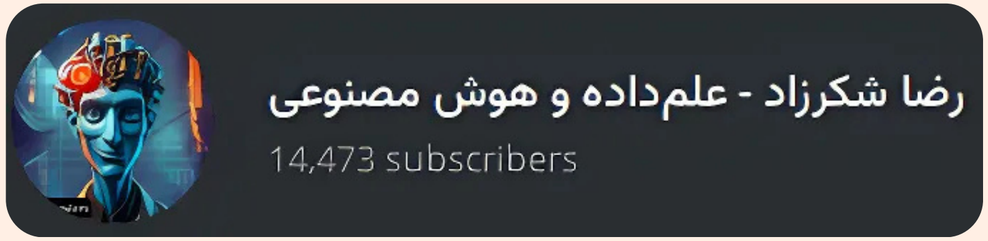
Reza Shokrzad

April 2025

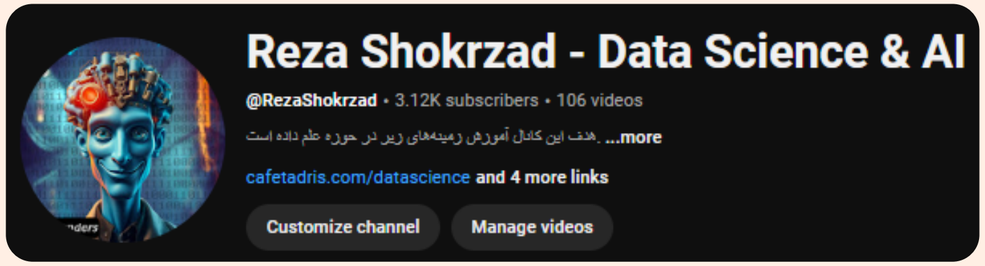
NLP Workshop

Session 6 - Large Language Models

Communities



@DSLanders



@RezaShokrzad

Content

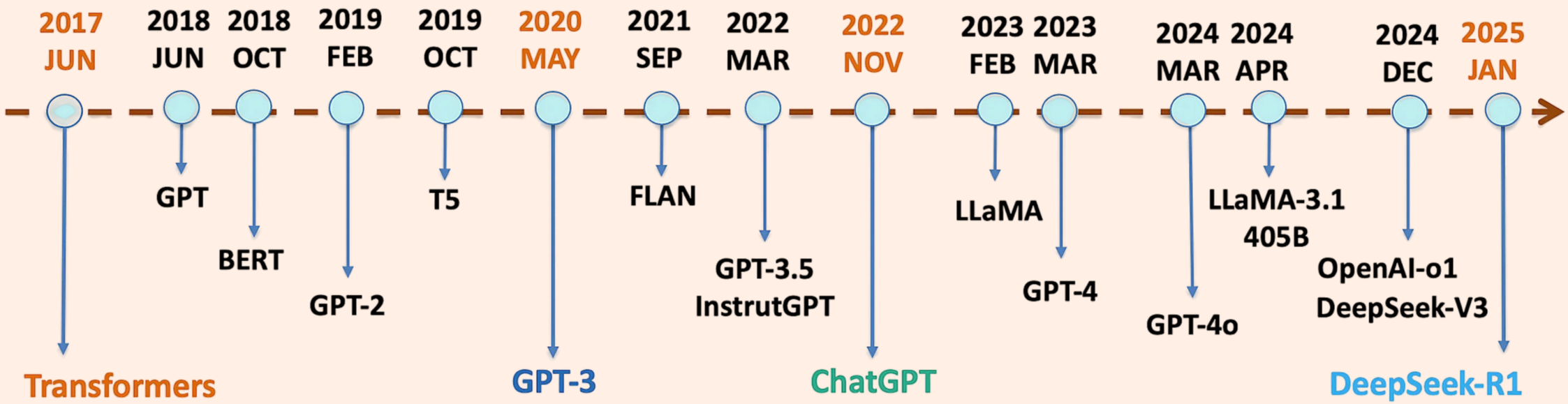
- Modern LLMs
- Model Architecture Categories
- The BERT Revolution and Its Descendants
- Open-Source LLMs
- Hugging Face
- Deployment Setup

