Multilinguality in Large Language Models

Nasredine SEMMAR

CEA-List – LASTI Laboratory, University of Paris-Saclay

Outline

- Language Modeling and Large Language Models (LLMs)
- Multilinguality in LLMs

■ Language Models

Language Models: Definition and Approaches

- **Statistical Language Modeling (Shannon, 1948; Miller & Selfridge, 1950; Maltese & Mancini, 1992)**
  - Model constructed from a large corpus (composed of sequence of words)
  - Estimates the probability of any given sequence W to occur
  - Approximates the probability of a word given its entire context

$$W = (w_1, w2, \ldots, w_n) \qquad w_i \in V$$

$$P(W) = \prod_{i=1}^{T} p(w_i|h_i)$$

$$h_i = (w_1, w_2, \ldots, w_{i-1})$$

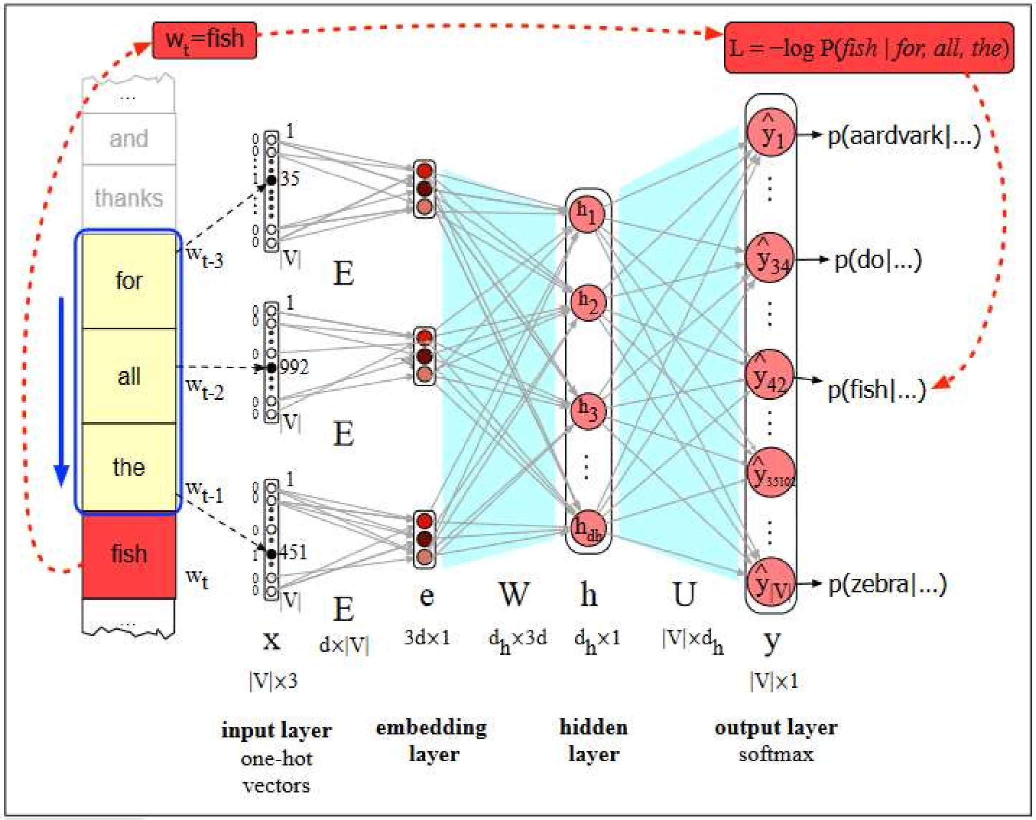
  - uni-gram probability: $p(w_i) = \dfrac{C(w_i)}{\sum_k C(w_k)}$
  - n-gram probability: $p(w_i|h_i^n) = \dfrac{C(h_i^n w_i)}{C(h_i^n)}$
  - Bi-gram model - probability of word pairs: $P(W) = p(w_1) \prod_{i=2}^{T} p(w_i|w_{i-1})$
  - Tri-gram model - probability of 3 words: $P(W) = p(w_1)p(w_2|w_1) \prod_{i=3}^{T} p(w_i|w_{i-1}, w_{i-2})$

Example of bi-grams: the cat is sitting on the mat ➔ the_cat, cat_is, is_sitting, sitting_on, on_the, the_mat

Language Models: Definition and Approaches

- **Neural Language Modeling (Bengio et al., 2003)**
  - Associates each word in the vocabulary with a distributed word feature vector
  - Expresses the joint probability function of word sequences in terms of the feature vectors of these words in the sequence
  - Learns simultaneously the word feature vector and the parameters of the probability function
  - *Learns to predict the next word from a given word sequence*
  - ➔ Neural language models represent words in this prior context by their embeddings, rather than just by their word identity as used in n-gram statistical language models
  - ➔ Using embeddings allows neural language models to generalize better to unseen data
  - ➔ Approximates the probability of a word given the entire prior context by approximating based on the N − 1 previous words:

$$P(w_t|w_1,\ldots,w_{t-1}) \approx P(w_t|w_{t-N+1},\ldots,w_{t-1})$$



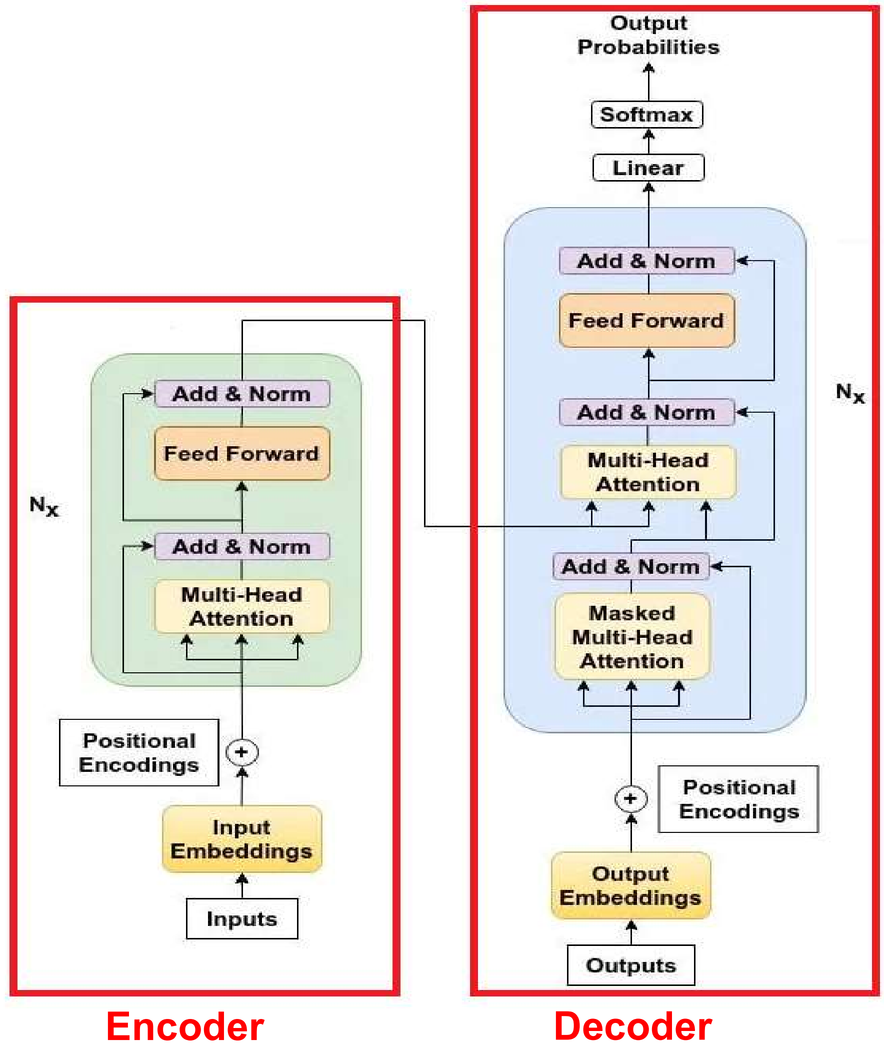
Source: (Jurafsky & Martin, 2023) - Neural Networks and Neural Language Models

Example: … thanks for all the fish ➔ … thanks for all the ____

The Rise of Transformers

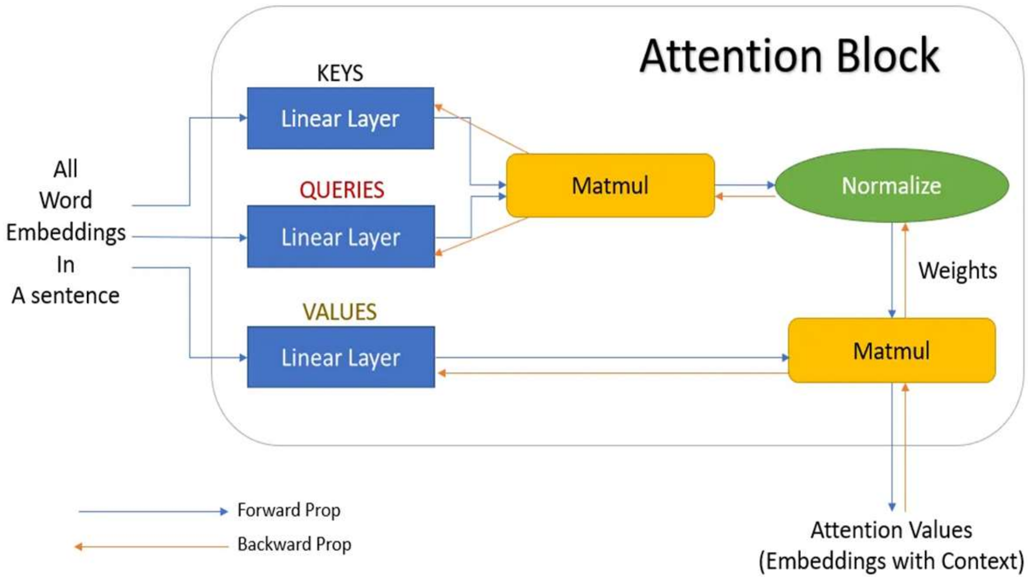
- **Transformer architecture (Vaswani et al., 2017)**
  - The encoder takes in a sequence of tokens and produces a fixed-size vector representation of the entire sequence
  - The decoder takes in a fixed-size vector representation of the context and uses it to generate a sequence of words one at a time, with each word being conditioned on the previously generated words
  - Based on the multi-head attention mechanism
  - The Transformer architecture is suitable for parallel processing of sequential data

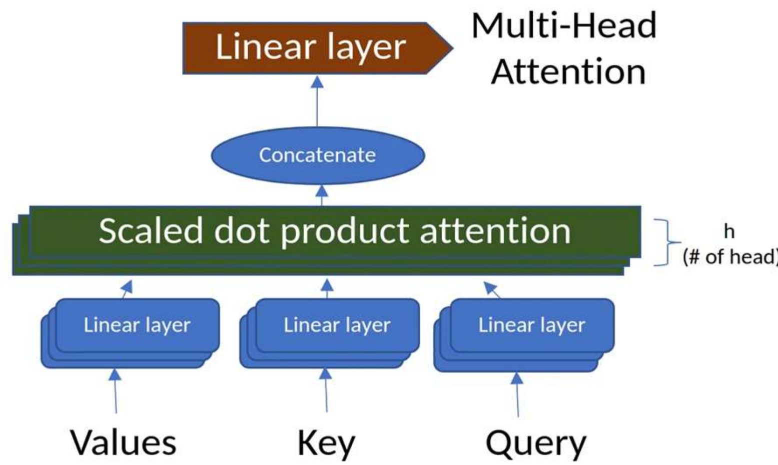
➔ **Faster training**



The Rise of Transformers

- **Attention Mechanism**
  - Allowing the model to focus on different parts of the input sequence independently of their position in the sequence
  - Assigning weights to different parts of the input sequence ➔ Enhancing the understanding of context and relationships



