Foundations of NLP

Lecture-12

Overview of Large Language Models

Mahindra™ University

Global Thinkers. Engaged Leaders.

Recap

- Masked Language Modeling
- Large Language Models/Survey
- Pipeline to build an LLM
- LLM Challenges

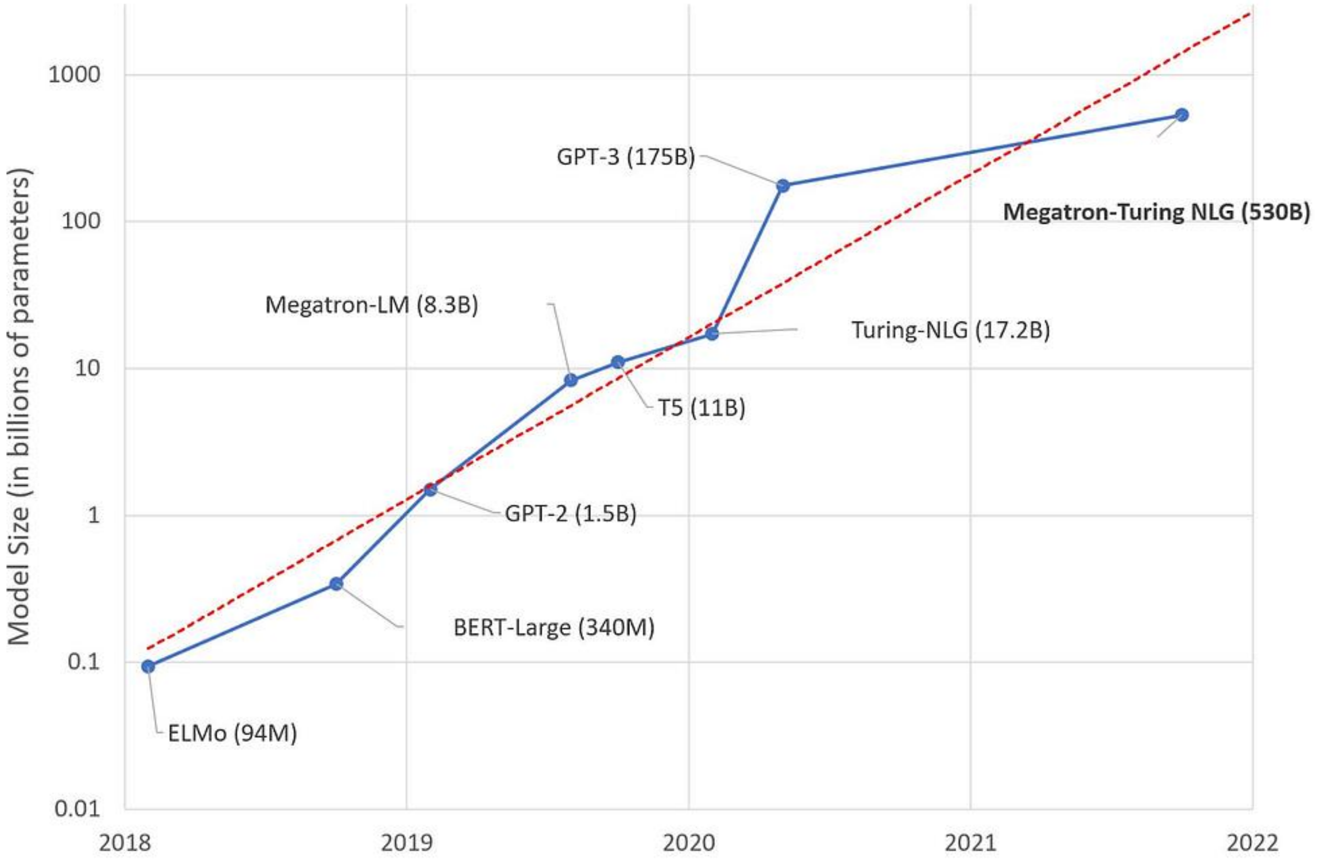
Asking
Questions

WHAT
to
LEARN?
dreamstime.com
ID 33913636 © Yuryz

Large Language Model Topics [Chapter 12]

- What is a large language model?
- Pretraining vs Fine-tuning
- Instruction Fine-tuning
- Greedy decoding
- Sampling
- Training
- Scaling Laws
- Potential harms with LLMs

Model Size with LLM's



Training them is more involved.

Think of it like compressing the internet.

