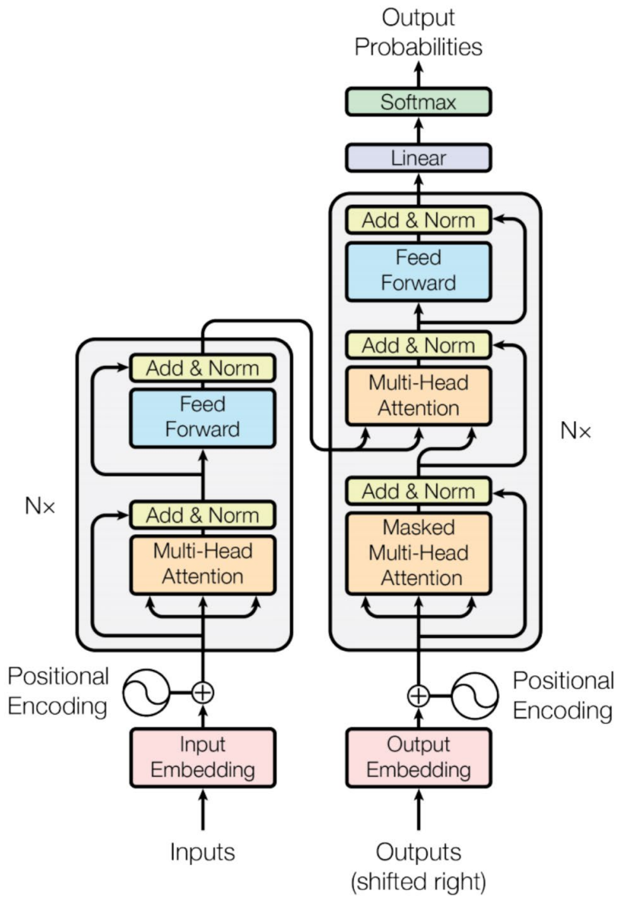
Lecture 11

BERT, GPT


Output
Probabilities
Softmax
Linear
Add & Norm
Feed
Forward
Add & Norm
Multi-Head
Attention
Add & Norm
Masked
Multi-Head
Attention
Nx
Add & Norm
Feed
Forward
Add & Norm
Multi-Head
Attention
Nx
Positional
Encoding
Positional
Encoding
Input
Embedding
Output
Embedding
Inputs
Outputs
(shifted right)

Generative Pre-trained Transformer (GPT)

- Improving Language Understanding by Generative Pre-Training (2018)

Alec Radford

Karthik Narasimhan

Tim Salimans

Ilya Sutskever

Bidirectional Encoder Representations from Transformers (BERT)

- BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (2018)

