Interpreting GPT-2’s Zero-Shot Sequence Completion Ability

Rahul Chowdhury, Prof. Dr. David Bau

Introduction

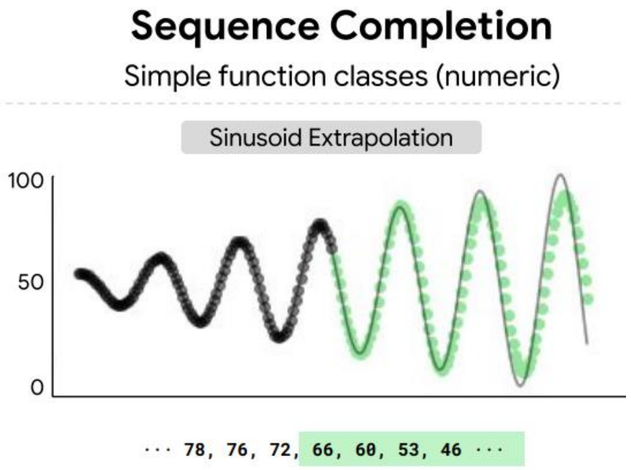


Fig 1 Sequence completion task performed on GPT-2 [1]

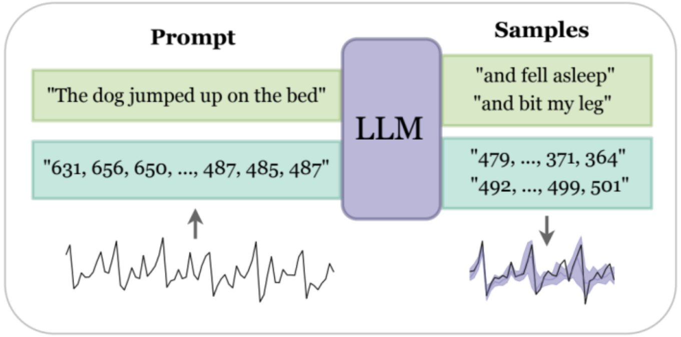
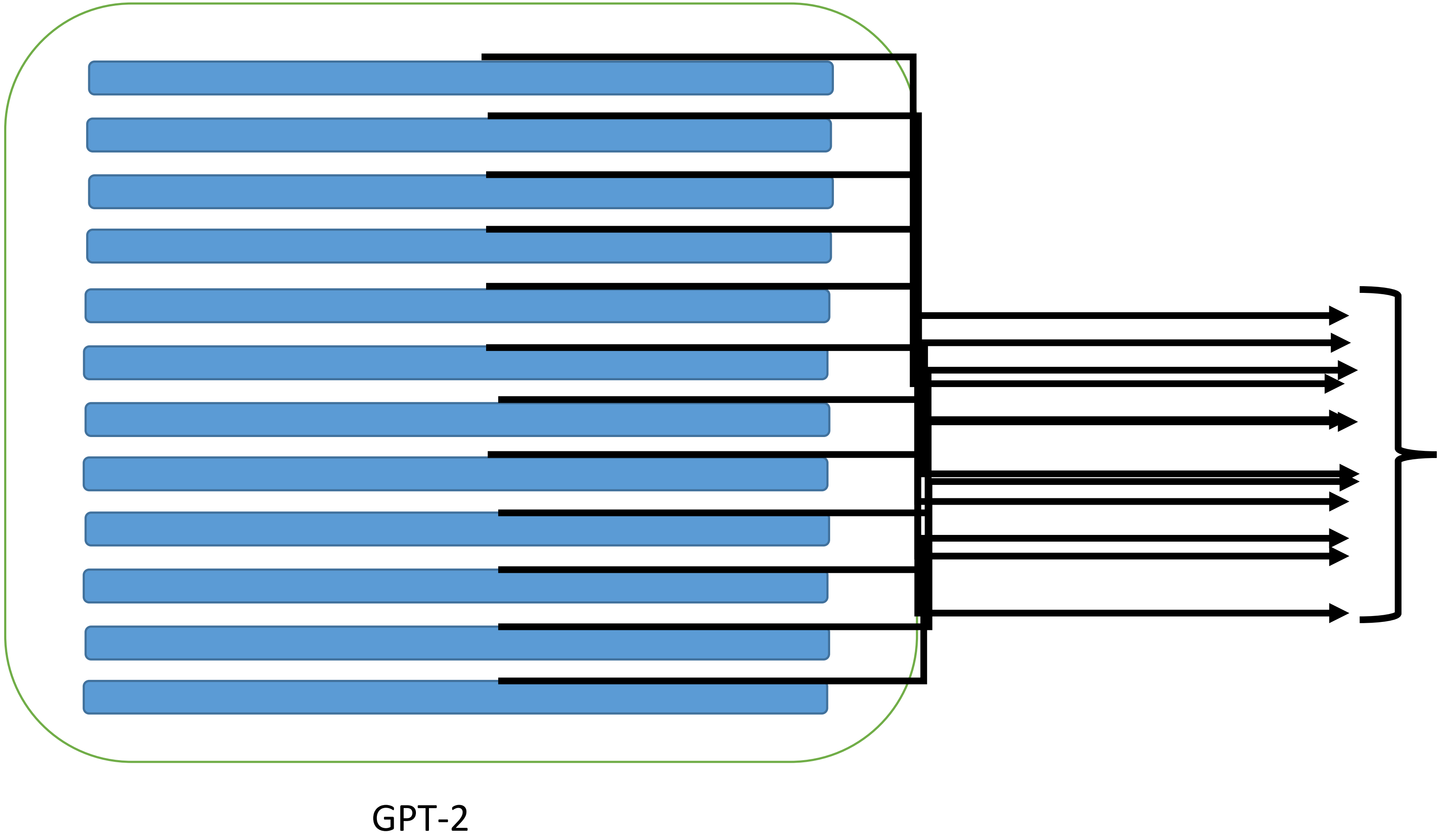