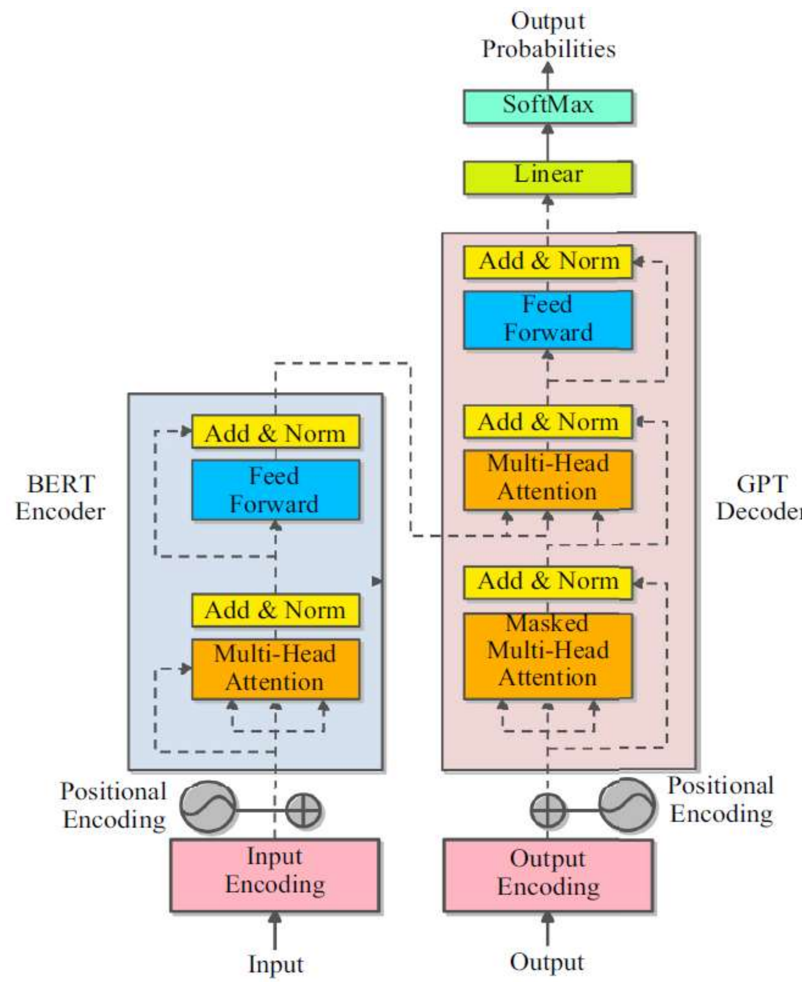


$$\mathrm{Attention}(Q,K,V) = \mathrm{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

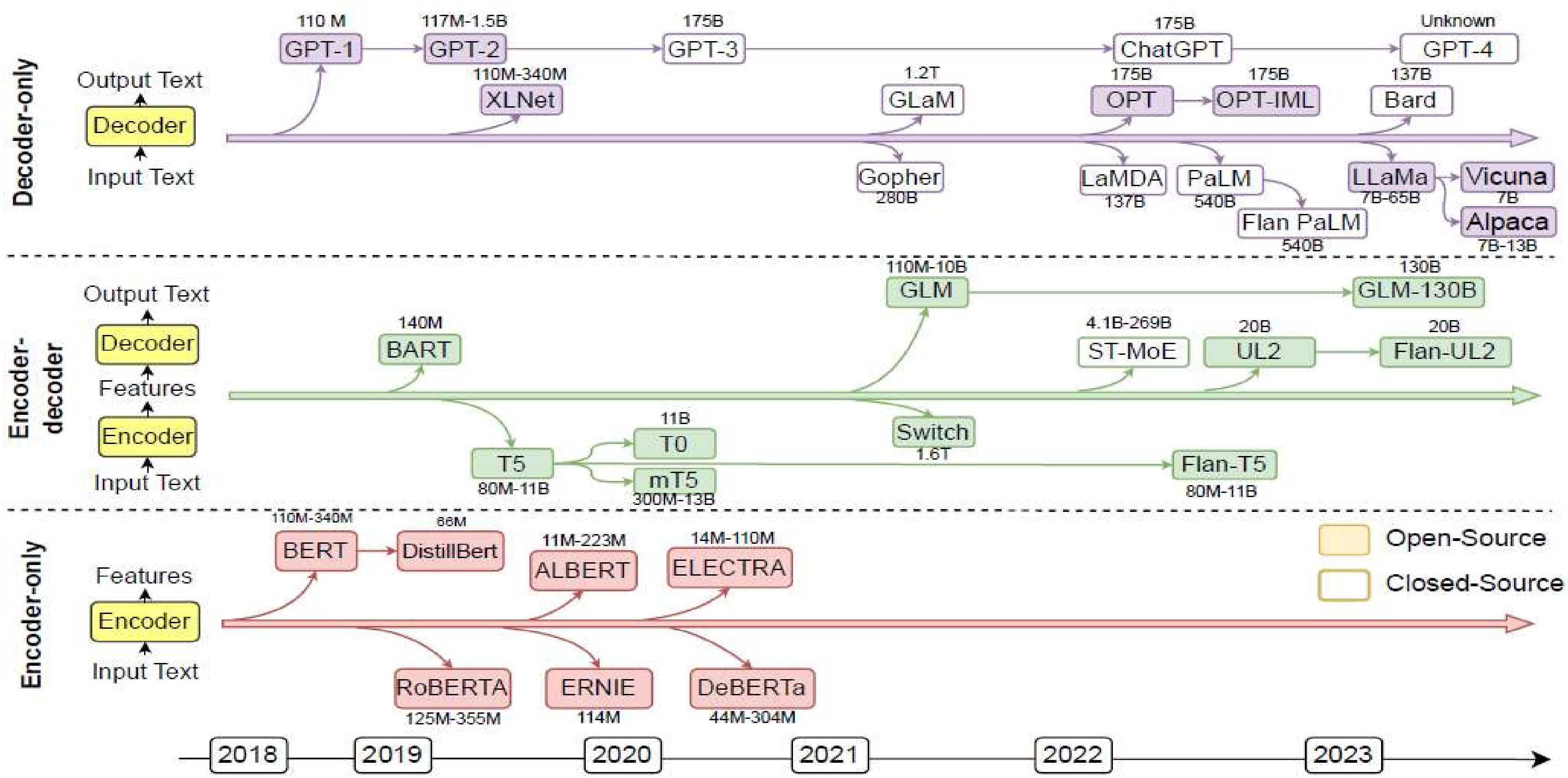
Source: (Arjun Sarkar, 2022) - All you need to know about Attention and Transformers

The Rise of Transformers

- Different language models based on Transformer
  - Encoder only (BERT)
    - Classification tasks: Sentiment analysis, Topic modeling, etc.
    - Sequence-to-sequence labeling tasks: Named entity recognition, Part-Of-Speech tagging, etc.
  - Decoder only (GPT)
    - Generation tasks: Dialogue, Text generation, etc.
  - Encoder-Decoder (T5-Text-to-Text Transfer Transformer, BART)
    - Text transformation: Machine translation, Text summarization, etc.

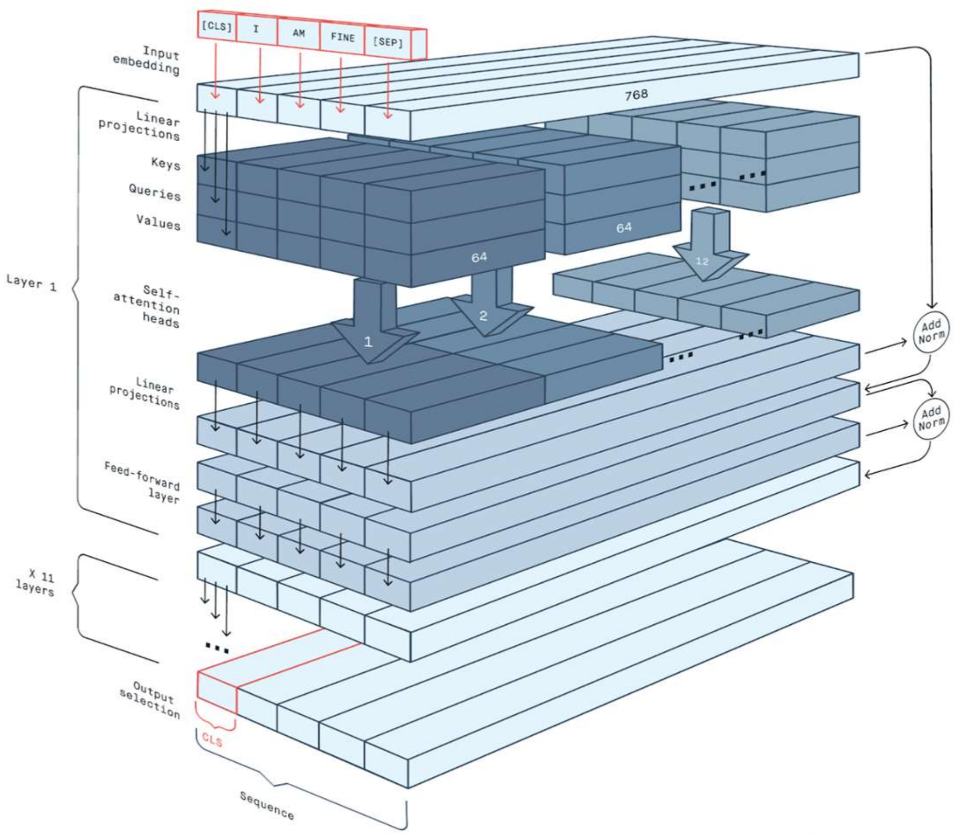


Evolution of PLMs and LLMs based on Transformers over the past five years



Source: (Pan et al., 2024) - Unifying Large Language Models and Knowledge Graphs: A Roadmap

Architecture of BERT (Bidirectional Encoder Representations from Transformers)



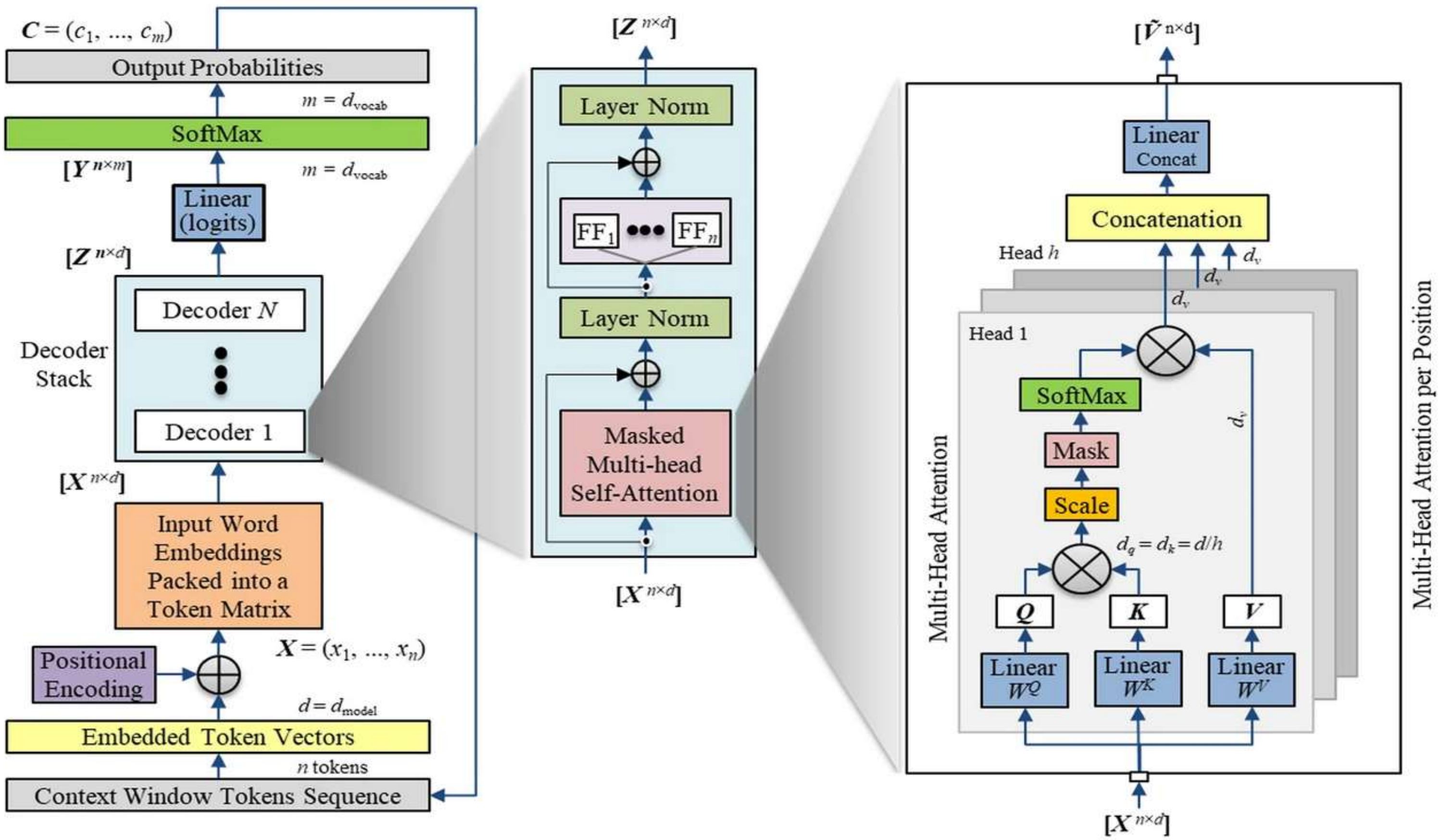
- BERT base
  - 12 Encoder layers
  - 12 Attention heads
  - 110M parameters

