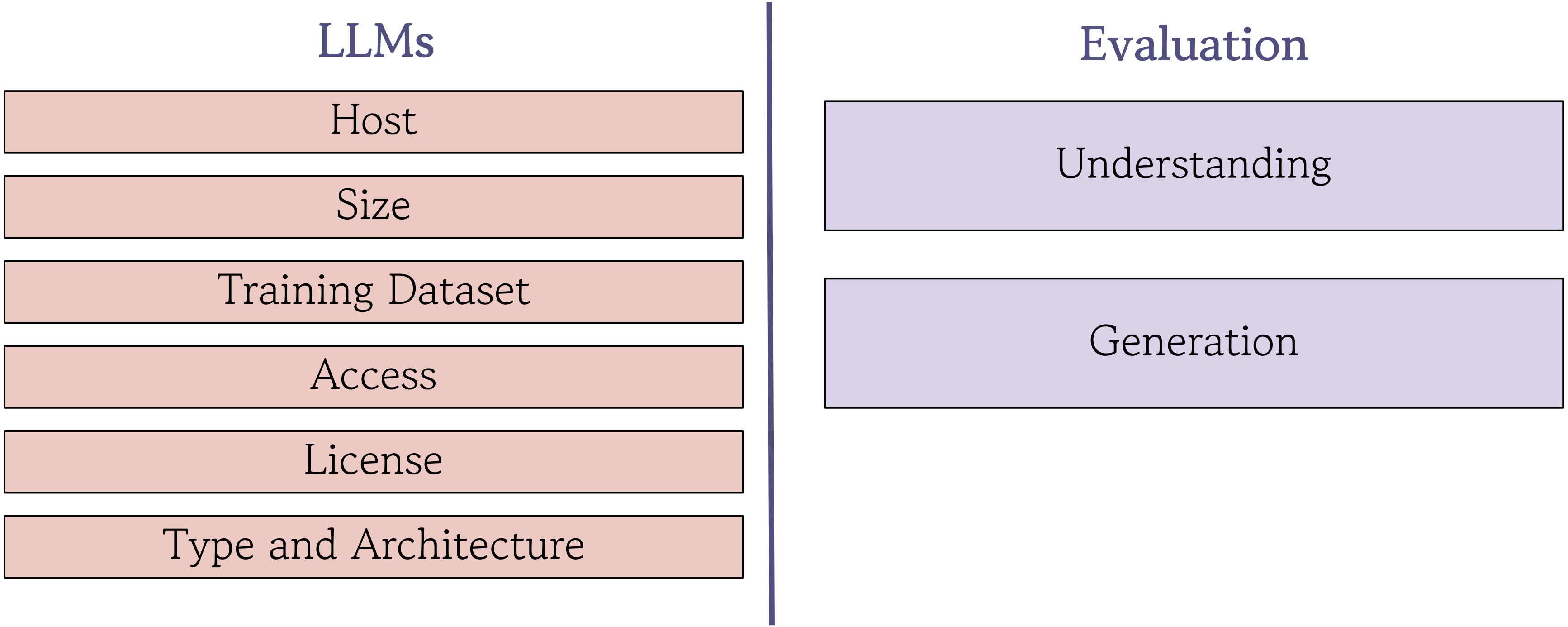
- Host
- Size
- Training Dataset
- Access
- License
- Type and Architecture

Evaluation

[1] - https://crfm.stanford.edu/ecosystem-graphs/

Methodology

We conduct analysis using existing frameworks proposed by Ecosystem Graphs (Bommasani et al 2023)[1] and by KJ et al. 2024



LLMs

Host

Size

Training Dataset

Access

License

Type and Architecture

Evaluation

Understanding

Generation

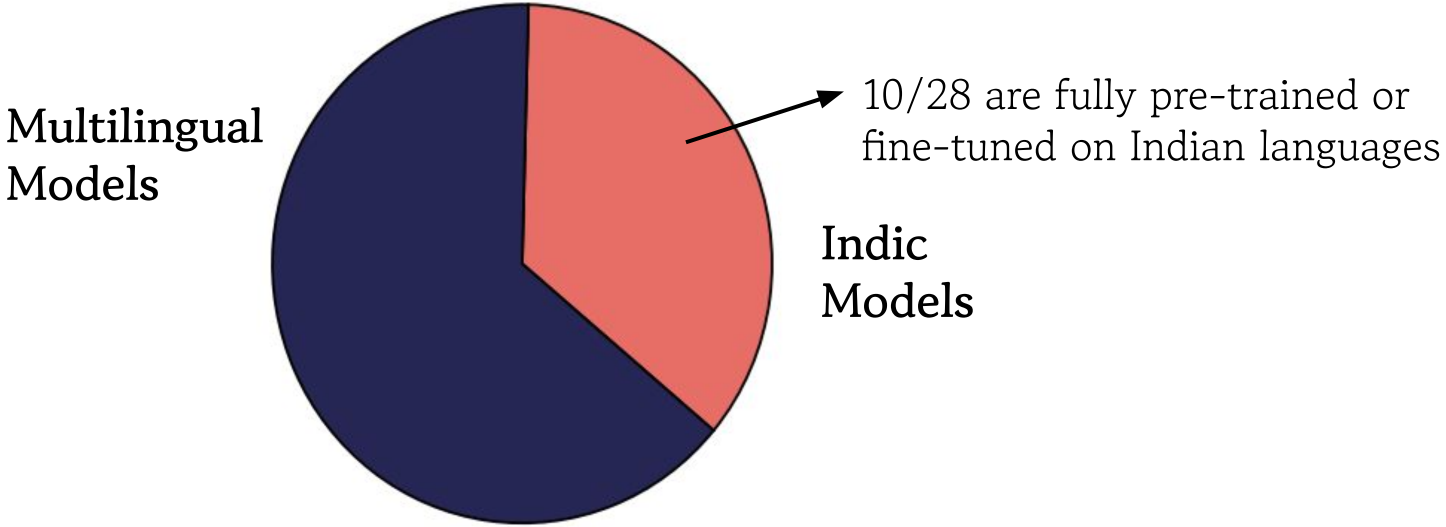
[1] - https://crfm.stanford.edu/ecosystem-graphs/

Analysis

- We analyze 28 models that support Indian languages using the methodology outlined before.
- Models trained on multilingual corpora with Indian language data in them were also included.

Analysis

- We analyze 28 models that support Indian languages using the methodology outlined before.
- Models trained on multilingual corpora with Indian language data in them were also included.