Fig 2 Sequence completion task as observed in Large Language Models [2]

Probing



1. Linear Regression Using The Activations
2. Logistic Regression Using The Activations

Probing for exponential, sinsuoids and mixture of both

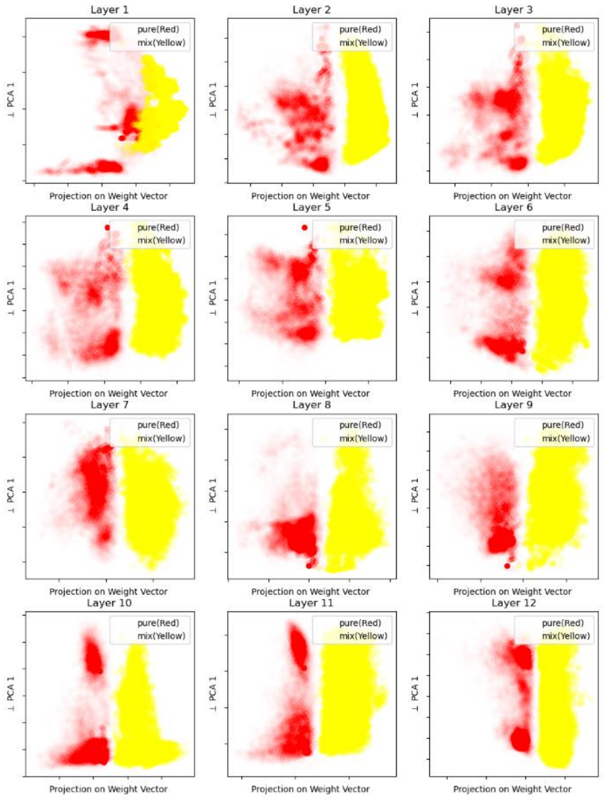


Figure 3. Linearly classifiable activation projection of mixture and pure sinusoids/ exponential signals at each layer of GPT-2

