INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD

Bhashik Language Resources

BhashaVerse

LTRC, IIIT Hyderabad

Why do large language models work?

Model Size

Training Data

Scaling compute improves loss

Possible Reason ??

Scaling Laws for Neural Language Models
https://arxiv.org/pdf/2001.08361

Datasets

The Pile is a dataset of 825GB of text collected from various sources (e.g., books, Web scrapes, open source code)

Common Crawl weighs in at some 250+TB.

Just 1% of that web data is usable text (it's likely much more)

it's still 2.5+TB.

Red Pajama-Data-v2

	# Documents	Estimated Token count (deduped)
en	14.5B	20.5T
de	1.9B	3.0T
fr	1.6B	2.7T
es	1.8B	2.8T
it	0.9B	1.5T
Total	20.8B	30.4T

The English Wikipedia is around 40 GB of text

Component	Raw Size	Weight	Epochs	Effective Size	Mean Document Size
Pile-CC	227.12 GiB	18.11%	1.0	227.12 GiB	4.33 KiB
PubMed Central	90.27 GiB	14.40%	2.0	180.55 GiB	30.55 KiB
Books3†	100.96 GiB	12.07%	1.5	151.44 GiB	538.36 KiB
OpenWebText2	62.77 GiB	10.01%	2.0	125.54 GiB	3.85 KiB
ArXiv	56.21 GiB	8.96%	2.0	112.42 GiB	46.61 KiB
Github	95.16 GiB	7.59%	1.0	95.16 GiB	5.25 KiB
FreeLaw	51.15 GiB	6.12%	1.5	76.73 GiB	15.06 KiB
Stack Exchange	32.20 GiB	5.13%	2.0	64.39 GiB	2.16 KiB
USPTO Backgrounds	22.90 GiB	3.65%	2.0	45.81 GiB	4.08 KiB
PubMed Abstracts	19.26 GiB	3.07%	2.0	38.53 GiB	1.30 KiB
Gutenberg (PG-19)†	10.88 GiB	2.17%	2.5	27.19 GiB	398.73 KiB
OpenSubtitles†	12.98 GiB	1.55%	1.5	19.47 GiB	30.48 KiB
Wikipedia (en)†	6.38 GiB	1.53%	3.0	19.13 GiB	1.11 KiB
DM Mathematics†	7.75 GiB	1.24%	2.0	15.49 GiB	8.00 KiB
Ubuntu IRC	5.52 GiB	0.88%	2.0	11.03 GiB	545.48 KiB
BookCorpus2	6.30 GiB	0.75%	1.5	9.45 GiB	369.87 KiB
EuroParl†	4.59 GiB	0.73%	2.0	9.17 GiB	68.87 KiB
HackerNews	3.90 GiB	0.62%	2.0	7.80 GiB	4.92 KiB
YoutubeSubtitles	3.73 GiB	0.60%	2.0	7.47 GiB	22.55 KiB
PhilPapers	2.38 GiB	0.38%	2.0	4.76 GiB	73.37 KiB
NIH ExPorter	1.89 GiB	0.30%	2.0	3.79 GiB	2.11 KiB
Enron Emails†	0.88 GiB	0.14%	2.0	1.76 GiB	1.78 KiB
The Pile	**825.18 GiB**			**1254.20 GiB**	**5.91 KiB**

https://arxiv.org/pdf/2101.00027

LLMs

GPT-3 training data[1]:9

Dataset	# tokens	Proportion within training
Common Crawl	410 billion	60%
WebText2	19 billion	22%
Books1	12 billion	8%
Books2	55 billion	8%
Wikipedia	3 billion	3%

Llama-2

2,000,000,000,000 Tokens

(2 trilian)