Jacob Devlin

Ming-Wei Chang

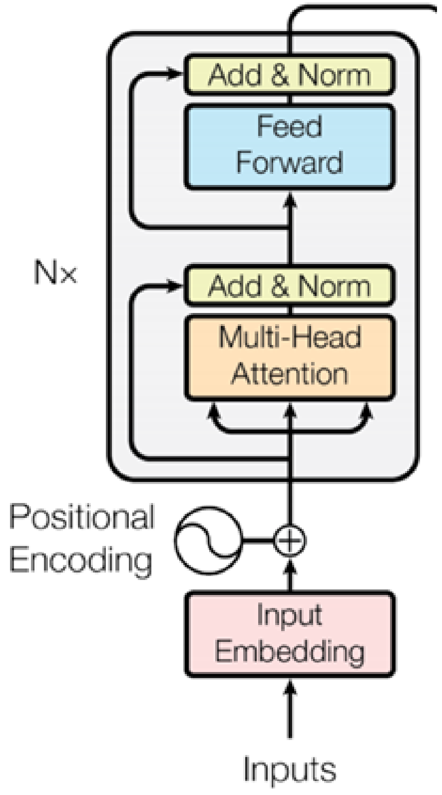
Kenton Lee

Kristina Toutanova

BERT and GPT

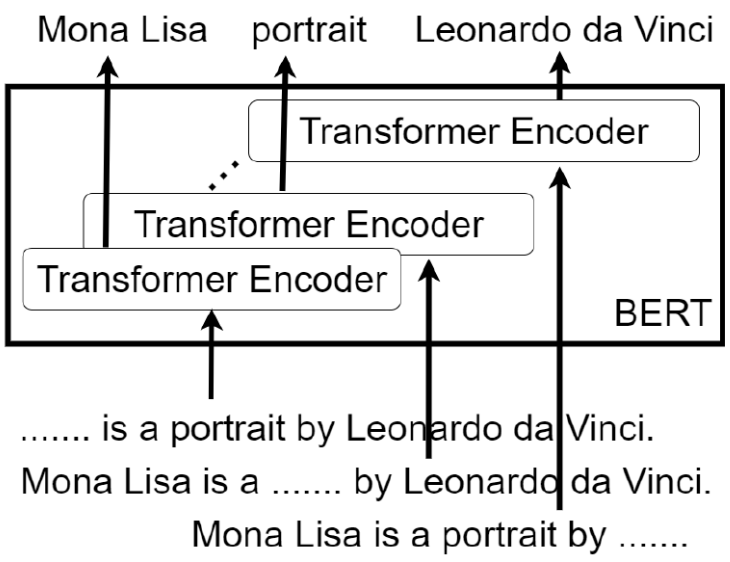
- The GPT is built using transformer decoder blocks.
- BERT is built using transformer encoder blocks.

BERT


Add & Norm
Feed Forward
Nx
Add & Norm
Multi-Head Attention
Positional Encoding
Input Embedding
Inputs

Masked language model

- Masks words in the input and asks the model to predict the missing word.



BERT: Bidirectional language model

- Masks words in the input and asks the model to predict the missing word.

- Additional task: Given two sentences (A and B), is B likely to be the sentence that follows A, or not?

BERT

- BERT is designed to pretrain bidirectional representations from unlabeled text.

BERT

- BERT is designed to pretrain bidirectional representations from unlabeled text.

- Jointly conditioning on both left and right context.

BERT

- BERT is designed to pretrain bidirectional representations from unlabeled text.

- Jointly conditioning on both left and right context.

- The pre-trained BERT model can be finetuned with just one additional output layer.

BERT

- BERT is designed to pretrain bidirectional representations from unlabeled text.
- Jointly conditioning on both left and right context.
- The pre-trained BERT model can be finetuned with just one additional output layer.
- It creates state-of-the-art models for a wide range of tasks, such as question answering and language inference.

[CLS] Token in BERT

- The [CLS] token is prepended to the input text and travels through the Transformer layers alongside other tokens.
- All tokens, including [CLS], gather contextual information from the entire sequence due to the self-attention mechanism.
- For sentence-level tasks, the final hidden state of the [CLS] token is used as the sentence representation.
- During fine-tuning on a specific task, the model learns to imbue the [CLS] token with a meaningful representation of the entire sentence, optimized for that task.
- Example Usage: In classification tasks, the [CLS] token representation is fed into a classifier to determine the sentence's class.

BERT

BERT is basically a trained Transformer Encoder stack

1. **Transformer**:
    1. Encoder Layers: 6
    2. FFNN Hidden Layer Units: 512
    3. Attention Heads: 8
2. **BERT Base**:
    1. Encoder Layers: 12
    2. FFNN Hidden Layer Units: 768
    3. Attention Heads: 12
    4. Total Parameters: 110 million
3. **BERT Large**:
    1. Encoder Layers: 24
    2. FFNN Hidden Layer Units: 1024
    3. Attention Heads: 16
    4. Total Parameters: 340 million

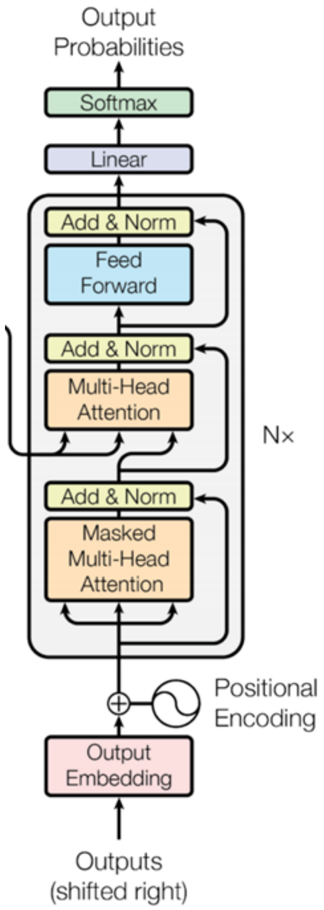
BERT

- **RoBERTa**:
  - Optimizes BERT's training process by using more data, larger batch sizes, and longer training times, resulting in improved performance on NLP tasks.
- **TinyBERT**:
  - A smaller and faster version of BERT designed for resource-constrained environments, retaining competitive performance with significantly fewer parameters.

