Approach

The Sherlock Holmes corpus was used to train an LSTM based model for next-token prediction. A total of ~500,000 individual tokens were seen with a unique word count of ~15k. A SentencePiece tokenizer was trained on the corpus, resulting in a vocabulary size of approximately 14,000 subword units. The text was split into sequences of 48 tokens, with the model trained to predict the subsequent token.

A hybrid model architecture was designed, combining an LSTM to capture sequential dependencies with a multi-head self-attention layer to refine contextual understanding. Residual connections, layer normalization, and dropout were incorporated for stability and regularization.

The training setup used the Adam optimizer with weight decay, gradient clipping, learning rate scheduling and early stopping with a patience of five epochs to combat overfitting. Key evaluation metrics included cross-entropy loss, accuracy, top-k accuracies (Top-1,Top-5), and perplexity, which together captured both predictive accuracy and sequence modeling performance.

Metrics:

Train - Loss: 3.7132, **Perplexity**: 40.99

Val - Loss: 4.9789, **Perplexity**: 145.32

Train Acc - Top-1 Accuracy: 0.252, Top-5 Accuracy : 0.501

Val Acc - Top-1 Accuracy : 0.205, Top-5 Accuracy : 0.421

Sample Generations (30-Word Generations)

Prompt: The mystery deepened when Holmes
Generated: *The mystery deepened when Holmes turned in the centre of the hall. “Yes, I think that you may have come from your family. I am afraid to know that I have*

Prompt: Inspector Lestrade examined the evidence
Generated: *Inspector Lestrade examined the evidence. “I am afraid that it is now,” said he. “I should like to see you, Watson, it is a very interesting one or*

Prompt: The chemical analysis revealed
Generated: *The chemical analysis revealed that he was not. “It is quite impossible,” said he. “I have seen some good enough. He has already been an old eye upon the*

Prompt: I saw Holmes

Generated: *I saw Holmes. “I am afraid that he has been a little cold-looking, and that it is not very much mistaken,” said Holmes. “I am*

Trace

1. **Current text:** `I saw Holmes.`
   Chosen: `.` (Rank: 1, Prob: 0.5108) — **SELECTED**
   Top-5:
   1. `.` (0.5108) — SELECTED
   2. `,` (0.3927)
   3. `and` (0.0459)
   4. `with` (0.0229)
   5. `in` (0.0138)
2. **Current text:** `I saw Holmes. “`
   Chosen: `“` (Rank: 1, Prob: 0.9998) — **SELECTED**
   Top-5:
   1. `“` (0.9998) — SELECTED
   2. `A` (0.0001)
   3. `The` (≈0.0)
   4. `“‘` (≈0.0)
   5. `He` (≈0.0)
3. **Current text:** `I saw Holmes. “I`
   Chosen: `I` (Rank: 1, Prob: 0.7735) — **SELECTED**
   Top-5:
   1. `I` (0.7735) — SELECTED
   2. `It` (0.1038)

3. `You` (0.0932)
4. `There` (0.0072)
5. `He` (0.0057)

4. **Current text:** `I saw Holmes. “I am`
   Chosen: `am` (Rank: 1, Prob: 0.7979) — **SELECTED**
   Top-5:
   1. `am` (0.7979) — SELECTED
   2. `have` (0.0707)
   3. `cannot` (0.0339)
   4. `think` (0.0325)
   5. `should` (0.0137)

5. **Current text:** `I saw Holmes. “I am afraid`
   Chosen: `afraid` (Rank: 1, Prob: 0.9903) — **SELECTED**
   Top-5:
   1. `afraid` (0.9903) — SELECTED
   2. `glad` (0.0027)
   3. `delighted` (0.0020)
   4. `sorry` (0.0019)
   5. `sure` (0.0019)

6. **Current text:** `I saw Holmes. “I am afraid that`
   Chosen: `that` (Rank: 1, Prob: 0.9706) — **SELECTED**
   Top-5:
   1. `that` (0.9706) — SELECTED
   2. `to` (0.0166)
   3. `of` (0.0023)

4. `,` (0.0018)

5. `now` (0.0014)

7. **Current text:** `I saw Holmes. “I am afraid that he`
   Chosen: `he` (Rank: 1, Prob: 0.4641) — **SELECTED**
   Top-5:

   1. `he` (0.4641) — SELECTED
   2. `I` (0.4228)
   3. `it` (0.0391)
   4. `you` (0.0291)
   5. `the` (0.0230)