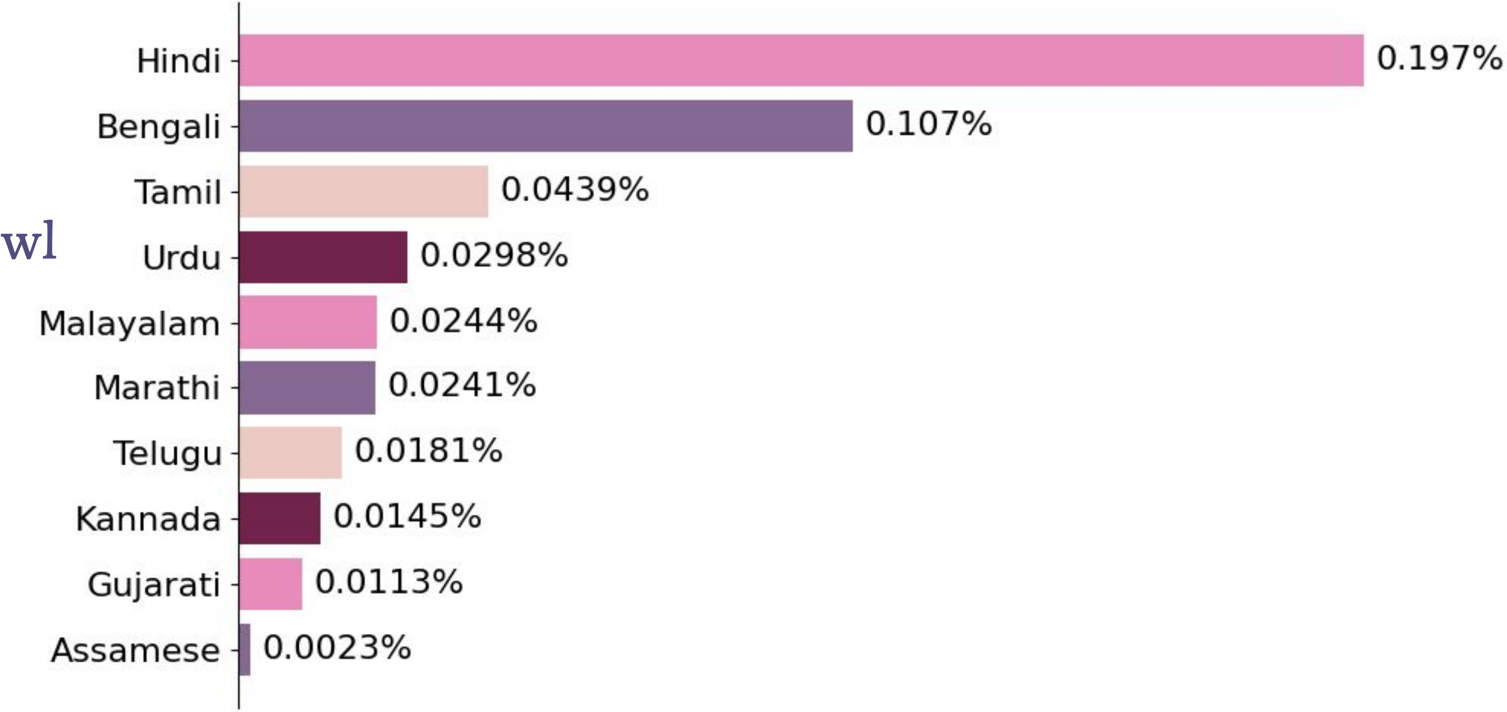


10/28 are fully pre-trained or fine-tuned on Indian languages

Training Corpora

Multilingual Models

Common Crawl

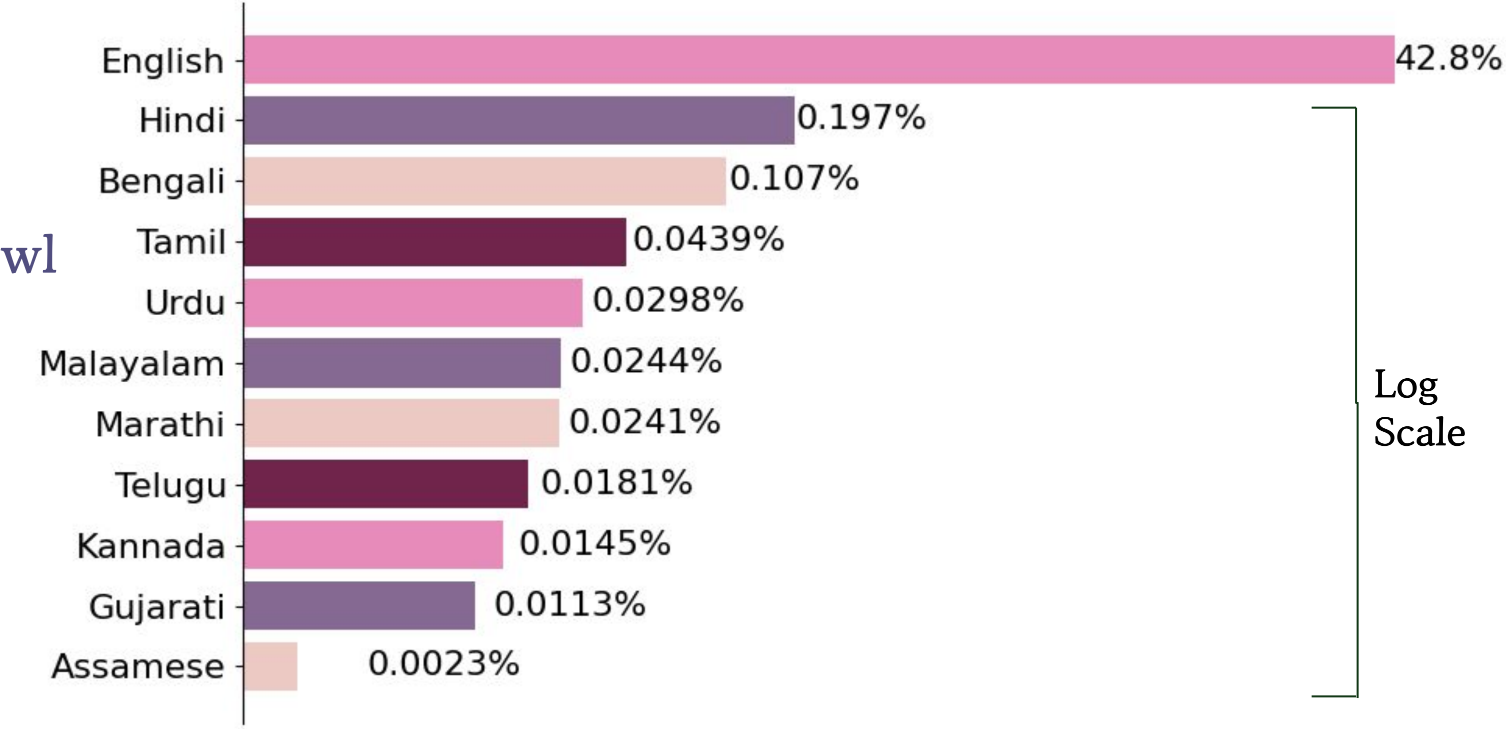


Source - https://commoncrawl.github.io/cc-crawl-statistics/plots/languages

Training Corpora

Multilingual Models

Common Crawl

