Mechanistic Interpretability

GPT-2 small
WT 2024/25

Frederick Riemenschneider

21.11.2024

GPT-2

https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

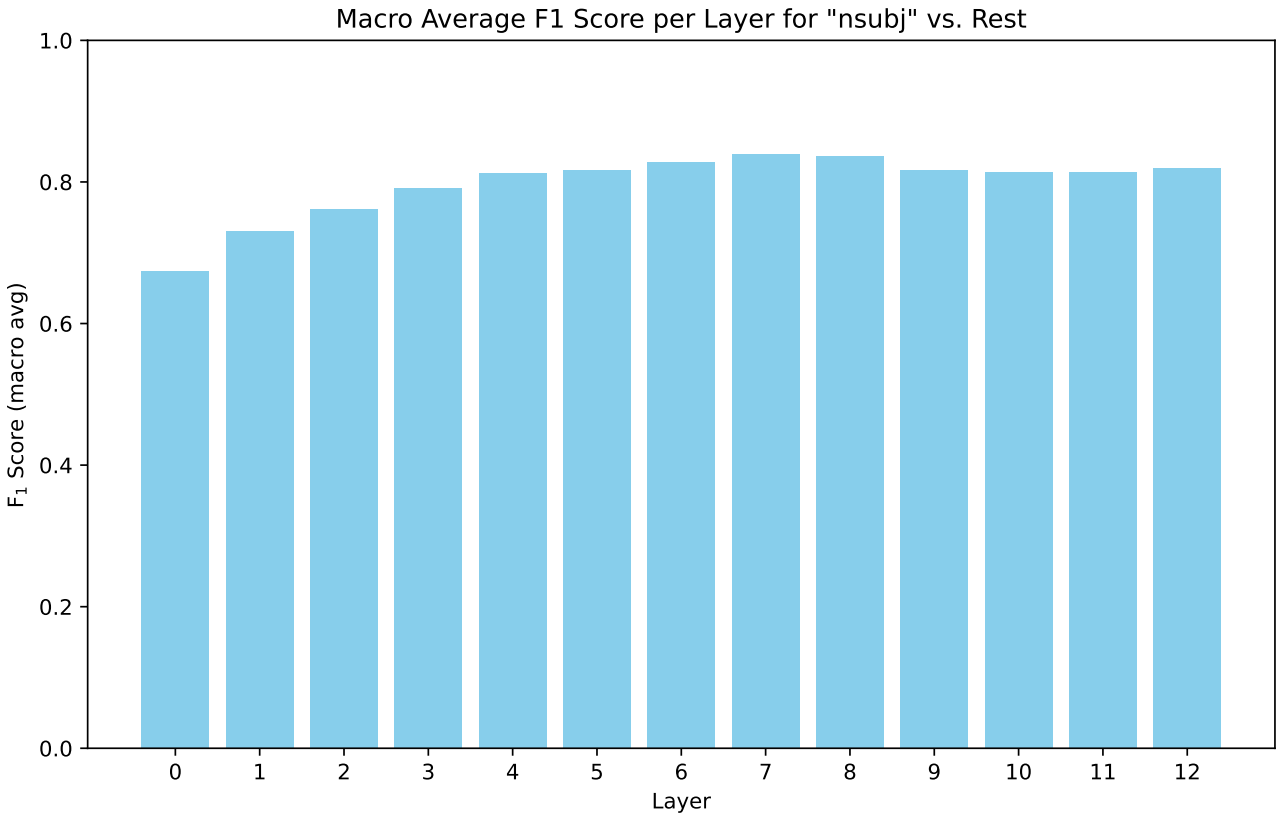
GPT-2

- 124M parameters (BERT base: 110M)
- 40GB pre-training data (BERT base: 16GB)