Chunk of the internet,
~10TB of text

6,000 GPUs for 12 days, ~$2M
~1e24 FLOPS

ZIP

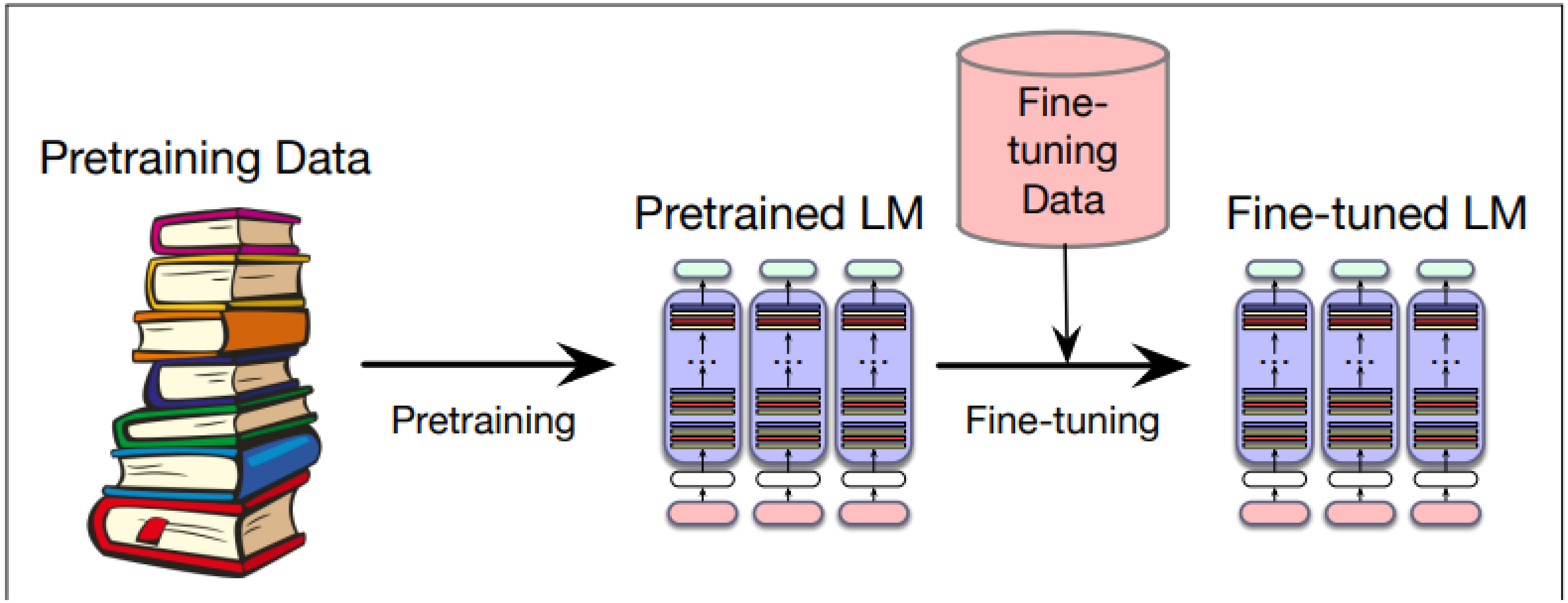
parameters.zip

~140GB file

*numbers for Llama 2 70B