BERT

1. **Multilingual BERT**:
    1. Trained on 104 different languages, capable of "zero-shot" adaptation to a new language domain.
2. **Domain Specific BERT Variants**:
    1. BioBERT: Retrained on a biomedical corpus.
    2. SciBERT: Trained on over one million published articles.
    3. BERTweet: A RoBERTa model trained on 850 million tweets.
    4. FinBERT: Adapted to the financial domain.

GPT


Output
Probabilities
Softmax
Linear
Add & Norm
Feed
Forward
Add & Norm
Multi-Head
Attention
Nx
Add & Norm
Masked
Multi-Head
Attention
Positional
Encoding
Output
Embedding
Outputs
(shifted right)

GPT

Predict the next word, given all of the previous words

- **Architecture**:
  - Stack of Transformer decoder blocks.
  - No encoder, hence no cross-attention module.
  - Components: Positional encoding, masked multihead self-attention, and feedforward network.
- **Directionality**:
  - Only considers previous (left) words in attention, not bidirectional like BERT.
  - Utilizes masked multihead self-attention for this purpose.

GPT 1

- Released: 2018
- Parameters: 117 Million
- Layers: 12
- Training Data: Books1 Corpus (7,000 unpublished books)
- Focus: Unsupervised pre-training, Transformer architecture, large-scale language modeling

GPT 2

- Released: 2019
- Parameters: ~1.5 Billion
- Layers: 48
- Training Data: 40GB (English)
- Focus: Transformer architecture, self-attention mechanism

GPT 3

- Released: 2020
- Parameters: 175 Billion
- Layers: 175
- Training Data: 570GB (Multilingual)
- Focus: Few-shot learning, prompt engineering, Python support

GPT 4

- Release: Not Yet
- Parameters: ~100 Trillion (speculative)
- Layers: Unknown
- Training Data: Larger, more diverse (speculative)
- Focus: GPT-4 is known to be a multimodal model, capable of processing both text and image inputs to generate text outputs. Advanced few-shot learning, improved NLU and NLG, reasoning and inference

Introduction to T5

- Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer by Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J. Liu.

- T5 stands for Text-to-Text Transfer Transformer.
- Introduced as a unified framework for addressing NLP tasks by converting them into a text-to-text format.
- Key Focus: Transfer learning - Pre-training on data-rich tasks followed by fine-tuning on downstream tasks.

T5 Architecture

- Operates on an encoder-decoder model.
- Pretrained on a multi-task mixture of both unsupervised and supervised tasks, each converted into a text-to-text format.
- Utilizes relative scalar embeddings; encoder input padding can be performed on both left and right.