- BERT Large
  - 24 Encoder layers
  - 16 Attention heads
  - 340M parameters

- Each input token is represented as a 768 long size vector which is dot multiplied with 12 Key, Query and Value embeddings
- BERT is pre-trained on 3200 million words (Wikipedia+Book)
- Two unsupervised learning objectives:
  - Masked Language Modelling (MLM)
  - Next Sentence Prediction (NSP)

Source: (Peltarion, 2020) - BERT Architecture

Architecture of GPT (Generative Pre-trained Transformer)

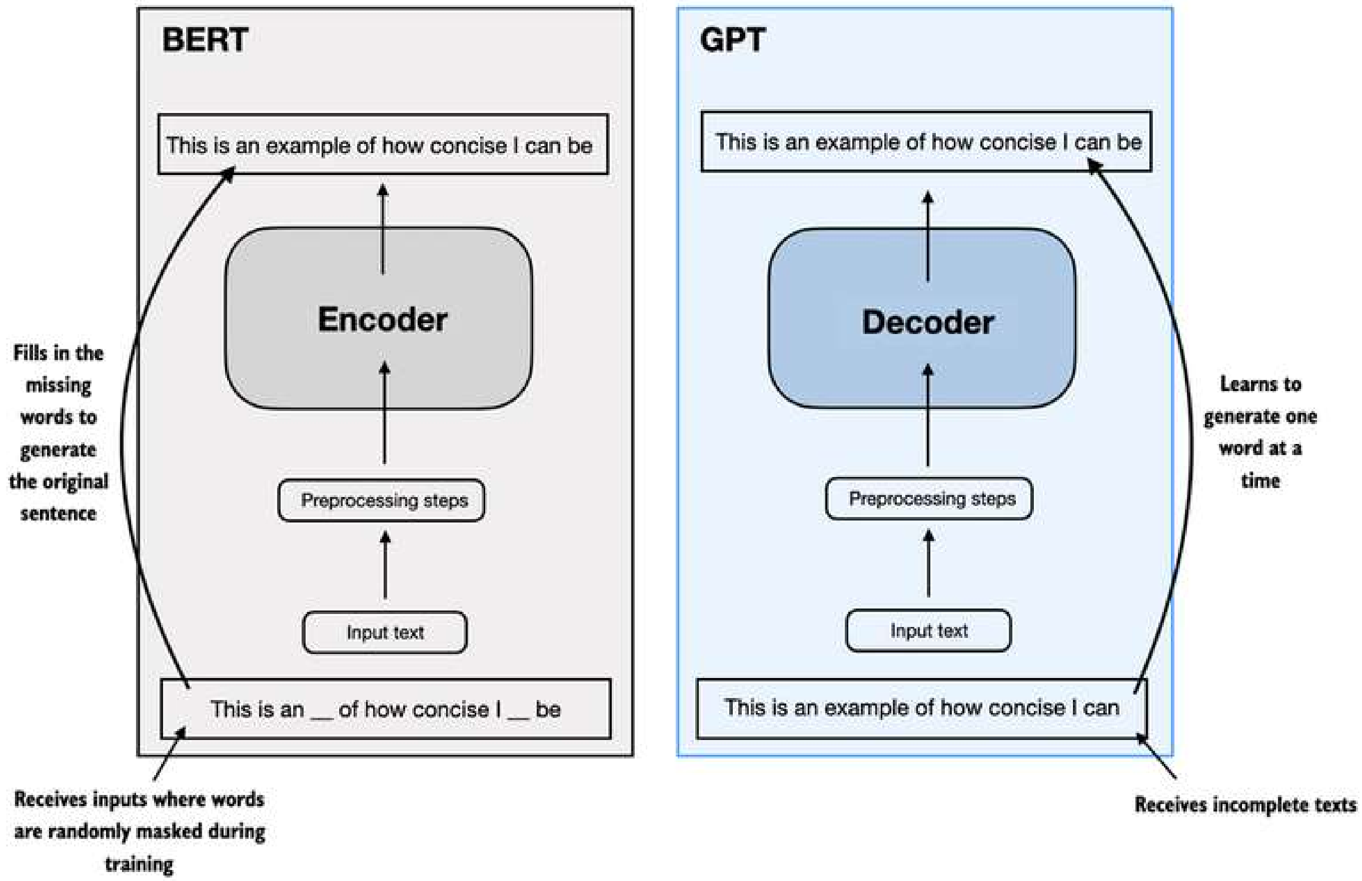


- **GPT-3**
  - Context window size: $n = 2048$
  - Dimension of each token vector: d = 12,288
  - Length of the vocabulary: m = 50,257
  - Multi-headed attention: more than one block of attention mechanism per decoder layer, h = 96 attention heads
  - Decoder: more than one decoder layer, N = 96 layers
  - Feed forward neural network: 2 hidden layers, each with 4 times the number of nodes, 4 × 12,288 = 49,152 nodes

Source: (Bridgelall, 2024) - Unraveling the mysteries of AI chatbots

BERT vs GPT

- **BERT**
  - The encoder focuses on masked word prediction
  - It is used for tasks such as text classification
- **GPT**
  - The decoder produces coherent text sequences
  - It is designed for generative tasks



Source: (Raschka, 2023) - Finetuning LLMs Efficiently with Adapters

BERT vs GPT

	BERT (PLM: Pre-trained Language Model)	GPT (LLM: Large Language Model)
Architecture	BERT is designed for bidirectional representation learning. It uses a masked language model objective, where it predicts missing words in a sentence based on both left and right context.	GPT, is designed for generative language modeling. It predicts the next word in a sentence given the preceding context, utilizing a unidirectional autoregressive approach.
Pre-training Objectives	BERT is pre-trained using a masked language model objective and next sentence prediction. It focuses on capturing bidirectional context and understanding relationships between words in a sentence.	GPT is pre-trained to predict the next word in a sentence, which encourages the model to learn a coherent representation of language and generate contextually relevant sequences.
Context Understanding	BERT is effective for tasks that require a deep understanding of context and relationships within a sentence, such as text classification, named entity recognition, and question-answering.	GPT is strong in generating coherent and contextually relevant text. It is often used in creative tasks, dialogue systems, and tasks requiring the generation of natural language sequences.
Task types and Use Cases	Commonly used in tasks like text classification, named entity recognition, sentiment analysis, and question-answering.	Applied to tasks such as text generation, dialogue systems, summarization, and creative writing.
Fine-tuning vs Few-Shot Learning	BERT is often fine-tuned on specific downstream tasks with labeled data to adapt its pre-trained representations to the task at hand.	GPT is designed to perform few-shot learning, where it can generalize to new tasks with minimal task-specific training data.

Source: Explanation of BERT Model- https://www.geeksforgeeks.org/explanation-of-bert-model-nlp/

Statistics on Large Language Models (1)

	Model	Release Time	Size (B)	Base Model	Adaptation		Pre-train Data Scale	Latest Data Timestamp	Hardware (GPUs / TPUs)	Training Time	Evaluation	
					IT	RLHF					ICL	CoT
	T5 [82]	Oct-2019	11	-	-	-	1T tokens	Apr-2019	1024 TPU v3	-	✓	-
	mT5 [83]	Oct-2020	13	-	-	-	1T tokens	-	-	-	✓	-
	PanGu-α [84]	Apr-2021	13*	-	-	-	1.1TB	-	2048 Ascend 910	-	✓	-
	CPM-2 [85]	Jun-2021	198	-	-	-	2.6TB	-	-	-	-	-
	T0 [28]	Oct-2021	11	T5	✓	-	-	-	512 TPU v3	27 h	✓	-
	CodeGen [86]	Mar-2022	16	-	-	-	577B tokens	-	-	-	✓	-
	GPT-NeoX-20B [87]	Apr-2022	20	-	-	-	825GB	-	96 40G A100	-	✓	-
	Tk-Instruct [88]	Apr-2022	11	T5	✓	-	-	-	256 TPU v3	4 h	✓	-
	UL2 [89]	May-2022	20	-	-	-	1T tokens	Apr-2019	512 TPU v4	-	✓	✓
	OPT [90]	May-2022	175	-	-	-	180B tokens	-	992 80G A100	-	✓	-
	NLLB [91]	Jul-2022	54.5	-	-	-	-	-	-	-	✓	-
	CodeGeeX [92]	Sep-2022	13	-	-	-	850B tokens	-	1536 Ascend 910	60 d	✓	-
	GLM [93]	Oct-2022	130	-	-	-	400B tokens	-	768 40G A100	60 d	✓	-
	Flan-T5 [69]	Oct-2022	11	T5	✓	-	-	-	-	-	✓	✓
	BLOOM [78]	Nov-2022	176	-	-	-	366B tokens	-	384 80G A100	105 d	✓	-
	mT0 [94]	Nov-2022	13	mT5	✓	-	-	-	-	-	✓	-
	Galactica [35]	Nov-2022	120	-	-	-	106B tokens	-	-	-	✓	✓
	BLOOMZ [94]	Nov-2022	176	BLOOM	✓	-	-	-	-	-	✓	-
Publicly	OPT-IML [95]	Dec-2022	175	OPT	✓	-	-	-	128 40G A100	-	✓	✓
Available	LLaMA [57]	Feb-2023	65	-	-	-	1.4T tokens	-	2048 80G A100	21 d	✓	-
	Pythia [96]	Apr-2023	12	-	-	-	300B tokens	-	256 40G A100	-	✓	-
	CodeGen2 [97]	May-2023	16	-	-	-	400B tokens	-	-	-	✓	-
	StarCoder [98]	May-2023	15.5	-	-	-	1T tokens	-	512 40G A100	-	✓	✓
	LLaMA2 [99]	Jul-2023	70	-	✓	✓	2T tokens	-	2000 80G A100	-	✓	-
	Baichuan2 [100]	Sep-2023	13	-	✓	✓	2.6T tokens	-	1024 A800	-	✓	-
	QWEN [101]	Sep-2023	14	-	✓	✓	3T tokens	-	-	-	✓	-
	FLM [102]	Sep-2023	101	-	✓	-	311B tokens	-	192 A800	22 d	✓	-
	Skywork [103]	Oct-2023	13	-	-	-	3.2T tokens	-	512 80G A800	-	✓	-

