Lecture 2:
LLM Concepts

SPRING 2025

MOHAMED FARAG

FARAG@CMU.EDU

Agenda

- NLP as Introductory Domain to LLMs
- What are LLMs?
- Key Components of LLMs
- Exercises
- Lecture Notebook: https://colab.research.google.com/drive/17UOIxhKWMAy9qg9r3nvZEgCmZSSO3erT?usp=sharing

NLP and LLMs

- LLMs (Large Language Models) are a type of model used in NLP (Natural Language Processing).

- NLP is a broader field focused on enabling computers to understand and generate human language.

Natural Language Processing (NLP)

- Statistical methods, large datasets, and deep learning led to ML-based NLP adoption in the 2000s and 2010s.
- ML-based NLP systems are used in customer support chatbots, virtual assistants, sentiment analysis, and machine translation.
- Late 2010s saw the emergence of pre-trained language models like ELMo, GPT, and BERT.
- These models which are pre-trained on large data and fine-tuned for specific NLP tasks have achieved top results in benchmarks.
- As a result, there has been a significant progress in language understanding, text generation, and other NLP tasks due to these developments.
- NLP became crucial in various modern applications and services.

Key NLP & LLM Concepts

- **Tokenization:** Breaking text into smaller units (words or sub-words) called tokens.
- **Part-Of-Speech (POS) tagging:** Assigning grammatical tags (noun, verb, adjective, etc.) to each word in a sentence.
- **Word embeddings:** Creating dense vector representations of words (e.g., Word2Vec, GloVe) that capture semantic relationships.
- **Stemming and lemmatization:** Reducing words to their base or root form (e.g., 'running' to 'run’)
- **Language models:** Predicting word sequence likelihood, crucial for tasks like machine translation and text generation.

Popular NLP Tasks

NLP aims to bridge human language and computer understanding, applied in various language tasks.

- **Text classification:** Labeling texts, like spam detection, sentiment analysis, and topic categorization.
- **Named Entity Recognition (NER):** Identifying and classifying entities within text (people, organizations, locations, dates).

Digression - NER

NER has a wide range of applications across various domains, e.g., information retrieval, question answering, and sentiment analysis.

Several Python libraries are used for NER.

- **spaCy**
- **NLTK**
- **Stanford NER**
- **AllenNLP**

Digression - NER Example – POS Tagging

POS tagging involves assigning grammatical labels (e.g., nouns, verbs, adjectives) to words in a sentence. The following are examples of tags.

Check the notebook for an example

Tag	Meaning	Examples
CC	Coordinating conjunction	and, but, or
CD	Cardinal number	one, two
DT	Determiner	the, a, an
EX	Existential there	there (as in "there is")
FW	Foreign word	d'hoevre, faux
TO	to	to (as in "to run")
NNP	Proper noun, singular	John, London

Popular NLP Tasks – Cont’d

- **Language translation:** Automatic translation between languages.
- **Text generation:** Creating human-like text for chatbots, autogenerated content, or summarization.
- **Speech recognition:** Converting spoken language into written text.
- **Text summarization:** Generating concise summaries of longer texts.
- **Question answering:** Providing answers to natural language questions.

These tasks form the foundation of current NLP applications.

What are Large Language Models?

LLMs are **large**,
general-purpose language models
that can be **pre-trained** and
then **fine-tuned** for specific purposes.

Large?

1. Large training datasets
2. Large number of parameters (billions of parameters).
    - Linear regression has 2 parameters!!