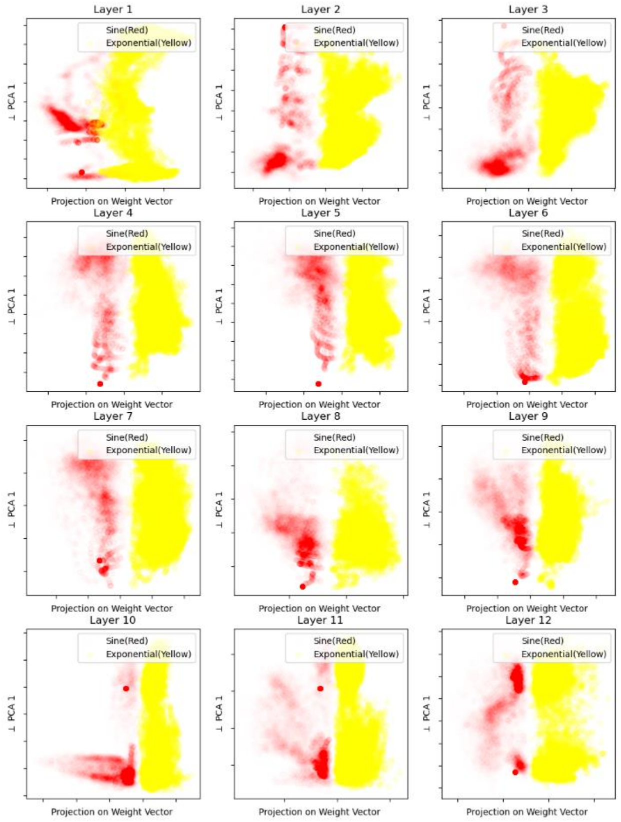
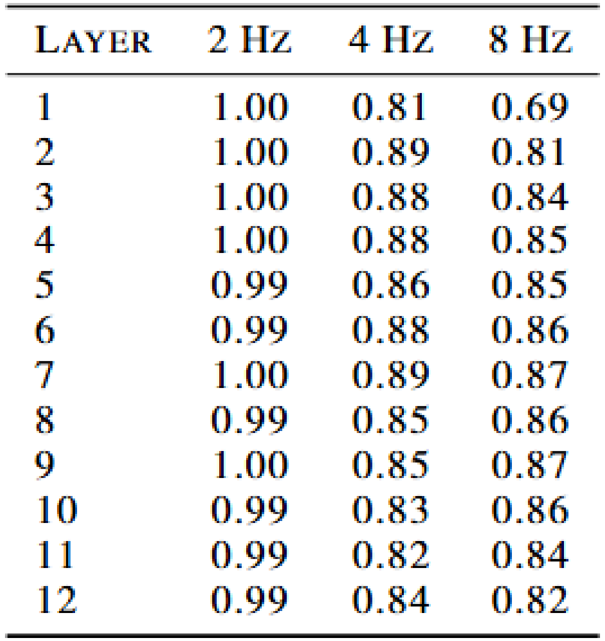


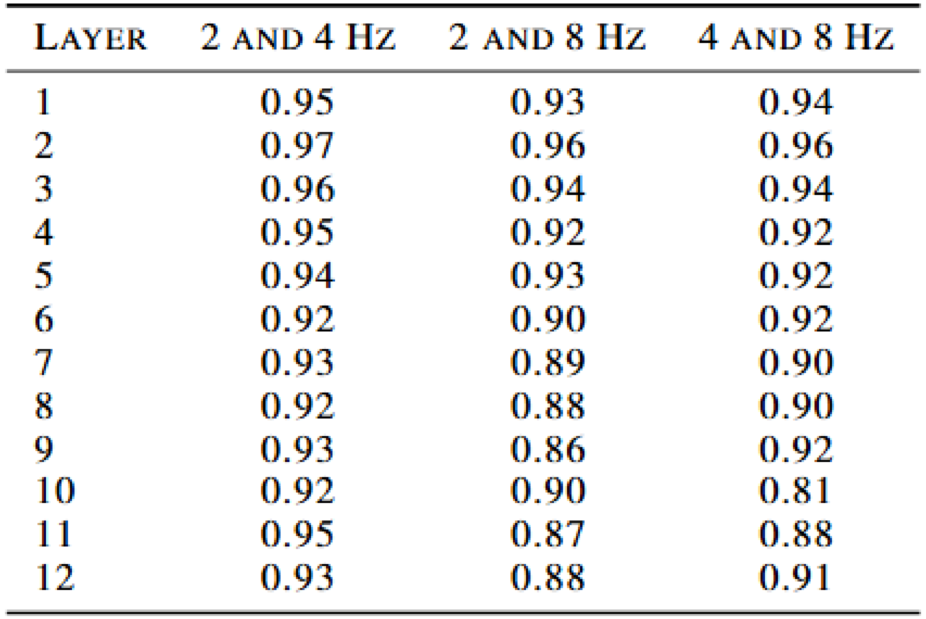
Figure 4. Linearly classifiable activation projection of sinusoidsand exponential signals at each layer of GPT-2

