Evaluation Results

First Part: Perplexity

Perplexity: 99145.3671875

We randomly chose 10 samples from openwebtext-10k as input and analyzed the production. The mean perplexity is noted.

Second Part: Token Probability Analysis

Using GPT-4o, we generated short questions and answers as follows:

Example Input: Question: What is the capital of Turkey? Answer: Ankara

Input to Model: What is the capital of Turkey?

Output Analysis:

- Real Answer: Ankara
- Answer Tokens: tensor([2025, 74, 3301], device='cuda:0')
- Shape of Answer Tokens: torch.Size([1, 3])
- Tokens: ['An', 'k', 'ara']

Token Probabilities:

```
[[[1.5493580e-10 1.1943685e-03 3.7615869e-06]]

  [[7.5981096e-11 6.2307419e-04 1.5106030e-05]]

  [[1.2372516e-09 6.2036997e-04 3.7345469e-06]]

  [[4.3968430e-11 5.4030988e-04 5.9681784e-06]]

  [[2.7153095e-09 5.0398626e-04 4.2269635e-06]]

  [[4.3304700e-11 3.8512098e-04 3.9767128e-06]]

  [[2.8124436e-09 5.0820532e-04 4.3839314e-06]]

  [[7.0158962e-11 4.9186411e-04 4.2624911e-06]]

  [[8.8572683e-10 8.1146474e-04 3.8328849e-06]]

  [[1.7995778e-10 1.2498297e-03 8.6838563e-06]]]
```

Observations:

- The model seems to give first token low probabilities. And Tokens which are not common in English, exhibit lower probabilities. This trend is especially visible in geographical names or non-English words, such as Hiroshima, Guangzhou, and Deutsch.

Third Part: LLM Sentence Quality Evaluation

Criteria:

- Grammar and Syntax: X/10
- Semantic Clarity: X/10
- Contextual Relevance: X/10
- Creativity and Style: X/10

Results for 10 Generated Sentences:

- Sentence 0: **4/40**
- Sentence 1: **7/40**
- Sentence 2: **4/40**
- Sentence 3: **6/40**
- Sentence 4: **4/40**
- Sentence 5: **7/40**
- Sentence 6: **4/40**
- Sentence 7: **4/40**
- Sentence 8: **4/40**
- Sentence 9: **4/40**