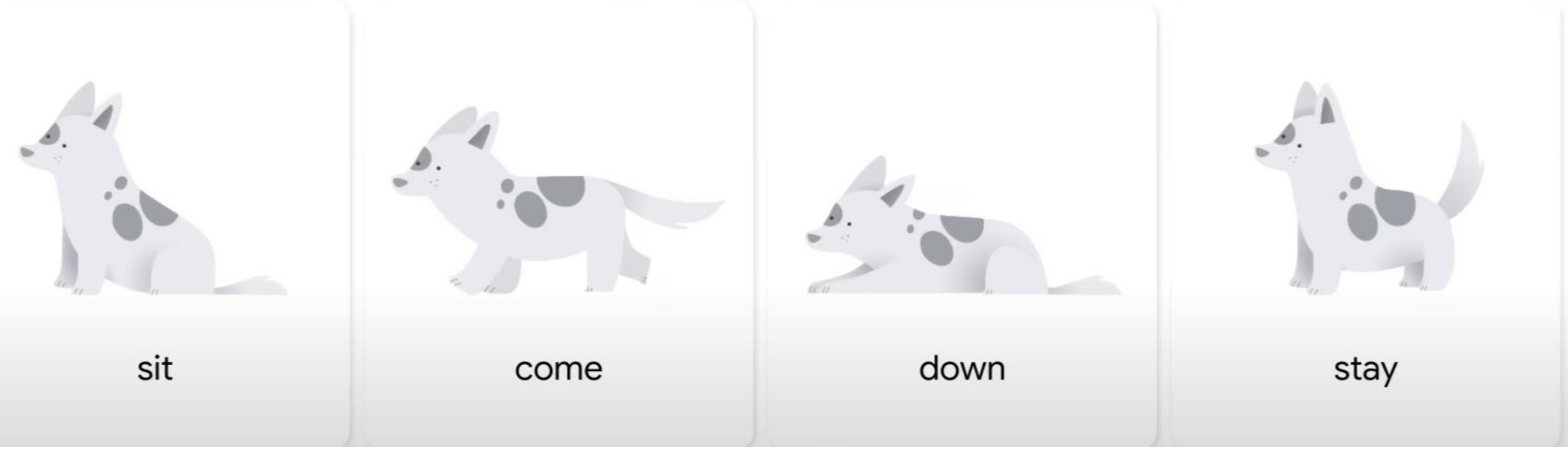
General Purpose?

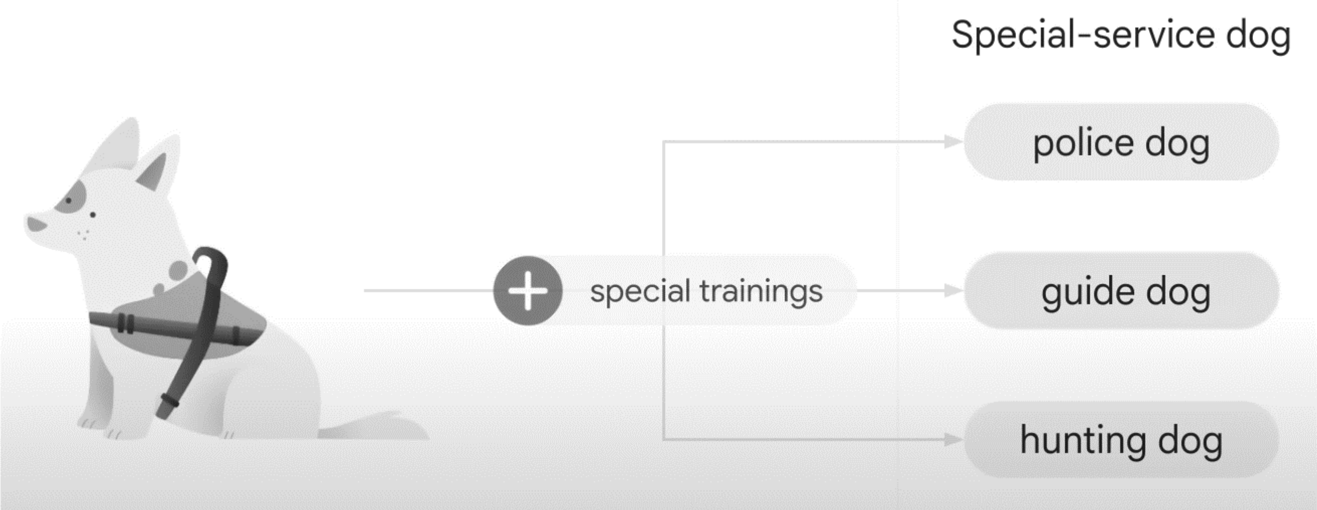
1. Commonality of human languages
2. Resource restrictions