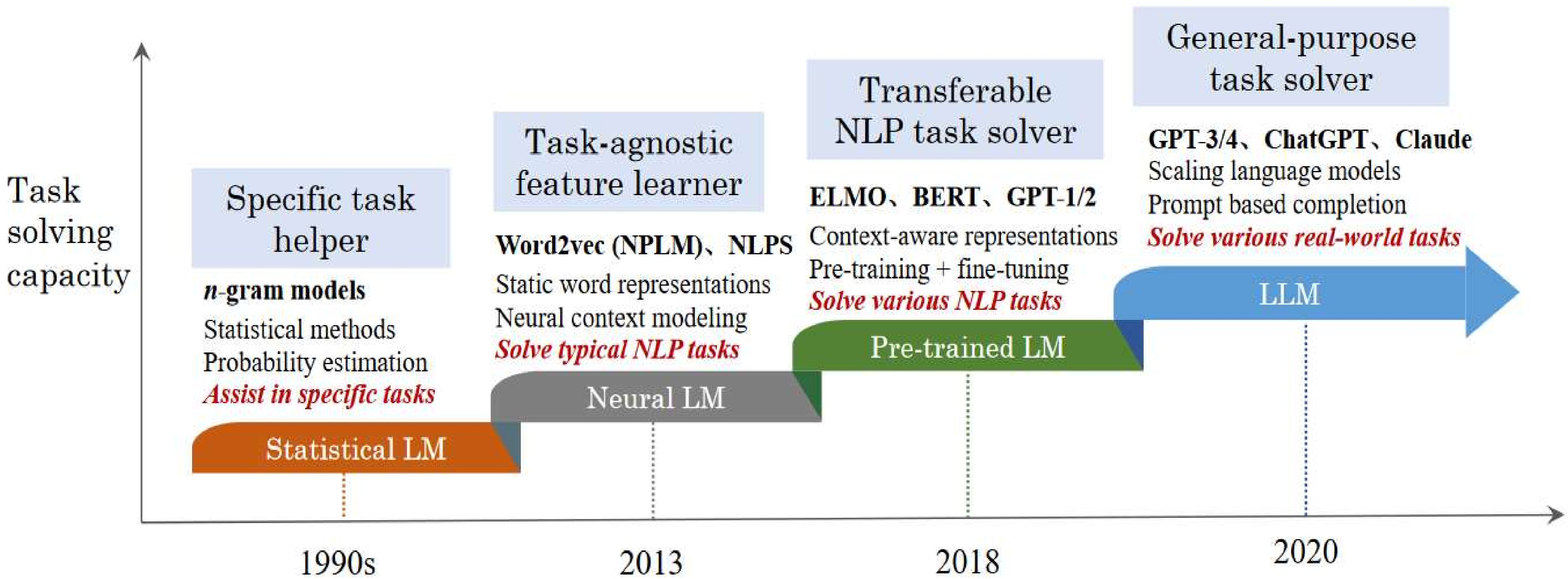
Source: (Zhao et al., 2023) - A Survey of Large Language Models

Statistics on Large Language Models (2)

	GPT-3 [55]	May-2020	175	-	-	-	300B tokens	-	-	-	✓	-
	GShard [104]	Jun-2020	600	-	-	-	1T tokens	-	2048 TPU v3	4 d	-	-
	Codex [105]	Jul-2021	12	GPT-3	-	-	100B tokens	May-2020	-	-	✓	-
	ERNIE 3.0 [106]	Jul-2021	10	-	-	-	375B tokens	-	384 V100	-	✓	-
	Jurassic-1 [107]	Aug-2021	178	-	-	-	300B tokens	-	800 GPU	-	✓	-
	HyperCLOVA [108]	Sep-2021	82	-	-	-	300B tokens	-	1024 A100	13.4 d	✓	-
	FLAN [67]	Sep-2021	137	LaMDA-PT	✓	-	-	-	128 TPU v3	60 h	✓	-
	Yuan 1.0 [109]	Oct-2021	245	-	-	-	180B tokens	-	2128 GPU	-	✓	-
	Anthropic [110]	Dec-2021	52	-	-	-	400B tokens	-	-	-	✓	-
	WebGPT [81]	Dec-2021	175	GPT-3	-	✓	-	-	-	-	✓	-
	Gopher [64]	Dec-2021	280	-	-	-	300B tokens	-	4096 TPU v3	920 h	✓	-
	ERNIE 3.0 Titan [111]	Dec-2021	260	-	-	-	-	-	-	-	✓	-
	GLaM [112]	Dec-2021	1200	-	-	-	280B tokens	-	1024 TPU v4	574 h	✓	-
	LaMDA [68]	Jan-2022	137	-	-	-	768B tokens	-	1024 TPU v3	57.7 d	-	-
Closed Source	MT-NLG [113]	Jan-2022	530	-	-	-	270B tokens	-	4480 80G A100	-	✓	-
	AlphaCode [114]	Feb-2022	41	-	-	-	967B tokens	Jul-2021	-	-	-	-
	InstructGPT [66]	Mar-2022	175	GPT-3	✓	✓	-	-	-	-	✓	-
	Chinchilla [34]	Mar-2022	70	-	-	-	1.4T tokens	-	-	-	✓	-
	PaLM [56]	Apr-2022	540	-	-	-	780B tokens	-	6144 TPU v4	-	✓	✓
	AlexaTM [115]	Aug-2022	20	-	-	-	1.3T tokens	-	128 A100	120 d	✓	✓
	Sparrow [116]	Sep-2022	70	-	-	✓	-	-	64 TPU v3	-	✓	-
	WeLM [117]	Sep-2022	10	-	-	-	300B tokens	-	128 A100 40G	24 d	✓	-
	U-PaLM [118]	Oct-2022	540	PaLM	-	-	-	-	512 TPU v4	5 d	✓	✓
	Flan-PaLM [69]	Oct-2022	540	PaLM	✓	-	-	-	512 TPU v4	37 h	✓	✓
	Flan-U-PaLM [69]	Oct-2022	540	U-PaLM	✓	-	-	-	-	-	✓	✓
	GPT-4 [46]	Mar-2023	-	-	✓	✓	-	-	-	-	✓	✓
	PanGu-Σ [119]	Mar-2023	1085	PanGu-α	-	-	329B tokens	-	512 Ascend 910	100 d	✓	-
	PaLM2 [120]	May-2023	16	-	✓	-	100B tokens	-	-	-	✓	✓

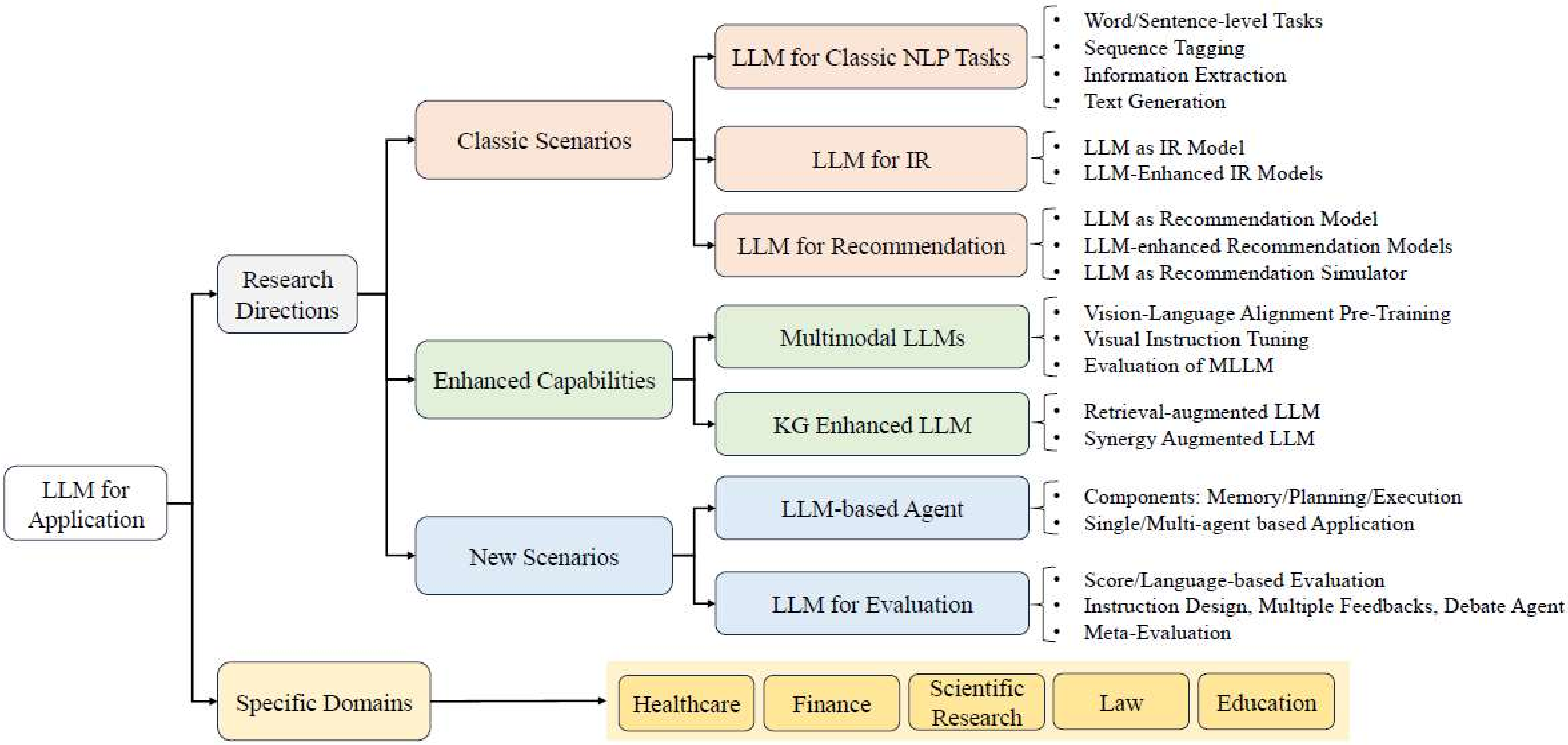
Source: (Zhao et al., 2023) - A Survey of Large Language Models

Four Generations of Language Models



Source: (Zhao et al., 2023) - A Survey of Large Language Models

Applications of Large Language Models

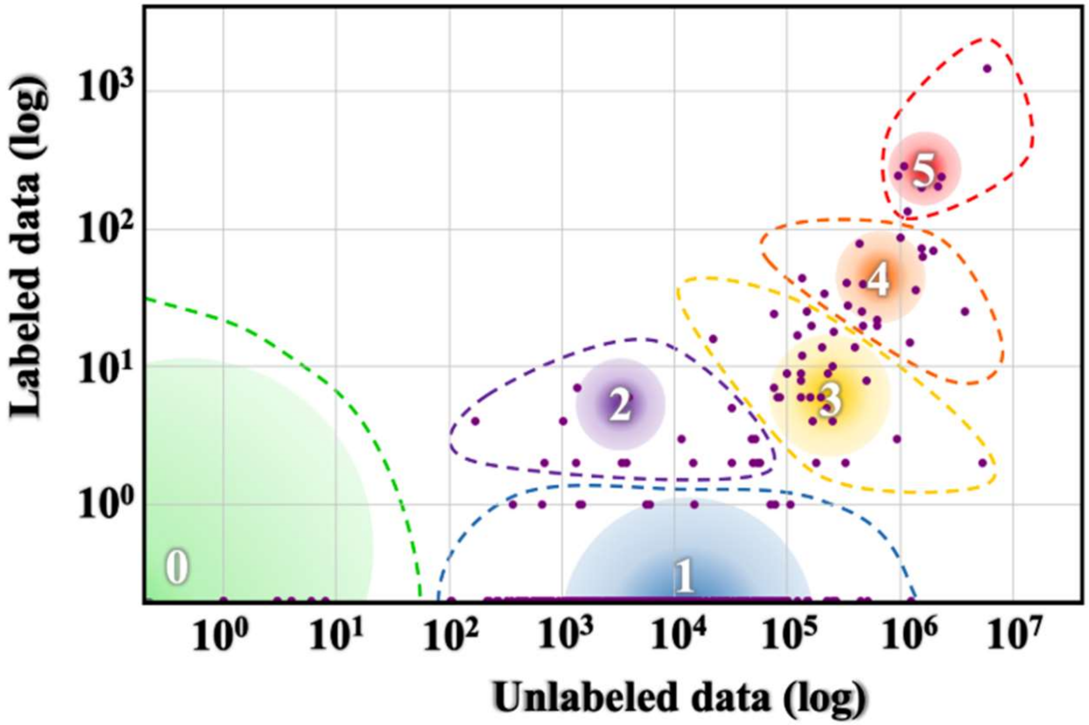


Source: (Zhao et al., 2023) - A Survey of Large Language Models

Multilinguality in Large Language Models

Multilinguality - Challenges

- **Limited Data (Joshi et al., 2020)**
  - The languages of the world are categorized into six different categories based on the amount of labeled and unlabeled data available in them
  - 88% of the world's languages are in resource group 0 with virtually no text data available
  - 5% of languages are in resource group 1 where there is very limited text data available



