LLM Best Practises

ODSC Conference: Module 0

Oct 31, 2023

Sanyam Bhutani, Sr Data Scientist H2O.ai

Mega-Agenda

- 👩‍🎓👨‍🎓 3x Theory Sessions
- 💻 4x Hands on Exercises
- ✍️ 3x Quizzes
- 📄 4x Case Studies

Ground Rules

- I am not world's top LLM Experts! When in doubt, interrupt!

- You are allowed to:

  - Interrupt for breaks
  - Interrupt for repetition
  - Interrupt for questions
  - Interrupt for clarifications
  - Interrupt for examples

- You are not allowed to:
  - Leave without understanding a topic
  - Staying quiet
  - Not Code
  - Having Coffee

Vague Goals

- Build a LLM App
- Have a working understanding of LLMs
- Become your organisation's lead R&D Contact for LLM
- Be able to read and understand LLM Apps
- Be able to digest LLM Papers

Concrete Goals

- Understand the spectrum of the field of LLMs
- Have a working understanding of LLM APIs
- Learn how to work with open source models
- Read and Understand 4 top papers
- Pass all quizzes
- Complete the 4 hands-on exercises

Agenda

- Introduction
- Ice Breaker
- [Hands-on] Small LM vs LLM
- History of LLMs
- What makes a Large Language Model?
- Understanding Current SOTA

Ice Breaker

Please take 30-60 seconds to answer (2 or all):

- Who are you?
- What is your goal for attending?
- What’s one problem you’re excited about?
- How can Sanyam make this session a success for you?

Democratizing AI and LLMs with H2O.ai

H2O.ai

50% OF THE FORTUNE 500 ❤ **H2O**

8 OF THE TOP 10 BANKS

7 OF THE TOP 10 INSURANCE COMPANIES

6 OF THE TOP 10 MANUFACTURING COMPANIES

Customer Obsession
Maker Culture

30+
Kaggle Grandmasters

World's #1, #3, #5, and #9

2.5M+
Community

100K+
h2ogpt requests per month

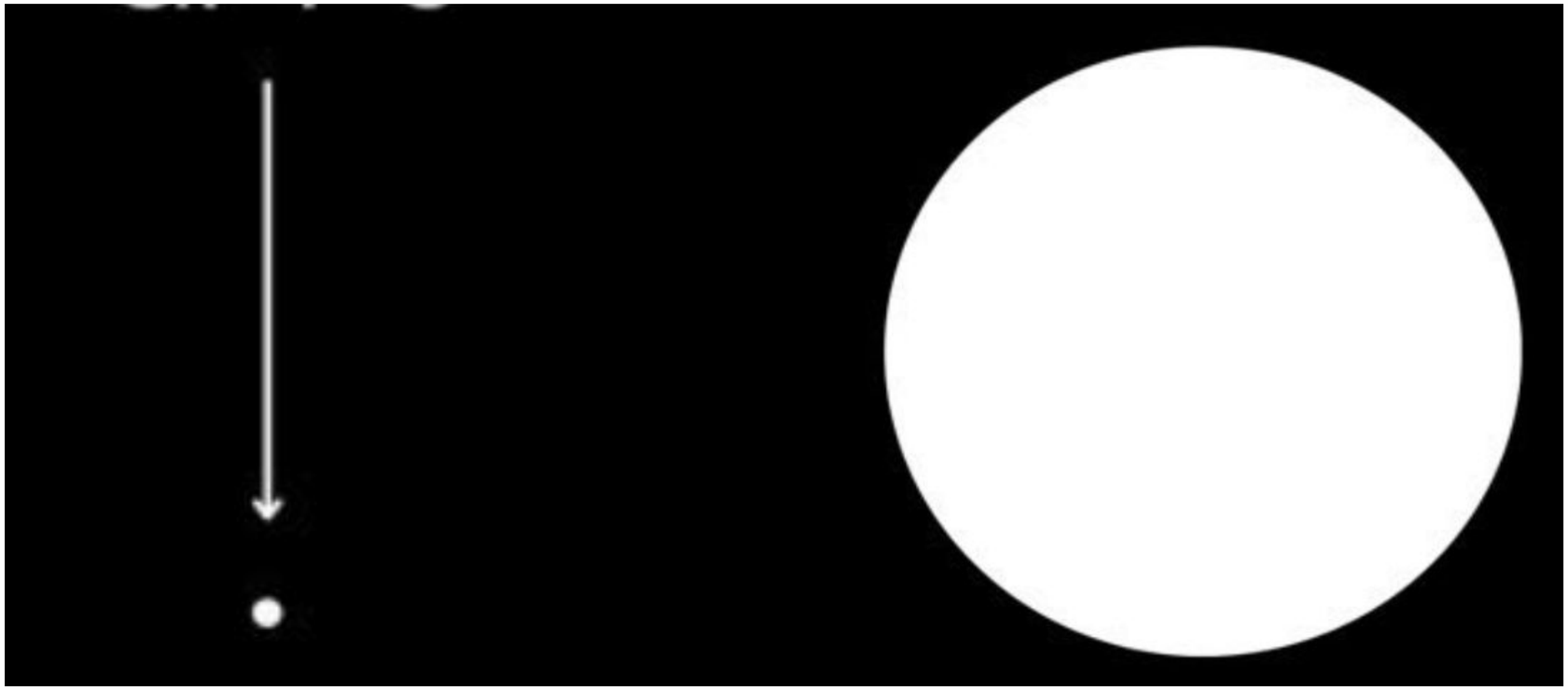
Innovation inspired and powered by World's Top 10% Data Scientist