Fine-tuning

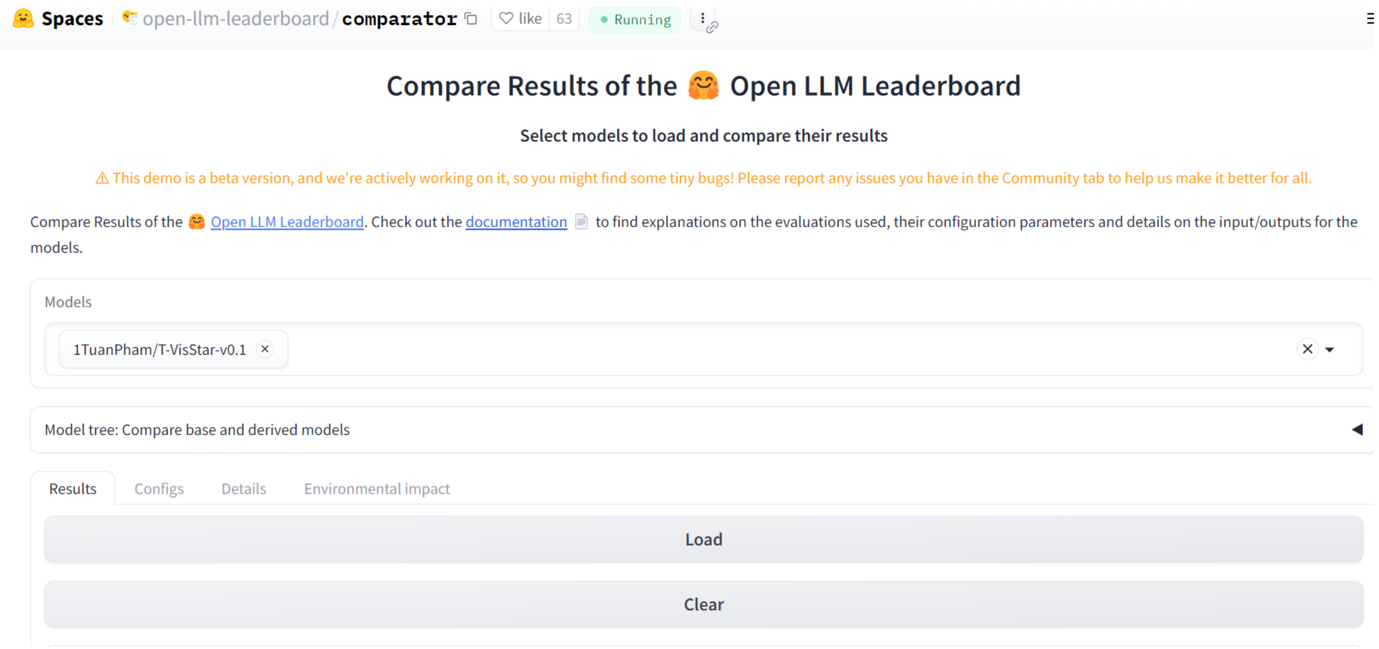
The process of taking a pretrained model and further adapting some or all of its parameters to some new data.