Source: (Zhao et al., 2023) - A Survey of Large Language Models

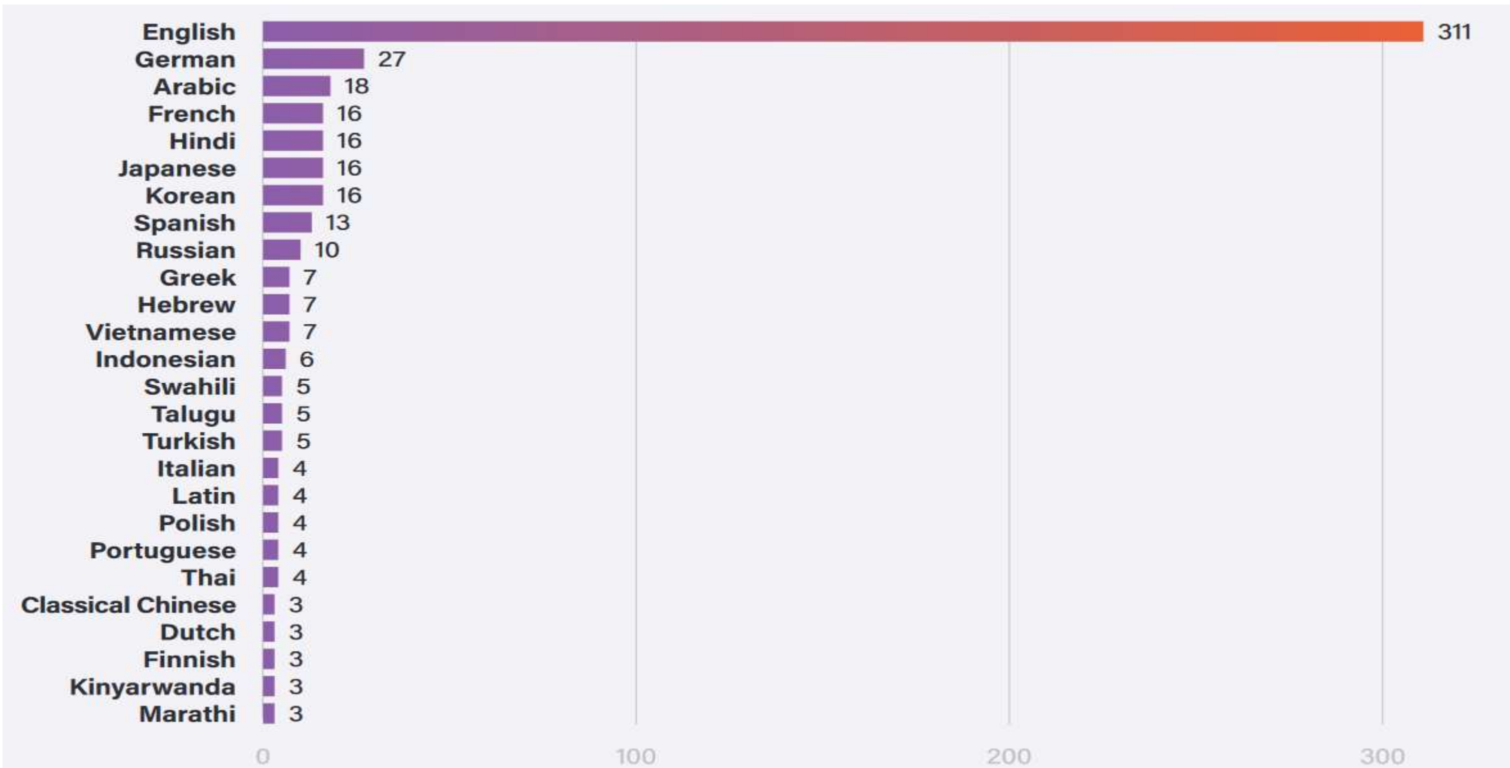
Multilinguality – Categories of language resourcedness

Resourcedness	Languages	Number of Languages	Number of Speakers
Extremely High Resource	English	1	1.1B
High Resource	Arabic, French, Japanese, German, Spanish, Mandarin	6	2.7B
Medium Resource	Dutch, Vietnamese, Korean, Portuguese, Hindi, Slovak, Hebrew, Indonesian, Afrikaans, Bengali, etc.	Dozens	2.7B
Low Resource	Haitian Creole, Tigrinya, Swahili, Bavarian, Cherokee, Zulu, Burmese, Telugu, Maltese, Amharic, etc.	Hundreds	0.5B
Extremely Low Resource	Dahalo, Warlpiri, Popoloca, Wallisian, Bora, etc.	Thousands	1.1B

Languages divided into different levels of resourcedness, according to labeled and unlabeled datasets

Source: (Joshi et al., 2020) - The State and Fate of Linguistic Diversity and Inclusion in the NLP World

Multilinguality – Languages mentioned in paper abstracts



Top most mentioned languages in abstracts of papers published by ACL (Association for Computational Linguistics) May 2022-January 2023 (Santy et al., 2023)

Multilinguality – Similarties between languages

Language	Sentence
English	The Emir of Kano turbaned Zhang who has spent 18 years in Nigeria
Amharic	የካኖ ኢምር በናይጀርያ ፲፰ ዓመት ያሳለፈውን ዛንግን ዋና መሪ አደረጉት
Hausa	Sarkin Kano yayi wa Zhang wanda yayi shekara 18 a Najeriya sarauta
Igbo	Onye Emir nke Kano kpubere Zhang okpu onye nke nọgoro afọ iri na asatọ na Naijirịa
Kinyarwanda	Emir w'i Kano yimitse Zhang wari umaze imyaka 18 muri Nijeriya
Luganda	Emir w'e Kano yatikkidde Zhang amaze emyaka 18 mu Nigeria
Luo	Emir mar Kano ne orwakone turban Zhang ma osedak Nigeria kwuom higni 18
Nigerian-Pidgin	Emir of Kano turban Zhang wey don spend 18 years for Nigeria
Swahili	Emir wa Kano alimvisha kilemba Zhang ambaye alikaa miaka 18 nchini Nigeria
Wolof	Emiiru Kanó dafa kaala kii di Zhang mii def Nigeria fukki at ak juróom ñett
Yorùbá	Ẹ́míà ìlú Kánò wé láwàní lé orí Zhang ẹni tí ó ti lo ọdún méjìdínlógún ní orílè-èdè Nàìjíríà

Named entity annotations in African languages (Adelani et al., 2021)

Multilinguality in Pre-trained Language Models (PLMs)

- How can the benefits of BERT-like pre-trained models be utilized for other languages of interest?

- For a given language, is a language-specific BERT better than a Multilingual PLM?

- Can the shared representations learned by Multilingual PLMs improve machine translation performance between two resource-rich languages?

Multilingual Pre-trained Language Models to Bridge the Resourcedness Gap

- **Monolingual Pre-trained Language Models**
  - Generate texts, one token at time
  - Compute dense representations

  Examples: BERT (English), AraBERT (Arabic), CamemBERT (French), AlBERTo (Italian), BERTje(Dutch), BERTeus(Basque), BERTu (Maltese), SwahBERT (Swahili), etc.

- **Multilingual Pre-trained Language Models**
  - Generate texts in multiple languages
  - Compute dense multilingual representations

  Examples: mBERT (104 languages), XLM-R (100 languages), BLOOM (46 languages), AfriBERTa (African languages), AfroXLM-R (African languages), IndoBERT (Indonesian languages), IndicBERT (Indian languages), etc.

➔ Cross-lingual training helps the model to generalize better: For the BLOOMZ models which were trained on machine translated corpora as well as original multi language documents, it actually performed better on even English tasks compared to its base BLOOM model.

Building Multilingual Pre-trained Language Models

- **Training Multilingual Pre-trained Language Models**
  - Basic requirements (Devlin et al., 2019; Conneau et al., 2020):
  - ➔ Multilingual corpora
  - ➔ Languge-independent representations: Multilingual Byte-Pair Encoding (BPE) or WordPiece Tokenization

Example (WordPiece):

Sentence: This is the Hugging Face course!

Tokenization: ['Th', '##i', '##s', 'is', 'th', '##e', 'Hugg', '##i', '##n', '##g', 'Fac', '##e', 'c', '##o', '##u', '##r', '##s', '##e', '[UNK]']

- **Training Objectives: Neighbor Word Prediction (NWP), Masked Language Model (MLM)**
  - Pay attention to the data distribution
  - Parallel corpora and dedicated losses are important factors for high performance can help (Ouyang et al., 2021; Chi et al., 2021)

Overview of some Pre-trained language models across languages based on BERT

Language	Model	Pre-training Corpus	#Tokens	Vocab	Params
Multi	mBERT	Wiki-100	3.3B	106K	167M
	XLM-R	CC-100	167B	250K	278M
English (EN)	BERT	Wikipedia, BookCorpus	3.3B	30K	109M
	RoBERTa	BookCorpus, CC-News, OpenWebText, Stories	40B	50K	125M
Chinese (ZH)	BERT	Wikipedia	0.4B	21K	103M
	RoBERTa	Wikipedia	0.4B	21K	102M
Spanish (ES)	BERT	Wikipedia, OPUS	3B	31K	110M
	RoBERTa	Web crawl	135B	50K	125M
French (FR)	BERT	Europeana	11B	32K	111M
	RoBERTa	Wikipedia, CC-100	59B	50K	124M
Hindi (HI)	BERT	L3Cube	0.3B	52K	126M
	RoBERTa	mC4, OSCAR, IndicNLP	1.5B	52K	83M

Source: (Zhao and Aletras, 2024) - Comparing Explanation Faithfulness between Multilingual and Monolingual Fine-tuned Language Models

Representative multilingual training corpora of LLMs

Model	Language	Language proportion	Source
mBERT [2]	104 languages	Unknown	Wikipedia
XLM-R [7]	100 languages	English (12.56%); Russian (11.61%); Others (63.89%) Indonesian (6.19%); Vietnamese (5.73%)	Generated using the open source; CC-Net repository
mT5 [4]	101 languages	English (5.67%); Russian (3.71%); Spanish (3.09 %); German (3.05%); Others (84.48%)	Common Crawl
GPT-3 [20]	95 languages	English (92.7%); French (1.8%); German (1.5%); Others (5.9%)	Common Crawl; Wikipedia; Books1; Books2; WebText2
Gopher [38]	51 languages	Over 99% English	MassiveWeb (48%); C4 (10%); News (10%); Books (27%); GitHub (3%);Wikipedia (2%)
LaMDA [30]	Unknown	Over 90% English	Public dialog data and other public web documents
InstructGPT [21]	Unknown	Over 96% English	Text prompts written by labelers or from the OpenAI API
PaLM [29]	Over 100 languages	English (77.98%); German (3.50%); French (3.25%); Spanish (2.11%); Others (13.15%)	Social media conversations (50%); Filtered webpages (27%); Books (13%); GitHub (5%); Wikipedia (4%); News (1%)
BLOOM [5]	46 languages	English (30.03%); Simplified Chinese (16.16%); French (12.9%); Spanish (10.85%); Portuguese (4.91%); Arabic (4.6%); Others (20.55%)	Web Crawl(38%); BigScience Catalogue Data(62%)
LLaMA [6]	Over 20 languages	Over 67% English	Common Crawl (67.0%); C4 (15.0%); Github (4.5%);Wikipedia (4.5%); Books (4.5%); ArXiv (2.5%); StackExchange (2.0%)
Vicuna [34]	Unknown	Unknown	User-shared conversations from ShareGPT.com
Falcon [85]	Over 100 languages	Excluding English: Russian (13.19%); German (10.81%); Spanish (9.45%); Others (66.55%)	Common Crawl
PaLM 2 [46]	Over 100 languages	Excluding English: Spanish (11.51%); Chinese (10.19%); Russian (8.73%); Others (69.57%)	Web documents; books; code; mathematics; conversational data
LLaMA 2 [47]	Over 100 languages	English (89.70%); Unknown (8.38%); German (0.17%); France (0.16%); Others (1.59%)	Publicly available sources excludes Meta user data

Comparison of predictive performance between mBERT and monolingual BERT across languages and tasks

Lg	Model	NER Test F1	SA Test Acc	QA Dev EM / F1	UDP Test UAS/LAS	POS Test Acc
Arabic AR	Monolingual	91.1	95.9	68.3/82.4	90.1/85.6	96.8
	mBERT	90	95.4	66.1/80.6	88.8/83.8	96.8
English	Monolingual	91.5	91.6	80.5/88.0	92.1/89.7	97
	mBERT	91.2	89.8	80.9/88.4	91.6/89.1	96.9
Finnish	Monolingual	92	-	69.9/81.6	95.9/94.4	98.4
	mBERT	88.2	-	66.6/77.6	91.9/88.7	96.2
Indonesian	Monolingual	91	96	66.8/78.1	85.3/78.1	92.1
	mBERT	93.5	91.4	71.2/82.1	85.9/79.3	93.5
Japanese	Monolingual	72.4	88	-	94.7/93.0	98.1
	mBERT	73.4	87.8	-	94.0/92.3	97.8
Korean	Monolingual	88.8	89.7	74.2/91.1	90.3/87.2	97
	mBERT	86.6	86.7	69.7/89.5	89.2/85.7	96
Russian	Monolingual	91	95.2	64.3/83.7	93.1/89.9	98.4
	mBERT	90	95	63.3/82.6	91.9/88.5	98.2
Turkish	Monolingual	92.8	88.8	60.6/78.1	79.8/73.2	96.9
	mBERT	93.8	86.4	57.9/76.4	74.5/67.4	95.7
Chinese	Monolingual	76.5	95.3	82.3/89.3	88.6/85.6	97.2
	mBERT	76.1	93.8	82.0/89.3	88.1/85.0	96.7
AVG	Monolingual	87.4	92.4	70.8/84.0	90.0/86.3	96.9
	mBERT	87	91	69.7/83.3	88.4/84.4	96.4

Source: (Rust et al., 2021) - How good is your tokenizer? on the monolingual performance of multilingual language models

Multilinguality in Large Language Models

- Are Multilingual LLMs better than monolingual models for a given language?
- Do Multilingual LLMs enable cross-lingual transfer?
- Do Multilingual LLMs learn universal/generalizable patterns across languages?