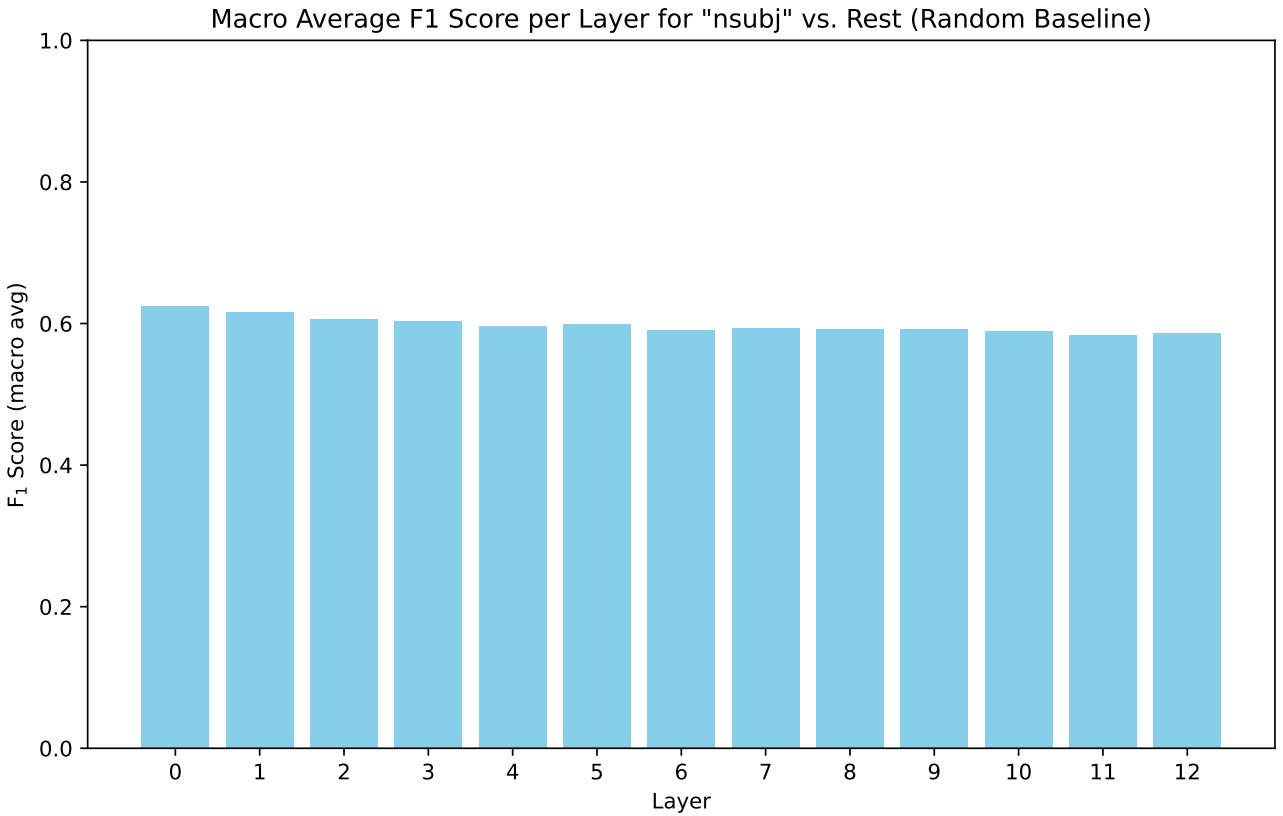
Experiment Setup

- English Web Treebank (Universal Dependencies)
- first 1000 sentences
  - $\rightarrow$ 25533 words
- random split, test size $= 0.2$
- logistic regression