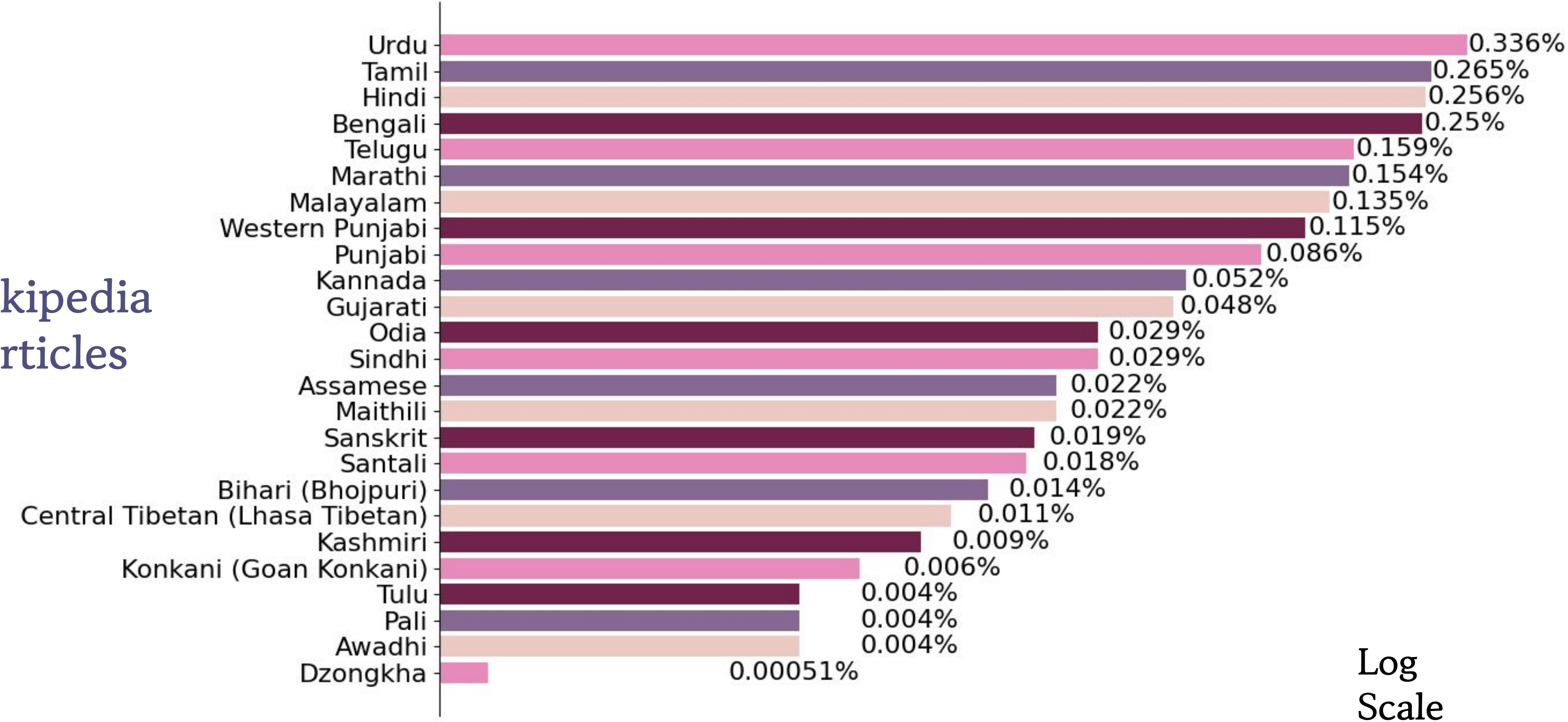


Log Scale

Source - https://commoncrawl.github.io/cc-crawl-statistics/plots/languages

Training Corpora

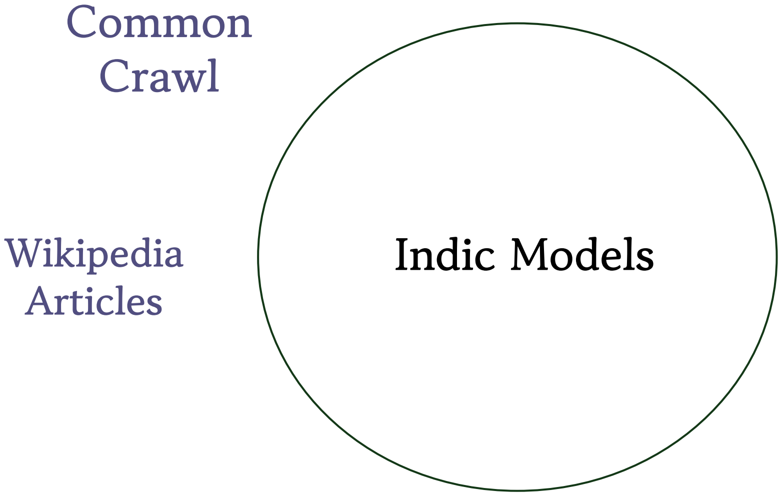
Wikipedia Articles



Log Scale

Source - https://en.wikipedia.org/wiki/List_of_Wikipedias
Source - https://internetlanguages.org/en/numbers/wikipedia-language-geography/