Probing For Basic Time Series Elements



Layer	2 Hz	4 Hz	8 Hz
1	1.00	0.81	0.69
2	1.00	0.89	0.81
3	1.00	0.88	0.84
4	1.00	0.88	0.85
5	0.99	0.86	0.85
6	0.99	0.88	0.86
7	1.00	0.89	0.87
8	0.99	0.85	0.86
9	1.00	0.85	0.87
10	0.99	0.83	0.86
11	0.99	0.82	0.84
12	0.99	0.84	0.82

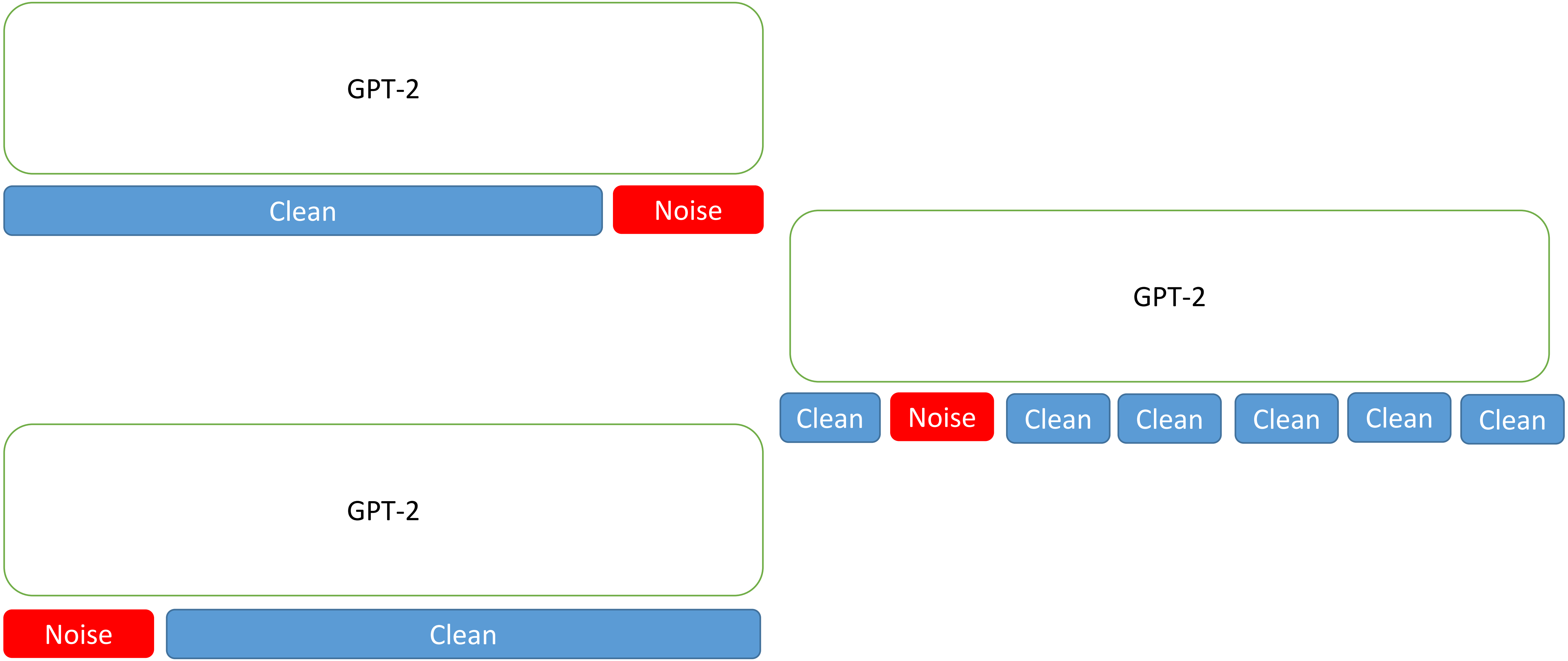
Table 1. Classification accuracies of three frequencies from waves containing pure and composition of different frequencies using activations of GPT-2 from each layer