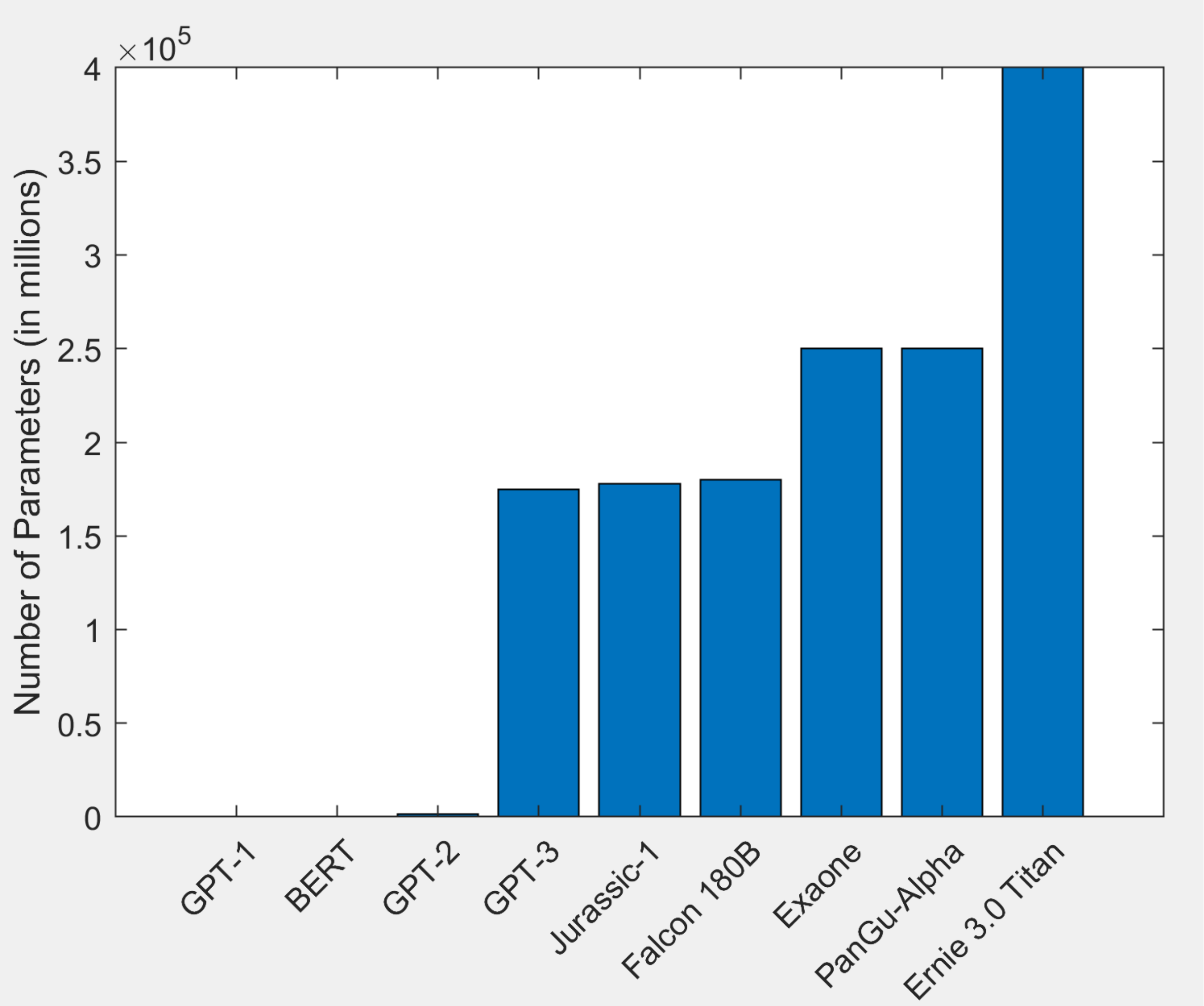
Pretraining Objectives

- Supervised training conducted on downstream tasks provided by GLUE and SuperGLUE benchmarks, reformulated into text-to-text tasks.
- Self-supervised training employs corrupted tokens: 15% of tokens are randomly removed and replaced with sentinel tokens.