Training Corpora

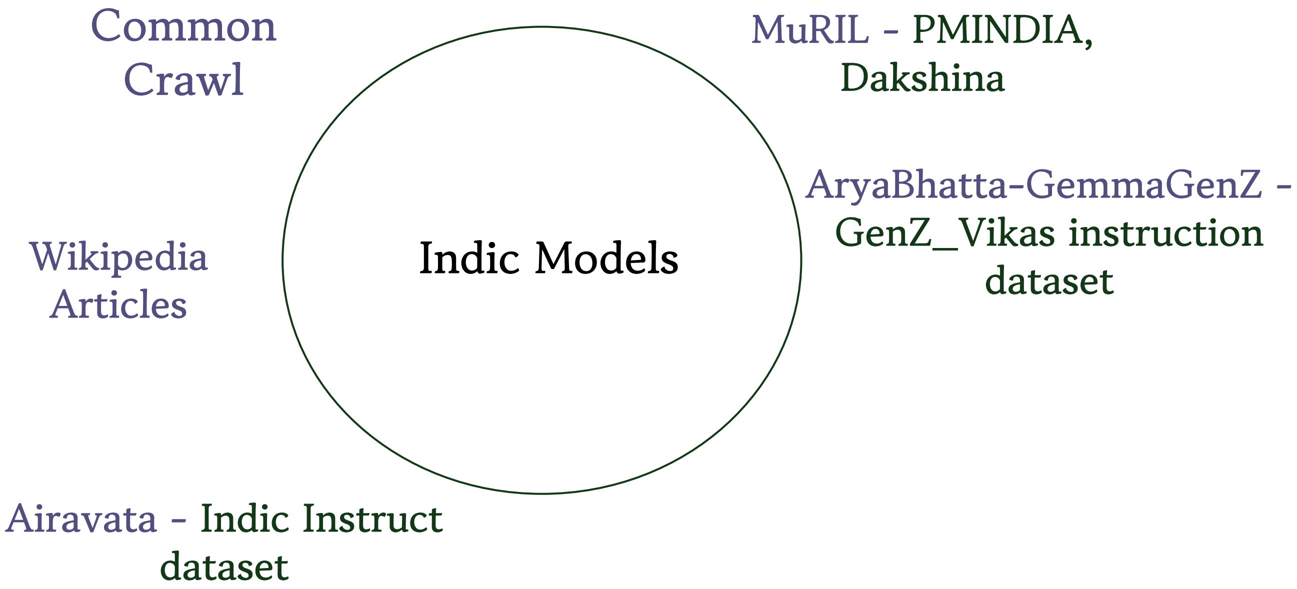


Common Crawl

Wikipedia Articles

Indic Models

Training Corpora



Indic Models

Common Crawl

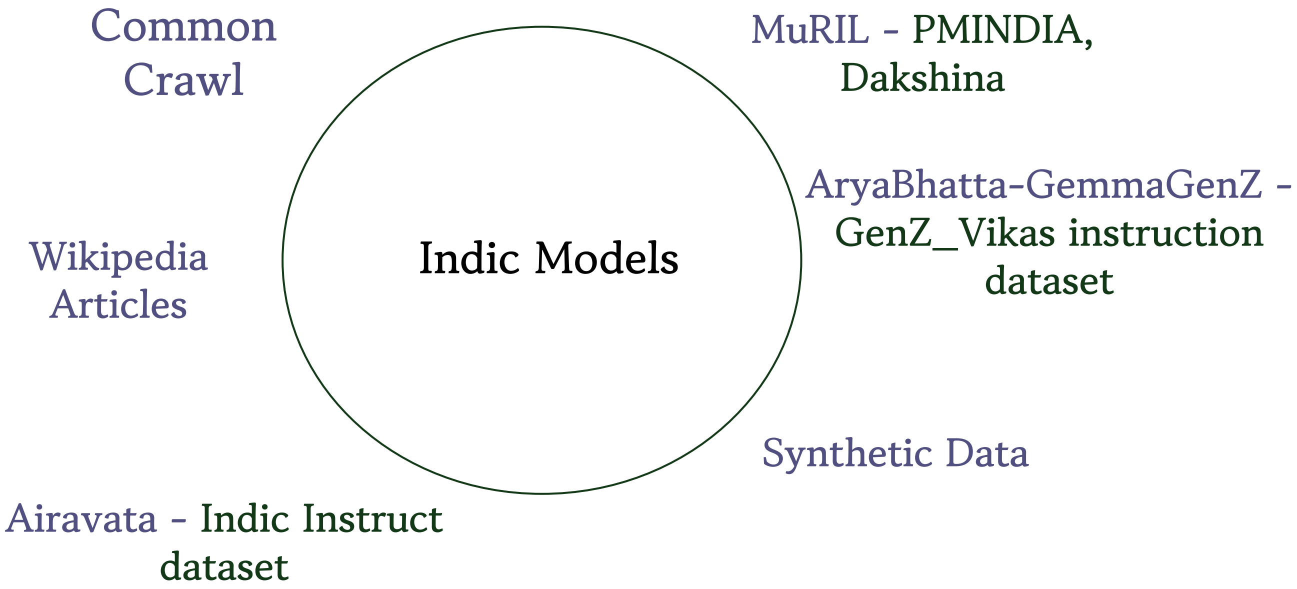
Wikipedia Articles

Airavata - Indic Instruct dataset

MuRIL - PMINDIA, Dakshina

AryaBhatta-GemmaGenZ - GenZ_Vikas instruction dataset

Training Corpora



Indic Models

Common Crawl

MuRIL - PMINDIA, Dakshina

Wikipedia Articles

AryaBhatta-GemmaGenZ - GenZ_Vikas instruction dataset

Airavata - Indic Instruct dataset

Synthetic Data

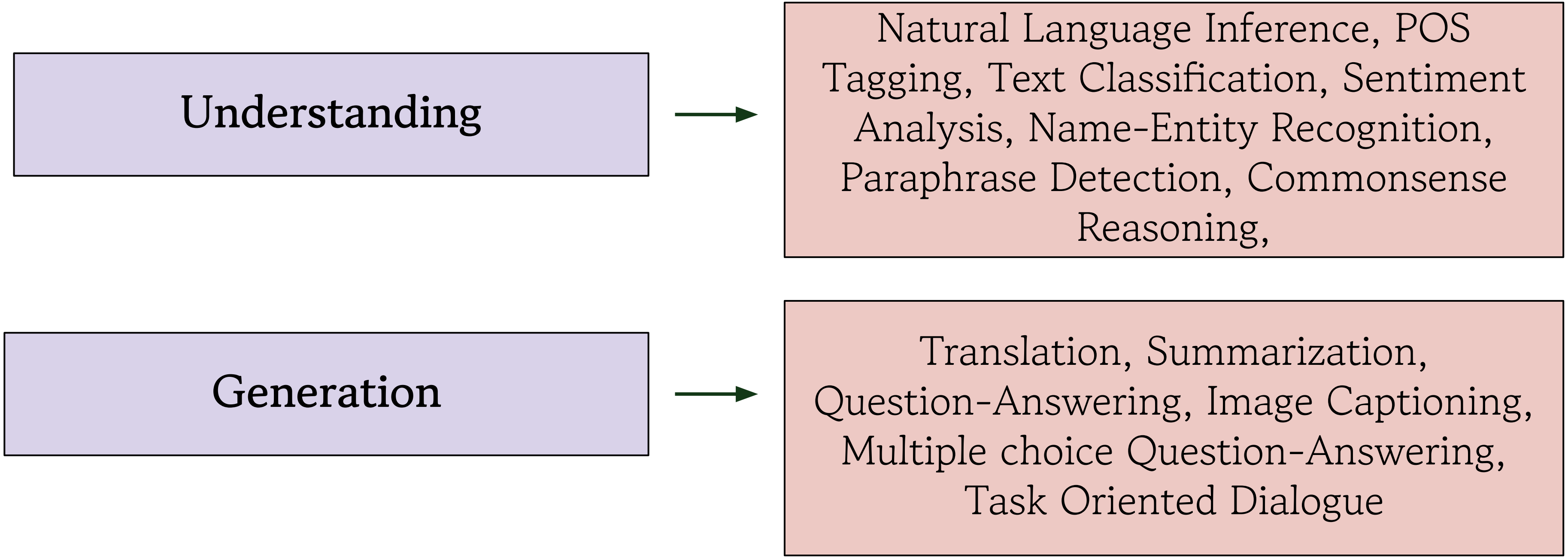
Evaluation

Understanding

Generation

(Ahuja et. al 2024) “MEGAVERSE: Benchmarking Large Language Models Across Languages, Modalities, Models and Tasks” - https://arxiv.org/abs/2311.07463

