Modern Language Models

Language Models

Why are language models important?

Many tasks can be expressed as a sequence prediction problem

- Programming Languages: Generate code
- Language of Mathematics: Solving equations

Language of Mathematics

$$49 + 10 =$$

Language of Mathematics

$$49 + 10 = 59$$

Language of Mathematics

49 10 59

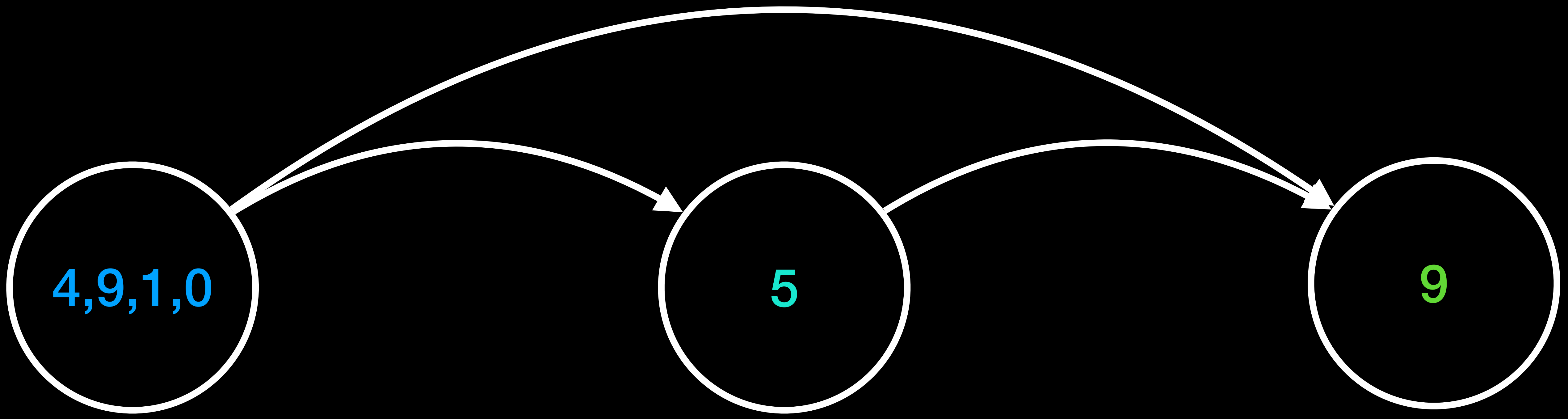
Language of Mathematics

4,9,1,0,5,9

Language Model

n^{th} order Markov assumption

49 + 10 = 59



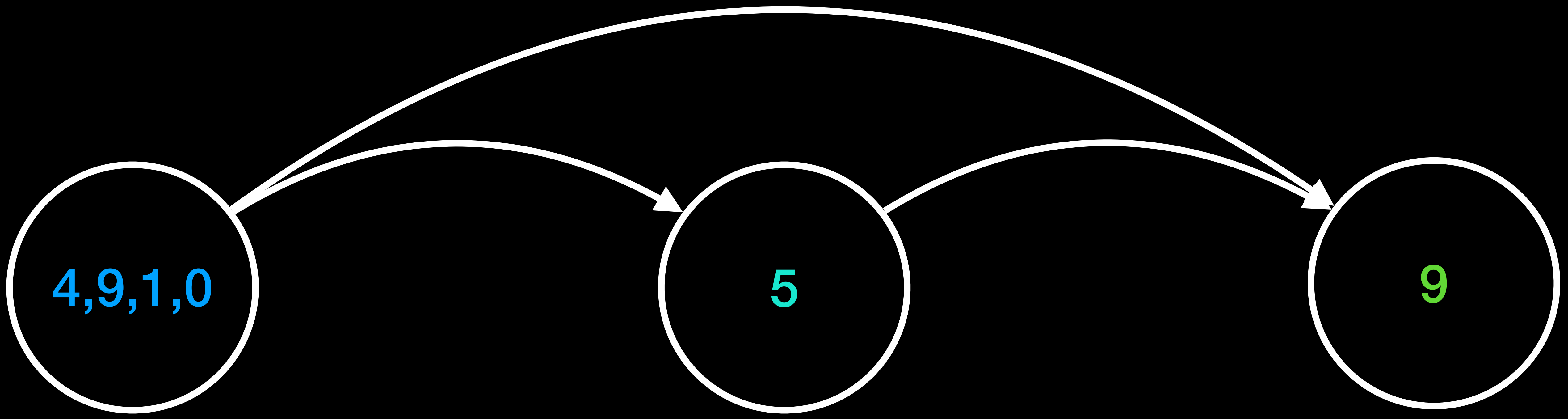
The next symbol 5 is influenced by all previous symbols

The next symbol 9 is influenced by all previous symbols

Language Model

Sequence Probability

49 + 10 = 59



p(5,9 | 4,9,1,0)