Types of Fine-tuning

- **Continued Pretraining:** We retrain all the parameters of the model on this new data, using the same method (word prediction) and loss function (cross-entropy loss) as for pretraining.
- **PEFT:** We efficiently select specific parameters to update when finetuning and leave the rest in their pretrained values.
- **Parameter-efficient:** The goal is to use a language model as a kind of classifier or labeler for a specific task.
- **SFT:** SFT is often used for instruction finetuning, in which we want a pretrained language model to learn to follow text instructions.

Open LLM Leaderboard



Compare Results of the 🤗 Open LLM Leaderboard

Select models to load and compare their results

⚠ This demo is a beta version, and we're actively working on it, so you might find some tiny bugs! Please report any issues you have in the Community tab to help us make it better for all.

Compare Results of the 🤗 Open LLM Leaderboard. Check out the documentation 📄 to find explanations on the evaluations used, their configuration parameters and details on the input/outputs for the models.

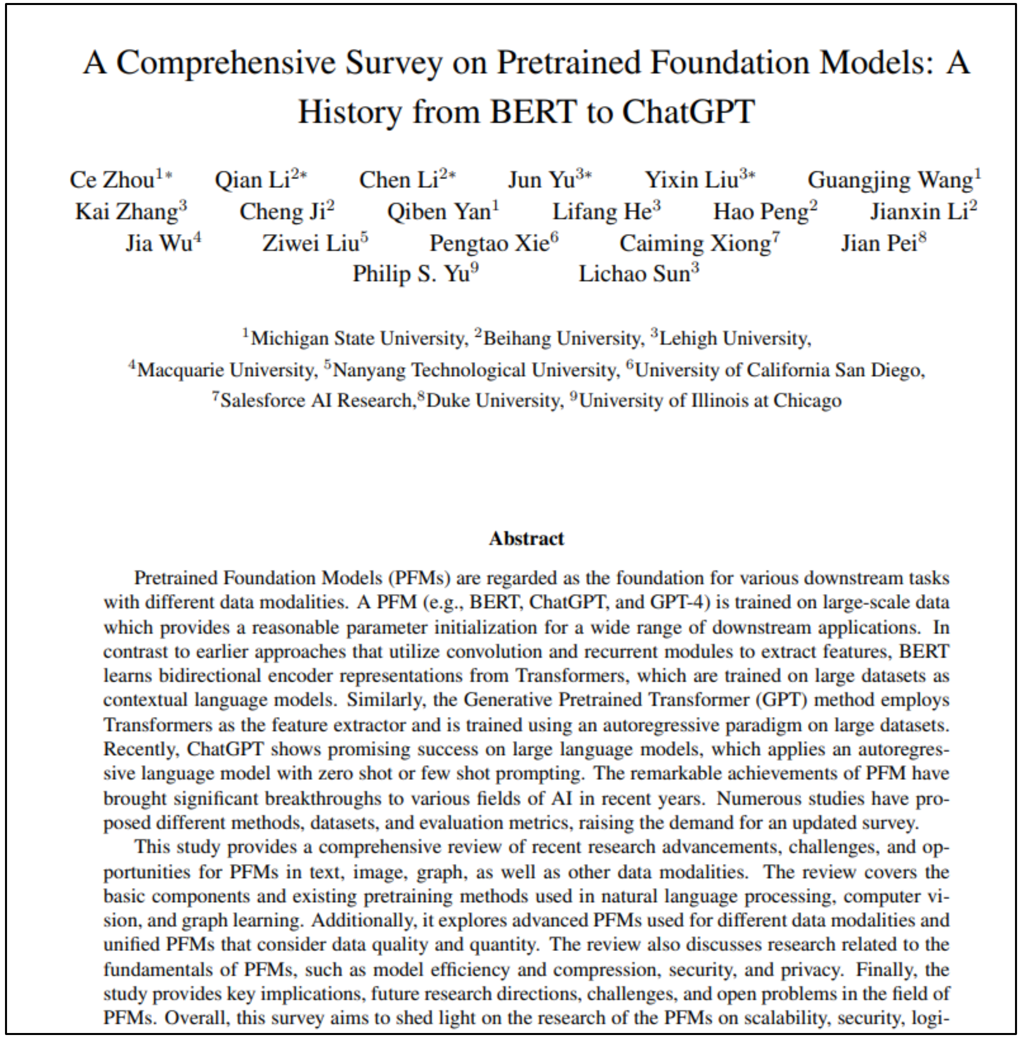
Models

1TuanPham/T-VisStar-v0.1 ×

Model tree: Compare base and derived models ◀

[Open LLM Leaderboard Model Comparator - a Hugging Face Space by open-llm-leaderboard]

History from BERT to ChatGPT



A Comprehensive Survey on Pretrained Foundation Models: A History from BERT to ChatGPT

Ce Zhou[1*] Qian Li[2*] Chen Li[2*] Jun Yu[3*] Yixin Liu[3*] Guangjing Wang[1]
Kai Zhang[3] Cheng Ji[2] Qiben Yan[1] Lifang He[3] Hao Peng[2] Jianxin Li[2]
Jia Wu[4] Ziwei Liu[5] Pengtao Xie[6] Caiming Xiong[7] Jian Pei[8]
Philip S. Yu[9] Lichao Sun[3]

[1]Michigan State University, [2]Beihang University, [3]Lehigh University,
[4]Macquarie University, [5]Nanyang Technological University, [6]University of California San Diego,
[7]Salesforce AI Research,[8]Duke University, [9]University of Illinois at Chicago