Evaluation



Understanding

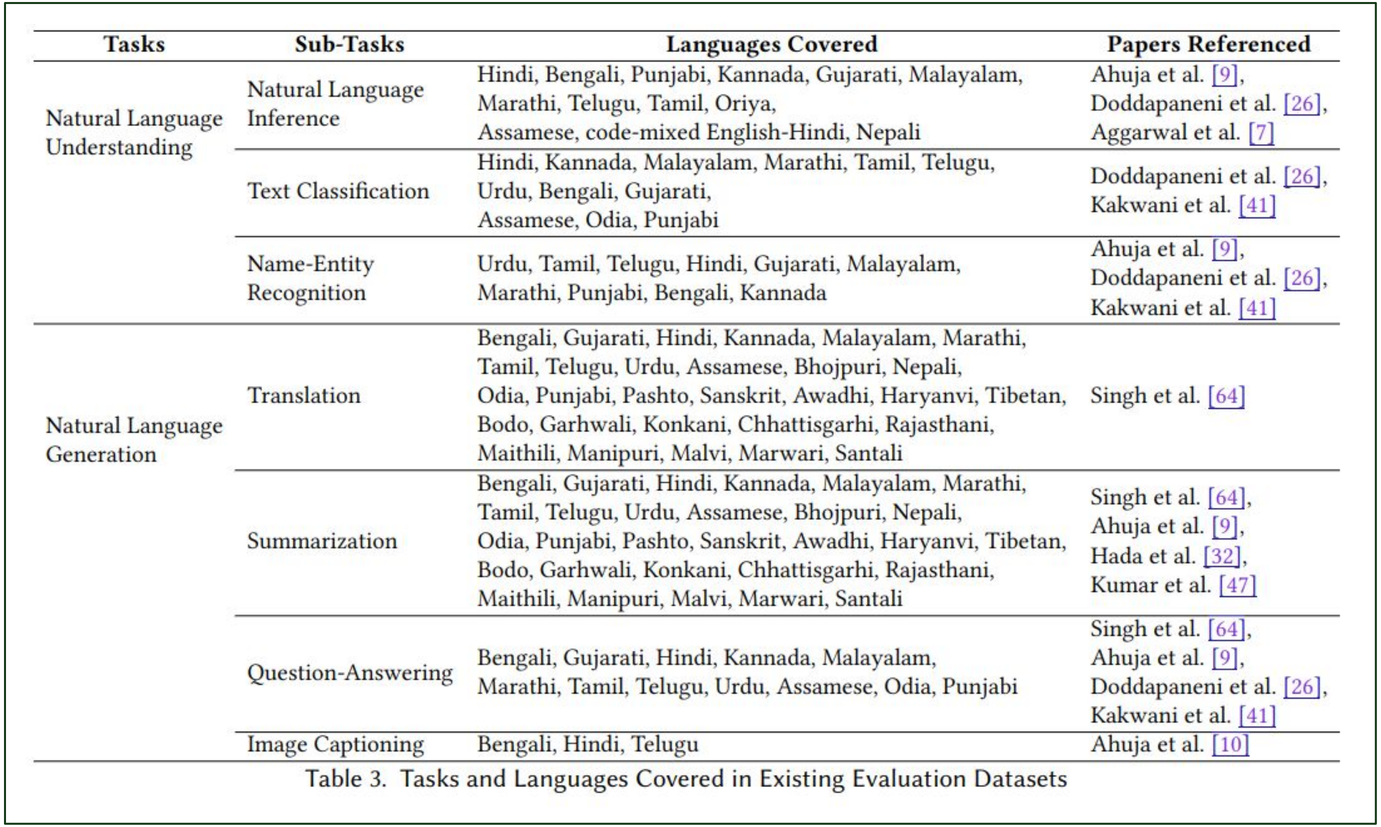
Natural Language Inference, POS Tagging, Text Classification, Sentiment Analysis, Name-Entity Recognition, Paraphrase Detection, Commonsense Reasoning,

Generation

Translation, Summarization, Question-Answering, Image Captioning, Multiple choice Question-Answering, Task Oriented Dialogue

(Ahuja et. al 2024) “MEGAVERSE: Benchmarking Large Language Models Across Languages, Modalities, Models and Tasks” - https://arxiv.org/abs/2311.07463