Pre-trained and fine-tuned

Pre-trained

Fine-tuned





Images from Google’s Introduction to Large Language Model

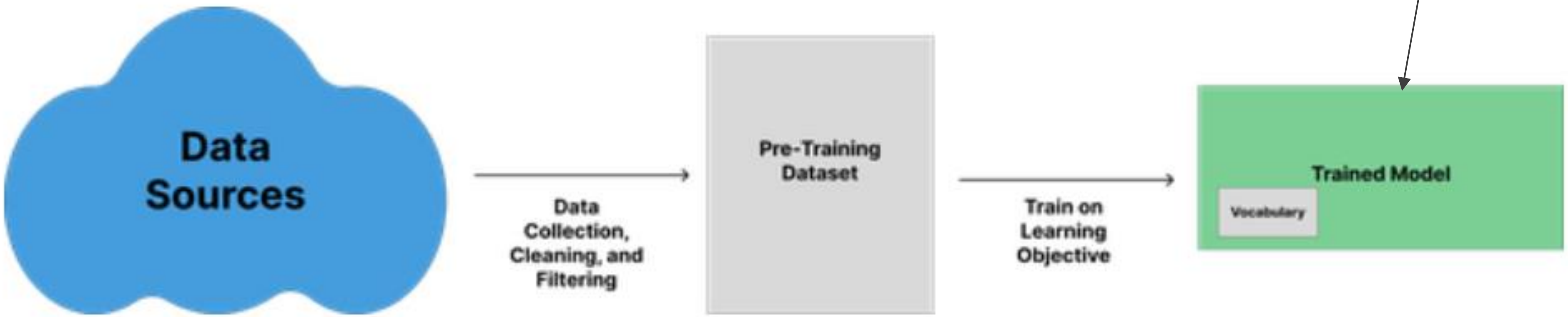
In your opinion, what is the difference between Prompt Tuning and Instruction Tuning?

Key Components of LLMs

1. Pre-training Data
2. Vocabulary and Tokenizer
3. Learning Objective
4. Architecture

We will focus on these two areas

Base Model


Data Sources
Data Collection, Cleaning, and Filtering
Pre-Training Dataset
Train on Learning Objective
Trained Model
Vocabulary

1. LLM Components: Pre-training Data

- Here, our goal is to answer the question: **“What’s it trained on?”**
- It’s import to use high-quality data to avoid “Garbage-in, Garbage-out”.
- Pre-trained data come from “Corpus”.
- A corpus is a large collection of text or utterances used for language analysis and model training.

LLM Components: Pre-training Data
Corpora Types

1. **Text Corpora**
   - Collection of written texts (books, articles, web pages, emails, social media posts).
   - Used for language modeling, sentiment analysis, text classification, information retrieval.
2. **Speech Corpora**
   - Contains audio recordings or transcriptions of spoken language.
   - Utilized in speech recognition, speaker identification, emotion detection.

LLM Components: Pre-training Data

Corpora Types (Cont’d)

3. **Parallel Corpora**
   - Text in multiple languages aligned at sentence or document level.
   - Employed for machine translation and cross-lingual tasks.
4. **Treebanks**
   - Annotated corpora with syntactic parse trees.
   - Used in parsing and syntax-based machine learning
5. **Multimodal Corpora**
   - Includes text and other modalities like images, videos, or audio.
   - Applied in tasks involving multiple modalities' understanding and generation.

In your opinion, which corpora type is used the most for LLMs?