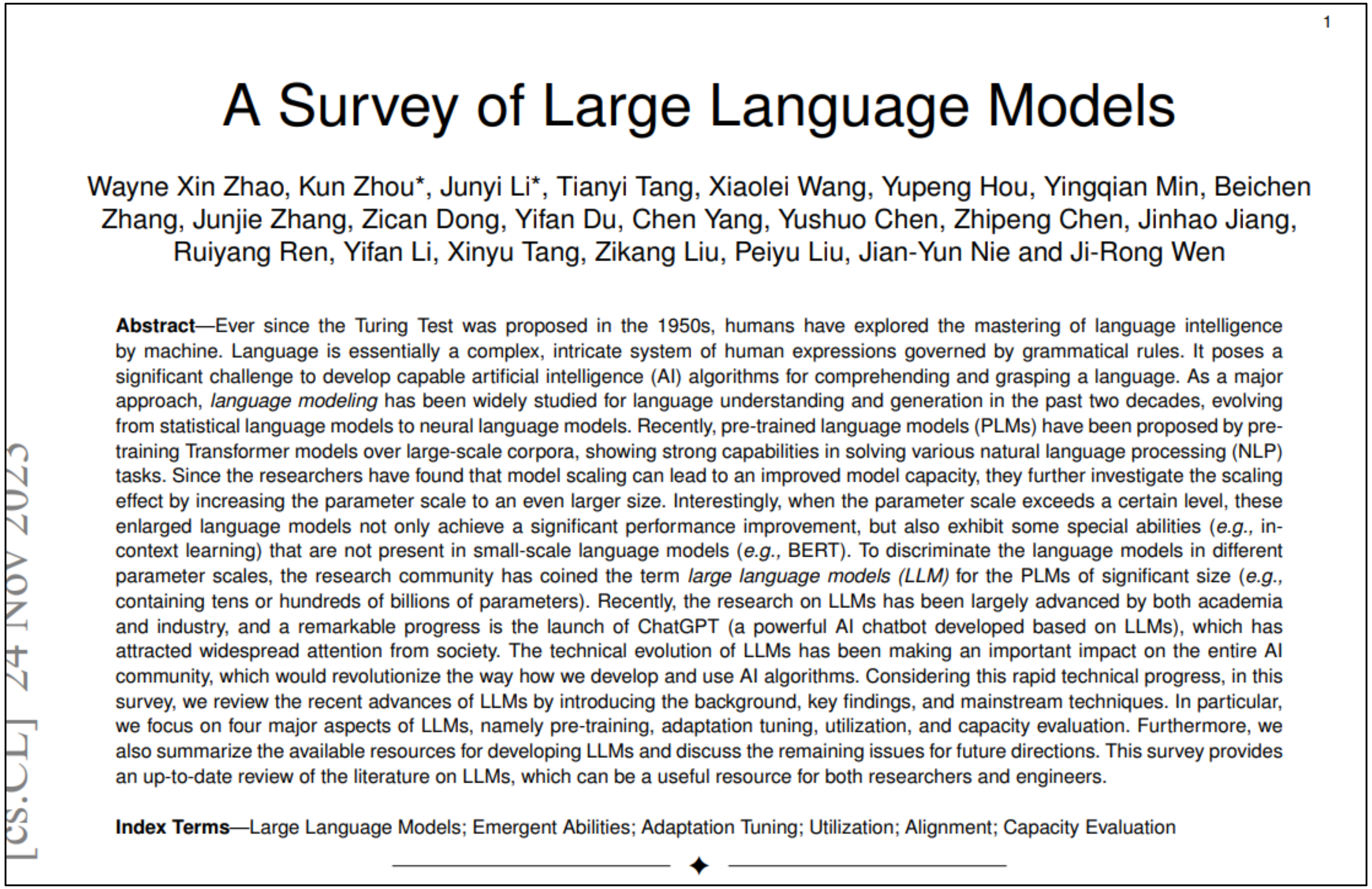
Abstract

Pretrained Foundation Models (PFMs) are regarded as the foundation for various downstream tasks with different data modalities. A PFM (e.g., BERT, ChatGPT, and GPT-4) is trained on large-scale data which provides a reasonable parameter initialization for a wide range of downstream applications. In contrast to earlier approaches that utilize convolution and recurrent modules to extract features, BERT learns bidirectional encoder representations from Transformers, which are trained on large datasets as contextual language models. Similarly, the Generative Pretrained Transformer (GPT) method employs Transformers as the feature extractor and is trained using an autoregressive paradigm on large datasets. Recently, ChatGPT shows promising success on large language models, which applies an autoregressive language model with zero shot or few shot prompting. The remarkable achievements of PFM have brought significant breakthroughs to various fields of AI in recent years. Numerous studies have proposed different methods, datasets, and evaluation metrics, raising the demand for an updated survey.

This study provides a comprehensive review of recent research advancements, challenges, and opportunities for PFMs in text, image, graph, as well as other data modalities. The review covers the basic components and existing pretraining methods used in natural language processing, computer vision, and graph learning. Additionally, it explores advanced PFMs used for different data modalities and unified PFMs that consider data quality and quantity. The review also discusses research related to the fundamentals of PFMs, such as model efficiency and compression, security, and privacy. Finally, the study provides key implications, future research directions, challenges, and open problems in the field of PFMs. Overall, this survey aims to shed light on the research of the PFMs on scalability, security, logi-

https://arxiv.org/pdf/2302.09419

Large Language models are on Hype!!!!!