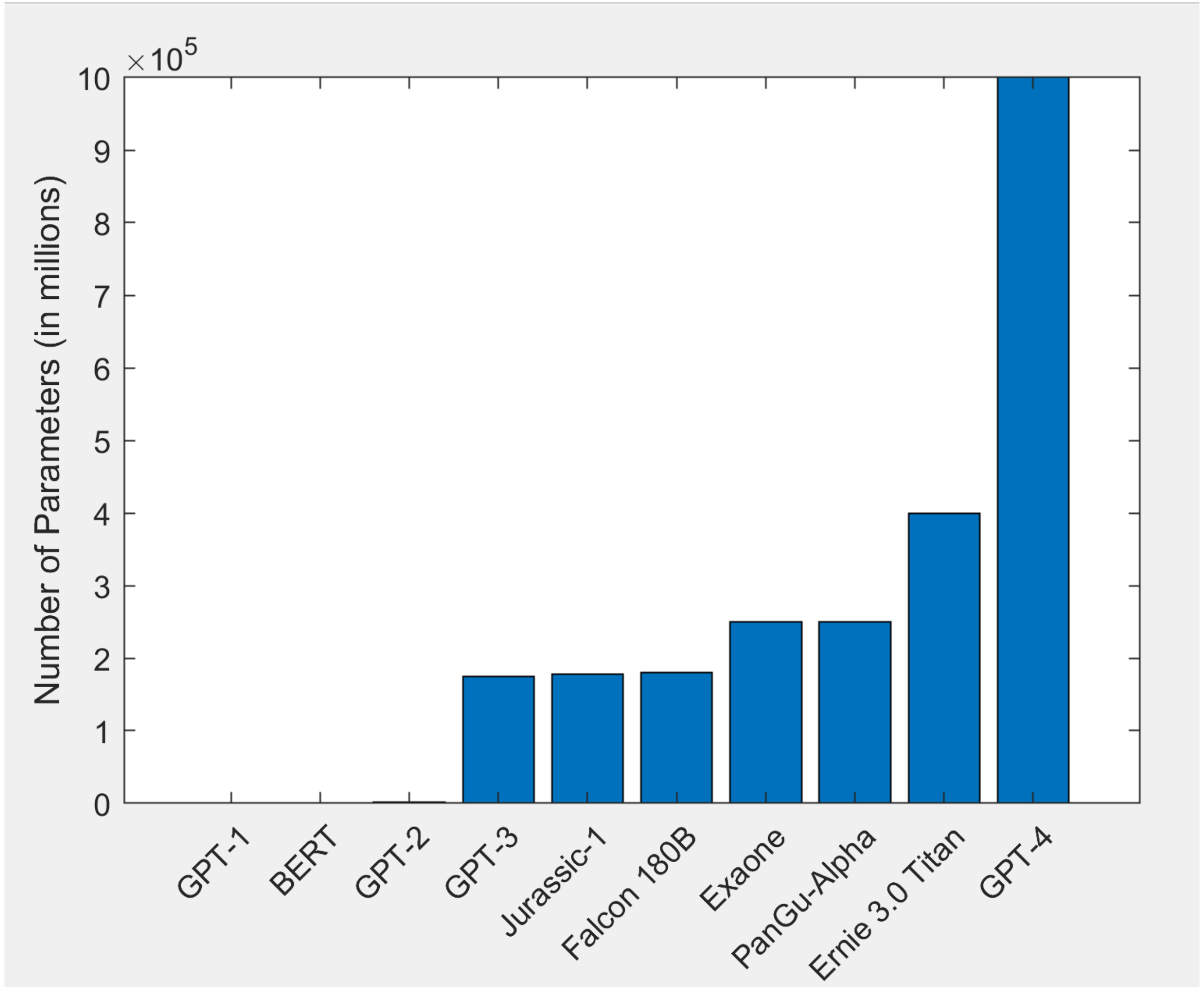
Task Adaptation

- T5 adapts to various tasks by prepending task-specific prefixes to the input, e.g., for translation: "translate English to German: …", for summarization: "summarize: …".
- Achieves very good results on many benchmarks including summarization, question answering, and text classification.