Large Language Models Life-cycle

Scope	Select	Adapt the model		Application integration	
Specify the task	Move with an existing model or build your own	Prompt engineering, fine tuning, human feedback	Evaluate	Optimize and deploy the model	Augment model and build LLM-powered applications

Source: (Xu et al., 2024) - A Survey on Multilingual Large Language Models: Corpora, Alignment, and Bias

Steps for Building a Multilingual Large Language Model (MLLM)

- **Preparing a balanced corpus of text in various languages**
  - Multilingual corpus
  - Data cleaning and Preprocessing
  - Data balancing

- **Training the Model**
  - Multilingual pretraining
  - Multilingual fine-tuning
  - Multi-task learning

- **Evaluation and Refinement**
  - Multilingual benchmarks
  - Error analysis and Bias Detection
  - Continuous improvement

State of the Art of LLMs for Low-resource Languages

- **Category 1: English-first models (ChatGPT, LLAMA2, etc.)**
  - Designed for English, transfer well for other languages

- **Category 2: Multilingual models (GXLM, BLOOM, mT0, XLM-R, etc.)**
  - Beat category 1 on some languages/tasks, even if they are not as powerful

- **Catgory 3: Low-resource language-first models (JASMINE, AraGPT, etc.)**
  - Category 1 / 2 models, modified and fine-tuned for the low-resource language (Arabic)

➔ **Datasets used in training and evaluation are English-centric even if the models are multilingual**

Adapting LLMs to Low-resource Languages

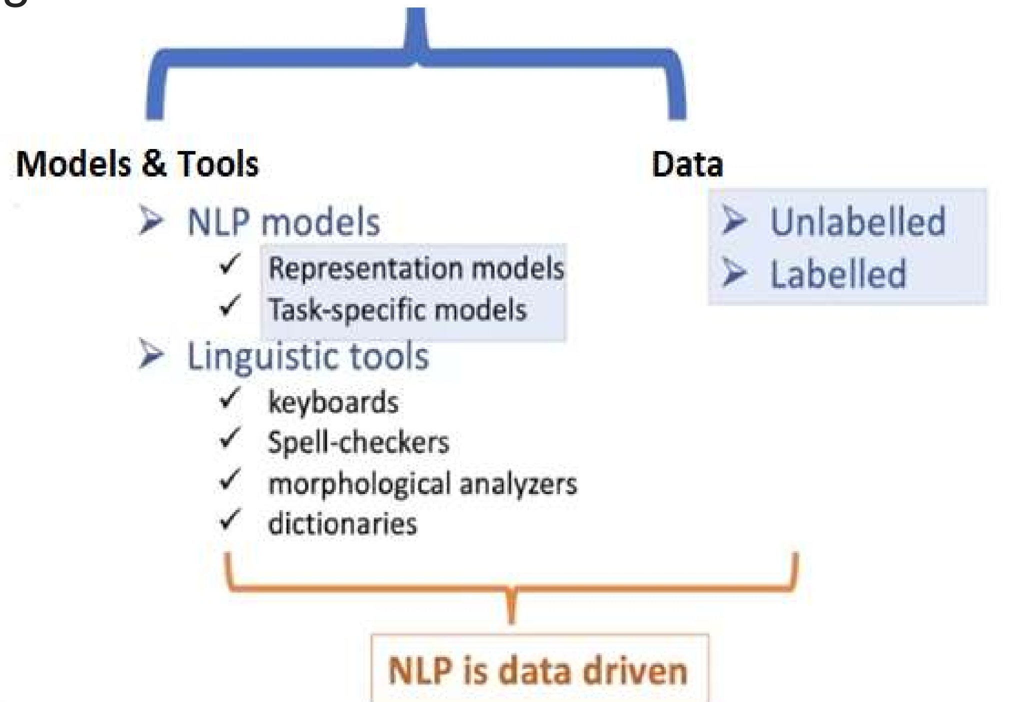
- **Data for Adapting LLMs**
  - Low-resource language

- **Architectural modification and training**
  - Training from scratch vs. Fine-tuning
  - ➔ Approaches: Pre-training / Instruction tuning / Human alignment
  - Vocabulary extensions
  - PEFT: Parameter-Efficient Fine-Tuning (LORA / Adapters)

- **Evaluation and deployment**
  - Availability of benchmarks
  - Different evaluation modes optimize different objectives

Adapting LLMs to Low-resource Languages

Low-resource languages = Languages with less recources



Issues when Adapting LLMs to Low-resource Languages

- Is it possible to adapt a Multilingual LLM (such as XGLM) from scratch to a low-resource language (such as Arabic)?
- Will it hold better word knowledge than another LLM (such as GPT, etc.)?
- Is it worth training from scratch? Where do we find the resources?

GPT-3			XGLM		
size	l	h	*size*	l	h
125M	12	768	—		
355M	24	1024	564M	24	1024
760M	24	1536	—		
1.3B	24	2048	1.7B	24	2048
2.7B	32	2560	2.9B	48	2048
6.7B	32	4096	7.5B	32	4096

Models' details – *size:* number of parameters, l: layers, h: hidden dimension

Source: (Linet al., 2022) - Few-shot Learning with Multilingual Generative Language Models

Data for Adapting LLMs to Low-resource Languages

- **Data is a key element in DL-based NLP**
  - Scaling law talks about number of tokens and number of parameters
  - Model analysis in 2019 was about architecture, hyper-parameter search
  - Model analysis in 2024 is about figuring out how the corpora CommonCrawl, C4 and Wikipedia are in the pre-training corpus

Dataset	Sampling prop.	Epochs	Disk size
CommonCrawl	67.0%	1.10	3.3 TB
C4	15.0%	1.06	783 GB
Github	4.5%	0.64	328 GB
Wikipedia	4.5%	2.45	83 GB
Books	4.5%	2.23	85 GB
ArXiv	2.5%	1.06	92 GB
StackExchange	2.0%	1.03	78 GB

Overview of datasets to train LLaMA

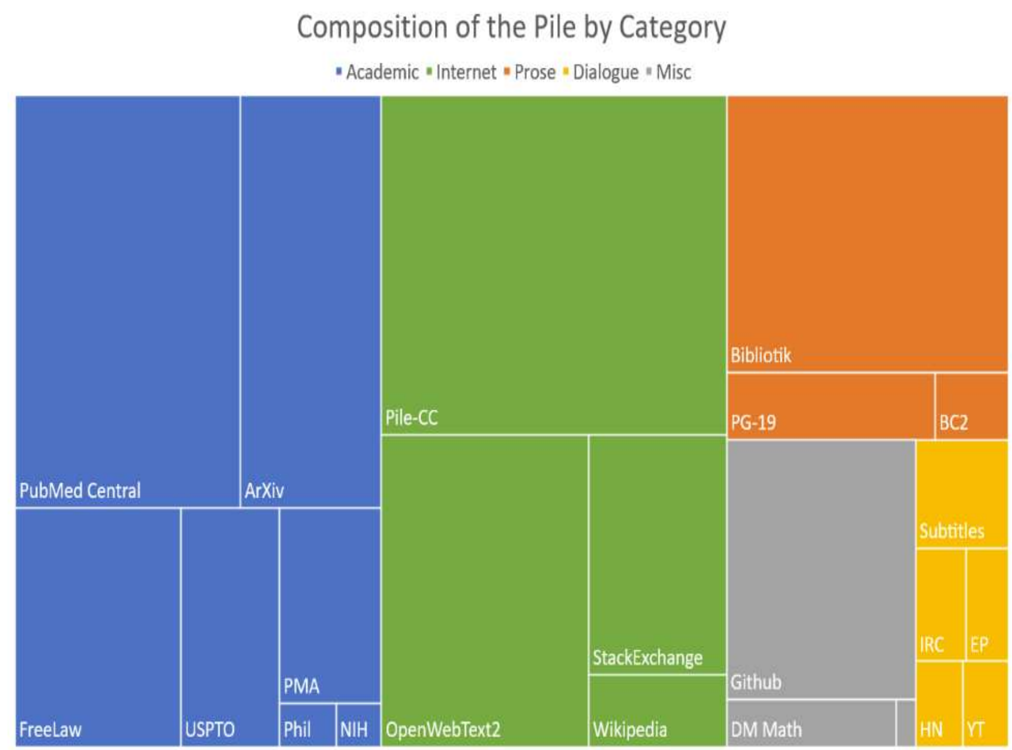
Source: (Touvron et al., 2023) - LLaMA: Open and Efficient Foundation Language Models

Data for Adapting LLMs to Low-resource Languages

- Data is a key element in DL-based NLP

Overview of datasets in the Pile (large, diverse, open source language modelling data composed of many combined smaller datasets)

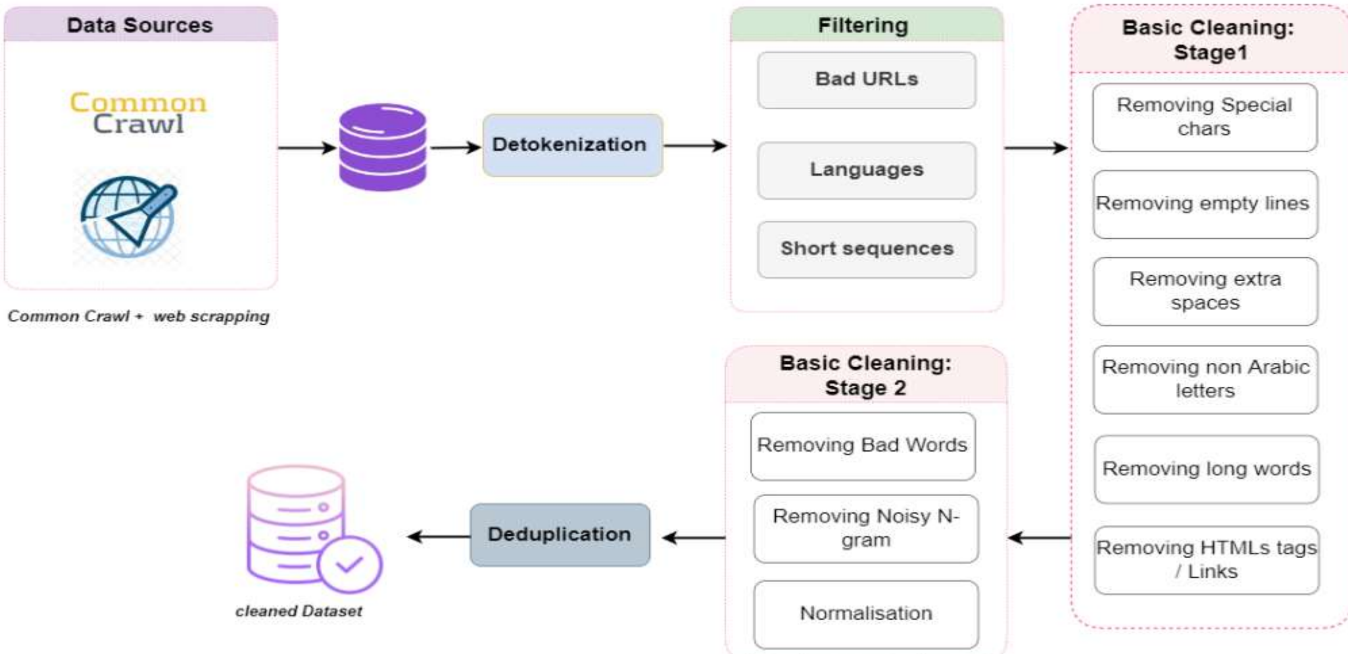
Component	Raw Size	Weight	Epochs	Effective Size	Mean Document Size
Pile-CC	227.12 GiB	18.11%	1.0	227.12 GiB	4.33 KiB
PubMed Central	90.27 GiB	14.40%	2.0	180.55 GiB	30.55 KiB
Books3†	100.96 GiB	12.07%	1.5	151.44 GiB	538.36 KiB
OpenWebText2	62.77 GiB	10.01%	2.0	125.54 GiB	3.85 KiB
ArXiv	56.21 GiB	8.96%	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**