A Survey of Large Language Models

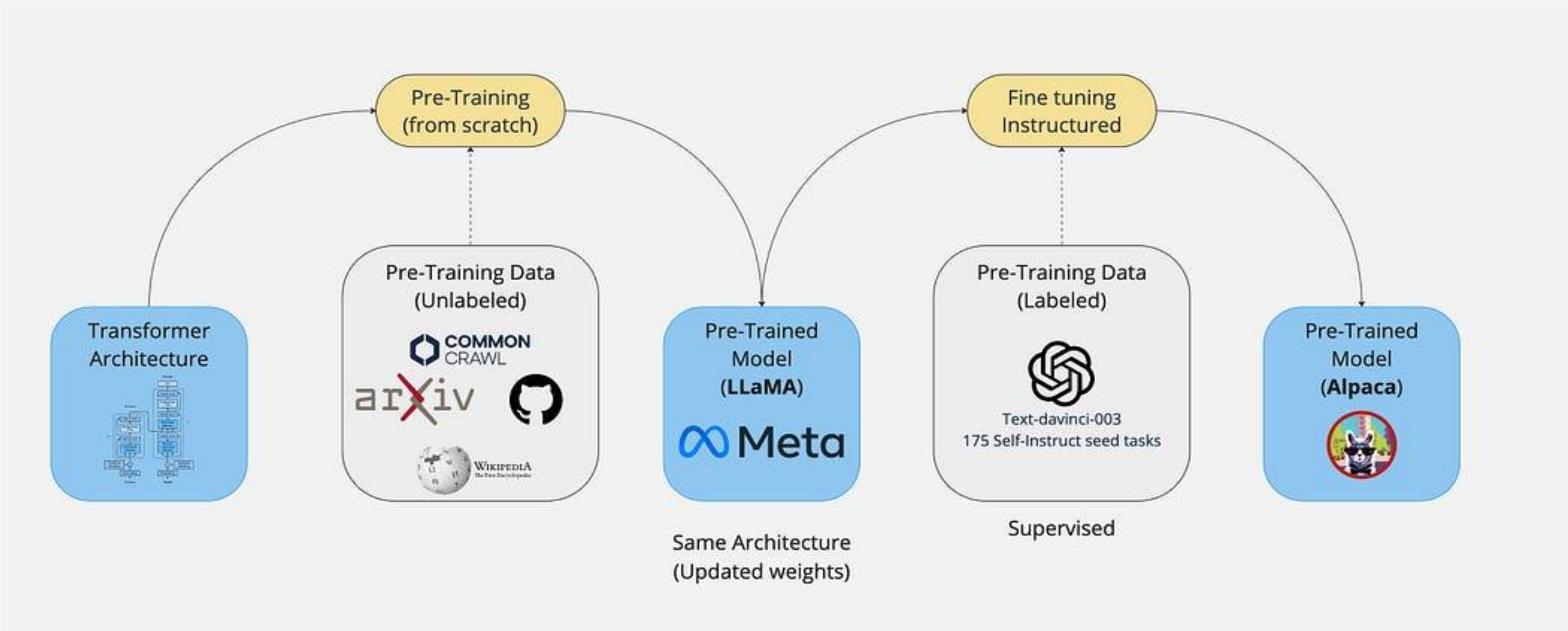
Wayne Xin Zhao, Kun Zhou*, Junyi Li*, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie and Ji-Rong Wen

Abstract—Ever since the Turing Test was proposed in the 1950s, humans have explored the mastering of language intelligence by machine. Language is essentially a complex, intricate system of human expressions governed by grammatical rules. It poses a significant challenge to develop capable artificial intelligence (AI) algorithms for comprehending and grasping a language. As a major approach, *language modeling* has been widely studied for language understanding and generation in the past two decades, evolving from statistical language models to neural language models. Recently, pre-trained language models (PLMs) have been proposed by pre-training Transformer models over large-scale corpora, showing strong capabilities in solving various natural language processing (NLP) tasks. Since the researchers have found that model scaling can lead to an improved model capacity, they further investigate the scaling effect by increasing the parameter scale to an even larger size. Interestingly, when the parameter scale exceeds a certain level, these enlarged language models not only achieve a significant performance improvement, but also exhibit some special abilities (*e.g.*, in-context learning) that are not present in small-scale language models (*e.g.*, BERT). To discriminate the language models in different parameter scales, the research community has coined the term *large language models (LLM)* for the PLMs of significant size (*e.g.*, containing tens or hundreds of billions of parameters). Recently, the research on LLMs has been largely advanced by both academia and industry, and a remarkable progress is the launch of ChatGPT (a powerful AI chatbot developed based on LLMs), which has attracted widespread attention from society. The technical evolution of LLMs has been making an important impact on the entire AI community, which would revolutionize the way how we develop and use AI algorithms. Considering this rapid technical progress, in this survey, we review the recent advances of LLMs by introducing the background, key findings, and mainstream techniques. In particular, we focus on four major aspects of LLMs, namely pre-training, adaptation tuning, utilization, and capacity evaluation. Furthermore, we also summarize the available resources for developing LLMs and discuss the remaining issues for future directions. This survey provides an up-to-date review of the literature on LLMs, which can be a useful resource for both researchers and engineers.

Index Terms—Large Language Models; Emergent Abilities; Adaptation Tuning; Utilization; Alignment; Capacity Evaluation