LLMs: Large Language Models

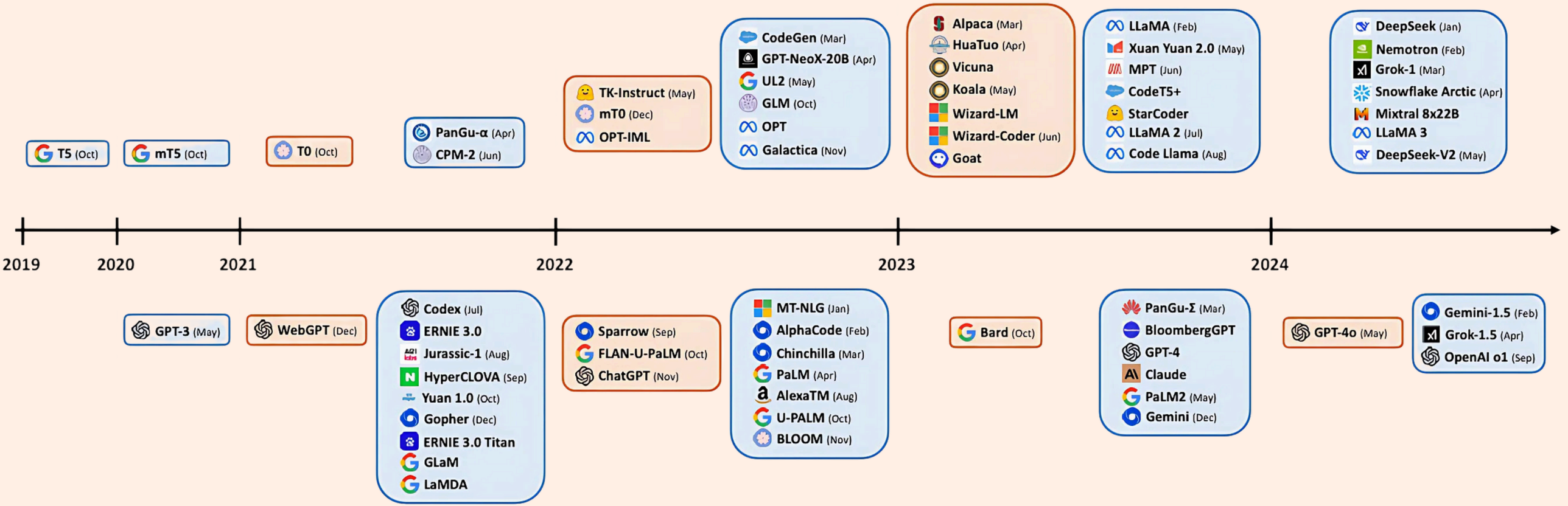
... are AI models trained on massive text datasets to understand and generate human-like text.

https://en.wikipedia.org/wiki/Large_language_model

LLMs Timeline



LLMs Timeline



LLMs Trends


LLM PAPERS' NUMBERS
160
140
120
100
80
60
40
20
0
10
58
120
2022
2023
2024
Papers number
10
58
120
PUBLISHING YEAR

Foundation Model Families

BERT Descendants

- **BERT:** The original bidirectional transformer that reads text in both directions
- **RoBERTa:** Optimized BERT with improved training methodology
- **DistilBERT:** Compressed BERT for efficiency while maintaining performance
- **ALBERT:** Parameter-efficient BERT that shares layers
- **DeBERTa:** Enhanced BERT with improved attention mechanisms

GPT Family

- **GPT-2:** Early powerful text generation model
- **GPT-3:** Scaled-up architecture with powerful few-shot learning abilities

Open-Source Models

- **LLaMA:** Meta's foundation models designed for research accessibility
- **Mistral:** Efficient architecture optimized for smaller deployment
- **Falcon:** Performance-focused open models with various size options
- **Phi:** Microsoft's compact but powerful reasoning models

