

BERT, GPT and the LLM Landscape

Two pretraining objectives, and what scale bought

Ramaseshan Ramachandran

LARGE LANGUAGE MODELS

- ▶ LLMs are typically trained on massive datasets of text and code
- ▶ Large language models transform natural language processing
- ▶ Capable of generating human-like text
- ▶ Learn patterns, context, and semantic relationships in data
- ▶ GPT (OpenAI), Claude (Anthropic), Gemini (Google), and Llama (Meta) are current state-of-the-art LLM families
- ▶ Revolutionizing various applications across industries

TYPICAL APPLICATIONS OF LLMS

- ▶ Natural language understanding
- ▶ Natural language generation
- ▶ Machine translation
- ▶ Question answering
- ▶ Text summarization
- ▶ code generation

LANGUAGE UNDERSTANDING

- ▶ Comprehends and interprets text with high accuracy
- ▶ Captures meaning, context, and nuances in language
- ▶ Understands complex queries and generates relevant responses
- ▶ Enables chatbots and virtual assistants to communicate effectively
- ▶ Enhances machine translation and language comprehension tasks

CREATIVE CONTENT GENERATION

- ▶ Generates coherent and contextually appropriate text.
- ▶ Creates engaging and original stories, articles, and scripts.
- ▶ Assists with content creation for marketing and advertising.
- ▶ Provides innovative solutions for writing, gaming, and entertainment.
- ▶ Expands creative possibilities for artists and content creators.

KNOWLEDGE EXPANSION

- ▶ Accesses vast amounts of information from the Internet.
- ▶ Processes and summarizes complex texts and articles.
- ▶ Helps in research, data analysis, and decision-making.
- ▶ Provides insights and answers to specific questions.
- ▶ Enables continuous learning and knowledge sharing

REAL WORLD APPLICATIONS

- ▶ Improves customer service through chatbots and support systems.
- ▶ Personalizes user experiences in e-commerce and recommendation systems.
- ▶ Enhances medical diagnostics and healthcare decision-making.
- ▶ Powers virtual assistants and voice-activated technologies.
- ▶ Transforms education and online learning platforms.

HOW DO LARGE LANGUAGE MODELS WORK?

- ▶ Uses transformer architecture
- ▶ Learns long-range dependencies between words
- ▶ Generalize the statistical regularities of human language

WORD EMBEDDING

1. Word embedding is fundamental to Deep Learning of NLP
2. Contextual word embedding is now used in most of the downstream applications

CLOZE TEST

According to a report in yesterday's newspaper (1)—— police dog was taken to Raj Bhavan (2)—— Monday. This was to trace the (3)—— of the "very important horse" which (4)—— reported missing on Sunday. The dog picked (5)—— the scent on some traces of (6)—— and ran a few yards before losing the (7)——. The police have launched a vigorous (8)—— into the whole affair. They have (9)—— the services of a forensic expert, (10)—— fingerprint expert and a photographer. (11)—— are now fourteen horses at Raj Bhavan (12)—— are kept in a large shed near the gate.

1	once	a	new
2	at	next	on
3	police	killers	dogs
...	...	...	...
11	There	We	So
12	who	were	which

The purpose of the Cloze test is to measure the reading comprehension of a student with respect to grammar, usage, vocabulary, and contextual understanding.

MASKED LM[1]

Objectives

- ▶ Predict the masked word using the context
- ▶ Combine left and right contexts to predict the masked word
- ▶ Use sentence pairs to predict the next sentence
- ▶ Use the trained model for token-level and sentence-level tasks
- ▶ Mask some of the tokens from the input
- ▶ Predict the original token using its context
- ▶ Use Bidirectional deep transformer model fuse the left and right context and develop a contextual representation

HOW DO WE MASK?

- ▶ 15% of the words are randomly masked
- ▶ Perform the following operation on the the i^{th} token
 1. Replace with a `<MASK>` 80% of the time
 2. Replace with a random token 10% of the time
 3. Retain the original 10% of the time