Number of Parameters in Various Language Models



Number of Parameters in Various Language Models



GPT-2

- **HUMAN-WRITTEN**

In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

- **MACHINE-WRITTEN**

The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

Open AI

GPT-2

- **HUMAN-WRITTEN**

In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

- **MACHINE-WRITTEN**

The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

Open AI

GPT-2

- **HUMAN-WRITTEN**

In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

- **MACHINE-WRITTEN**

The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

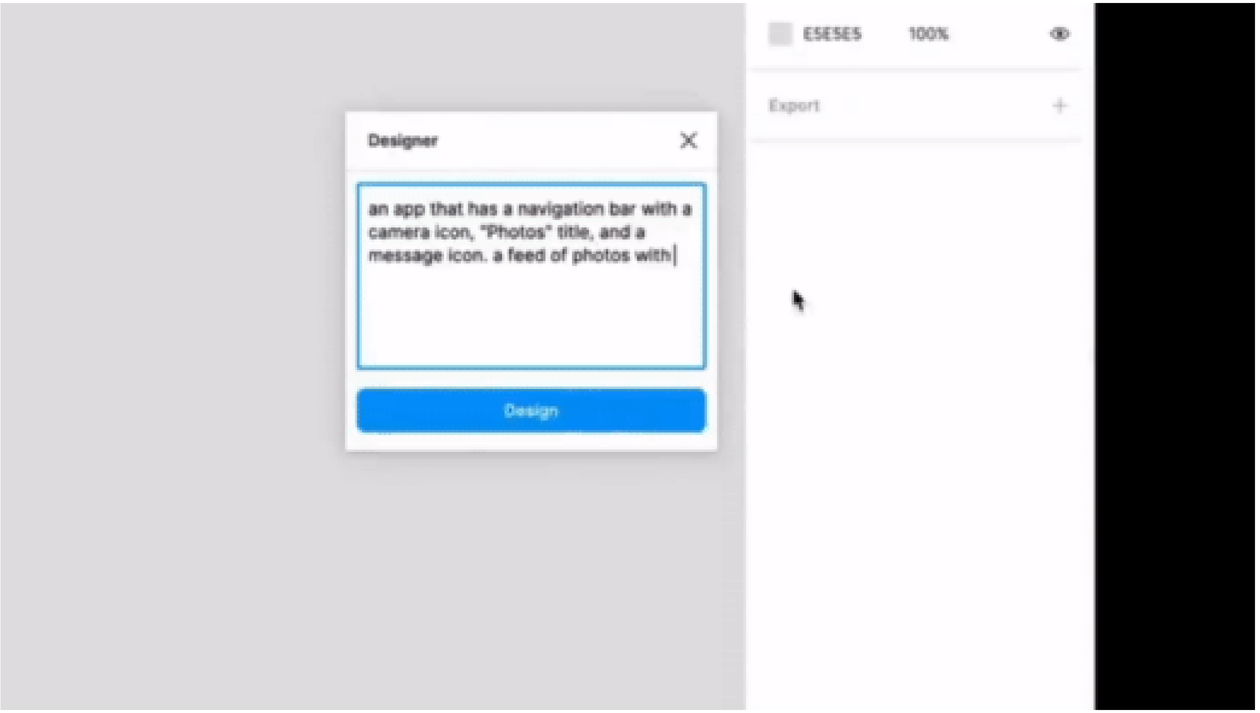
Open AI

GPT-3

- It can
  - Create fiction so close to human level
  - Write Poem
  - Chat similar to humans
  - Write computer codes
  - Design applications
  - Summarize texts
  - Answer questions
  - ….

GPT-3 (Applications)

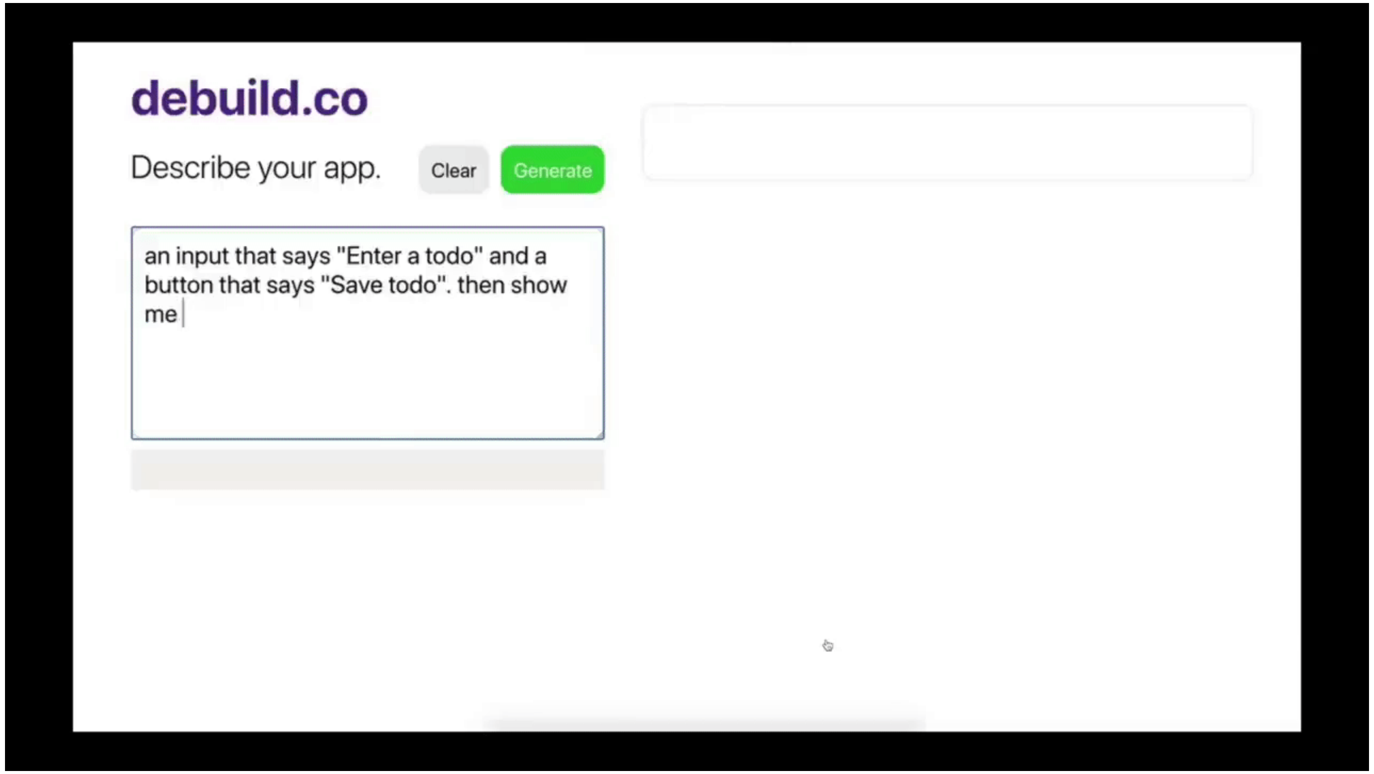
Design application (Figma): Enter the description of an application. Then the Software can design an app for that description.