Example: BERT Pre-training Data Sources

1. **English Wikipedia:**
   - Contains articles from the English Wikipedia.
   - Diverse topics and writing styles, representing English language well.
   - Size: 2.5 billion words.
2. **The BookCorpus:**
   - Large collection of fiction and non-fiction books scraped from the web.
   - Includes various genres like romance, mystery, science fiction, and history.
   - Books have a minimum of 2000 words and are written by verified authors.
   - Size: 800 million words.

LLM Components: Pretrained Data Open Topics

1. **There is an ongoing debate on whether the text data are sufficient to teach the model on logical reasoning**

   - Only around 12% of information we understand from text is explicitly mentioned in text

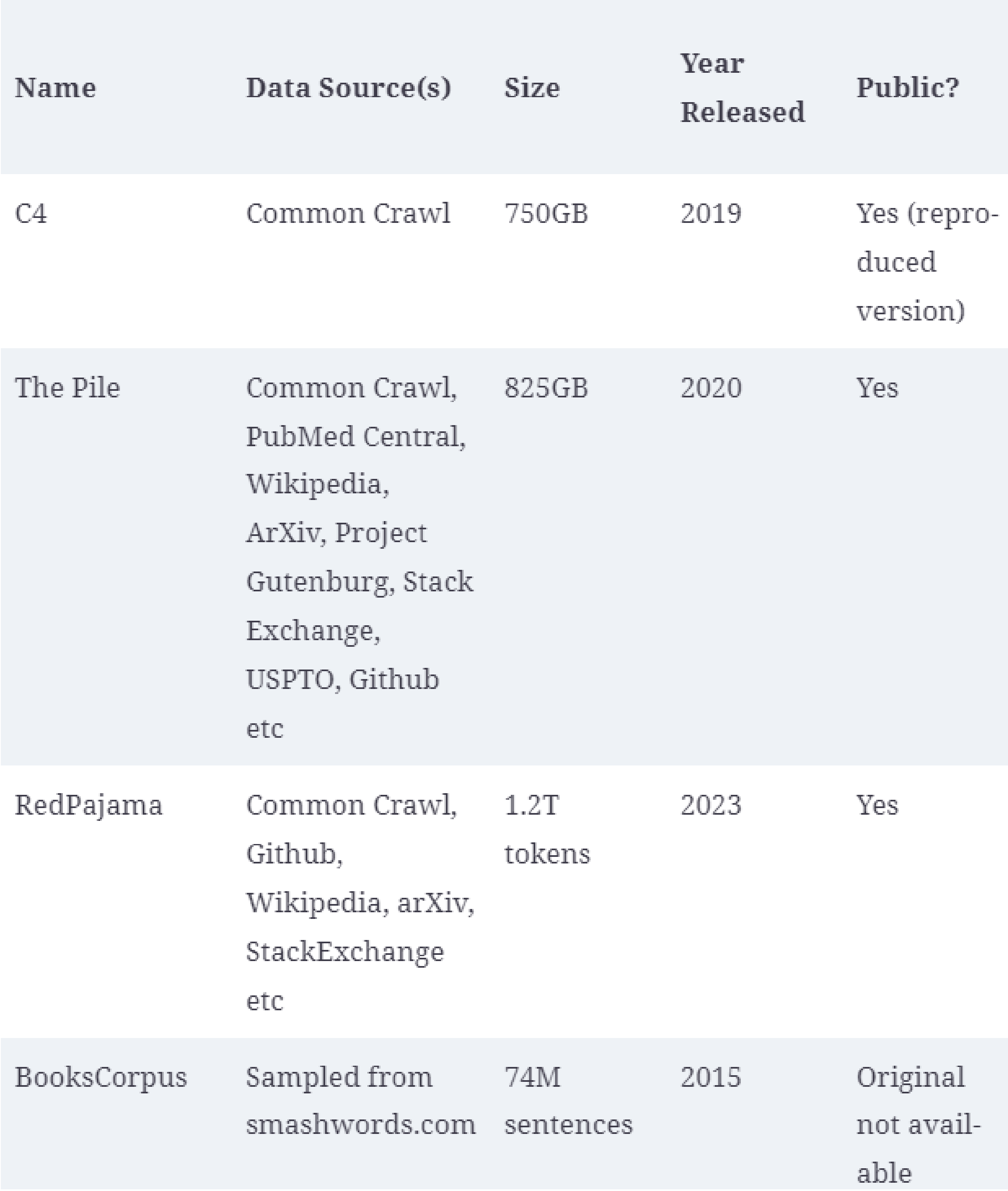
   - Multimodal models combine different modalities like image, video, speech, and text. They are becoming a promising avenue of research and are likely to see more widespread usage in the coming years.