Source: (Gao et al., 2020) - The Pile: An 800GB Dataset of Diverse Text for Language Modeling

Example of Adapting LLMs to a Low-resource Language (Arabic)

- **Pre-training**
  - Pre-training with the high quality of data: Books, Wikipedia, News, etc.
  - Content of CommonCrawl + Web scraping for Arabic (Forums, Blogs, etc.)
  - ➔ Filtering Pipeline to remove noise, rectify errors, and ensure data integrity



➔ **There is limited benefit to massive pre-training in Arabic, due to computers resources, data quality, versus fine-tuning from open source checkpoints**

Source: (Aloui et al., 2024) - 101 Billion Arabic Words Dataset

Data for Adapting LLMs to Low-resource Languages

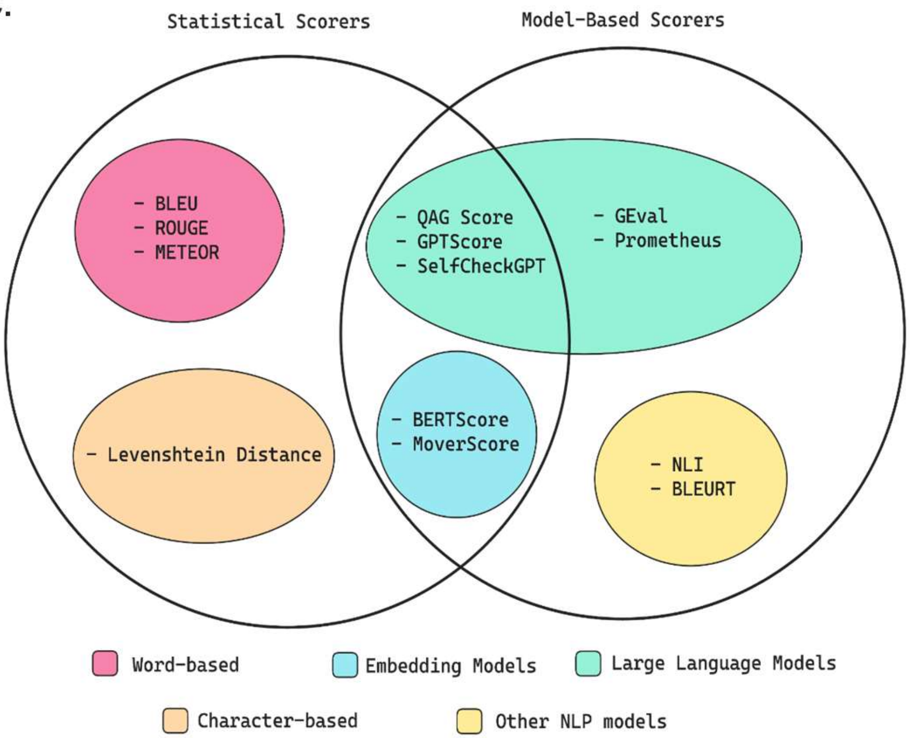
- **Instruction Tuning (IT) and Human Alignment**
  - Translating IT / Human Alignment data from English to the low-resource language
    + Much more available IT / Human Alignment data in English
    - Translation quality may be very poor
  - Carrefully curated, high quality human alignment data of the low-resource language
    + High performance
    - Does this exist for each low-resource language?
  - Multilingual Instruction Tuning (BLOOMZ, mT0: a Multitask prompted finetuning variant of mT5)
  - Translation data
    - Gold standard: crowdsourced translations – extremely expensive
    - Web-crawled: more available (Wikipedia, News, etc.)
      ➔ Issue: Text isn't necessary aligned

Model	Language	Average	ARC (25-shot)	HellaSwag (10-shot)	MMLU (5-shot)	TruthfulQA (0-shot)
Bloom-7b1	Multilingual	36.2	31.4	43.3	27.5	42.6
Llama-7B	Multilingual	32.1	24.6	30.9	28.0	45.1
ArabianGPT-0.3B	Arabic	32.7	24.3	28.4	25.7	52.5
ArabianGPT-0.1B	Arabic	31.9	24.0	26.6	25.4	51.8
AraGPT-Base	Arabic	31.7	24.6	27.5	25.1	49.5
AraGPT-Medium	Arabic	32.2	23.9	28.5	26.3	50.0

Zero and Few-shot Evaluation Scores of Multilingual LLMs and Arabic GPT

Evaluation and deployment

- **Some NLP benchmarks**
  - Tasks: QA, Summarization, Translation, Natural Language Inference, Math, etc.
  - Metrics: F1, BLEU, ROUGE, etc.

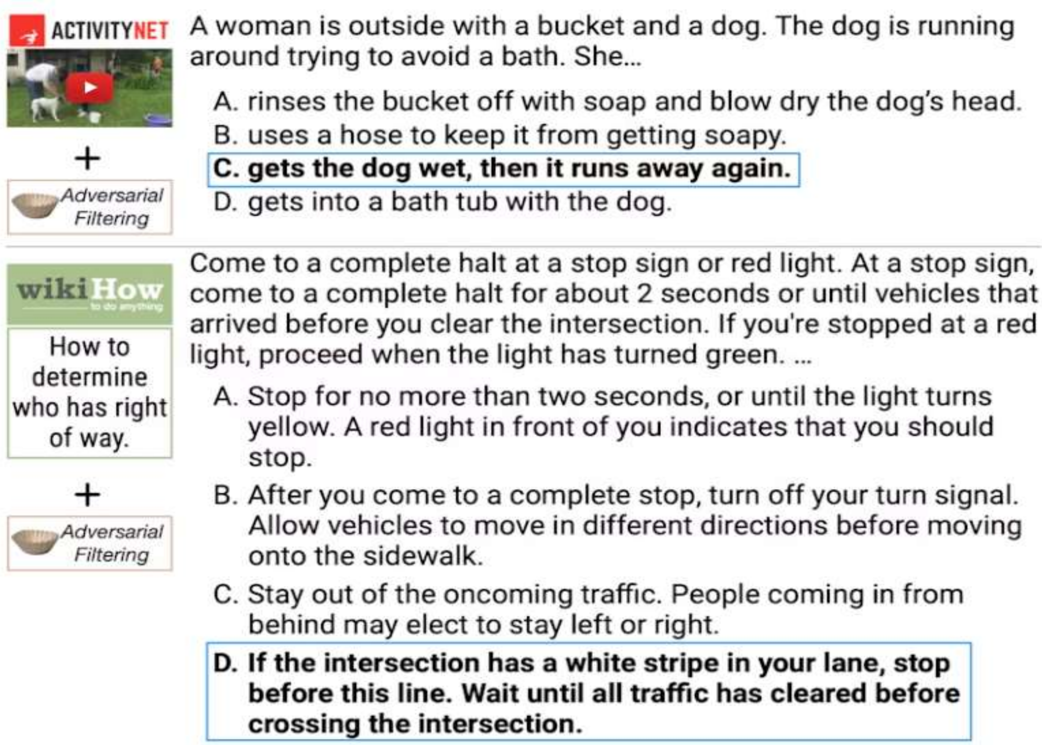


Types of metric scorers

Source: (Jeffrey Ip, 2024) - LLM Evaluation Metrics: The Ultimate LLM Evaluation Guide

Evaluation and deployment

- **Some NLP benchmarks**
  - Dataset: HellaSwag (Commonsense Reasoning)



Example of HellaSwag context and it corresponding completion option

Source: (AI & Engineering, 2023) - HellaSwag: Understanding the LLM Benchmark for Commonsense Reasoning

Evaluation and deployment

- **Evaluation of State-Of-the-Art models**
  - Dataset: HellasWag (Commonsense Reasoning)
  - State-Of-the-Art models: OpenAI GPT, BERT-Base and BERT-Large, ESIM+ELMo, fastText
  - ➔ Humans significantly outperform all models

Model	Overall		In-Domain		Zero-Shot		ActivityNet		WikiHow	
	Val	Test	Val	Test	Val	Test	Val	Test	Val	Test
Split Size→	10K	10K	5K	5K	5K	5K	3.2K	3.5K	6.8K	6.5K
Chance	25.0									
fastText	30.9	31.6	33.8	32.9	28.0	30.2	27.7	28.4	32.4	33.3
LSTM+GloVe	31.9	31.7	34.3	32.9	29.5	30.4	34.3	33.8	30.7	30.5
LSTM+ELMo	31.7	31.4	33.2	32.8	30.4	30.0	33.8	33.3	30.8	30.4
LSTM+BERT-Base	35.9	36.2	38.7	38.2	33.2	34.1	40.5	40.5	33.7	33.8
ESIM+ELMo	33.6	33.3	35.7	34.2	31.5	32.3	37.7	36.6	31.6	31.5
OpenAI GPT	41.9	41.7	45.3	44.0	38.6	39.3	46.4	43.8	39.8	40.5
BERT-Base	39.5	40.5	42.9	42.8	36.1	38.3	48.9	45.7	34.9	37.7
BERT-Large	**46.7**	**47.3**	**50.2**	**49.7**	**43.3**	**45.0**	**54.7**	**51.7**	**42.9**	**45.0**
Human	95.7	95.6	95.6	95.6	95.8	95.7	94.0	94.0	96.5	96.5

Performance of State-Of-the-Art models on HellaSwag dataset

Source: (AI & Engineering, 2023) - HellaSwag: Understanding the LLM Benchmark for Commonsense Reasoning

Evaluation and deployment

- **Evaluation of PaLM models**
  - Dataset: Multilingual TyDi QA (Question/Answering)

➔ Even the smallest PaLM 2 variant achieves performance competitive with the much larger PaLM 540B

➔ PaLM 2-M outperforms PaLM consistently

	Gold Passage				No-context			
Language	PaLM	PaLM 2-S	PaLM 2-M	PaLM 2-L	PaLM	PaLM 2-S	PaLM 2-M	PaLM 2-L
Arabic	67.2	**73.8**	73.5	72.8	34.5	36.4	40.2	**42.6**
Bengali	74.0	**75.4**	72.9	73.3	27.6	29.5	36.7	**41.6**
English	69.3	**73.4**	**73.4**	72.4	38.3	38.0	42.0	**43.7**
Finnish	68.1	**71.9**	71.7	71.0	38.3	36.8	38.8	**45.5**
Indonesian	75.7	79.5	80.2	**81.5**	35.5	37.7	41.3	**46.4**
Korean	70.6	71.4	72.3	**73.3**	35.0	38.7	41.7	**46.9**
Russian	57.6	**59.1**	58.6	58.1	24.6	26.0	29.2	**33.5**
Swahili	77.3	79.7	81.8	**82.5**	39.7	39.9	45.1	**50.3**
Telugu	68.0	75.7	75.5	**77.3**	9.6	9.2	10.5	**12.2**
Average	69.8	73.3	73.3	**73.6**	31.5	32.5	36.2	**40.3**

F1 scores in a 1-shot setting: Evaluation in the Gold Passage and a no-context setting (the model has to answer the question solely based on the knowledge stored in its parameters)

Source: (Google, 2023) - PaLM 2 Technical Report

Evaluation and deployment - Summary

- **NLP benchmarks metrics (F1, BLEU, ROUGE, etc.) reward concise, extractive answers**
  - ➔ Short answers (1, 2 words): not great LLM

- **Human evaluation is subjective, but values conversational responses**
  - ➔ Does not perform well on research benchmarks

- **Customer applications values accurate information**
  - ➔ NLP metrics do not apply
  - ➔ IR (Information Retrieval) / RAG (Retrieval Augmented Generation): What proportion of this task is Information Retrieval vs. Text Generation?

