|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | **GLUE Score** | **SuperGLUE Score** | **MMLU Score** | **Code Generation Score** | **Overall Score** | **Parameter** | **Training Data Size (Tokens)** |
| LLaMA 3 - 70b | 86.5 | 84.5 | 83 | 81.5 | 83.9 | 70 billion | 15 Trillion |
| PaLM 2 | 86 | 84 | 82.5 | 80 | 83.1 | 340 billion | 3.6 trillion |
| LLaMA 2 - 70b | 85 | 83.5 | 82 | 79 | 82.4 | 70 billion | 2 trillion |
| Vicuna | 84 | 82.5 | 81 | 78 | 81.4 | 13 billion | 2048 |
| LLaMA 3 - 8b | 82 | 80.5 | 79 | 76 | 79.4 | 8 billion | 15 Trillion |
| LLaMA 2 - 13b | 82.5 | 81 | 79.5 | 76.5 | 79.9 | 13 billion | 2 trillion |

A screenshot of a phone

Description automatically generated

A screenshot of a screenshot of a graph

Description automatically generated

A screenshot of a graph

Description automatically generated

A screenshot of a graph

Description automatically generated

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | **GLUE Score** | **SuperGLUE Score** | **MMLU Score** | **Code Generation Score** | **Overall Score** | **Parameter** | **Training Data Size (Tokens)** |
| LLaMA 2 - 7b | 80.5 | 79.5 | 78 | 75 | 78.3 | 7 billion |  |
| LLaMA 2 - 7b-instruct | 80 | 79 | 77.5 | 74.5 | 77.8 | 7 billion (fine-tuned) |  |
| Electra | 80 | 78.5 | 77 | Not ranked | 78.5 | 110 million |  |
| XLNet | 81 | 80 | 78.5 | Not ranked | 79.8 | 340 million |  |
| RoBERTa | 82.5 | 81.5 | 80.5 | Not ranked | 81.5 | 355 million |  |
| DeBERTa | 82 | 81 | 79.5 | Not ranked | 80.8 | 340 million |  |
| UniLM | 81 | 80 | 78.5 | Not ranked | 79.8 | 340 million |  |
| GPT-2 | 82 | 80.5 | 79 | Not ranked | 80.5 | 1.5 billion |  |
| CTRL | 79.5 | 78 | 76.5 | Not ranked | 78 | 110 million | 140GB Text Data |
| ERNIE | 80 | 78.5 | 77.5 | Not ranked | 78.7 | 340 million |  |
| BERT | 80.5 | 79.5 | 78 | Not ranked | 79.3 | 110 million |  |
| StableLM | 79 | 77.5 | 76 | 73 | 76.4 | 7 billion | 1.5 Trillion |
| Flan-T5 | 81.5 | 80 | 78.5 | 75.5 | 78.9 | 11 billion |  |
| BLOOM | 79 | 77.5 | 76 | 73 | 76.4 | 8 billion |  |
| T5 | 82 | 81 | 80 | Not ranked | 81 | 11 billion |  |
| ALBERT | 81.5 | 80.5 | 79.5 | Not ranked | 80.5 | 11 billion |  |

A diagram of a diagram

Description automatically generated

A screenshot of a phone

Description automatically generated

**Parameter Count:**

The number of parameters determines the size and complexity of the model. More parameters indicate a larger model.

Larger models can represent more information and learn more complex relationships.

However, more parameters also require more computational power and memory.

**Performance and Generalization:**

More parameters generally lead to better performance, especially with large datasets.

However, there is a risk of overfitting. Too many parameters can cause the model to memorize the training data and reduce generalization ability.

**Computational Power and Memory:**

More parameters require more computational power and memory. This is crucial during training and inference.

Larger models demand more CPU and memory resources.

**Censorship and Constraints:**

Fewer parameters may result in less censorship and fewer constraints. This allows the model to produce freer responses.

--------------------------------------------------------------------------------------

**Token Count:**

The number of tokens is an important factor that affects a language models performance and generalization ability.

**Few Tokens:**

Having fewer tokens results in faster inference times. The model can respond more quickly.

However, a smaller context is represented, containing less information. This may lead the model to provide narrower responses.

**Many Tokens:**

More tokens represent more information. This helps the model understand a broader context.

However, having too many tokens increases computational power and memory requirements. Larger models demand more resources.

**Risk of Overfitting:**

More tokens can increase the risk of overfitting to the training data. The model might memorize the training data and reduce generalization ability.

**Data Coverage:**

More tokens represent a wider range of data. This helps the model have knowledge across various topics.

In summary, finding the right balance of token count is essential for performance, speed, and generalization ability. The ideal token count depends on the specific use case and available resources