Evaluation

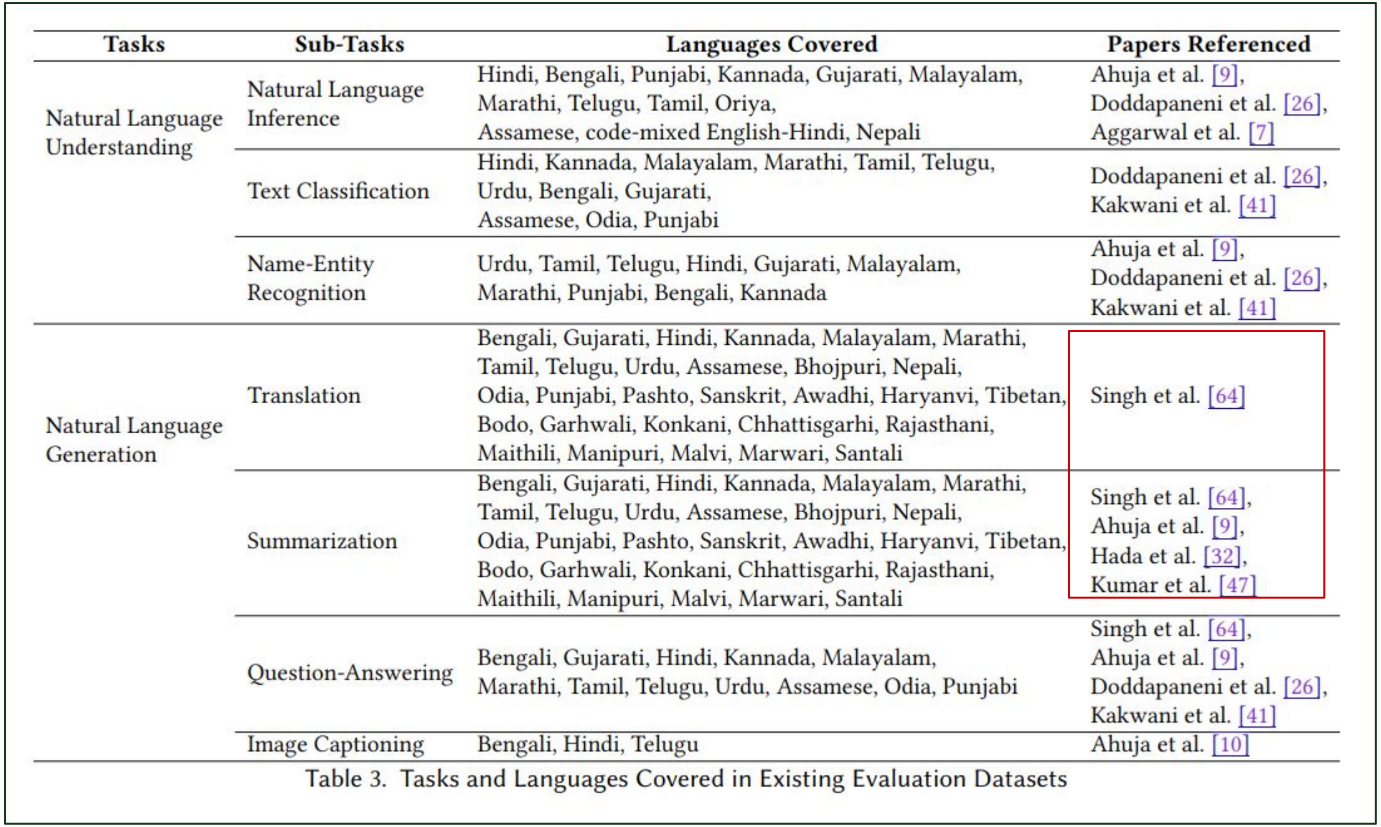


Tasks	Sub-Tasks	Languages Covered	Papers Referenced
Natural Language Understanding	Natural Language Inference	Hindi, Bengali, Punjabi, Kannada, Gujarati, Malayalam, Marathi, Telugu, Tamil, Oriya, Assamese, code-mixed English-Hindi, Nepali	Ahuja et al. [9], Doddapaneni et al. [26], Aggarwal et al. [7]
	Text Classification	Hindi, Kannada, Malayalam, Marathi, Tamil, Telugu, Urdu, Bengali, Gujarati, Assamese, Odia, Punjabi	Doddapaneni et al. [26], Kakwani et al. [41]
	Name-Entity Recognition	Urdu, Tamil, Telugu, Hindi, Gujarati, Malayalam, Marathi, Punjabi, Bengali, Kannada	Ahuja et al. [9], Doddapaneni et al. [26], Kakwani et al. [41]
Natural Language Generation	Translation	Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Tamil, Telugu, Urdu, Assamese, Bhojpuri, Nepali, Odia, Punjabi, Pashto, Sanskrit, Awadhi, Haryanvi, Tibetan, Bodo, Garhwali, Konkani, Chhattisgarhi, Rajasthani, Maithili, Manipuri, Malvi, Marwari, Santali	Singh et al. [64]
	Summarization	Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Tamil, Telugu, Urdu, Assamese, Bhojpuri, Nepali, Odia, Punjabi, Pashto, Sanskrit, Awadhi, Haryanvi, Tibetan, Bodo, Garhwali, Konkani, Chhattisgarhi, Rajasthani, Maithili, Manipuri, Malvi, Marwari, Santali	Singh et al. [64], Ahuja et al. [9], Hada et al. [32], Kumar et al. [47]
	Question-Answering	Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Tamil, Telugu, Urdu, Assamese, Odia, Punjabi	Singh et al. [64], Ahuja et al. [9], Doddapaneni et al. [26], Kakwani et al. [41]
	Image Captioning	Bengali, Hindi, Telugu	Ahuja et al. [10]

Table 3. Tasks and Languages Covered in Existing Evaluation Datasets

Evaluation

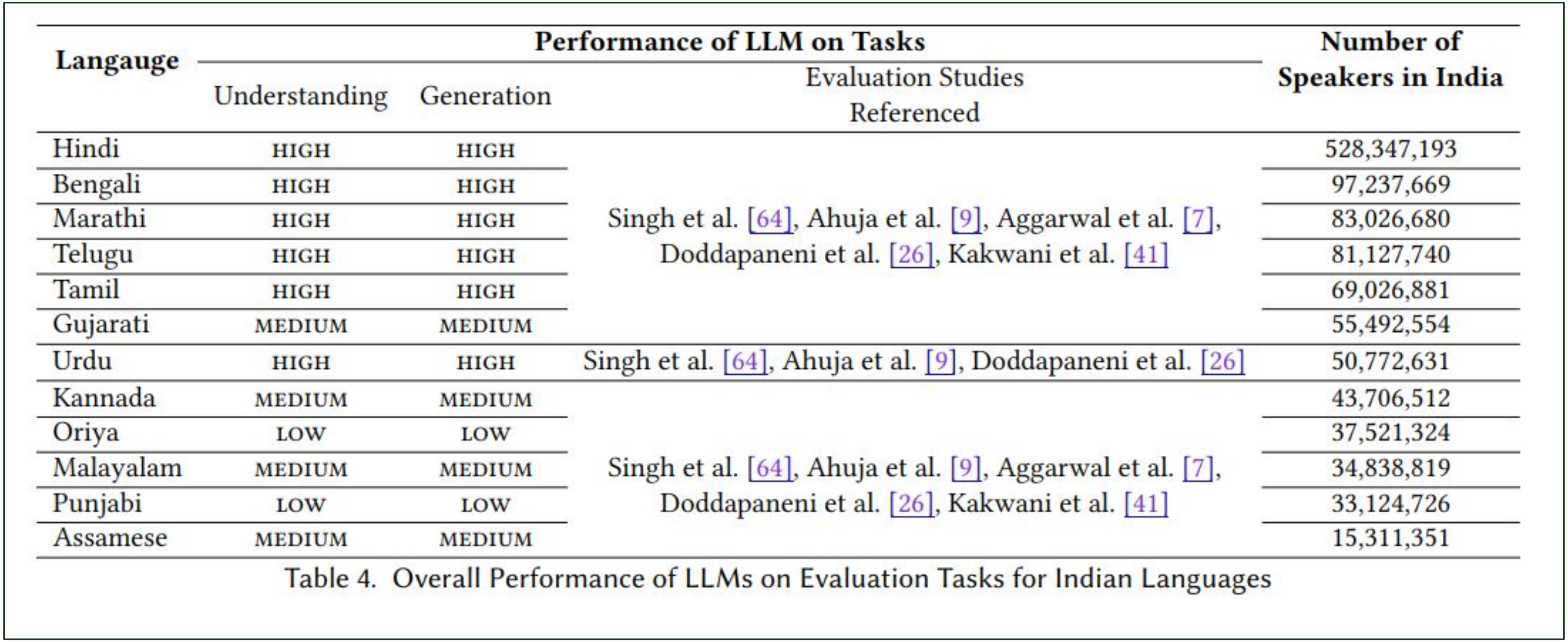
Majority of Evaluation Datasets support
12 most widely spoken Indian Languages