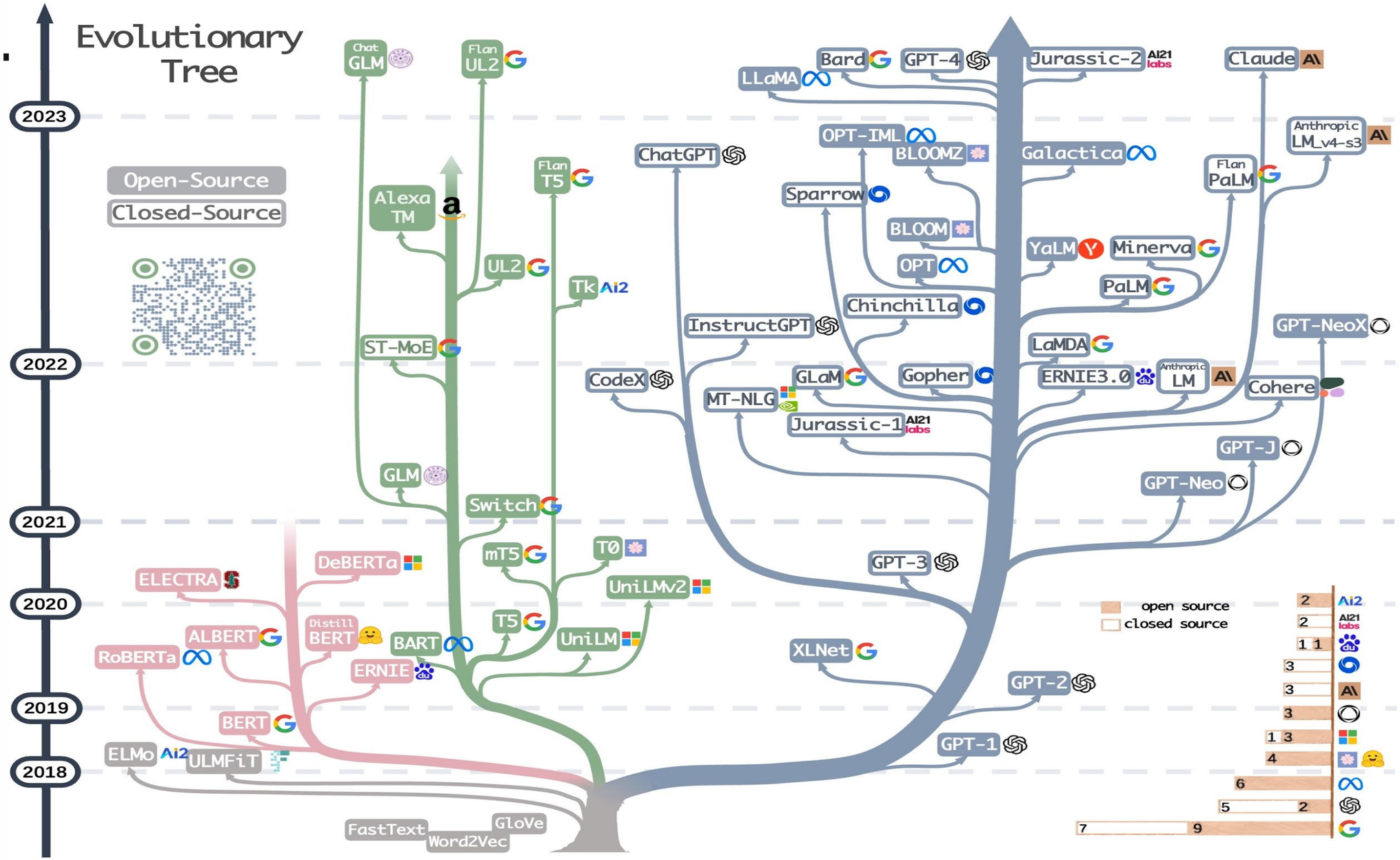
Your projects are backed by 10% of the World's Data Science Grandmasters and a Team of Experts who are relentless in solving your critical problems.

What is a LLM?

Language Model

Large Language Model





Credit: https://github.com/Mooler0410/LLMsPracticalGuide

Emergent Abilities

S Order me a pizza

I got you bro.

S Can you do my homework?

I got you bro.

S Can you solve world hunger?

I got you bro.

S Can you do Math?

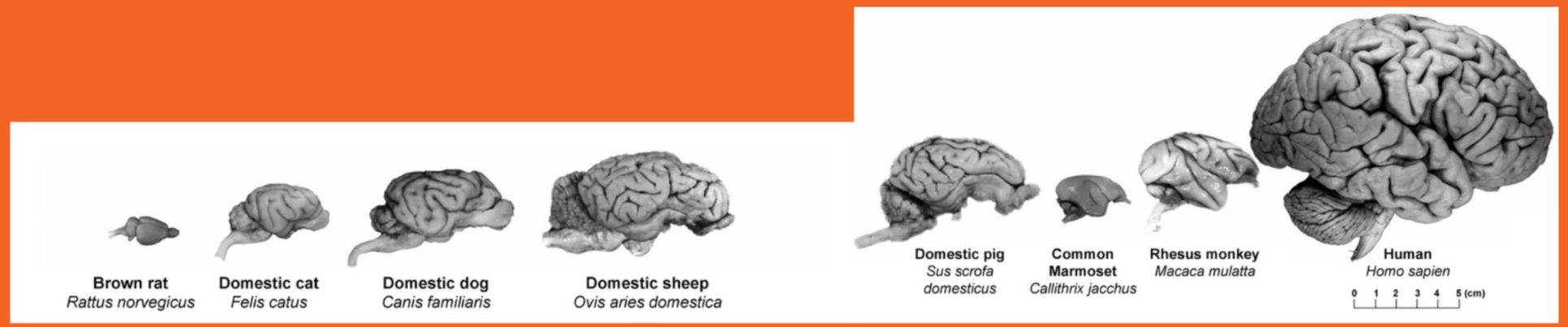
I got you bro.