demo from: https://www.youtube.com/watch?v=8psgEDhT1MM

GPT-3 (Applications)

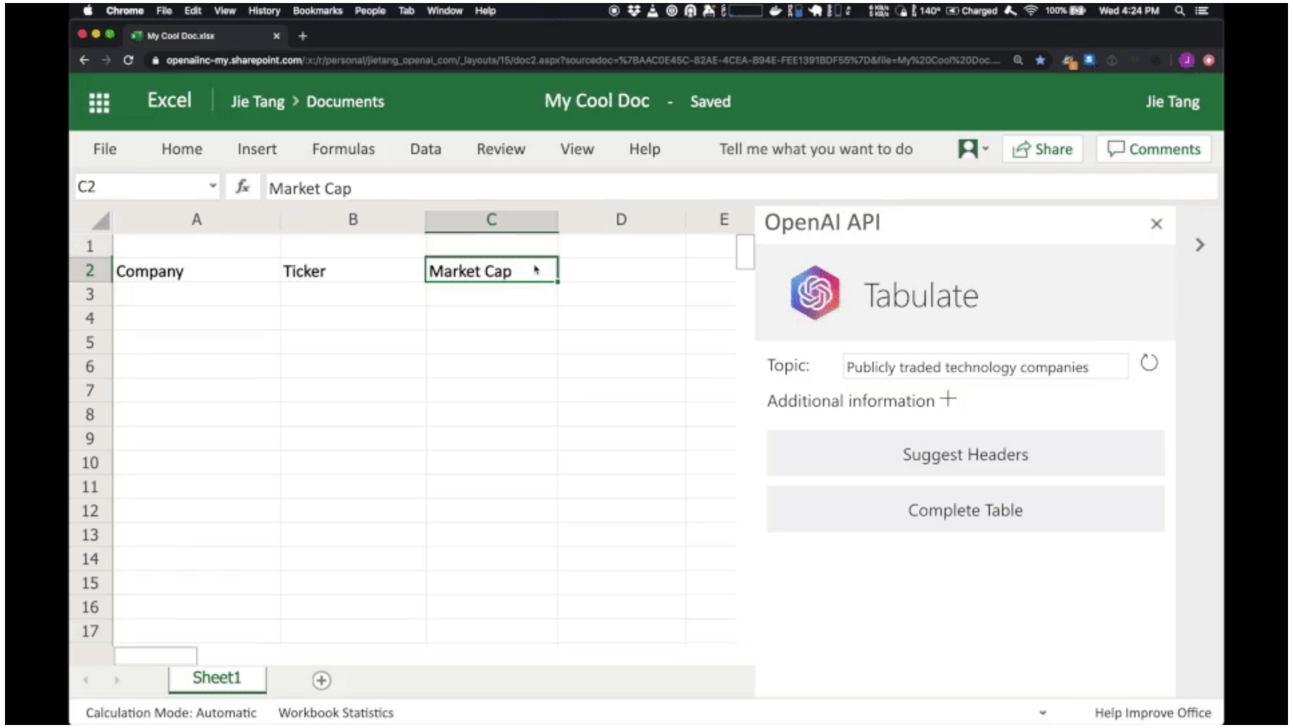
Write code (debuild.co): Enter the description of an application. Then the Software will generate react code for described app