BIDIRECTIONAL ENCODER REPRESENTATIONS FROM TRANSFORMERS (BERT) - ARCHITECTURE

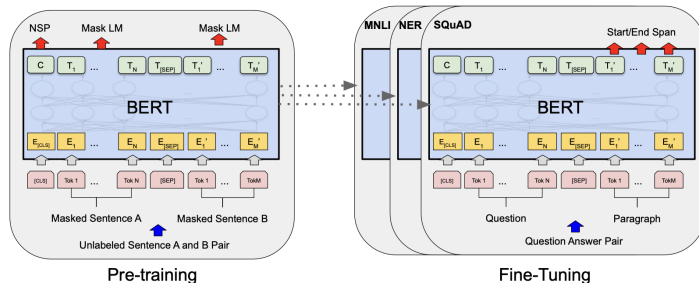
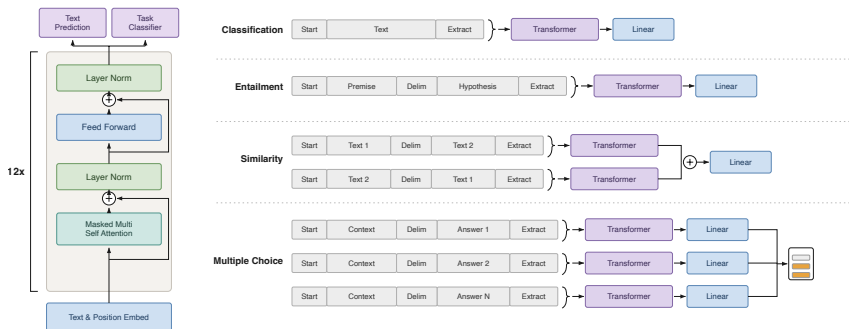


Figure: Overall pre-training and fine-tuning procedures for BERT. Apart from output layers, the same architectures are used in both pre-training and fine-tuning. The same pre-trained model parameters are used to initialize models for different downstream tasks. During fine-tuning, all parameters are fine-tuned. [CLS] is a special symbol added in front of every input example, and [SEP] is a special separator token (e.g. separating questions/answers). Next Sentence Prediction (NSP) And Perturbation is a method used to assess or enhance the performance of BERT-like models by altering or analyzing their training dynamics, specifically in tasks involving next sentence prediction.

DOES BERT PREDICT THE "NEXT WORD"?

- ▶ **Unlike GPT-style models, BERT does not predict the next word in the sequence**
- ▶ Instead, BERT predicts the **original word at the masked position**, leveraging:
 - ▶ **Bidirectional context** from both preceding and following tokens.
- ▶ **Example:**
 - ▶ Input: The [MASK] sat on the mat.
 - ▶ BERT considers both sides of [MASK]:
 - ▶ Context: "The" and "sat on the mat."
 - ▶ Prediction: "cat."
- ▶ **Why is this effective?**
 - ▶ Bidirectional prediction enables **contextual understanding**.
 - ▶ Excels in tasks like **question answering** and **natural language inference**.

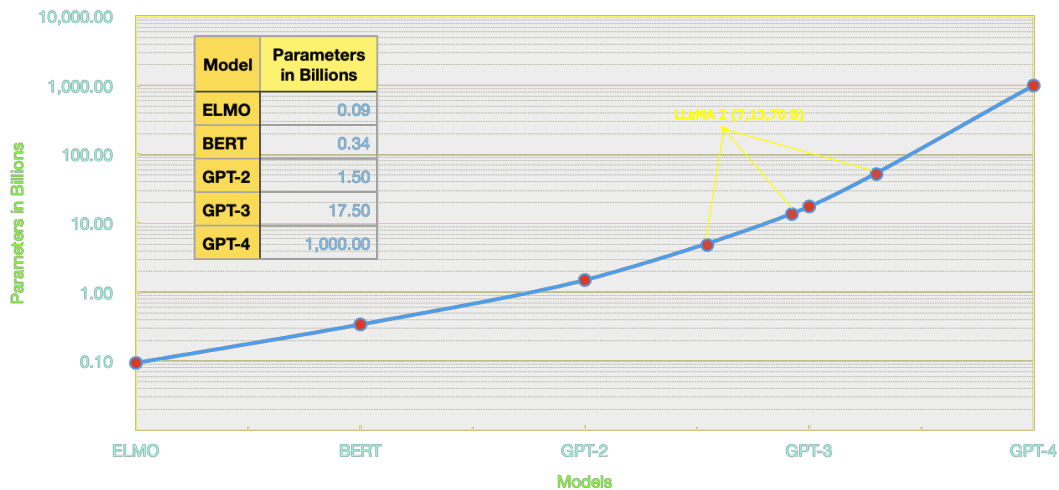
GENERATIVE PRE-TRAINED TRANSFORMER - GPT



After Radford et al., Improving Language Understanding by Generative Pre-Training (2018)