Layer	2 and 4 Hz	2 and 8 Hz	4 and 8 Hz
1	0.95	0.93	0.94
2	0.97	0.96	0.96
3	0.96	0.94	0.94
4	0.95	0.92	0.92
5	0.94	0.93	0.92
6	0.92	0.90	0.92
7	0.93	0.89	0.90
8	0.92	0.88	0.90
9	0.93	0.86	0.92
10	0.92	0.90	0.81
11	0.95	0.87	0.88
12	0.93	0.88	0.91

Table 2. Regression scores of peaks of waves composed of multiplefrequencies using activations of GPT-2 from each layer

Counterfactual Inputs



GPT-2

Clean Noise

GPT-2

Clean Noise Clean Clean Clean Clean Clean

GPT-2

Noise Clean

Counterfactuals-Contiguous Noise From First Token (left column) and From Last Token (right column)

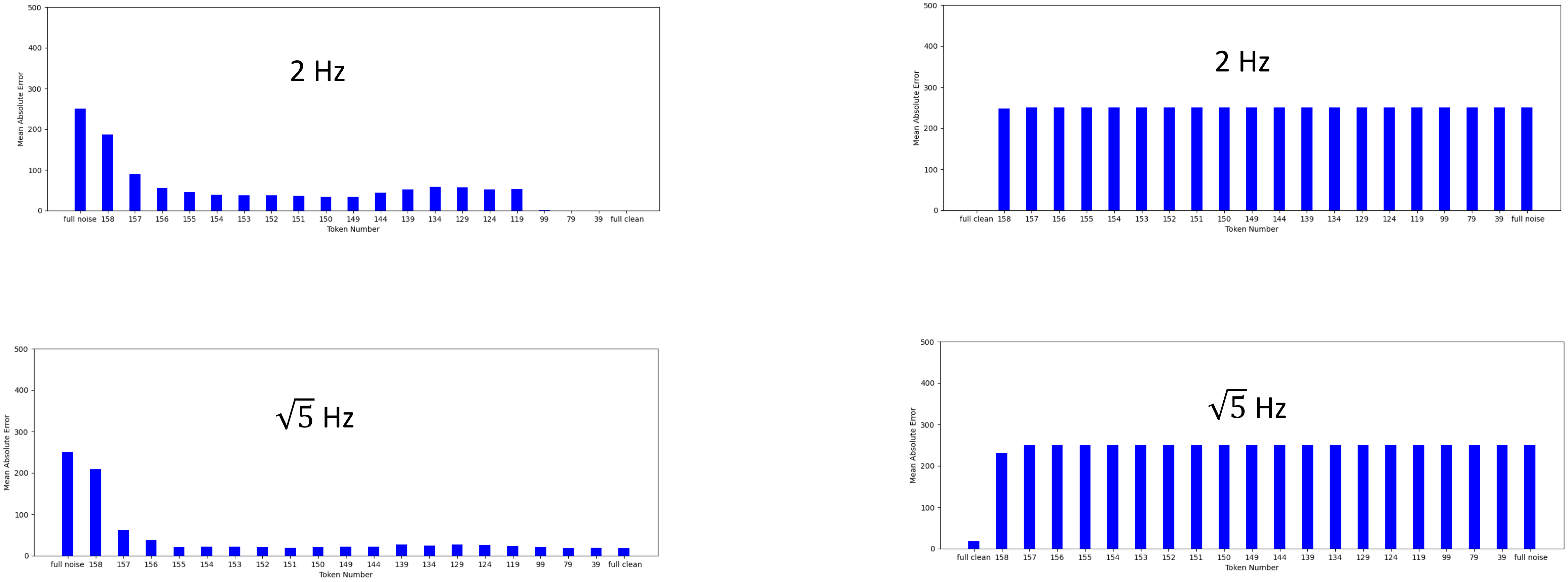


Fig 5 Shows the mean absolute error with respect to the original output after counterfactual signals are fed along with partial noise and vice verse

