General Language Models

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **ID** | **Model** | **Origin** | **Release Time** | **Publication** | **Comment** |
|  | Transformer | Google | 2017/06 | Attention is All You Need | The encoder-decoder structure inspires the development of the models such as BERT，GPT, GPT-2, -3, and ChatGPT. |
|  | GPT | OpenAI | 2018/06 | Improving Language Understanding by Generative Pre-Training | **Model**: the pretrained base model is a 12-layer decoder-only Transformer with masked self-attention heads; Besides unsupervised pre-training, it also adopts supervised fine-tuning.  **Datasets**: BooksCorpus with 7000 unique books; 1B Word Benchmark; ELMo |
|  | BERT | Google | 2018/10, 2019/05 | Pre-training of deep bidirectional transformers for language understanding |  |
|  | GPT-2 | OpenAI | 2019/02 | Language Models are Unsupervised Multitask Learners. | **Scale**: 1.5B parameters  **Model**: follows the GPT model with modifications  **Datasets**: WebText with over 8 million webpages for 40GB of texts.  Perform a wide range of tasks in a zero-shot setting. |
|  | GPT-3 | OpenAI | 2020/05, 2020/07 | Language Models are Few-Shot Learners | **Scale**: an autoregressive model with 175B parameters and 96 layers  **Model**: The model is similar to GPT-2 but with larger model size, dataset size, diversity, and length of training.  **Datasets**: a filtered CommonCrawl dataset together with an expanded version of WebText dataset, two books corpora, and English-language Wikipedia.  Test different levels of task-specific training: fine-tuning, few-shot, one-shot, and zero-shot. |
|  | GPT-4 | OpenAI | 2023/03 | Technical Report  https://openai.com/research/gpt-4?ref=emergentmind | Predictable scaling: predict GPT-4’s final loss by extrapolating from models trained using the same methodology but much less computing resource.  Fine-tune the base model using reinforcement learning with human feedback (RLHF)  Receive a prompt of text and images |
|  | BLOOM | BigScience  (Open-source) | 2023/03 | BLOOM: A 176B-Parameter Open-Access Multilingual Language Model | **Scale**: 176B parameters;  **Model**: The decoder-only architecture, with ALiBi positional embedding instead of positional encoding and with an additional layer normalization after the embedding layer.  **Dataset** for pretraining & finetuning: the ROOTS corpus (498 Hugging Face datasets amounting to 1.6TB texts that span 46 natural languages and 13 programming languages); xP3 (a subset of Public Pool of Prompts plus datasets in languages other than English and new tasks) |
|  | LLaMA | Meta AI  (Open-source) | 2023/02 | LLaMA: Open and Efficient Foundation Language. Models | **Scale**: a series of models ranging from 7B to 65B parameters  **Model**: LLaMA-\*B are pretrained base models; LLaMA-I is a fine-tuned one specific for the task of MMLU.  **Datatsets**: English CommonCrawl, C4 (Filtered CommonCrawl), Github, Wiki covering 20 languages, Gutenberg and Books3 (book corpora), ArXiv latex, Stack Exchange. |
|  | Chinchilla |  |  |  |  |
|  | PaLM |  |  |  |  |

LLM for Code

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **ID** | **Model** | **Origin** | **Release Time** | **Publication** | **Comment** |
|  | Codex | OpenAI | 2021/07 | Evaluating Large Language Models Trained on Code | Scale: 12B parameters;  Model: a GPT model fine-tuned on publicly available code from GitHub.  Training datasets: 179GB of unique Python files under 1MB from 54 million public software repositories on GitHub. (159GB after filtering) |
|  | StarCoder |  |  |  |  |