8. **Current text:** `I saw Holmes. “I am afraid that he has`
   Chosen: `has` (Rank: 4, Prob: 0.1086) — **SELECTED**
   Top-5:

   1. `had` (0.3027)
   2. `is` (0.2979)
   3. `was` (0.2036)
   4. `has` (0.1086) — SELECTED
   5. `might` (0.0455)

9. **Current text:** `I saw Holmes. “I am afraid that he has been`
   Chosen: `been` (Rank: 1, Prob: 0.9307) — **SELECTED**
   Top-5:

   1. `been` (0.9307) — SELECTED
   2. `done` (0.0361)
   3. `a` (0.0109)
   4. `already` (0.0043)

5. `heard` (0.0025)

10. **Current text:** `I saw Holmes. “I am afraid that he has been a`
    Chosen: `a` (Rank: 1, Prob: 0.5613) — **SELECTED**
    Top-5:
    1. `a` (0.5613) — SELECTED
    2. `no` (0.0990)
    3. `done` (0.0953)
    4. `taken` (0.0438)
    5. `in` (0.0197)
11. **Current text:** `I saw Holmes. “I am afraid that he has been a little`
    Chosen: `little` (Rank: 2, Prob: 0.0884) — **SELECTED**
    Top-5:
    1. `very` (0.6354)
    2. `little` (0.0884) — SELECTED
    3. `good` (0.0638)
    4. `considerable` (0.0565)
    5. `small` (0.0365)
12. **Current text:** `I saw Holmes. “I am afraid that he has been a little cold`
    Chosen: `cold` (Rank: Outside top-5) — **SELECTED**
    Top-5:
    1. `more` (0.4745)
    2. `,”` (0.1763)
    3. `very` (0.0951)
    4. `problem` (0.0781)

5. `enough` (0.0422)

13. **Current text:** `I saw Holmes. “I am afraid that he has been a little cold-`
    Chosen: `-` (Rank: 2, Prob: 0.1291) — **SELECTED**
    Top-5:

    1. `,”` (0.8063)
    2. `-` (0.1291) — SELECTED
    3. `,` (0.0191)
    4. `to` (0.0130)
    5. `for` (0.0076)

14. **Current text:** `I saw Holmes. “I am afraid that he has been a little cold-looking`
    Chosen: `looking` (Rank: 2, Prob: 0.2077) — **SELECTED**
    Top-5:

    1. `headed` (0.3456)
    2. `looking` (0.2077) — SELECTED
    3. `room` (0.1259)
    4. `handed` (0.0671)
    5. `twenty` (0.0364)

15. **Current text:** `I saw Holmes. “I am afraid that he has been a little cold-looking,`
    Chosen: `,` (Rank: 3, Prob: 0.1088) — **SELECTED**
    Top-5:

    1. `man` (0.7021)
    2. `,”` (0.1381)
    3. `,` (0.1088) — SELECTED

4. `.` (0.0165)
5. `-` (0.0084)

16. **Current text:** `I saw Holmes. “I am afraid that he has been a little cold-looking, and`
    Chosen: `and` (Rank: 1, Prob: 0.8569) — **SELECTED**
    Top-5:
    1. `and` (0.8569) — SELECTED
    2. `but` (0.0324)
    3. `however` (0.0229)
    4. `Watson` (0.0226)
    5. `for` (0.0069)
17. **Current text:** `I saw Holmes. “I am afraid that he has been a little cold-looking, and that`
    Chosen: `that` (Rank: 3, Prob: 0.1341) — **SELECTED**
    Top-5:
    1. `yet` (0.3803)
    2. `he` (0.1587)
    3. `that` (0.1341) — SELECTED
    4. `I` (0.1040)
    5. `the` (0.0954)
18. **Current text:** `I saw Holmes. “I am afraid that he has been a little cold-looking, and that it`
    Chosen: `it` (Rank: 3, Prob: 0.0330) — **SELECTED**
    Top-5:
    1. `he` (0.8163)
    2. `I` (0.0484)

3. `it` (0.0330) — SELECTED
4. `is` (0.0304)
5. `the` (0.0238)

19. **Current text:** `I saw Holmes. “I am afraid that he has been a little cold-looking, and that it is`
    Chosen: `is` (Rank: 1, Prob: 0.8485) — **SELECTED**
    Top-5:
    1. `is` (0.8485) — SELECTED
    2. `was` (0.1316)
    3. `has` (0.0158)
    4. `would` (0.0019)
    5. `must` (0.0007)