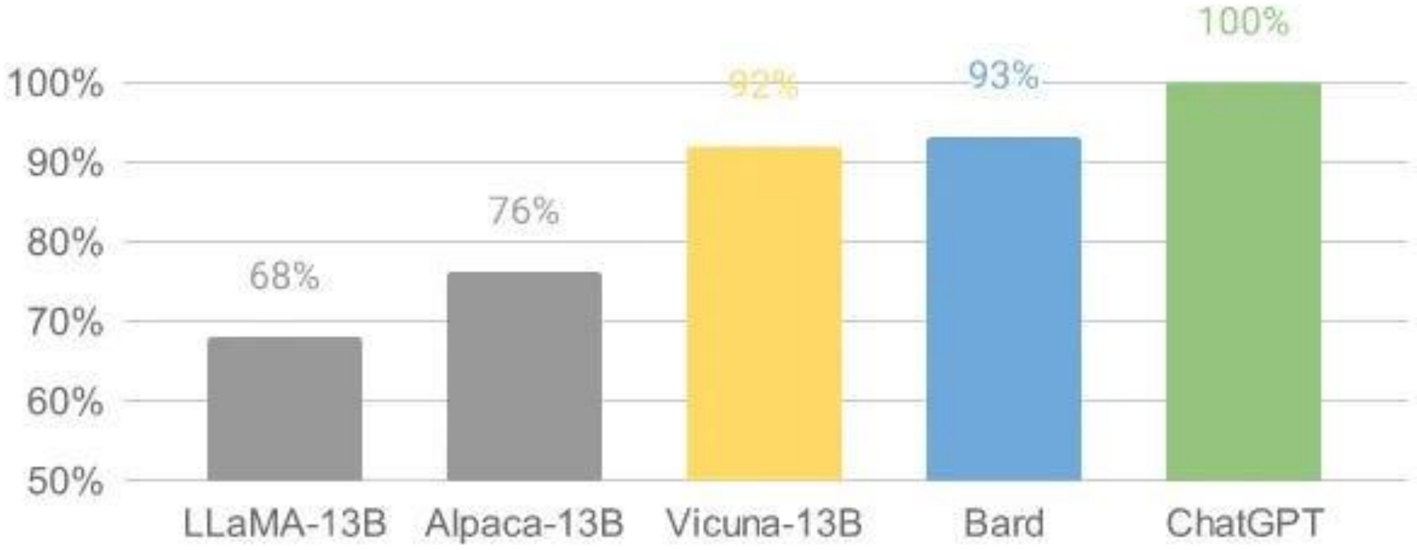
https://arxiv.org/abs/2303.18223

Example of fine-tuning a LLaMA-based model



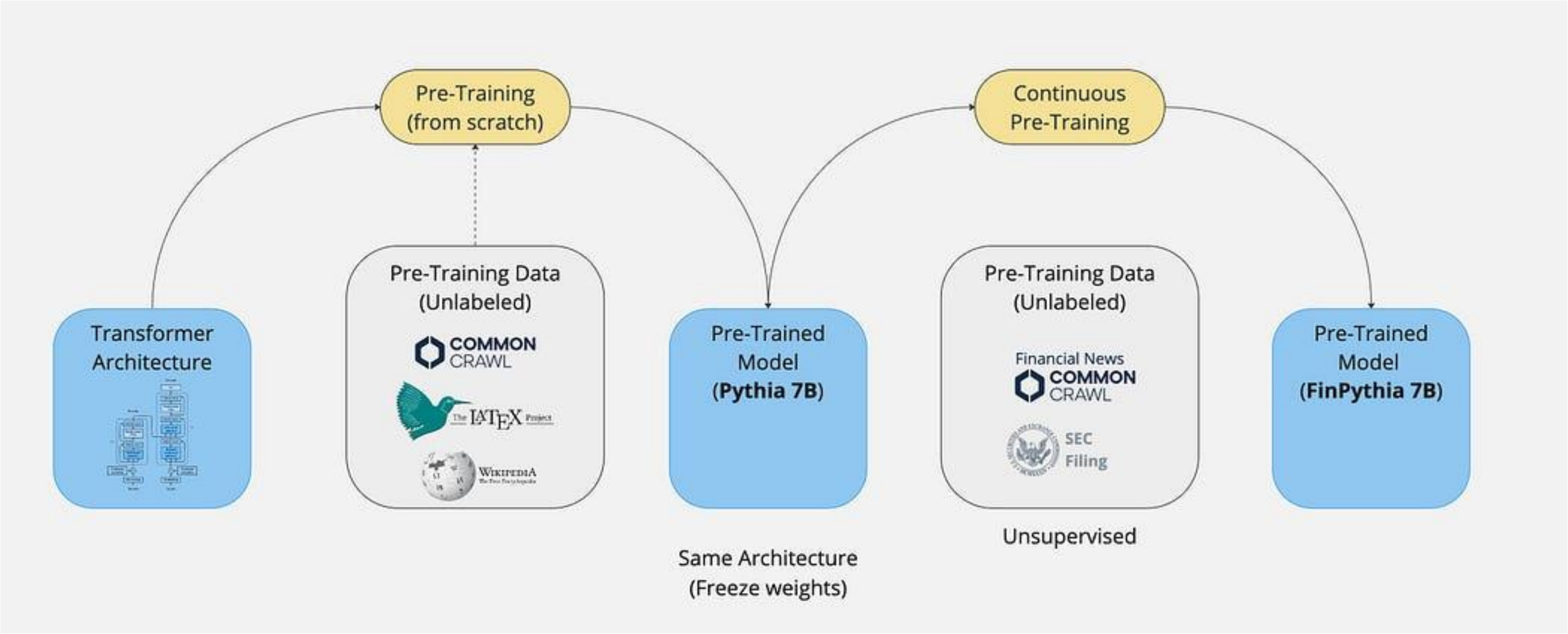
Example of fine-tuning a LLaMA-based model