Figure: (left) Transformer architecture and training objectives used in this work. (right) Input transformations for fine-tuning on different tasks. We convert all structured inputs into token sequences to be processed by our pre-trained model, followed by a linear+softmax layer

PARAMETERS OF LARGE LANGUAGE MODEL



SAMPLE TEXT GENERATION - MINIGPT

```
const LEARNING_RATE: f64 = 0.0003;
const BLOCK_SIZE: i64 = 128;
const BATCH_SIZE: i64 = 64;
const EPOCHS: i64 = 100;
const SAMPLING_LEN: i64 = 4096;
```

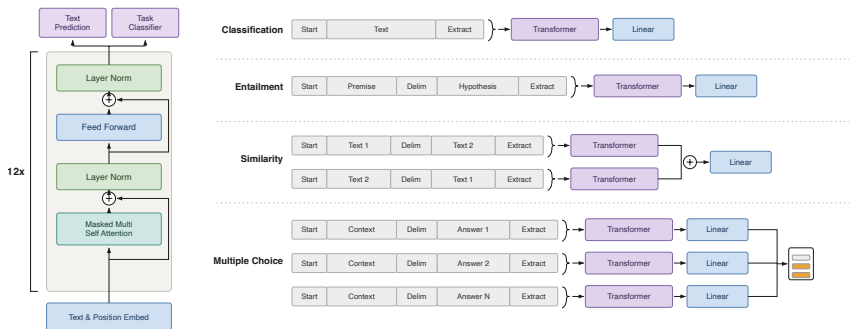
```
Config {
    vocab_size: 130,
    n_embd: 128,
    n_head: 8,
    n_layer: 8,
    block_size: 128,
    attn_pdrop: 0.1,
    resid_pdrop: 0.1,
    embd_pdrop: 0.1,
};
```

for the gbr genome determines a comprehensive comparison from existing quality was commonly important to a lung w fusion if they can be a compared x-rayed, with the blood scenario of the lesion of t-cell support probability.in this later, the next for the first room of responses to materials and infection has classified a retrospective response of younger genes [138] .

FROM GPT TO MODERN LLMS

- ▶ **GPT (2018)** was a decoder-only transformer, with generative pre-training plus task-specific fine-tuning
- ▶ **GPT-2 (2019)** was larger and showed strong zero-shot generation
- ▶ **GPT-3 (2020)** had 175B parameters and introduced *in-context* / *few-shot learning* without weight updates
- ▶ **InstructGPT / ChatGPT (2022)** used reinforcement learning from human feedback (RLHF) to align models with instructions
- ▶ **GPT-4 / GPT-4o / GPT-5 and peers (Claude, Gemini, Llama, DeepSeek)** add multimodal input, long context windows, tool use, and reasoning
- ▶ Common thread: the same next-token objective, scaled up in data, parameters, and alignment

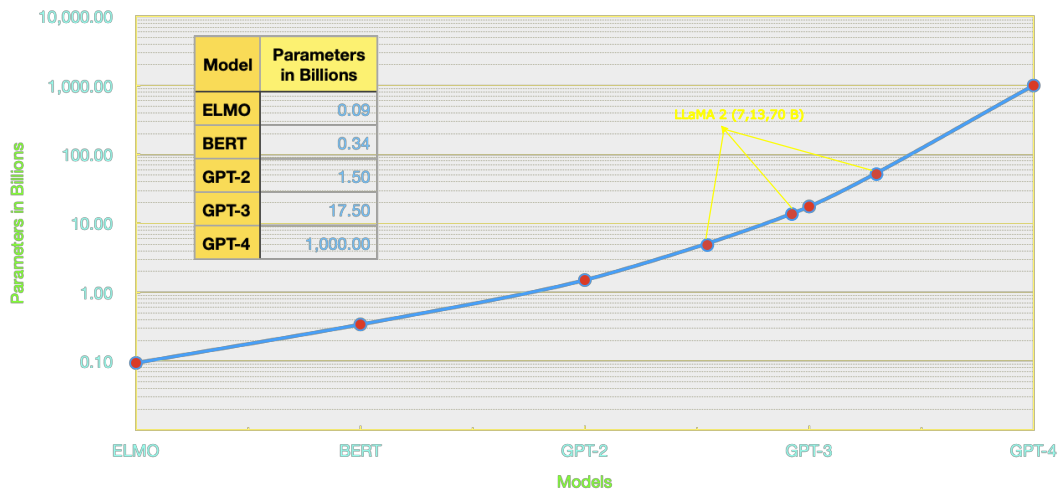
GENERATIVE PRE-TRAINED TRANSFORMER - GPT