2. **LLMs are usually trained using one epoch and they are considered underfit**. Recently, some research shows that LLMs can be trained with about 5 epochs.

LLM Components:

Examples of Popular Text Corpora

[Check C4 Dataset]



Name	Data Source(s)	Size	Year Released	Public?
C4	Common Crawl	750GB	2019	Yes (reproduced version)
The Pile	Common Crawl, PubMed Central, Wikipedia, ArXiv, Project Gutenburg, Stack Exchange, USPTO, Github etc	825GB	2020	Yes
RedPajama	Common Crawl, Github, Wikipedia, arXiv, StackExchange etc	1.2T tokens	2023	Yes
BooksCorpus	Sampled from smashwords.com	74M sentences	2015	Original not available

2. LLM Components: Vocabulary and Tokenizer

- Here, our goal is to answer the question: **“What’s it trained over?”**
- We need to determine the language's vocabulary and tokenization rules.
- Humans process language in terms of meaning-bearing words and sentences while Language models process language in terms of **tokens**.
- The term **token** refers to the smallest unit of semantic meaning created by breaking down a sentence or piece of text into smaller units and are the basic inputs for an LLM.
- Tokens can be words but also can be “sub-words”

Tokenization Types

- Word-based: Splitting text by spaces
  - "I love AI" → ["I", "love", "AI"])
- Subword-based: Breaking words into meaningful fragments
  - "unbelievable" → ["un", "believable"])
- Character-based: Treating each character as a token
  - "AI" → ["A", "I"]

LLaMa 2 Tokenization

- LLaMa 2 utilizes a BPE tokenizer that divides numbers into separate digits and decomposes unfamiliar UTF-8 characters into bytes.
- It has a total vocabulary of 32,000 tokens

Exercise: Find Number of Tokens in a Sentence (Check Lecture’s Colab notebook)

How many tokens are in

```
"what is 937 + 934?"
```

LLM Components: Vocabulary

- A vocabulary in NLP refers to the set of unique words or tokens present in a corpus of text.
- Vocabulary is a fundamental component of language processing, as it defines the complete list of words that a model or system can understand and work with

LLM Components: Vocabulary Creation

1. **Tokenization:**

    - Splitting text into individual tokens (words, subwords, or characters).
    - Depends on the chosen tokenization strategy.

2. **Filtering and Normalization:**

    - Common steps include converting text to lowercase, removing punctuation.
    - Filtering out stop-words to clean data and reduce vocabulary size.

LLM Components: Vocabulary Creation (Cont’d)

3. **Building Vocabulary:**

    - Collecting unique tokens post-tokenization and preprocessing.
    - Assigning each token a unique numerical index for model representation or encoding.

- In many LLM models, words are represented as dense vectors (**word embeddings**) where each word’s embedding is indexed using its integer representation in the vocabulary.

Exercise: Calculate the Embeddings for a given word in Context.

Refer to the Colab Notebook

Quiz-1 Google Form for Waitlisted Students

Waitlisted Students

- All materials for first two weeks will be uploaded here