7B to 70B

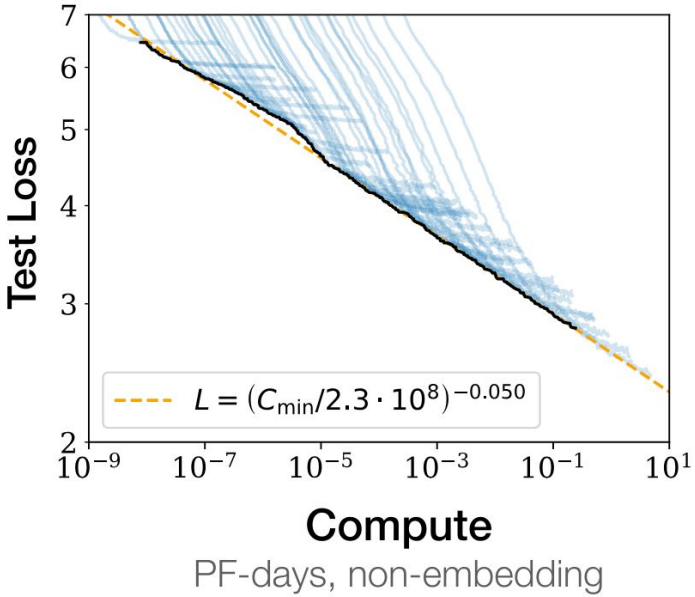
Llama-3

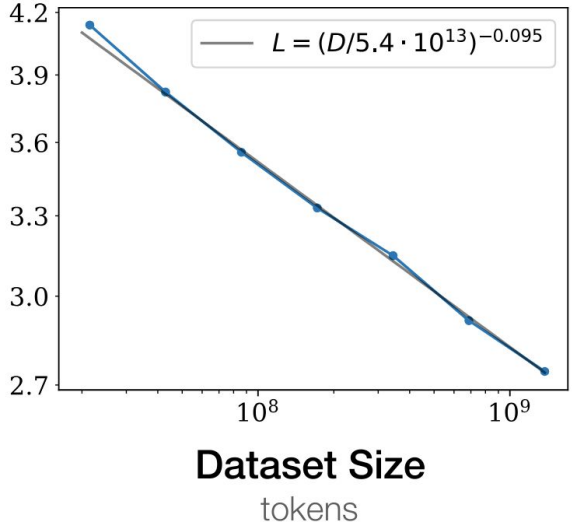
15,000,000,000,000 Tokens

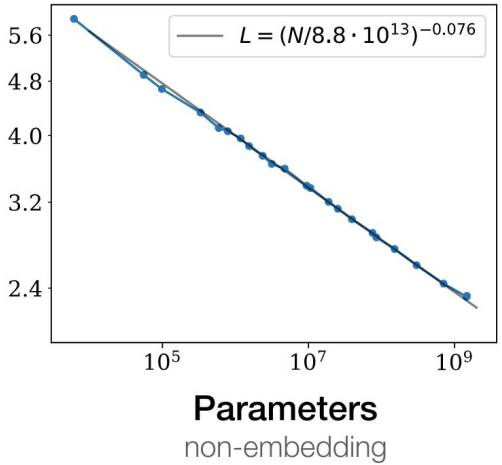
(15 trilian)

8B and 70B

Why do large language models work? Scaling







Scaling Laws for Neural Language Models
https://arxiv.org/pdf/2001.08361

Indian Languages ?

Indian Languages ?

Language Family	Indo-Aryan														
Language	**asm**	**ban**	**kas**	**snd**	**urd**	**doi**	**hin**	**gom**	**mai**	**mar**	**nep**	**san**	**guj**	**odi**	**pan**
Language Script	Bangla		Perso-Arabic			Devanagari							Gujarati	Odia	Gurmukhi
No of Letters in Unicode	96		256			128							91	91	80
Models (Vocab)															
BLOOM (250680)	(48,48)		(49,207)			(67,61)							(57,34)	(56,35)	(55,25)
FALCON (65024)	(00,96)		(12,244)			(2,126)							(00,91)	(00,91)	(00,72)
LLAMA-1,2 (32024)	(24,72)		(45,211)			(38,90)							(01,90)	(00,91)	(04,76)
MISTRAL (32052)	(34,62)		(47,209)			(43,85)							(05,86)	(00,91)	(02,78)
MPT (50277)	(05,91)		(35,221)			(22,106)							(02,89)	(00,91)	(00,80)
OPT (50265)	(00,96)		(13,243)			(1,127)							(00,91)	(00,91)	(00,80)

Dravidian				Sino-Tibetan		Austroasiatic
kan	**mal**	**tam**	**tel**	**mni**	**brx**	**sat**
Kannada	Malayalam	Tamil	Telugu	Meitei	Devanagari	Ol Chik
91	118	72	100	56	96	48
(62,29)	(66,52)	(46,26)	(61,39)	(00,56)	(67,29)	(00,48)
(0,100)	(00,56)	(02,70)	(04,96)	(00,56)	(02,94)	(00,48)
(02,89)	(33,155)	(19,53)	(01,99)	(00,56)	(38,90)	(00,48)
(18,73)	(04,116)	(22,50)	(11,89)	(00,56)	(43,53)	(00,48)
(00,91)	(01,117)	(05,67)	(03,97)	(00,56)	(22,106)	(00,48)
(00,91)	(0,118)	(00,72)	(0,100)	(00,56)	(01,95)	(00,48)