Emergent Abilities

- Ability to act on things that model was not trained on
- Model learns this at a certain stage during “pre-training”
- LLMs exhibit this behaviour
- Model size where this emerges: Open Question
- Recently Unclear: If Emergent Abilities is true phenomenon or not

of Parameters

Tiny Stories 0.3B	Pythia 1.3B	Red Pajama 7B	Falcon 40B	llama-2 70B	GPT-3.5 175B	GPT-4 220B*8



Credit:
https://www.researchgate.net/figure/Gross-comparative-neuroanatomy-of-various-large-animal-species-used-to-model-cerebral_fig2_323764775

LLMs: A Timeline

- GPT-3 was released
- GPT-3.5, ChatGPT was published: Sept 2022
- LLaMA: March, 2023
- GPT-4: April, 2023
- **“Totally Hopeless”: May, 2023**
- Llama-2: July 2023
- 50B+ models released: April-Today
- Blink again! We have a new SOTA!

LLMs: Timeline

- GPT-3 was released
- GPT-3.5, ChatGPT was published: Sept 2022
- LLaMA: March, 2023
- GPT-4: April, 2023
- 50B+ models released: April-Today
- While you blink, two more models are released
- Blink again!

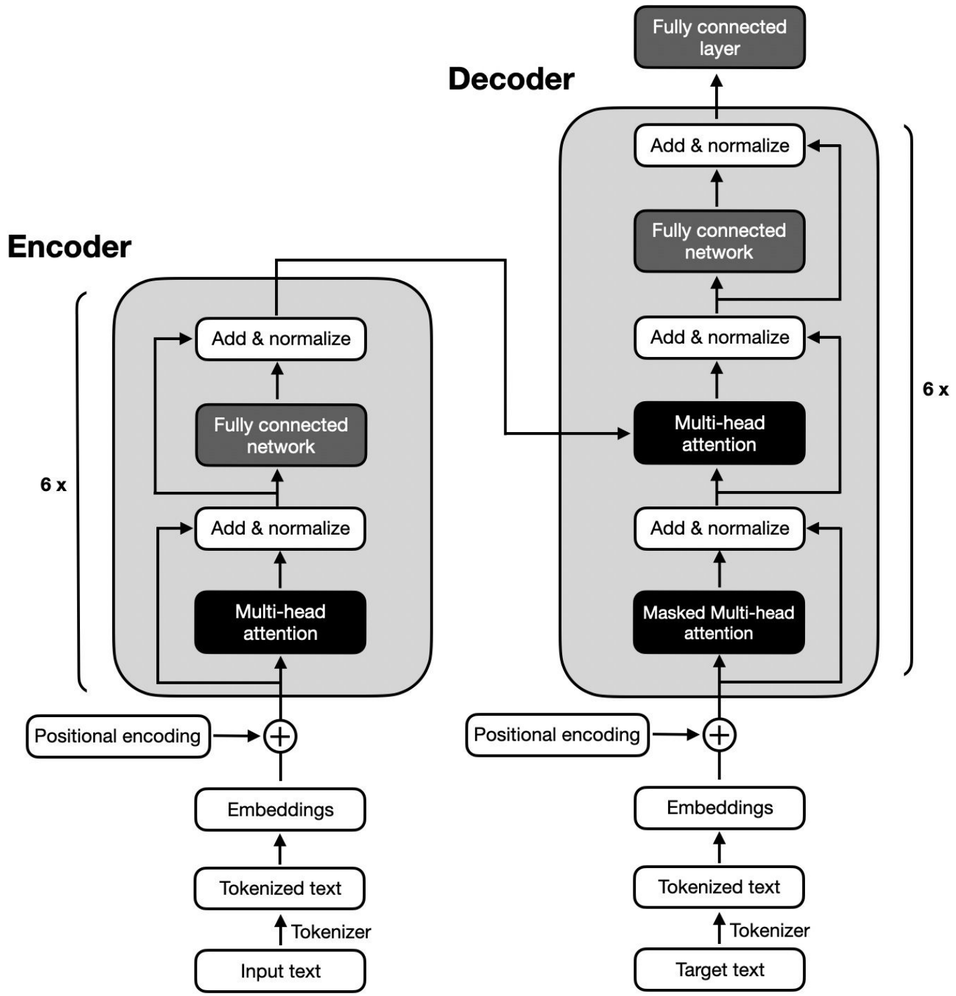
LLMs: Application

- In-Context Learning
- Training/Fine-Tuning
- External APIs

Encoder v/s Decoder Based Models

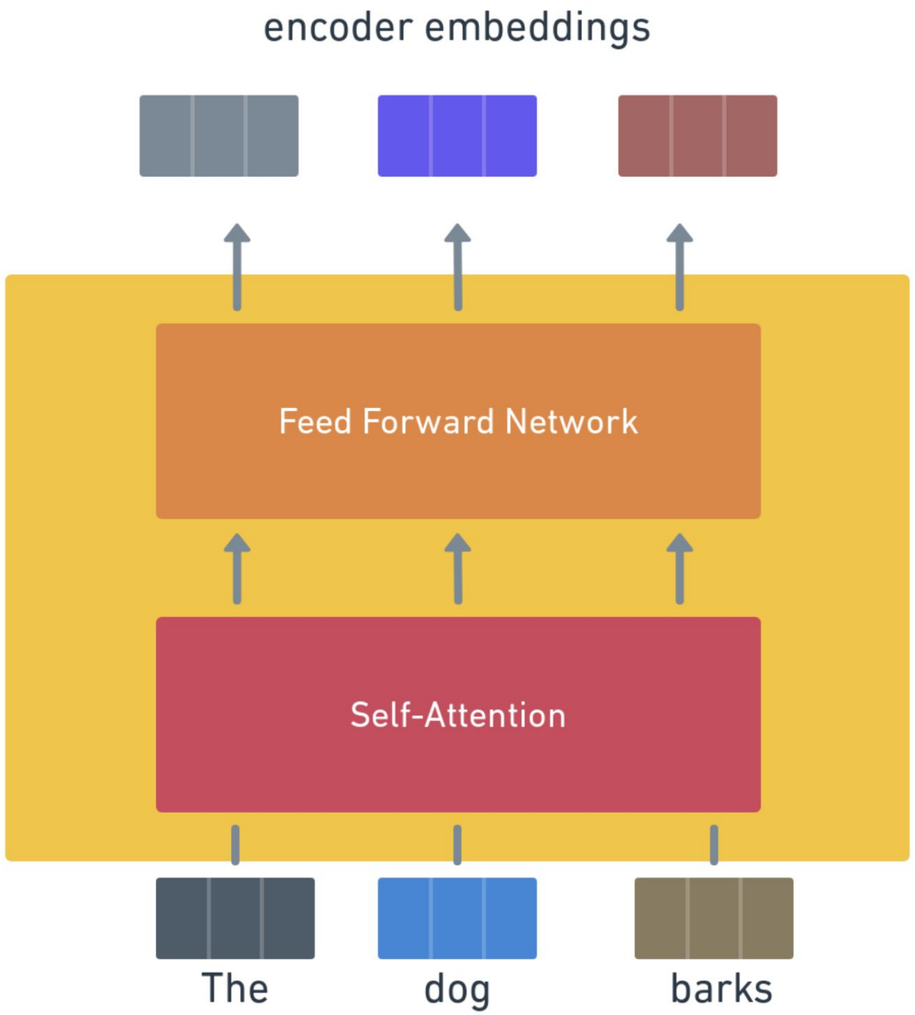
Any NLP Task can be solved with either:

- Encoder based models
- Decoder based models
- Encoder-Decoder models

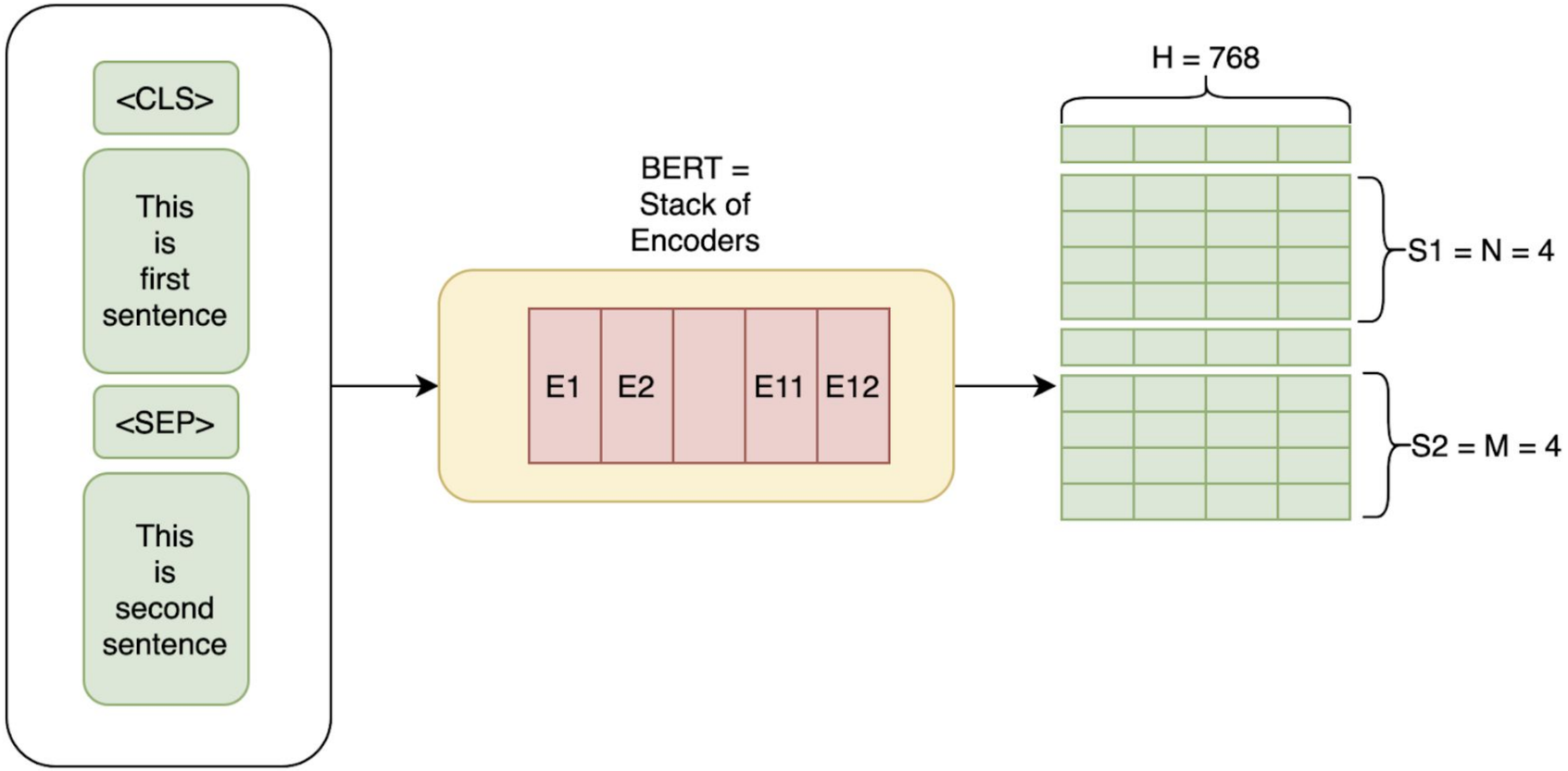


Encoder based models

- Processes input Sequence
- Generate embeddings based on each input token
- Bidirectional Self-Attention
- Can be finetuned for classification, NER, QA, Sentiment analysis,etc.



Encoder based models - BERT

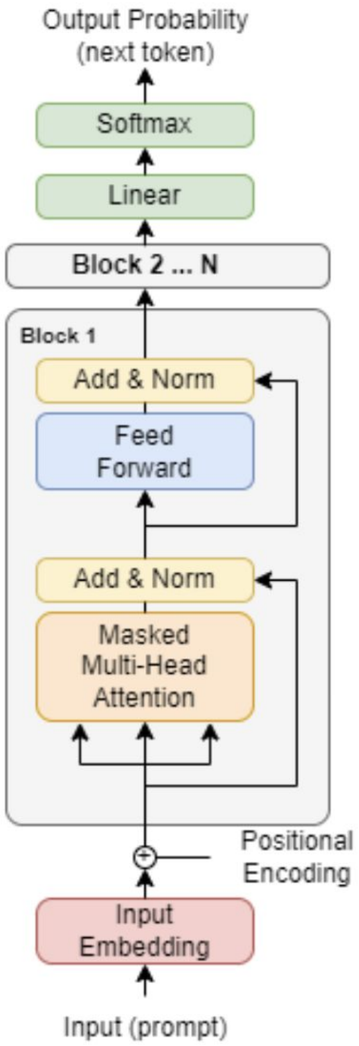


Encoder based models: Why BERT <<Roberta

- Trained on larger corpus of text
- More iterations, large batch size, better hyperparameters tuning during pretraining
- Removed NSP task
- Dynamically changing masking