Counterfactuals-Single Token Noise

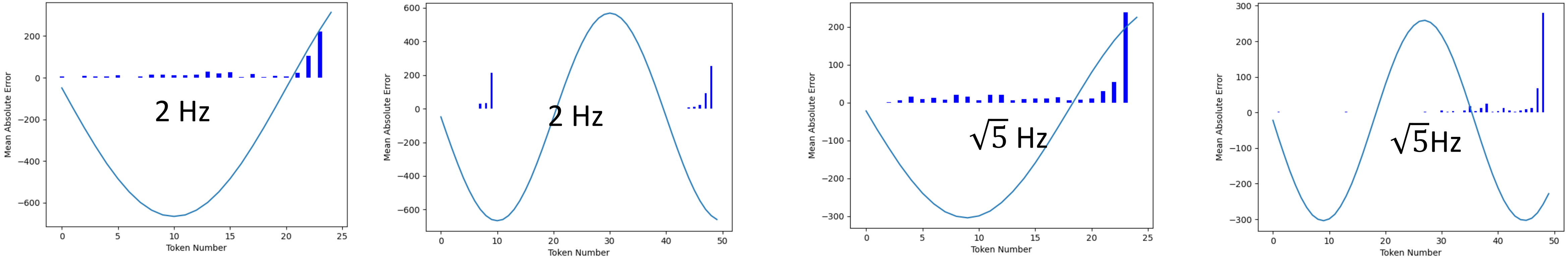


Fig 6 Shows the mean absolute error with respect to the output of the GPT-2 after counterfactual signal is fed at each token position