Llama 3.2

has all Indic vocab

Large Language Model for Machine Translation ecosystem

- Machine Translation ?
- Machine Translation Evaluation
- Machine Translation Post Editing
- Machine Translation Error Identification

36 Indian Subcontinent languages

Assamese, Awadhi, Bengali, Bhojpuri, Braj, Bodo, Dogri, English, Konkani, Gondi, Gujarati, Hindi, Hinglish, Ho, Kannada, Kangri, Kashmiri (Arabic and Devanagari), Khasi, Mizo, Magahi, Maithili, Malayalam, Marathi, Manipuri (Bengali and Meitei), Nepali, Oriya, Punjabi, Sanskrit, Santali, Sinhala, Sindhi (Arabic and Devanagari), Tamil, Tulu, Telugu, and Urdu

What about Indian Languages?

Monolingual Corpora ?

Task Specific Corpora ?

What about Indian Languages?

Monolingual Corpora ?

Up to 2,3,4T tokens

Task Specific Corpora ?

Parallel Corpora

		Existing					BPCC (Newly Added)			
		Mined		Human			Mined		Human	
Name	*Language*	Samanantar	NLLB	NLLB	ILCI	MASSIVE	Monolingual	Comparable	Wiki	Daily
Assamese	*asm_Beng*	58.8	506.3	-	82.1	-	712.5	37.8	44.7	11.3
Bengali	*ben_Beng*	2,946.3	13,580.5	-	123.8	16.5	16,055.1	258.2	48.0	8.5
Bodo	*brx_Deva*	-	-	-	83.2	-	-	<1	22.7	10.3
Dogri	*doi_Deva*	-	-	-	-	-	-	-	18.7	5.5
Konkani	*gom_Deva*	-	-	-	74.5	-	-	-	18.3	4.8
Gujarati	*guj_Gujr*	1,379.2	7,090.3	-	107.4	-	11,630.3	573.0	25.0	3.2
Hindi	*hin_Deva*	4,416.7	6,646.7	-	165.6	16.5	27,187.8	853.3	40.3	8.4
Kannada	*kan_Knda*	1,692.2	8,871.1	-	76.4	16.5	12,501.0	380.2	32.2	8.5
Kashmiri	*kas_Arab*	-	124.9	6.2	-	-	-	-	15.5	4.3
	kas_Deva	-	194.0	6.2	-	-	-	-	-	-
Maithili	*mai_Deva*	-	62.2	-	-	-	-	<1	24.4	4.2
Malayalam	*mal_Mlym*	2,029.2	8,818.2	-	87.9	16.5	12,378.6	356.4	41.6	8.4
Marathi	*mar_Deva*	1,366.1	6,393.2	-	117.0	-	10,806.0	432.4	54.3	4.6
Manipuri	*mni_Beng*	-	346.9	6.2	13.1	-	-	20.1	-	<1
	mni_Mtei	-	-	-	16.0	-	-	-	19.9	6.8
Nepali	*npi_Deva*	-	1,583.5	-	28.6	-	10.5	6.2	45.9	10.9
Odia	*ory_Orya*	514.9	2,382.6	-	-	-	2,863.1	121.5	33.7	3.2
Punjabi	*pan_Guru*	1,418.3	1,978.3	-	71.5	-	6,275.8	207.2	6.3	3.2
Sanskrit	*san_Deva*	-	244.1	-	-	-	-	<1	27.7	5.4
Santali	*sat_Olck*	-	-	-	-	-	-	-	22.5	1.8
Sindhi	*snd_Arab*	-	2,128.4	-	-	-	-	-	-	-
	snd_Deva	-	-	-	-	-	-	-	10.5	-
Tamil	*tam_Taml*	1,833.2	8,665.2	-	120.7	16.5	9,690.3	452.8	21.0	8.6
Telugu	*tel_Telu*	1,780.5	10,062.8	-	73.6	16.5	11,100.0	437.2	29.7	8.5
Urdu	*urd_Arab*	-	5,321.0	-	101.0	16.5	484.9	225.3	41.3	8.4
# Total		19,435.4	84,998.3	18.6	1,342.6	115.4	121,695.8	4,353.1	644.3	139.7