Relative Response Quality Assessed by GPT-4



Relative Response Quality Assessed by GPT-4

Pre-train a Pythia based model



Example of further pre-train a Pythia based model

Potential harm with LLM

- Hallucination
- Toxic language
- Misinformation

What's Next? /Topics to cover

- Chapter 17-
  - Sequence labeling-POS tagging
  - Named Entity Recognition/NER

Acknowledgments

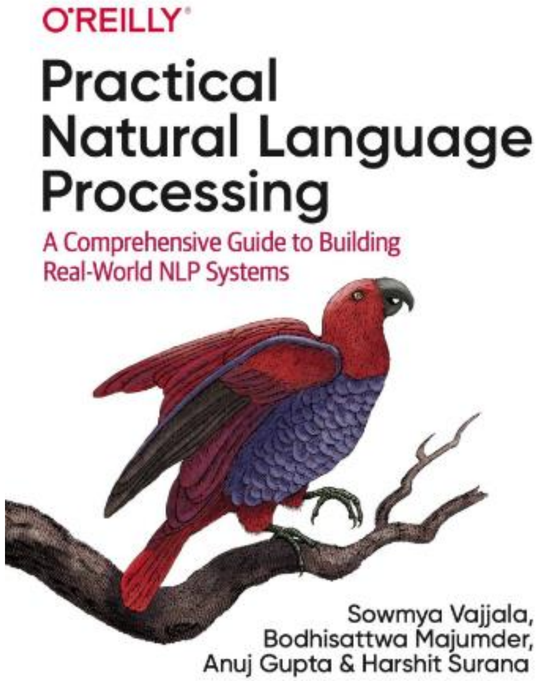
- These slides were adapted from the book SPEECH and LANGUAGE PROCESSING: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition
- Some modifications from CS224N presentations and resources found in the WEB by several scholars.

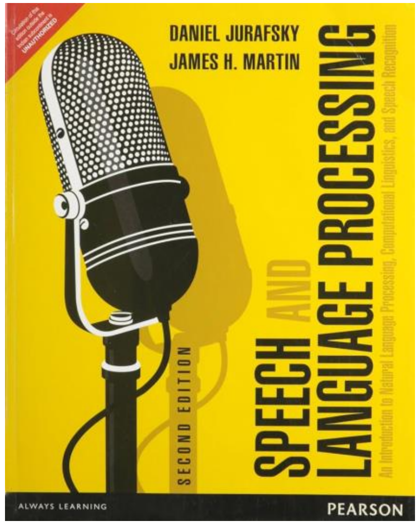
Reference materials

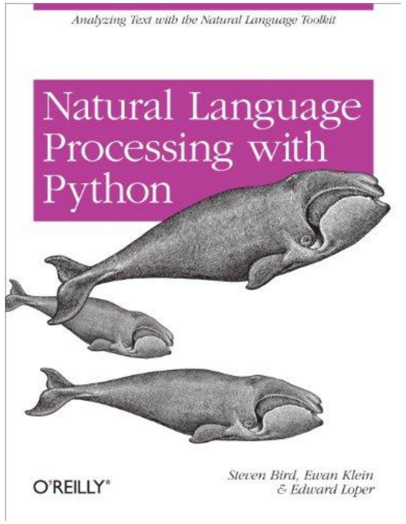
- https://lab.vlanc.co.in/teachings/fall2024-AI-schedule.html

- Lecture notes

- (A) Speech and Language Processing by Daniel Jurafsky and James H. Martin
- (B) Natural Language Processing with Python. (updated edition based on Python 3 and NLTK 3) Steven Bird et al. O’Reilly Media


O'REILLY
Practical Natural Language Processing
A Comprehensive Guide to Building Real-World NLP Systems
Sowmya Vajjala, Bodhisattwa Majumder, Anuj Gupta & Harshit Surana


DANIEL JURAFSKY
JAMES H. MARTIN
SECOND EDITION
SPEECH AND LANGUAGE PROCESSING
An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition
ALWAYS LEARNING
PEARSON


Analyzing Text with the Natural Language Toolkit
Natural Language Processing with Python
O'REILLY
Steven Bird, Ewan Klein & Edward Loper