Architecture Pattern

Encoder-Only Models

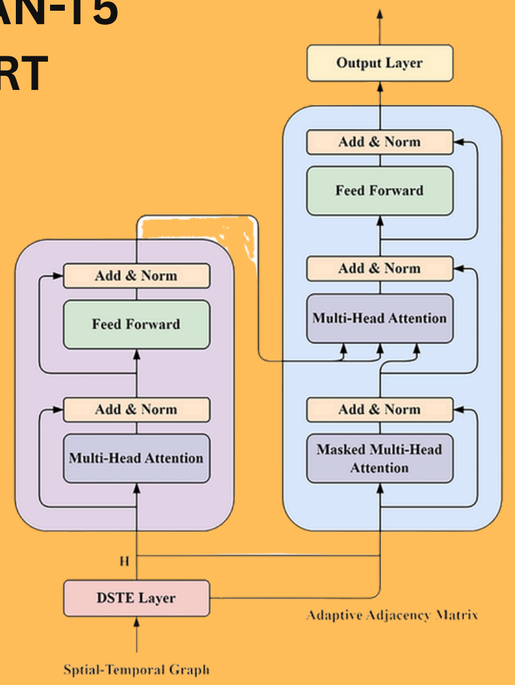
- BERT
- RoBERTa
- ALBERT
- DistilBERT
- ELECTRA

Decoder-Only Models

- GPT-2
- GPT-3
- LLaMA
- Mistral
- Falcon
- Phi

Encoder-Decoder

- T5
- FLAN-T5
- BART



Core Algorithms in LLMs

- **BPE (Byte-Pair Encoding):** Tokenization through subword merging.
- **WordPiece:** Google's vocabulary-building tokenization approach.
- **SentencePiece:** Unsupervised text tokenizer/detokenizer.
- **Transformers:** Self-attention based architecture foundation.
- **Rotary Embedding:** Position encoding via rotation.
- **KV-Cache:** Computation optimization through caching.
- **Quantization:** Reduces precision for efficiency.
- **RLHF:** Aligns models with human feedback.
- **Beam Search:** Efficient text generation strategy.

Open-source LLMs

- Open-source LLMs enable enhanced **security**, **customization**, and **cost-efficiency**.
- On-premises deployments now control more than **half of market**.
- Open-source LLMs come in **pre-trained** and **fine-tuned** variants.

Open-source LLMs

Advantages

- **Ownership:**
  - Complete control over model applications.
- **Accuracy:**
  - Flexible local parameter customization possible.
- **Longevity:**
  - Self-hosted models don't become obsolete.
- **Costs:**
  - Predictable infrastructure expenses versus usage.
- **Flexibility:**
  - Choose optimal software/hardware combinations.
- **Community:**
  - Enables optimization and efficient deployment.

Drawbacks

- **Quality:**
  - May lack large corporation resources.
- **Security:**
  - Vulnerable to input manipulation attacks.
- **Licensing:**
  - Varies from permissive to restricted.