Tasks	Sub-Tasks	Languages Covered	Papers Referenced
Natural Language Understanding	Natural Language Inference	Hindi, Bengali, Punjabi, Kannada, Gujarati, Malayalam, Marathi, Telugu, Tamil, Oriya, Assamese, code-mixed English-Hindi, Nepali	Ahuja et al. [9], Doddapaneni et al. [26], Aggarwal et al. [7]
	Text Classification	Hindi, Kannada, Malayalam, Marathi, Tamil, Telugu, Urdu, Bengali, Gujarati, Assamese, Odia, Punjabi	Doddapaneni et al. [26], Kakwani et al. [41]
	Name-Entity Recognition	Urdu, Tamil, Telugu, Hindi, Gujarati, Malayalam, Marathi, Punjabi, Bengali, Kannada	Ahuja et al. [9], Doddapaneni et al. [26], Kakwani et al. [41]
Natural Language Generation	Translation	Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Tamil, Telugu, Urdu, Assamese, Bhojpuri, Nepali, Odia, Punjabi, Pashto, Sanskrit, Awadhi, Haryanvi, Tibetan, Bodo, Garhwali, Konkani, Chhattisgarhi, Rajasthani, Maithili, Manipuri, Malvi, Marwari, Santali	Singh et al. [64]
	Summarization	Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Tamil, Telugu, Urdu, Assamese, Bhojpuri, Nepali, Odia, Punjabi, Pashto, Sanskrit, Awadhi, Haryanvi, Tibetan, Bodo, Garhwali, Konkani, Chhattisgarhi, Rajasthani, Maithili, Manipuri, Malvi, Marwari, Santali	Singh et al. [64], Ahuja et al. [9], Hada et al. [32], Kumar et al. [47]
	Question-Answering	Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Tamil, Telugu, Urdu, Assamese, Odia, Punjabi	Singh et al. [64], Ahuja et al. [9], Doddapaneni et al. [26], Kakwani et al. [41]
	Image Captioning	Bengali, Hindi, Telugu	Ahuja et al. [10]

Table 3. Tasks and Languages Covered in Existing Evaluation Datasets

Evaluation



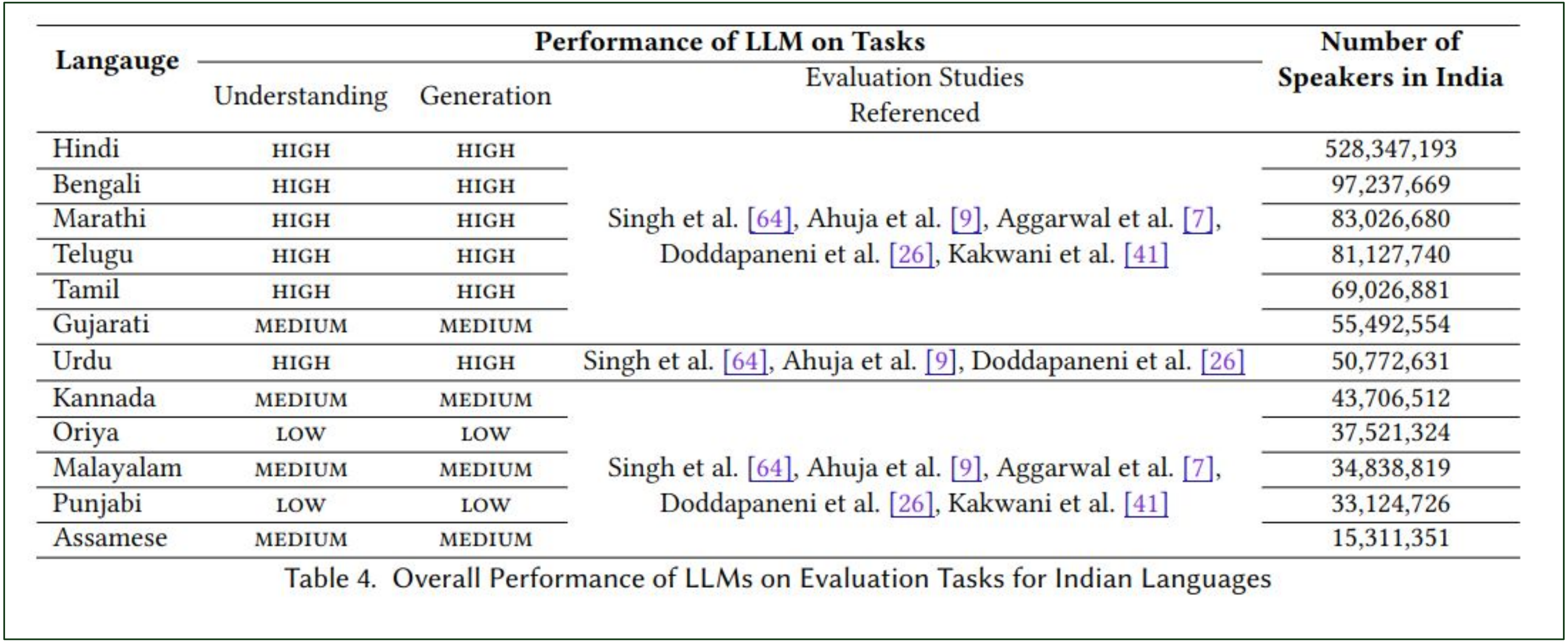
Langauge	Performance of LLM on Tasks			Number of Speakers in India
	Understanding	Generation	Evaluation Studies Referenced	
Hindi	HIGH	HIGH	Singh et al. [64], Ahuja et al. [9], Aggarwal et al. [7], Doddapaneni et al. [26], Kakwani et al. [41]	528,347,193
Bengali	HIGH	HIGH		97,237,669
Marathi	HIGH	HIGH		83,026,680
Telugu	HIGH	HIGH		81,127,740
Tamil	HIGH	HIGH		69,026,881
Gujarati	MEDIUM	MEDIUM		55,492,554
Urdu	HIGH	HIGH	Singh et al. [64], Ahuja et al. [9], Doddapaneni et al. [26]	50,772,631
Kannada	MEDIUM	MEDIUM	Singh et al. [64], Ahuja et al. [9], Aggarwal et al. [7], Doddapaneni et al. [26], Kakwani et al. [41]	43,706,512
Oriya	LOW	LOW		37,521,324
Malayalam	MEDIUM	MEDIUM		34,838,819
Punjabi	LOW	LOW		33,124,726
Assamese	MEDIUM	MEDIUM		15,311,351