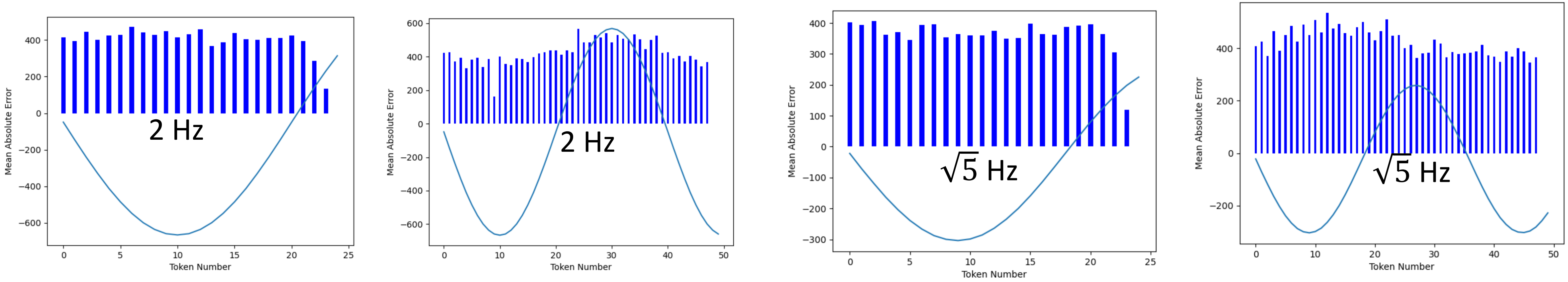
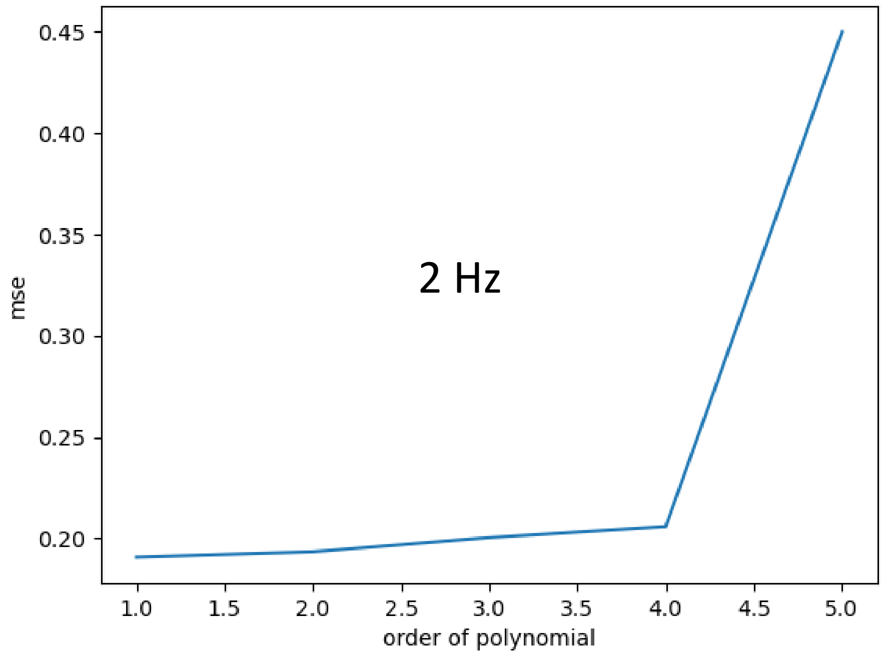
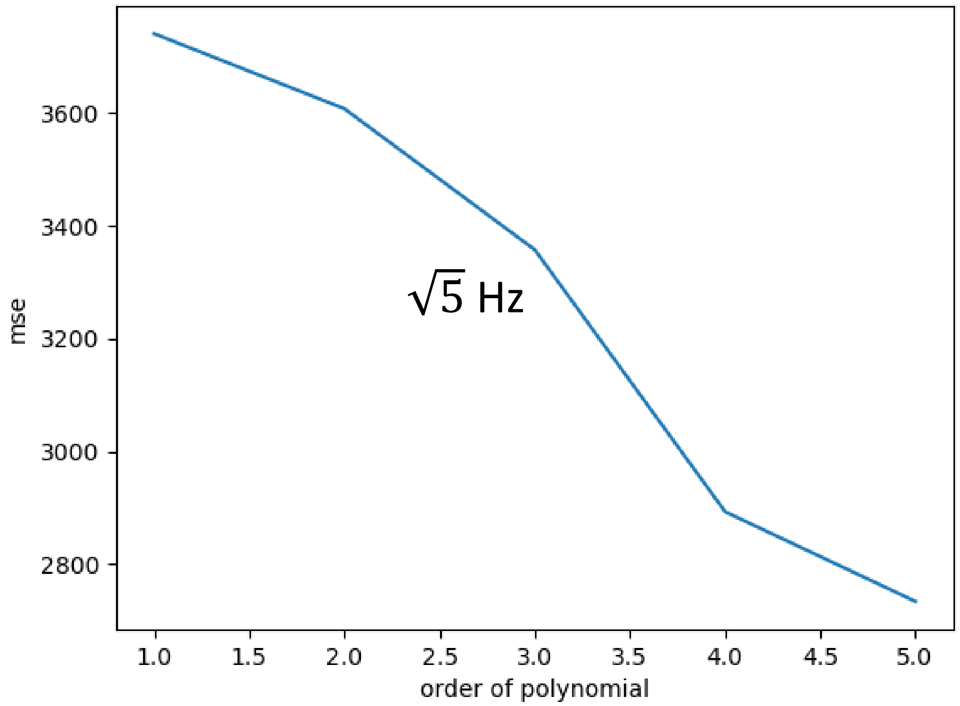


Fig 7 Shows the mean absolute error with respect to the counterfactual input to the GPT-2 at each token position to measure the alignment of the output to that noisy token

What Model Could GPT-2 Be Using?

For a length of 50 tokens the 46-49th tokens (last 3 tokens) were used to train a linear model to estimate a model of GPT-2's prediction





A Toy Experiment

References

[1] Mirchandani, S., Xia, F., Florence, P., Ichter, B., Driess, D., Arenas, M. G., Rao, K., Sadigh, D., and Zeng, A. Largelanguage models as general pattern machines. arXiv preprint arXiv:2307.04721, 2023.

[2] Gruver, N., Finzi, M., Qiu, S., and Wilson, A. G. Large language models are zero-shot time series forecasters. Advances inNeural Information Processing Systems, 36, 2024.