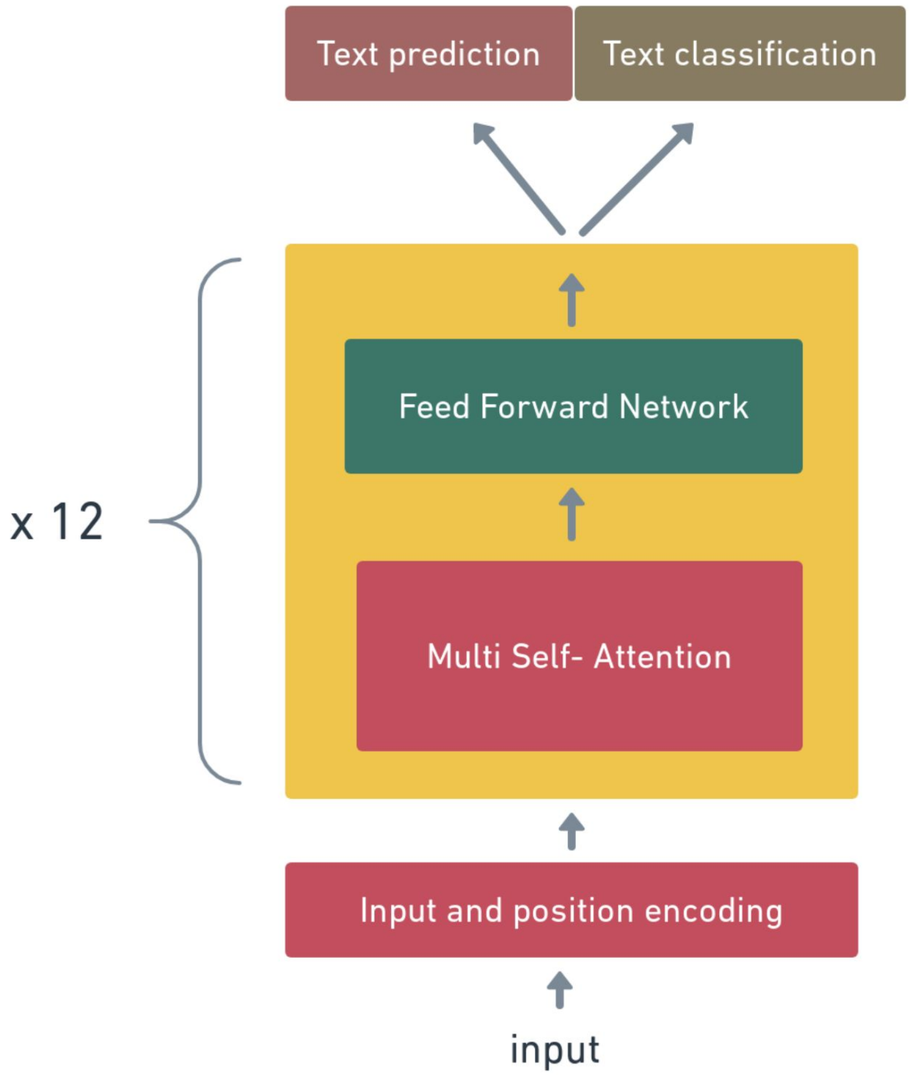
Decoder based models

- Generates output text sequences
- **AutoRegressive**
- Predict next tokens
- GPT, BARD and other generative models are decoder based models.



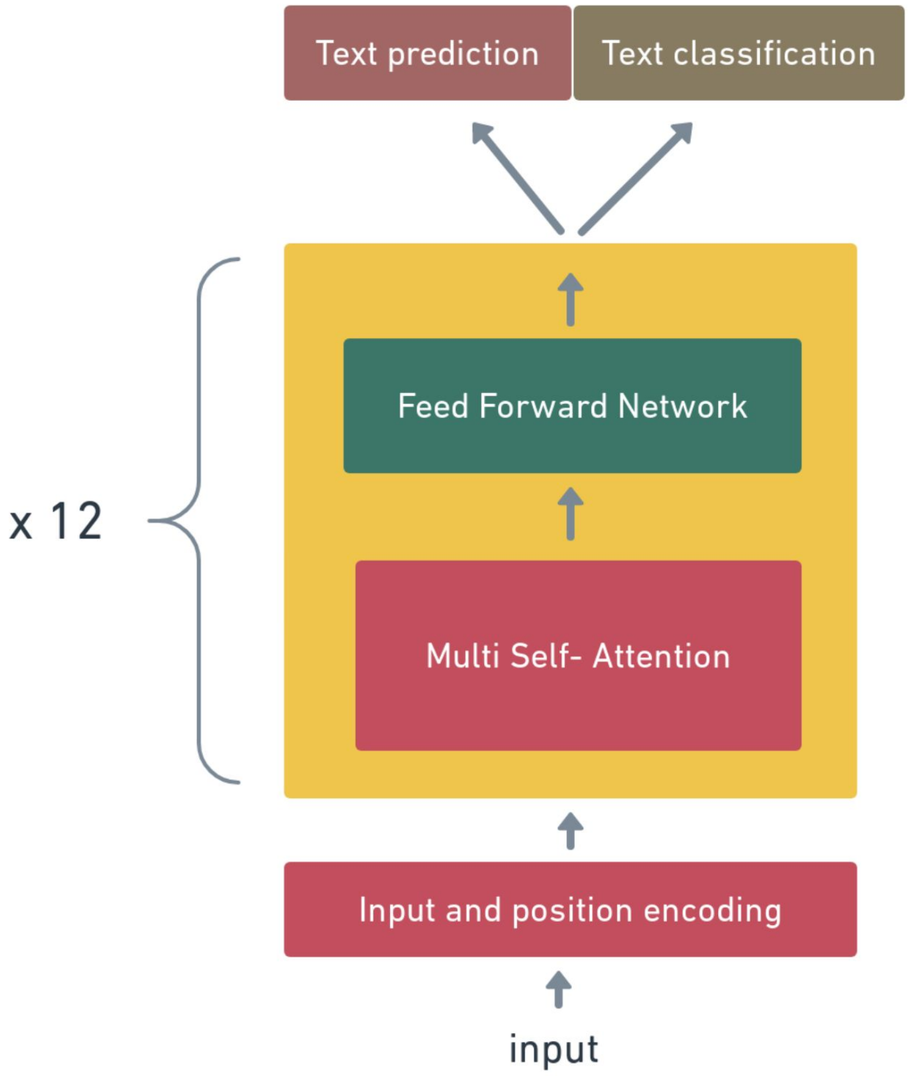
Decoder Based Models - GPT 1.0

- Released in 2018 by OpenAI
- Pretrained for 2 tasks
- Predict next tokens
- 12 Decoder blocks
- 117 M parameters