Table 4. Overall Performance of LLMs on Evaluation Tasks for Indian Languages

No: of Speakers Source - https://en.wikipedia.org/wiki/Languages_of_India

Evaluation

⚠ This ranking is relative to each other and NOT universal



Langauge	Performance of LLM on Tasks			Number of Speakers in India
	Understanding	Generation	Evaluation Studies Referenced	
Hindi	HIGH	HIGH	Singh et al. [64], Ahuja et al. [9], Aggarwal et al. [7], Doddapaneni et al. [26], Kakwani et al. [41]	528,347,193
Bengali	HIGH	HIGH		97,237,669
Marathi	HIGH	HIGH		83,026,680
Telugu	HIGH	HIGH		81,127,740
Tamil	HIGH	HIGH		69,026,881
Gujarati	MEDIUM	MEDIUM		55,492,554
Urdu	HIGH	HIGH	Singh et al. [64], Ahuja et al. [9], Doddapaneni et al. [26]	50,772,631
Kannada	MEDIUM	MEDIUM	Singh et al. [64], Ahuja et al. [9], Aggarwal et al. [7], Doddapaneni et al. [26], Kakwani et al. [41]	43,706,512
Oriya	LOW	LOW		37,521,324
Malayalam	MEDIUM	MEDIUM		34,838,819
Punjabi	LOW	LOW		33,124,726
Assamese	MEDIUM	MEDIUM		15,311,351

Table 4. Overall Performance of LLMs on Evaluation Tasks for Indian Languages

No: of Speakers Source - https://en.wikipedia.org/wiki/Languages_of_India

Takeaways

- **Evaluation for low-resource languages is task, language and context specific.**
  - Difficult to generalise about coverage across tasks and datasets.

Takeaways

- **Evaluation for low-resource languages is task, language and context specific.**
  - Difficult to generalise about coverage across tasks and datasets.

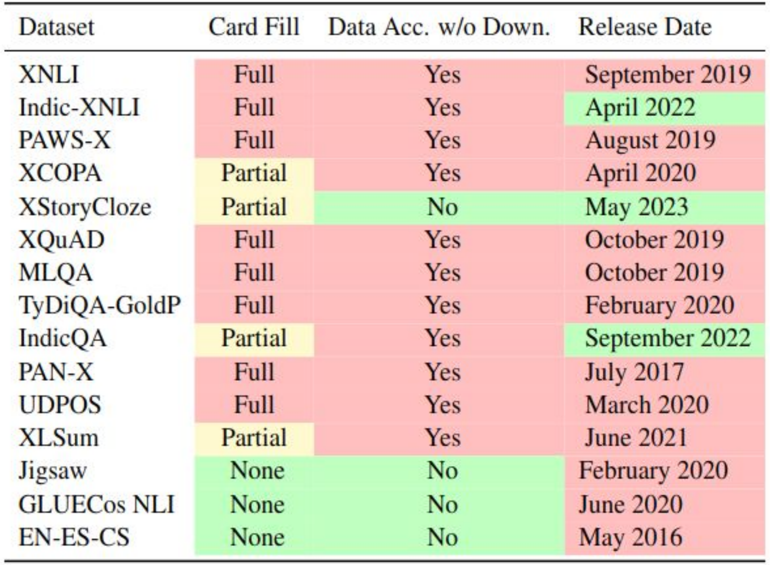
- **The performance of models in a language is not directly correlated with the number of people who speak the language.**

Takeaways

- **Evaluation for low-resource languages is task, language and context specific.**
    - Difficult to generalise about coverage across tasks and datasets.

- **The performance of models in a language is not directly correlated with the number of people who speak the language.**

- **Data Contamination** ([Ahuja et al. 2023](https://arxiv.org/pdf/2303.12528))



Dataset	Card Fill	Data Acc. w/o Down.	Release Date
XNLI	Full	Yes	September 2019
Indic-XNLI	Full	Yes	April 2022
PAWS-X	Full	Yes	August 2019
XCOPA	Partial	Yes	April 2020
XStoryCloze	Partial	No	May 2023
XQuAD	Full	Yes	October 2019
MLQA	Full	Yes	October 2019
TyDiQA-GoldP	Full	Yes	February 2020
IndicQA	Partial	Yes	September 2022
PAN-X	Full	Yes	July 2017
UDPOS	Full	Yes	March 2020
XLSum	Partial	Yes	June 2021
Jigsaw	None	No	February 2020
GLUECos NLI	None	No	June 2020
EN-ES-CS	None	No	May 2016

Conference on Language Modeling, “*Multilinguality and LLMs Special Session*”
Table Image Source - “*MEGA: Multilingual Evaluation of Generative AI* -
”https://arxiv.org/pdf/2303.12528

Takeaways

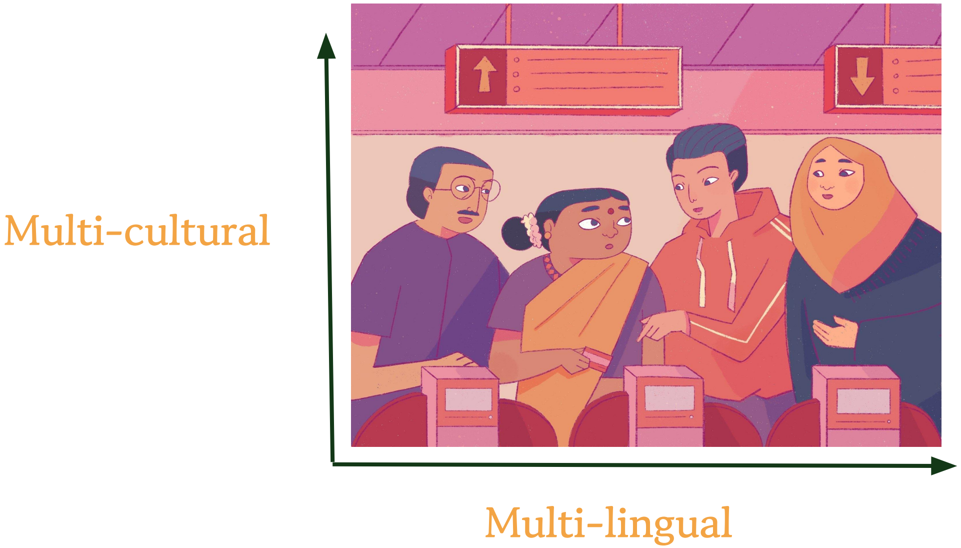
- Evaluation has been 1-D so far

Multi-lingual

Takeaways

- Evaluating Indian language performance effectively requires a multi-cultural and contextual approach

Multi-lingual



Multi-cultural

Multi-lingual

Community-led participatory approach

Acknowledgments

Tattle

Supported and Funded By

ML Commons

arXiv

Analysis of Indic Language Capabilities in LLMs