**Recommended Open-Source/Unpaid LLM Models**

**1. BLOOM by BigScience**

* **Description**: BLOOM (BigScience Large Open-science Open-access Multilingual Language Model) is a multilingual model designed to perform well across various languages and tasks.
* **SuperGLUE Score**: ~78.5
* **MMLU Score**: ~70.0
* **Strengths**: Multilingual capabilities, open-access nature, strong performance in summarization tasks.
* **Link**: BLOOM on Hugging Face

**2. GPT-NeoX by EleutherAI**

* **Description**: GPT-NeoX is an open-source LLM developed by EleutherAI, known for its performance in generating human-like text.
* **SuperGLUE Score**: ~72.3
* **MMLU Score**: ~63.2
* **Strengths**: Good performance on various NLP tasks, strong community support, scalability.
* **Link**: GPT-NeoX on Hugging Face

**3. GPT-J by EleutherAI**

* **Description**: GPT-J is an open-source model from EleutherAI with strong performance in various language tasks.
* **SuperGLUE Score**: ~67.3
* **MMLU Score**: ~55.6
* **Strengths**: Strong performance in generating coherent text, cost-effective.
* **Link**: GPT-J on Hugging Face

**4. T5 by Google**

* **Description**: T5 (Text-to-Text Transfer Transformer) is a versatile model that can be fine-tuned for various NLP tasks, including summarization.
* **SuperGLUE Score**: ~89.3 (for T5-XXL, the largest variant)
* **Strengths**: High performance, versatility in handling different tasks, effective in summarization.
* **Link**: T5 on Hugging Face

**6. RoBERTa by Facebook AI**

* **Description**: RoBERTa is an optimized version of BERT.
* **SuperGLUE Score**: ~89.9 (for RoBERTa-large)
* **Link**: RoBERTa on Hugging Face

**7. ALBERT by Google Research**

* **Description**: ALBERT is a smaller and more efficient version of BERT.
* **SuperGLUE Score**: ~89.3 (for ALBERT-xxlarge)
* **Link**: ALBERT on Hugging Face

**8. FLAN-T5 by Google Research**

* **Description**: FLAN-T5 is a family of models trained on a mixture of natural language understanding and generation tasks.
* **SuperGLUE Score**: ~89.3 (for FLAN-T5-XXL)
* **Link**: FLAN-T5 on Hugging Face

**Analysis**

* **Performance**: T5,RoBERTa,ALBERT, especially the larger variants, offers the highest performance but requires more computational resources. BLOOM and GPT-NeoX provide a balance of performance and accessibility.
* **Cost and Availability**: All recommended models are free and open-source, ensuring no cost for API usage. Computational costs can be managed by using efficient cloud services or local deployment.
* **Suitability for Summarization**: T5 and BLOOM are particularly strong in text generation and summarization tasks, making them ideal for our use case.

**Conclusion**

For the educational application aimed at summarizing student performance, the following models are recommended based on their performance, cost-effectiveness, and suitability:

1. T5
2. RoBERTa
3. ALBERT

**Recommended Cloud**

[**https://github.com/shreedhar13/OpenLLM**](https://github.com/shreedhar13/OpenLLM)

**Cloud – K8S , BentoCloud**