Indictrans2
Ai4bharat data

Bhashik:
Corpora for Indian Languages

Machine Translation

- Bhashik Generic
  - 10B Parallel Corpora for 36*36 language pairs
  - Quality Synthetic Corpora
  - Also, for ILs to ILs
- Bhashik Education (Human)
  - 2M Parallel Corpora for 17 domains for English and 5 ILs
  - 3 more languages (coming soon)
- Bhashik Health (Human)
  - 0.5M Parallel Corpora for Medical Domain for English and 8 ILs
- Bhashik HimangY IL-IL (Human)(coming soon)
  - 0.8M Parallel corpora for Hindi to 11 Language pairs

Bhashik:
Corpora for Indian Languages

Machine Translation Ecosystem

- Bhashik Automatic post editing
  - 2M corpora for English and 5 ILs (Human)
  - 8M pseudo APE corpora for English and 22 ILs

- Bhashik Machine Translation Evaluation (with/without reference)
  - 8M pseudo corpora for Direct evaluation
  - 0.1M corpora for Direct evaluation (Human) (Coming Soon)
  - 0.2M HimangY direct assessment corpora (Human) (Coming Soon)

BhashaVerse : Translation Ecosystem for Indian Subcontinent Languages
Vandan Mujadia and Dipti Misra Sharma

Bhashik:
Corpora for Indian Languages

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD

Machine Translation Ecosystem

- Bhashik MT Error Identification
  - 2M Parallel Corpora for English and 5 ILs (Human)
  - 8M pseudo APE corpora for English and 22 ILs
  - 100K MQM HimangY corpora (Human)(Coming Soon)

BhashaVerse : Translation Ecosystem for Indian Subcontinent Languages
Vandan Mujadia and Dipti Misra Sharma

How !

- Web crawl
  - 36 Indian Languages
  - News, wiki, books, parallel corpora, cleaned corpora
  - 2-3T tokens
- Pretraining
  - Encoder-Decoder
  - Decoder (Mixtral)
- Full Fine Tuning
  -

How !

- Data
  - 36 Indian Languages
  - News, wiki, books, parallel corpora, cleaned corpora
  - 2T-3T tokens
- Pretraining
  - Encoder-Decoder
    - Purtabation
      - Add/Delete/Replace Random Token
      - Change Pronoun
      - Change Prepositions or Postpositions
  - Decoder (Mixtral)
    - Next token prediction
- Full Fine Tuning
  -

How !

BhashaVerse : Translation Ecosystem for Indian Subcontinent Languages

- https://arxiv.org/pdf/2412.04351
- https://ssmt.iiit.ac.in/bhashaverse

BhashaVerse:
Model for Indian Language Translation Ecosystem

Single Multitask Model

- 36*36 language Machine Translation
  - Supports Discourse Translation (1024 token context)
- Machine Translation Evaluation (Direct Assessment; with and without reference)
- Machine Translation Error Identification
- Automatic Post Editing

- From scratch; ~ 2B parameters; Encoder-Decoder model
- Runs on a smaller GPU with SOTA performance
- Can be used for fine tuning

BhashaVerse : Translation Ecosystem for Indian Subcontinent Languages
Vandan Mujadia and Dipti Misra Sharma

Powered by : PARAM Siddhi-AI, Bhashini, MEITY, GOI

BhashaVerse:
BhashaverseLLM for 36 Indian Languages

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY HYDERABAD

Pretrained LLM

- 36*36 language Machine Translation
  - Supports Discourse Translation (2048 token context)
- Automatic Post Editing

- From scratch Mixtral; ~ 4x4B parameters; Up to ~2-3T tokens
- Decoder only model
- Can be used for Indian Language Generation Tasks with finetuning

BhashaVerse : Translation Ecosystem for Indian Subcontinent Languages
Vandan Mujadia and Dipti Misra Sharma

Powered by : PARAM Siddhi-AI, Bhashini, MEITY, GOI

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD

Releasing Bhashik Language Resources

Corpora

Bhashik Translation Corpora

By LTRC, IIIT Hyderabad
https://huggingface.co/ltrciiith

Model

Bhashaverse MultiTask Models

By LTRC, IIIT Hyderabad
https://ssmt.iiit.ac.in/bhashaverse
https://github.com/ltrc/onemtbhashaverse