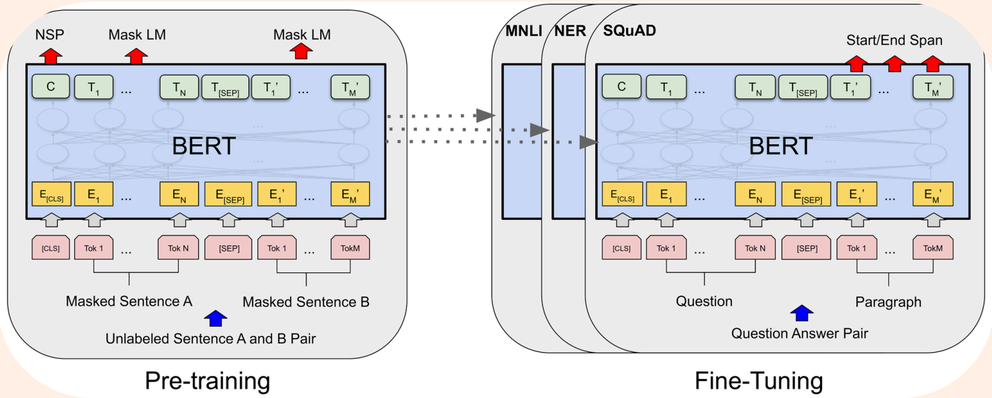
Open-source LLM leaderboard

Model Family	Developer	Params	Context window	Use-cases	License
Llama 3	Meta	1B, 3B, 8B, 70B, 405B	8k, 128k	General text generation, Multilingual tasks, Code generation, Long-form content, Fine-tuning for specific domains	Llama Community License
Mistral	Mistral AI	3B-124B	32k-128k	High-complexity tasks, Multilingual processing, Code generation, Image understanding, Edge computing, On-device AI, Function calling, Efficient large-scale processing	Apache 2.0 Mistral Research License Commercial License
Falcon 3	TII	1B, 3B, 7B, 10B	8k-32k	General text generation, Code generation, Mathematical tasks, Scientific knowledge, Multilingual applications, Fine-tuning for specific domains	TII Falcon License
Gemma 2	Google	2B, 9B, 27B	8k	General text generation, Question answering, Summarization, Code generation, Fine-tuning for specific domains	Gemma license
Phi-3.x / 4	Microsoft	3.8B (mini) 7B (small) 14B (medium) 42B (MoE)	4k, 8k, 128k 16k (Phi-4)	General text generation, Multi-lingual tasks, Code understanding, Math reasoning, Image understanding (vision model), On-device inference	Microsoft Research License
Command R	Cohere	7B, 35B, 104B	128k	Conversational AI, RAG, Tool use, Multilingual tasks, Long-form content generation	CC-BY-NC 4.0
StableLM 2	Stability AI	1.6B, 3B, 12B	Up to 16k	Multilingual text generation, Code generation and understanding, Fine-tuning for specific tasks, Research and commercial applications	Stability AI Community and Enterprise licenses
StarCoder2	BigCode	3B, 7B, 15B	16k	Code completion, Multi-language programming, Code understanding, Fine-tuning for specific tasks	Apache 2.0
Yi	01.AI	6B, 9B, 34B	4k, 8k, 200k	Bilingual text generation, Code understanding and generation, Math and reasoning tasks, Fine-tuning for specific domains	Apache 2.0
Qwen2.5	Alibaba	0.5B to 72B	128K	General text generation, Multilingual tasks, Code generation, Mathematical reasoning, Structured data processing	Qwen license (3B and 72B size models) Apache 2.0 (others)
DeepSeek-V2.x/V3	DeepSeek AI	16B, 236B, 671B for V3 (2.4B-37B activated)	32k-128k	General text generation, Multilingual tasks, Code generation, Fine-tuning, Advanced reasoning (V3)	DeepSeek License

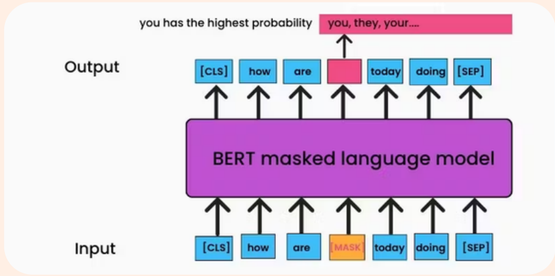
https://blog.n8n.io/open-source-llm/

BERT

- **Transformer Encoder:** multiple self-attention layers
- **Bidirectional Attention:** context from both directions
- **WordPiece Tokenization:** subword segmentation method
- **Position Embeddings:** encode token order
- **Segment Embeddings:** add sentence identifiers
- **Masked Language Modeling:** predict masked tokens
- **Next Sentence Prediction:** binary sentence relation
- **Layer Normalization:** stabilize layer activations
- **Pre-training Strategy:** dual unsupervised tasks
- **Fine-tuning Process:** adapt model to tasks



