General Language Models

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ID | Model | Origin | Release Time | Publication | Reviewed | 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 | √ | It adopts a two-stage framework: unsupervised pre-training and supervised fine-tuning.  Pre-training model is a 12-layer decoder-only Transformer with masked self-attention heads.  Training sets: 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: 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. |
|  | ChatGPT | OpenAI | 2023/02 | Technical Report | √ |  |
|  | BLOOM  (open-source) | HuggingFace | 2023/03 | BLOOM: A 176B-Parameter Open-Access Multilingual Language Model |  | Scale: 176B parameters;  Training sets: 1.6TB texts that span 46 natural languages and 13 programming languages |

LLM for Auto Coding

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ID | Model | Origin | Release Time | Publication | Reviewed | 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) |