After Radford et al., Improving Language Understanding by Generative Pre-Training (2018)

Figure: (left) Transformer architecture and training objectives used in this work. (right) Input transformations for fine-tuning on different tasks. We convert all structured inputs into token sequences to be processed by our pre-trained model, followed by a linear+softmax layer[2]

PARAMETERS OF LARGE LANGUAGE MODEL



LLAMA 2 OVERVIEW

- ▶ Open-source foundation model: Llama 2[3]
- ▶ Fine-tuned chat models for conversational AI
- ▶ Improves upon Llama 1 with new architectures and training methods
- ▶ Released by Meta AI
- ▶ Goals: Advance conversational AI, promote open research

KEY CONTRIBUTIONS

- ▶ Architecture: decoder-only Transformer with pre-normalization (RMSNorm), SwiGLU activations, and rotary positional embeddings (RoPE)
- ▶ Enhanced training: supervised fine-tuning followed by RLHF
- ▶ Large-scale pre-training: 2 trillion tokens; models of 7B, 13B, and 70B parameters; 4K-token context
- ▶ Grouped-query attention (in the 70B model) for faster inference
- ▶ Open-weights release: model weights and usage license (later Llama 3 models scaled to 15T training tokens and longer contexts)

TRAINING OF LLAMA 2-CHAT

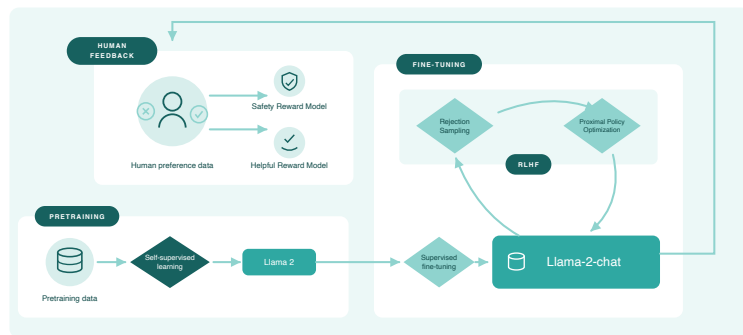


Figure: This process begins with the pretraining of Llama 2 using publicly available online sources. Following this, we create an initial version of Llama 2-Chat through the application of supervised fine-tuning. Subsequently, the model is iteratively refined using Reinforcement Learning with Human Feedback (RLHF) methodologies, specifically through rejection sampling and Proximal Policy Optimization (PPO). Throughout the RLHF stage, the accumulation of iterative reward modeling data in parallel with model enhancements is crucial to ensure the reward models remain within distribution [3].

IMPACT AND RESULTS

- ▶ State-of-the-art performance on conversational benchmarks
- ▶ Improved conversational flow and coherence
- ▶ Enhanced ability to understand context and nuance
- ▶ Demonstrates potential for real-world applications
- ▶ Opens doors for future research in conversational AI

SAMPLE TEXT GENERATION - MINIGPT

```
const LEARNING_RATE: f64 = 0.0003;  
const BLOCK_SIZE: i64 = 128;  
const BATCH_SIZE: i64 = 64;  
const EPOCHS: i64 = 100;  
const SAMPLING_LEN: i64 = 4096;
```

```
Config {  
    vocab_size: 130,  
    n_embd: 128,  
    n_head: 8,  
    n_layer: 8,  
    block_size: 128,  
    attn_pdrop: 0.1,  
    resid_pdrop: 0.1,  
    embd_pdrop: 0.1,  
};
```

Generated Text

for the gbr genome determines a comprehensive comparison from existing quality was commonly important to a lung w fusion if they can be a compared x-rayed, with the blood scenario of the lesion of t-cell support probability.in this later, the next for the first room of responses to materials and infection has classified a retrospective response of younger genes [138] .

REFERENCES I

- [1] Jacob Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2019. arXiv: [1810.04805 \[cs.CL\]](#). URL: <https://arxiv.org/abs/1810.04805>.
- [2] Alec Radford et al. “Improving language understanding by generative pre-training”. In: (2018).
- [3] Hugo Touvron et al. *Llama 2: Open Foundation and Fine-Tuned Chat Models*. 2023. arXiv: [2307.09288 \[cs.CL\]](#). URL: <https://arxiv.org/abs/2307.09288>.