MLM



NPS

Sentence A	The man went to the store.
Sentence B	He bought a gallon of milk.
Label	IsNextSentence

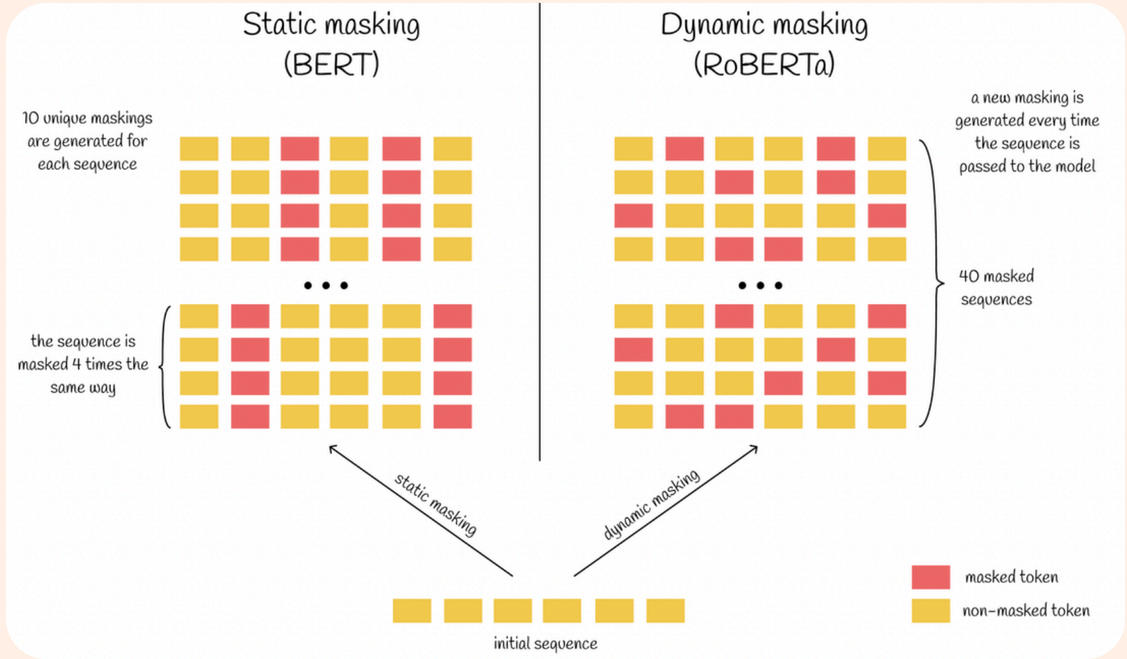
Sentence A	The man went to the store.
Sentence B	Penguins are flightless.
Label	NotNextSentence

https://arxiv.org/abs/1810.04805

https://en.wikipedia.org/wiki/BERT_(language_model)

RoBERTa

- **Dynamic Masking:** new masks each epoch
- **Bigger Batches:** larger batch sizes
- **No NSP Objective:** dropped sentence prediction



https://arxiv.org/pdf/1907.11692

BERT Family

- **BERT:** Masked bidirectional transformer
- **RoBERTa:** Robustly optimized BERT
- **ALBERT:** Parameter-reduced BERT variant
- **DistilBERT:** Distilled lightweight BERT
- **TinyBERT:** Miniaturized BERT distillation
- **ELECTRA:** Replaced token detection
- **SpanBERT:** Span-level masking improvement
- **BioBERT:** Biomedical domain adaptation
- **SciBERT:** Scientific text specialization
- **FinBERT:** Financial domain tuning

GPT

- **GPT:** foundational generative transformer
- **GPT-2:** scaled unsupervised model
- **GPT-3:** few-shot learning powerhouse
- **GPT-3.5:** optimized context handling
- **GPT-4:** multimodal advanced reasoning
- **Codex:** code-generation specialist
- **ChatGPT:** conversational fine-tuned GPT

https://en.wikipedia.org/wiki/Generative_pre-trained_transformer

BART

- **Bidirectional and Auto-Regressive Transformers**
- **Denoising Autoencoder:** reconstruct corrupted text
- **Noising Strategies:** token, span, sentence deletion
- **Encoder-Decoder:** bidirectional + autoregressive

BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension

Mike Lewis*, Yinhan Liu*, Naman Goyal*, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, Luke Zettlemoyer
Facebook AI
{mikelewis,yinhanliu,naman}@fb.com

https://arxiv.org/abs/1910.13461

T5

- **Text-to-Text Transfer Transformer:** unified input-output format
- **Encoder-Decoder:** bidirectional encoder, autoregressive decoder
- **Span Corruption:** fill-in-the-span pretraining
- **C4 (Colossal Clean Crawled Corpus):** massive web-derived corpus
- **Scale Variants:** small to XXL sizes
- **Multi-task:** diverse task mixture
- **Fine-tunable:** easily adapted downstream

https://arxiv.org/abs/1910.10683

Mistral AI

- **Mistral AI:** French startup, founded 2023, raized $428M, valued $2B
- **Mistral 7B:** 7 B dense transformer
- **Mistral 7B Instruct:** instruction-tuned variant
- **Mistral 7B v0.3:** extended vocab & function calling
- **Mixtral 8x7B:** sparse mixture-of-experts

https://en.wikipedia.org/wiki/Mistral_AI

Qwen

- **Alibaba** Cloud’s LLM family launched 2023
- **Variants:**
  - 1.8B–72B parameter models
- **Benchmark Leader:**
  - top Chinese, 3rd globally

https://en.wikipedia.org/wiki/Qwen

DeepSeek

- **Chinese** AI under High-Flyer
- **Mixture-of-Experts:**
  - selective expert activation
- **DeepSeek-R1:**
  - reasoning model, cost-efficient

https://en.wikipedia.org/wiki/DeepSeek

Hugging Face 

- **Model Hub:** share pre-trained models
- **Datasets Hub:** access diverse datasets
- **Transformers:** high-level model library
- **Inference API:** serve models easily
- **Community Forum:** discuss and collaborate

https://huggingface.co

Locate your HF cache

- **Linux / macOS / WSL**

```
export HF_HOME="${HOME}/.cache/huggingface"
export TRANSFORMERS_CACHE="${HF_HOME}/transformers"
```

- **Windows PowerShell**

```
$env:HF_HOME = "$env:USERPROFILE\.cache\huggingface"
$env:TRANSFORMERS_CACHE = "$env:HF_HOME\transformers"
```

List downloaded models

- **Linux / macOS / WSL**

```
# All repo folders in the HF hub cache:
ls ~/.cache/huggingface/hub/models--*
```

- **Windows PowerShell**

```
# All repo folders in the HF hub cache:
Get-ChildItem "$env:USERPROFILE\.cache\huggingface\hub\models--*" -Directory
```

Remove models (or wipe the cache)