ChatGPT: Evaluation in a Multilingual Setting

- **Multilinguality in ChatGPT**
  - ChatGPT is trained on a mix of training data from multiple languages
  - English is the majority

- **Evaluation of the performance of ChatGPT (Lai et al., EMNLP 2023)**
  - Multiple languages: 37 diverse languages, characterizing high-, medium-, low-, and extremely low-resource languages
  - Different NLP tasks:
    - Natural Language Inference (NLI)
    - Question Answering
    - Common Sense Reasoning
    - Part-of-Speech (POS) Tagging
    - Named Entity Recognition (NER)
    - Relation Extraction
    - Summarization

ChatGPT: Evaluation in a Multilingual Setting
Part-of-Speech (POS) Tagging

- **Part-of-Speech (POS) Tagging** is a coarse-grained word classification task whose goal is to label the syntactic information of the words in a sentence

Language	Code	Cat.	XLM-R	ChatGPT (en)	ChatGPT (spc)
English	en	H	96.2	88.5	89.6
Russian	ru	H	86.9	91.6	59.1
German	de	H	92.2	90.2	89.9
Chinese	zh	H	60.4	76.5	75.3
French	fr	H	89.9	93.2	93.5
Spanish	es	H	89.0	92.2	91.9
Italian	it	H	92.6	92.6	93.4
Dutch	nl	H	88.5	88.1	88.3
Polish	pl	H	85.4	90.4	64.5
Vietnamese	vi	H	55.2	64.8	65.9
Turkish	tr	M	72.7	78.6	69.6
Arabic	ar	M	67.3	81.0	80.9
Greek	el	M	88.2	87.1	79.8
Thai	th	M	57.9	68.5	69.1
Bulgarian	bg	M	88.8	91.2	92.3
Hindi	hi	M	74.5	83.1	72.8
Urdu	ur	L	62.1	78.4	80.7
Average			79.3	84.5	79.8

Accuracy of ChatGPT (zero-shot learning) and XLM-R (supervised learning) on the test sets of XGLUE-POS. ChatGPT is evaluated with both English (en) and language-specific (spc) task descriptions

ChatGPT: Evaluation in a Multilingual Setting
Named Entity Recognition

- **Named Entity Recognition (NER)** aims to identify spans and semantic types of names (e.g., person, organization) in text.

Language	Code	Cat.	DAMO	ChatGPT	
				(en)	(spc)
English	en	H	91.2	37.2	37.2
Russian	ru	H	91.5	27.4	22.0
German	de	H	90.7	37.1	32.8
Chinese	zh	H	81.7	18.8	19.8
Spanish	es	H	89.9	34.7	33.2
Dutch	nl	H	90.5	35.7	37.5
Turkish	tr	M	88.7	31.9	29.1
Persian	fa	M	89.7	25.9	21.9
Korean	ko	M	88.6	30.0	32.2
Hindi	hi	M	86.2	27.3	26.1
Bengali	bn	L	84.2	23.3	16.4
Average			88.4	29.9	28.0

Performance (F1 scores) of ChatGPT (zero-shot learning) and DAMO (supervised learning) on the test sets of MultiCoNER. ChatGPT is evaluated with both English (en) and language-specific (spc) task descriptions

ChatGPT: Evaluation in a Multilingual Setting
Relation Extraction

- **Relation Extraction (RE)** aims to identify and classify semantic relations between two entity mentions in an input text.

Language	Code	Cat.	mT5-IL	ChatGPT (en)	ChatGPT (spc)
English	en	H	96.0	61.9	61.8
Russian	ru	H	83.3	78.8	77.5
German	de	H	94.0	71.1	71.8
French	fr	H	97.2	72.4	73.9
Spanish	es	H	70.5	67.5	65.8
Italian	it	H	97.0	74.4	74.6
Dutch	nl	H	93.5	66.8	66.6
Polish	pl	H	93.0	63.4	65.8
Portuguese	pt	H	85.2	64.8	66.3
Arabic	ar	M	94.1	84.9	90.1
Persian	fa	M	73.1	58.9	63.8
Korean	ko	M	83.2	65.3	70.1
Swedish	sv	M	58.7	64.2	65.4
Ukrainian	uk	M	71.8	76.5	68.8
Average			85.0	69.4	70.2

Performance (F1 scores) of ChatGPT (zero-shot learning) and mT5-IL (supervised learning) on the test sets of SMiLER. ChatGPT is evaluated with both English (en) and language-specific (spc) task descriptions

ChatGPT: Evaluation in a Multilingual Setting
Natural Language Inference

- **Natural Language Inference (NLI)** aims to predict the entailment/contradiction relations between two input sentences, i.e., a premise and a hypothesis.

Language	Code	Cat.	mT5-XXL	ChatGPT (en)	ChatGPT (spc)
English	en	H	92.4	70.2	70.2
Russian	ru	H	86.4	60.8	45.4
German	de	H	89.2	64.5	51.1
Chinese	zh	H	86.2	58.2	35.5
French	fr	H	88.7	64.8	42.2
Spanish	es	H	89.4	65.8	47.4
Vietnamese	vi	H	86.6	55.4	44.8
Turkish	tr	M	86.4	57.1	37.1
Arabic	ar	M	87.1	55.3	22.3
Greek	el	M	88.7	55.9	54.5
Thai	th	M	84.5	44.7	11.5
Bulgarian	bg	M	88.7	59.7	44.6
Hindi	hi	M	85.3	48.8	5.6
Urdu	ur	L	82.9	43.7	6.3
Swahili	sw	X	83.4	50.3	40.8
Average			87.1	57.0	37.3

Accuracy of ChatGPT (zero-shot learning) and mT5-XXL (supervised learning with English and translated data) on the development set of XNLI. ChatGPT is evaluated with both English (en) and language-specific (spc) task descriptions

ChatGPT: Evaluation in a Multilingual Setting
Question Answering

- Given a context passage and a question, a **Question Answering (QA)** model needs to return the answer for the question, which should be a span of text in the input passage.

Language	Code	Cat.	mT5-XXL		ChatGPT(en)		ChatGPT(spc)	
			EM	F1	EM	F1	EM	F1
English	en	H	80.3	91.3	56.0	74.9	56.0	74.9
Russian	ru	H	70.4	85.2	30.2	49.1	22.4	52.6
German	de	H	68.2	85.0	45.9	65.8	44.7	65.8
Chinese	zh	H	80.0	85.7	37.1	42.3	20.5	20.8
Spanish	es	H	70.8	87.4	41.8	65.8	40.5	69.1
Vietnamese	vi	H	67.1	85.3	36.1	57.3	26.8	60.8
Turkish	tr	M	67.7	84.4	34.5	56.4	18.3	52.8
Arabic	ar	M	68.2	83.4	32.0	50.3	24.1	49.9
Greek	el	M	68.9	85.9	29.7	45.0	17.7	39.1
Thai	th	M	74.5	80.2	31.2	43.4	1.5	13.1
Hindi	hi	M	68.2	83.7	17.5	37.8	0.6	22.9
Average			71.3	85.2	35.6	53.5	21.7	47.4

Performance of ChatGPT (zero-shot learning) and mT5-XXL (supervised learning with translated data) on the XQuAD dataset. (en) and (spc) indicate whether ChatGPT uses English or target language prompts. The performance is computed using exact match (EM) and F1 scores.

ChatGPT: Evaluation in a Multilingual Setting
Common Sense Reasoning

- **Common Sense Reasoning (CSR)** evaluates the reasoning of the models via multiple-choice questions. The inputs for the models involve a question and a few choices for the answer, and the models need to select one of the choices.

Language	Code	Cat.	TRT	ChatGPT	
				(en)	(tgt)
English	en	H	70.0	75.0	75.0
Russian	ru	H	59.8	50.2	53.5
German	de	H	61.7	52.6	61.0
Chinese	zh	H	59.6	50.2	42.5
Japanese	jp	H	54.3	41.9	43.0
French	fr	H	60.9	50.5	61.7
Spanish	es	H	61.1	53.3	62.5
Italy	it	H	61.2	50.6	55.9
Dutch	nl	H	59.8	52.9	60.4
Polish	pl	H	59.7	35.2	51.1
Portugese	pt	H	60.5	49.5	59.2
Vietnamese	vi	H	59.3	42.3	47.9
Arabic	ar	M	58.1	49.4	47.3
Hindi	hi	M	53.8	41.1	38.6
Urdu	ur	L	52.8	34.7	24.5
Swahili	sw	X	51.8	35.6	46.6
Average			59.0	47.8	51.9

Accuracy of ChatGPT (zero-shot learning) and TRT (supervised learning) on the dev set of X-CSQA dataset. (en) and (spc) indicate whether ChatGPT uses English or language-specific prompts

ChatGPT Evaluation - Conclusions

- ChatGPT exhibits significantly worse performance than state-of-the-art supervised models for most of considered NLP tasks in different languages
- It is more reasonable to build smaller task-specific models for NLP problems in different languages that can be hosted locally to serve at lower costs
- It seems evident that data size might not be the only factor that dictates the resource level and performance for a task of a language with ChatGPT and LLMs
- The superior performance of ChatGPT with English task descriptions over a majority of problems and languages suggests that ChatGPT might better understand the tasks with English prompts to lead to improved abilities to generate responses with accurate outputs

Challenges of Multilingual LLMs

- **Data Quantity**
  - Multilingual models require a larger vocabulary to represent tokens in many languages than monolingual models, but many languages lack large-scale datasets

- **Data Quality Concerns**
  - Models must train and fine-tune with meticulous attention to linguistic and cultural nuances to avoid biases and inaccuracies

- **Resource Limitations**
  - Training and running multilingual models require substantial computational resources such as powerful GPUs

- **Model Architecture**
  - Models must be able to handle languages with different word orders, morphological variations, and writing systems while maintaining high performance and efficiency

- **Evaluation Complexities**
  - Evaluating the performance of multilingual LLMs beyond English benchmarks is critical for measuring their true effectiveness, it requires considering cultural nuances, linguistic peculiarities, and domain-specific requirements

Adapting LLMs to Low-resource Languages - Summary

- The state-of-the art is in Fine-tuning
- Relevant Instruction tuning and Human alignment is a key
- High quality of the low-resource language data is crucial
- Architecture and training considerations affect efficiency more that accuracy

References

- Vladislav Lialin, Vijeta Deshpande, Anna Rumshisky (2023). Scaling Down to Scale Up: A Guide to Parameter-Efficient Fine-Tuning, arXiv.
- Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, Sai Qian Zhang (2024). Parameter-Efficient Fine-Tuning for Large Models: A Comprehensive Survey, arXiv.
- Maltese, G., and Mancini, F. (1992). "An automatic technique to include grammatical and morphological information in a trigram-based statistical language model." In Proceedings, IEEE International Conference on Acoustics, Speech and Signal Processing. San Francisco CA, March 1992, 1-157-I-160
- Shannon, C. E. 1948. A mathematical theory of communication. Bell System Technical Journal, 27(3):379–423
- Miller, G. A. and J. A. Selfridge. 1950. Verbal context and the recall of meaningful material. American Journal of Psychology, 63:176–185.
- Bengio, Y., R. Ducharme, P. Vincent, and C. Jauvin. 2003. A neural probabilistic language model. JMLR, 3:1137–1155
- Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In Proceedings of ACL 2020.
- Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In Proceedings of the First Workshop on Neural Machine Translation, pages 28–39, Vancouver. Association for Computational Linguistics.
- ME van der Wees et al. 2017. What's in a domain?: Towards fine-grained adaptation for machine translation.
- Danielle Saunders. 2022. Domain adaptation and multidomain adaptation for neural machine translation: A survey. Journal of Artificial Intelligence Research, 75:351–424.
- Barbara Plank. 2016. What to do about non-standard (or non-canonical) language in nlp. arXiv preprint arXiv:1608.07836.