demo from: https://www.youtube.com/watch?v=8psgEDhT1MM

GPT-3 (Applications)

Fill a table: Make an empty table with some column names. Then the model fills the table by predicting what the entries of the table should be.



demo from: https://openai.com/blog/openai-api/

GPT-3 (Applications)

Semantic search: Go to a web page like Wikipedia and ask a question. Then the model can show you the paragraph where the answer of your question can be found there.

demo from: https://openai.com/blog/openai-api/

Core Challenge with GPT Models:

- **GPT models predict the next token based on historical data, lacking an innate ability to follow instructions.**

- GPT (Generative Pre-trained Transformer) models, at their core, predict the next word or token in a sequence based on the probabilities derived from pre-training on extensive text corpora.
- They don't inherently "understand" instructions or follow commands but generate what's statistically likely to come next, given their training.

The Goal of Alignment:

- **The aim is to align GPT's responses with specific user instructions and ethical standards, beyond just generating probable text.**

- The primary objective is to bridge the gap between these statistical predictions and meaningful adherence to instructions provided by users.
- This involves ensuring that the AI's responses are not just contextually appropriate or conversationally relevant but also aligned with the specific intentions, ethical expectations, and task-oriented goals of the user.

Why It Matters:

- **This alignment is crucial for enhancing GPT's reliability, ensuring it respects user intent and ethical norms.**

- Without this alignment, while a model like GPT might produce grammatically correct and contextually relevant content, it might diverge from user instructions or produce content that's inappropriate or misaligned with the user's ethical, cultural, or personal expectations.
- The goal is to enhance the model's reliability in following directives accurately, respecting ethical boundaries, and fulfilling the user's actual needs, thereby making the technology more trustworthy and effective in real-world applications.

Key Areas of Focus in AI Alignment

1.Learning from Human Feedback:

- Tailoring AI through human interaction.

2. Training to Follow Instructions:

- Teaching AI specific task adherence.

3. Evaluating AI Harms:

- Identifying risks in AI outputs.

4. Modifying Behavior to Mitigate Harms:

- Adjusting AI to prevent negative impacts.

Key Areas of Focus in AI Alignment

1.Learning from Human Feedback:

- Tailoring AI through human interaction.

2. Training to Follow Instructions:

- Teaching AI specific task adherence.

3. Evaluating AI Harms:

- Identifying risks in AI outputs.

4. Modifying Behavior to Mitigate Harms:

- Adjusting AI to prevent negative impacts.

Addressing GPT Challenges - Training Strategy Overview

- To address inherent challenges with GPT models, ChatGPT's training strategy mirrors the "Instruct GPT" approach, itself an amalgamation of strategies from preceding works.
- **The Three-Phased Training Approach:**

1. **Supervised Fine-Tuning (SFT):**
    1. "Refines a pre-trained GPT-3 model's responses for specific tasks or guidelines, enhancing its understanding and output relevance."
2. **Training a Reward Model (RM):**
    1. "Develops a system that assesses the quality of text generated by the model, guiding it towards human-preferred responses."
3. **Reinforcement Learning Training (RL):**

The Three-Phased Training Approach

1. **Supervised Fine-Tuning (SFT):**
   - Refines a pre-trained GPT-3 model's responses for specific tasks or guidelines, enhancing its understanding and output relevance.
2. **Training a Reward Model (RM):**
   - Develops a system that assesses the quality of text generated by the model, guiding it towards human-preferred responses.
3. **Reinforcement Learning from Human Feedback (RLHF)**
   - Refining AI behavior through direct human feedback.