- **Linux / macOS / WSL**

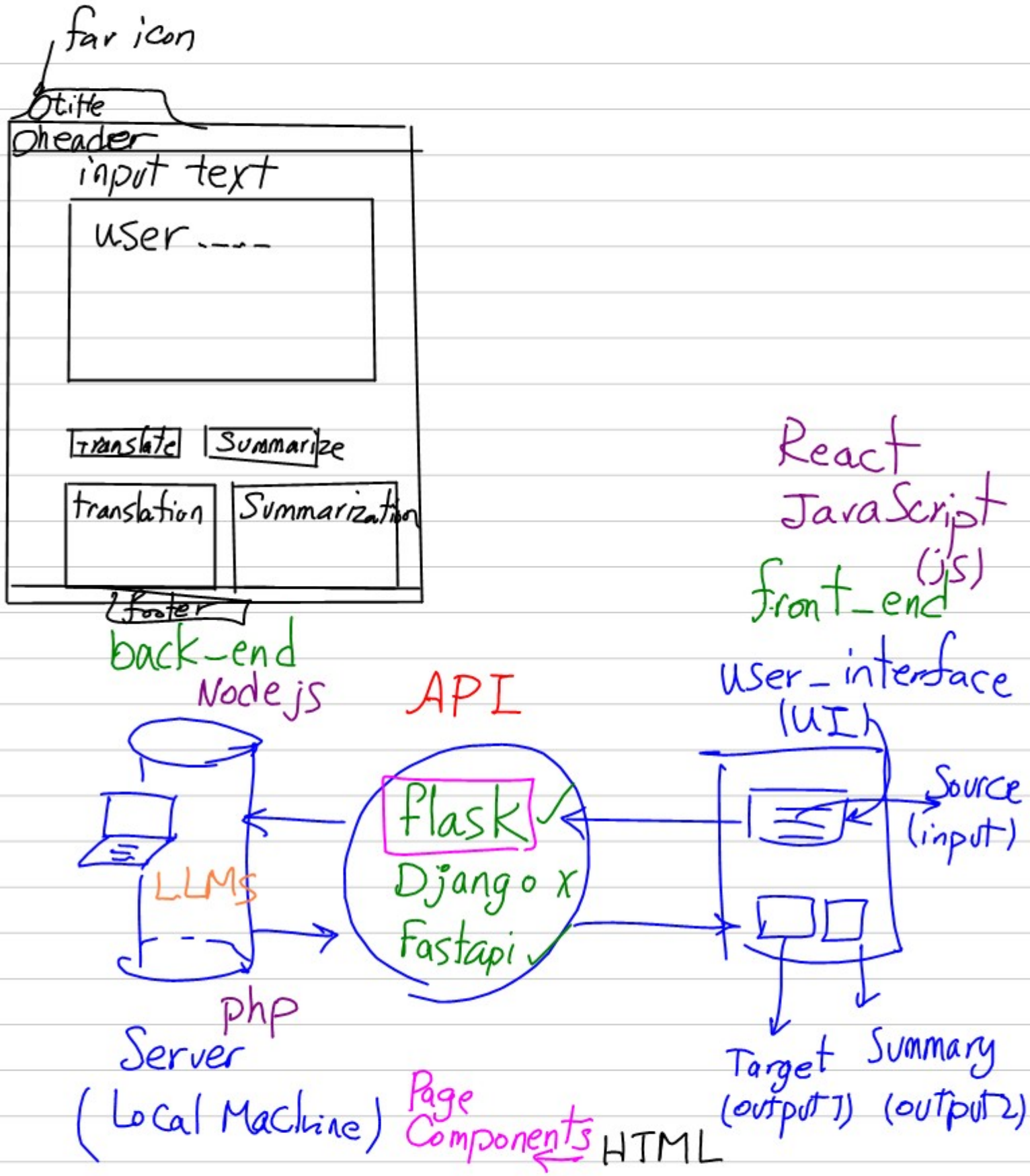
```
rm -rf ~/.cache/huggingface/hub/models--google--flan-t5-large*
```

- **Windows PowerShell**

```
Remove-Item -Recurse -Force "$env:USERPROFILE\.cache\huggingface\hub\models--google--flan-t5-large*"
```

Today's Project Structure

```
language_tools_project/
├── app.py                      # Flask application
├── static/
│   ├── css/
│   │   └── style.css           # Styling for the web app
│   ├── js/
│   │   └── script.js           # JavaScript for interactivity
│   └── images/
│       └── logo.svg            # Logo for the header
├── templates/
│   └── index.html              # Main HTML template
├── models/
│   ├── translator.py           # Translation functionality
│   └── summarizer.py           # Summarization functionality
└── requirements.txt            # Dependencies
```


far icon
title
header
input text
user ....
translate
Summarize
translation
Summarization
footer
back-end
Node.js
API
front_end
React
JavaScript
(JS)
user_interface
(UI)
Flask
Django x
Fastapi
LLMs
php
Server
(Local Machine)
Source
(input)
Target
(output1)
Summary
(output2)
Page
Components
HTML
host
(Server)
Page style
CSS
action
JS
client