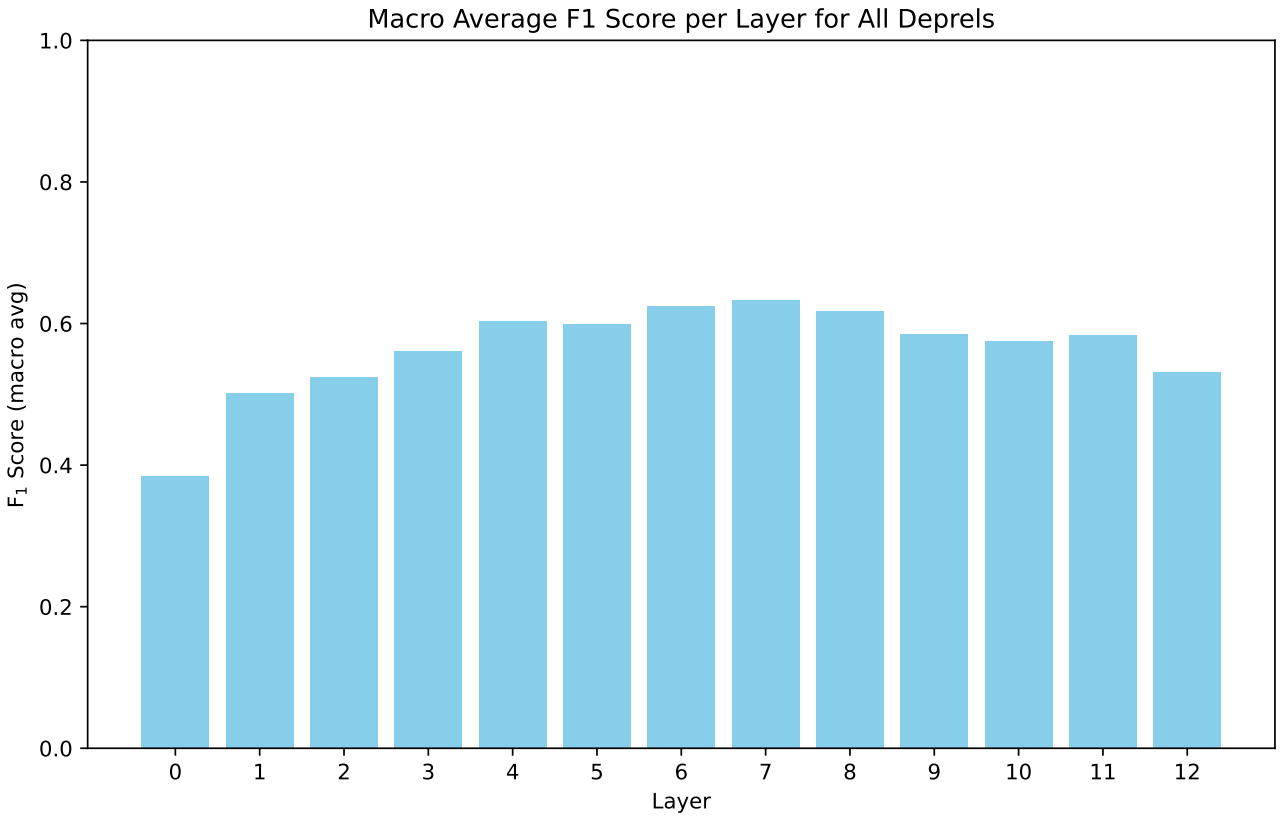
GPT-2 small


Macro Average F1 Score per Layer for "nsubj" vs. Rest
F1 Score (macro avg)
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
6
7
8
9
10
11
12
Layer

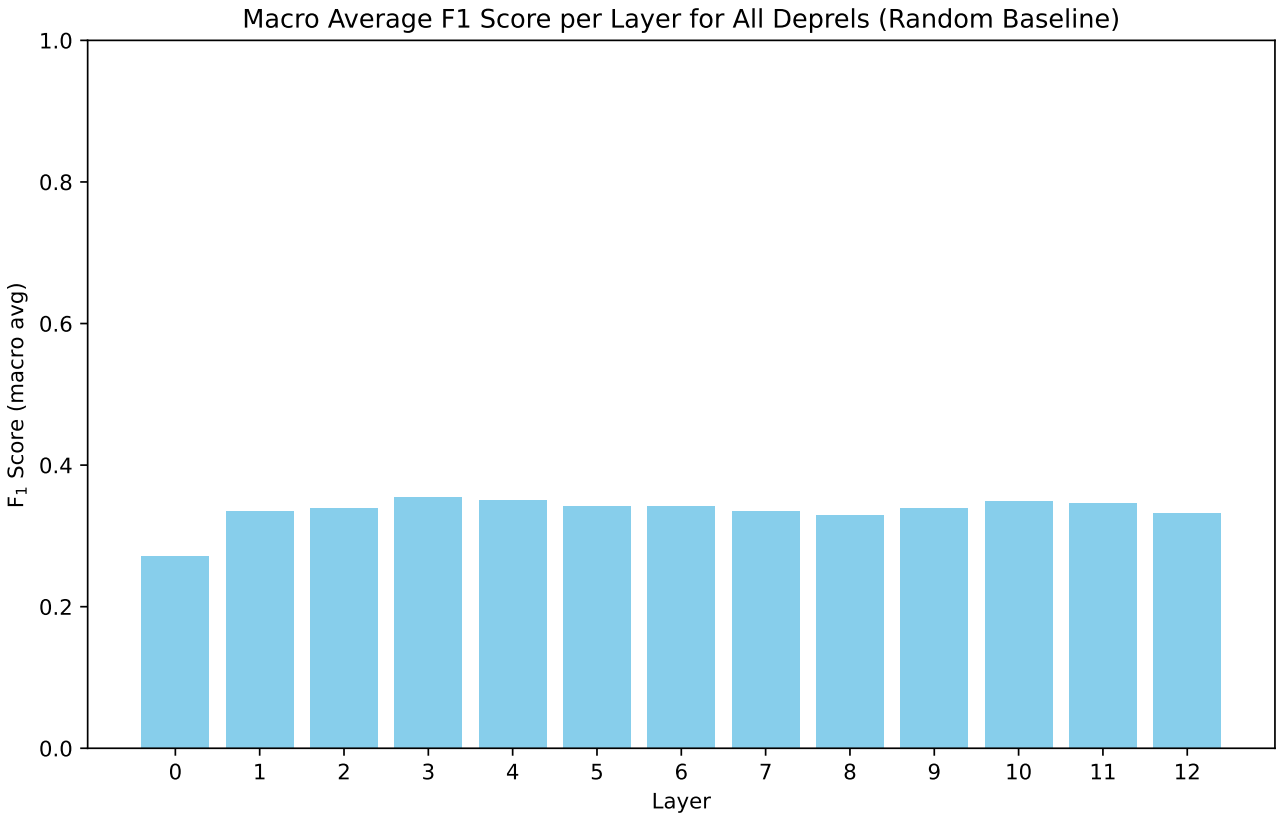
GPT-2 small


Macro Average F1 Score per Layer for "nsubj" vs. Rest (Random Baseline)
F1 Score (macro avg)
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
6
7
8
9
10
11
12
Layer

GPT-2 small


Macro Average F1 Score per Layer for All Deprels
F1 Score (macro avg)
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
6
7
8
9
10
11
12
Layer

GPT-2 small


Macro Average F1 Score per Layer for All Deprels (Random Baseline)
F1 Score (macro avg)
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
6
7
8
9
10
11
12
Layer