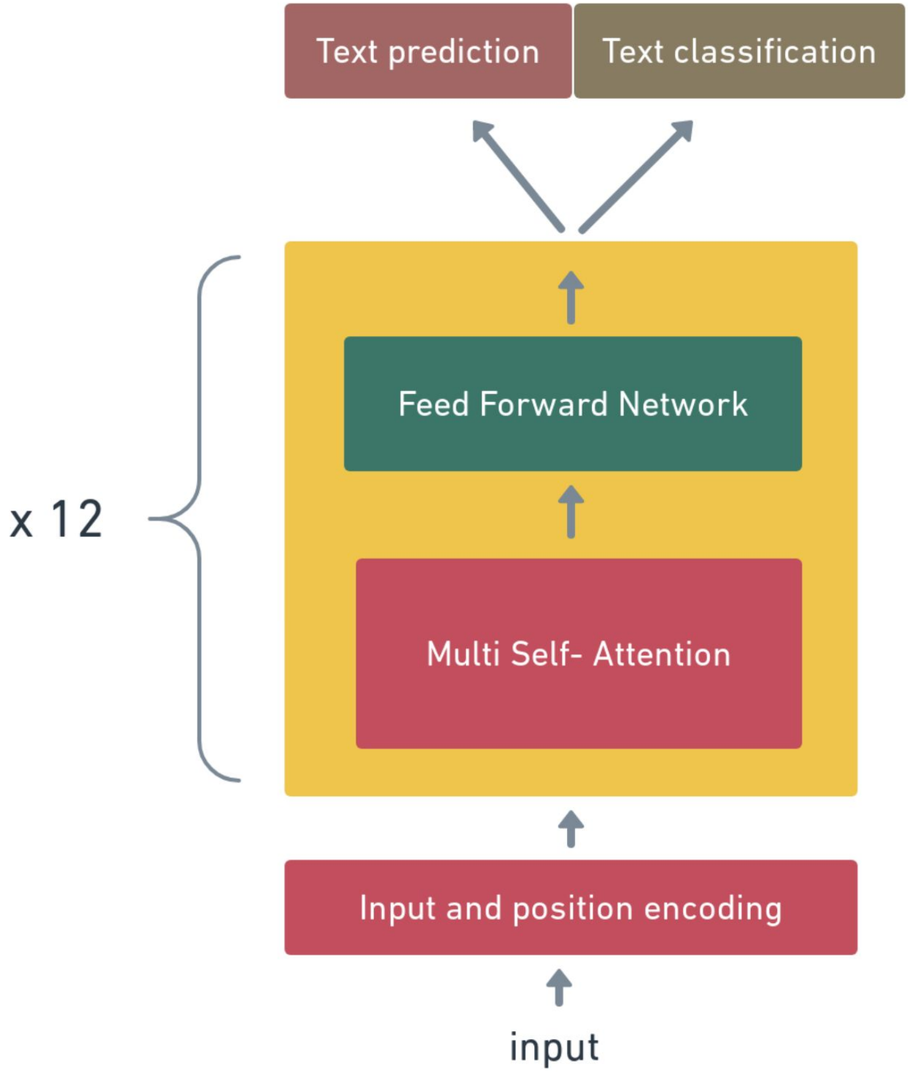
Decoder Based Models - GPT 2.0

- Large Training Data & Model size
- 1.5 Billion Params (~10x larger than GPT 1.0)
- Much better zero shot learning
- 48 Decoder blocks