20. **Current text:** `I saw Holmes. “I am afraid that he has been a little cold-looking, and that it is not`
    Chosen: `not` (Rank: 2, Prob: 0.2994) — **SELECTED**
    Top-5:
    1. `a` (0.3342)
    2. `not` (0.2994) — SELECTED
    3. `quite` (0.1238)
    4. `very` (0.0525)
    5. `no` (0.0229)

21. **Current text:** `I saw Holmes. “I am afraid that he has been a little cold-looking, and that it is not very`
    Chosen: `very` (Rank: 2, Prob: 0.2683) — **SELECTED**
    Top-5:
    1. `a` (0.3299)

2. `very` (0.2683) — SELECTED
3. `quite` (0.1417)
4. `yet` (0.0801)
5. `an` (0.0198)

22. **Current text:** `I saw Holmes. “I am afraid that he has been a little cold-looking, and that it is not very much`
    Chosen: `much` (Rank: 1, Prob: 0.8641) — **SELECTED**
    Top-5:
    1. `much` (0.8641) — SELECTED
    2. `well` (0.1053)
    3. `interesting` (0.0080)
    4. `nicely` (0.0037)
    5. `clear` (0.0036)

23. **Current text:** `I saw Holmes. “I am afraid that he has been a little cold-looking, and that it is not very much mistaken`
    Chosen: `mistaken` (Rank: Outside top-5) — **SELECTED**
    Top-5:
    1. `for` (0.2713)
    2. `obliged` (0.2310)
    3. `to` (0.0963)
    4. `enough` (0.0601)
    5. `,”` (0.0569)

24. **Current text:** `I saw Holmes. “I am afraid that he has been a little cold-looking, and that it is not very much mistaken,”`
    Chosen: `,”` (Rank: 1, Prob: 0.4069) — **SELECTED**
    Top-5:

1. `,”` (0.4069) — SELECTED
2. `,` (0.2440)
3. `.` (0.1161)
4. `to` (0.0843)
5. `for` (0.0670)

25. **Current text:** `I saw Holmes. “I am afraid that he has been a little cold-looking, and that it is not very much mistaken,” said`
    Chosen: `said` (Rank: 1, Prob: 0.9843) — **SELECTED**
    Top-5:
    1. `said` (0.9843) — SELECTED
    2. `he` (0.0147)
    3. `remarked` (0.0005)
    4. `I` (0.0002)
    5. `cried` (0.0001)

26. **Current text:** `I saw Holmes. “I am afraid that he has been a little cold-looking, and that it is not very much mistaken,” said Holmes`
    Chosen: `Holmes` (Rank: 1, Prob: 0.6264) — **SELECTED**
    Top-5:
    1. `Holmes` (0.6264) — SELECTED
    2. `he` (0.3643)
    3. `I` (0.0061)
    4. `the` (0.0020)
    5. `she` (0.0006)

27. **Current text:** `I saw Holmes. “I am afraid that he has been a little cold-looking, and that it is not very much mistaken,” said Holmes.`

Chosen: `.` (Rank: 1, Prob: 0.6404) — **SELECTED**
Top-5:

1. `.` (0.6404) — SELECTED
2. `,` (0.3588)
3. `;` (0.0005)
4. `as` (0.0002)
5. `."` (≈0.0)

28. **Current text:** `I saw Holmes. “I am afraid that he has been a little cold-looking, and that it is not very much mistaken,” said Holmes. “`
Chosen: “ (Rank: 1, Prob: 0.9999) — **SELECTED**
Top-5:

1. “ (0.9999) — SELECTED
2. “ ‘ (≈0.0)
3. `He` (≈0.0)
4. `A` (≈0.0)
5. `The` (≈0.0)

29. **Current text:** `I saw Holmes. “I am afraid that he has been a little cold-looking, and that it is not very much mistaken,” said Holmes. “I`
Chosen: `I` (Rank: 1, Prob: 0.8682) — **SELECTED**
Top-5:

1. `I` (0.8682) — SELECTED
2. `It` (0.1000)
3. `You` (0.0155)
4. `There` (0.0032)

5. `Well` (0.0030)

30. **Current text:** `I saw Holmes. “I am afraid that he has been a little cold-looking, and that it is not very much mistaken,” said Holmes. “I am`
    Chosen: `am` (Rank: 1, Prob: 0.7606) — **SELECTED**
    Top-5:

    1. `am` (0.7606) — SELECTED
    2. `have` (0.1017)
    3. `cannot` (0.0488)
    4. `should` (0.0261)
    5. `will` (0.0221)