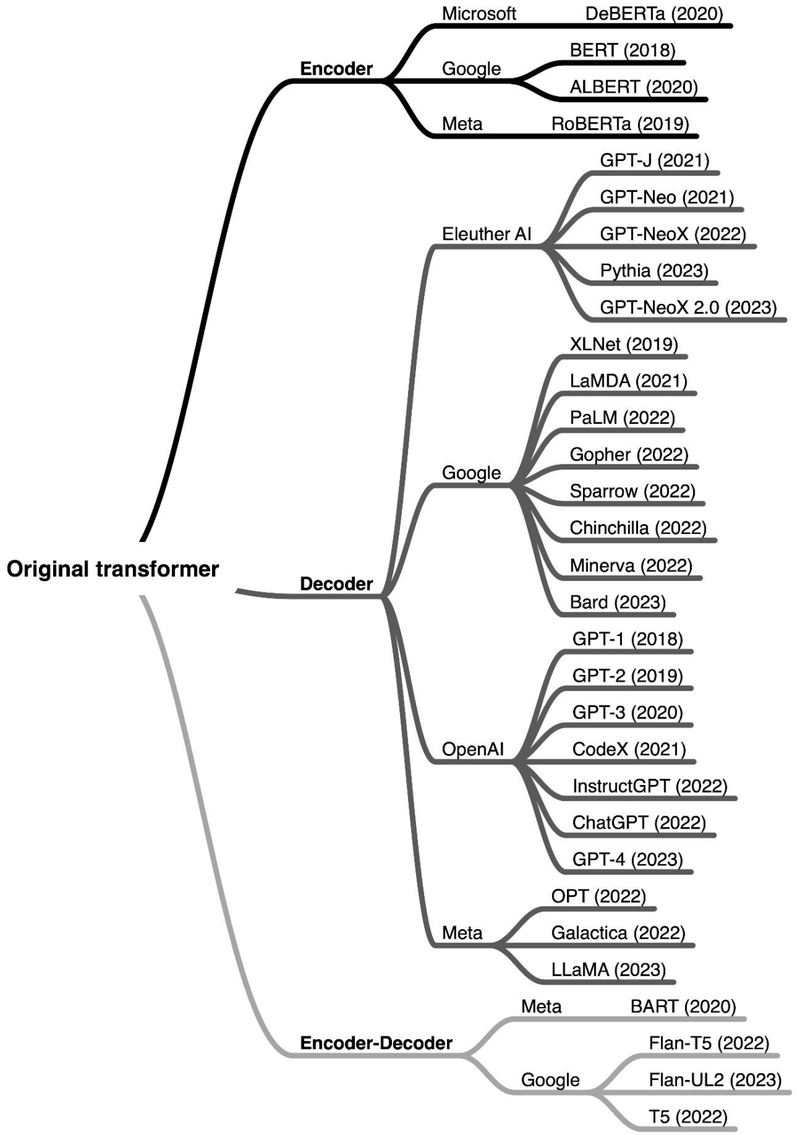


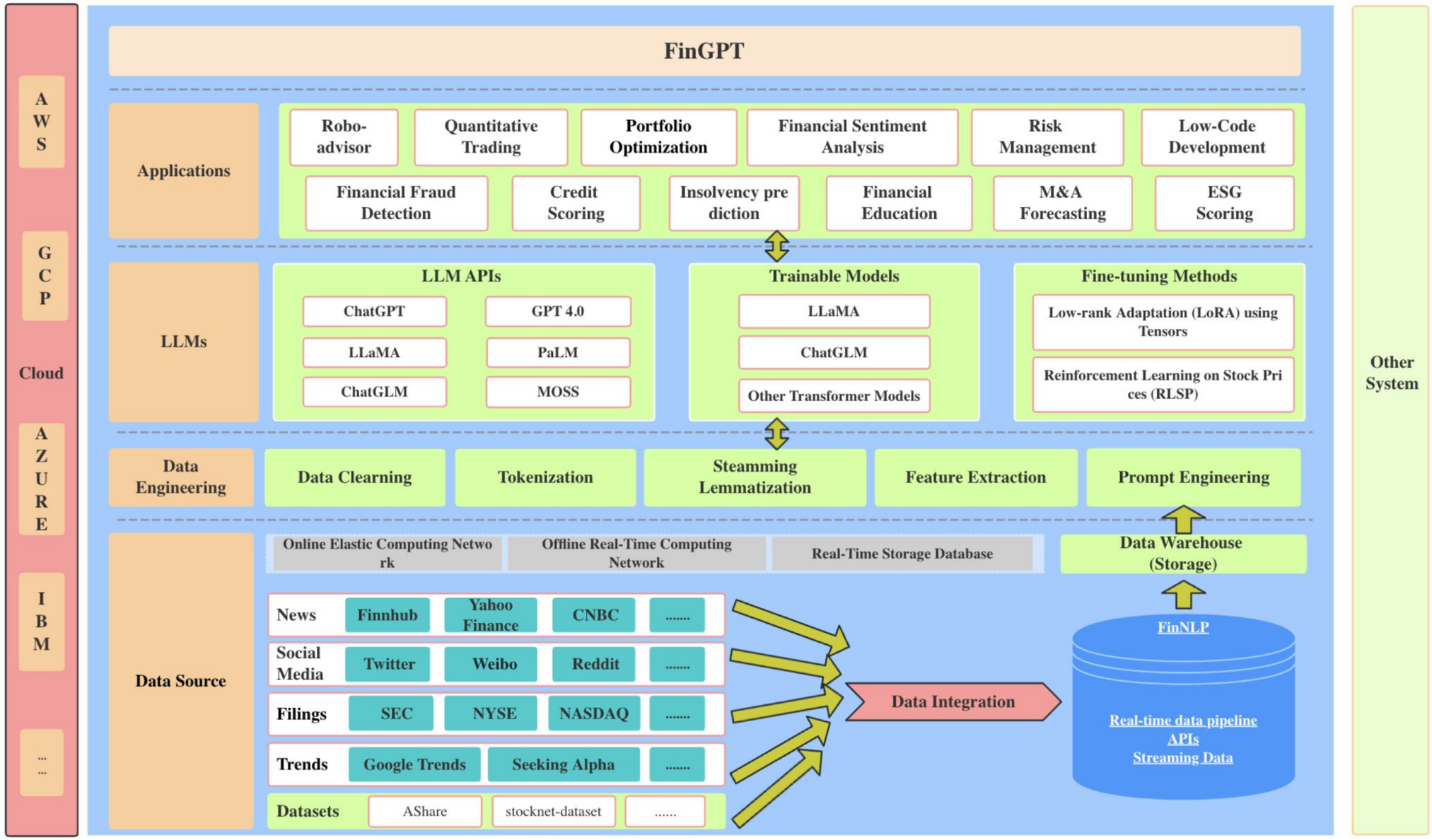
Decoder Based Models - GPT 3.0

- Large training data & Multilingual
- 175 Billion Params (~10x larger than GPT 2.0)
- 3200 GPUs used for training
- 48 Decoder blocks



Encoder Vs Decoder




AWS
GCP
Cloud
AZURE
IBM
...
FinGPT
Applications
Robo-advisor
Quantitative Trading
Portfolio Optimization
Financial Sentiment Analysis
Risk Management
Low-Code Development
Financial Fraud Detection
Credit Scoring
Insolvency pre diction
Financial Education
M&A Forecasting
ESG Scoring
LLMs
LLM APIs
ChatGPT
GPT 4.0
LLaMA
PaLM
ChatGLM
MOSS
Trainable Models
LLaMA
ChatGLM
Other Transformer Models
Fine-tuning Methods
Low-rank Adaptation (LoRA) using Tensors
Reinforcement Learning on Stock Pri ces (RLSP)
Data Engineering
Data Clearning
Tokenization
Steamming Lemmatization
Feature Extraction
Prompt Engineering
Data Source
Online Elastic Computing Netwo rk
Offline Real-Time Computing Network
Real-Time Storage Database
Data Warehouse (Storage)
News
Finnhub
Yahoo Finance
CNBC
.....
Social Media
Twitter
Weibo
Reddit
.....
Filings
SEC
NYSE
NASDAQ
.....
Trends
Google Trends
Seeking Alpha
.....
Datasets
AShare
stocknet-dataset
.....
Data Integration
FinNLP
Real-time data pipeline
APIs
Streaming Data
Other System

Quiz Time!

https://tinyurl.com/ODSCLLMQ0

Demo in Practise

Thank You