Language Model

Predict a Sequence

$$p(5,8 \mid 4,9,1,0) = 0$$

$$p(5,9 \mid 4,9,1,0) = 1$$

$$p(6,0 \mid 4,9,1,0) = 0$$

Language Model

Predict a Sequence

$$p(5,8 \mid 4,9,1,0) = 0$$

$$p(5,9 \mid 4,9,1,0) = 1$$

$$p(6,0 \mid 4,9,1,0) = 0$$

Sequence Probability

Combinations

49 + 10 = 59
27 + 30 = 57
00 + 26 = 26
40 + 47 = 87
03 + 32 = 35
…

Sequence Probability

Combinations

49 + 10 = 59

27 + 30 = 57

00 + 26 = 26

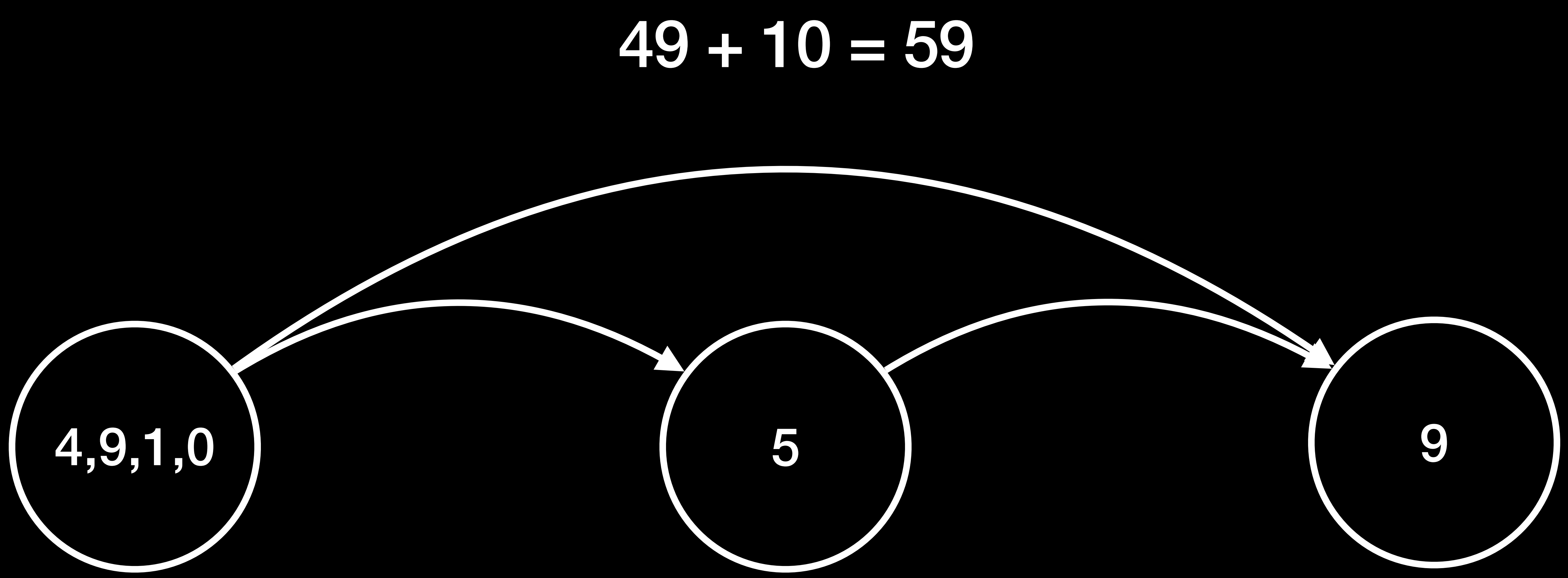
40 + 47 = 87

03 + 32 = 35

…

Language Model

Sequence Probability

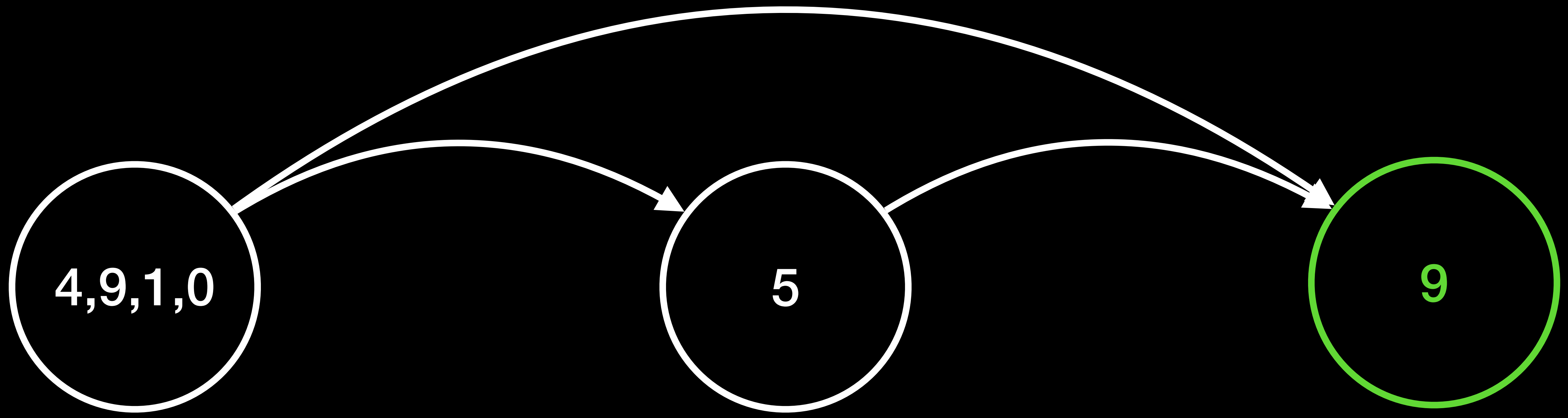


$p(5 \mid 4,9,1,0) \cdot p(9 \mid 4,9,1,0,5) = p(5,9 \mid 4,9,1,0)$

Language Model

Sequence Probability

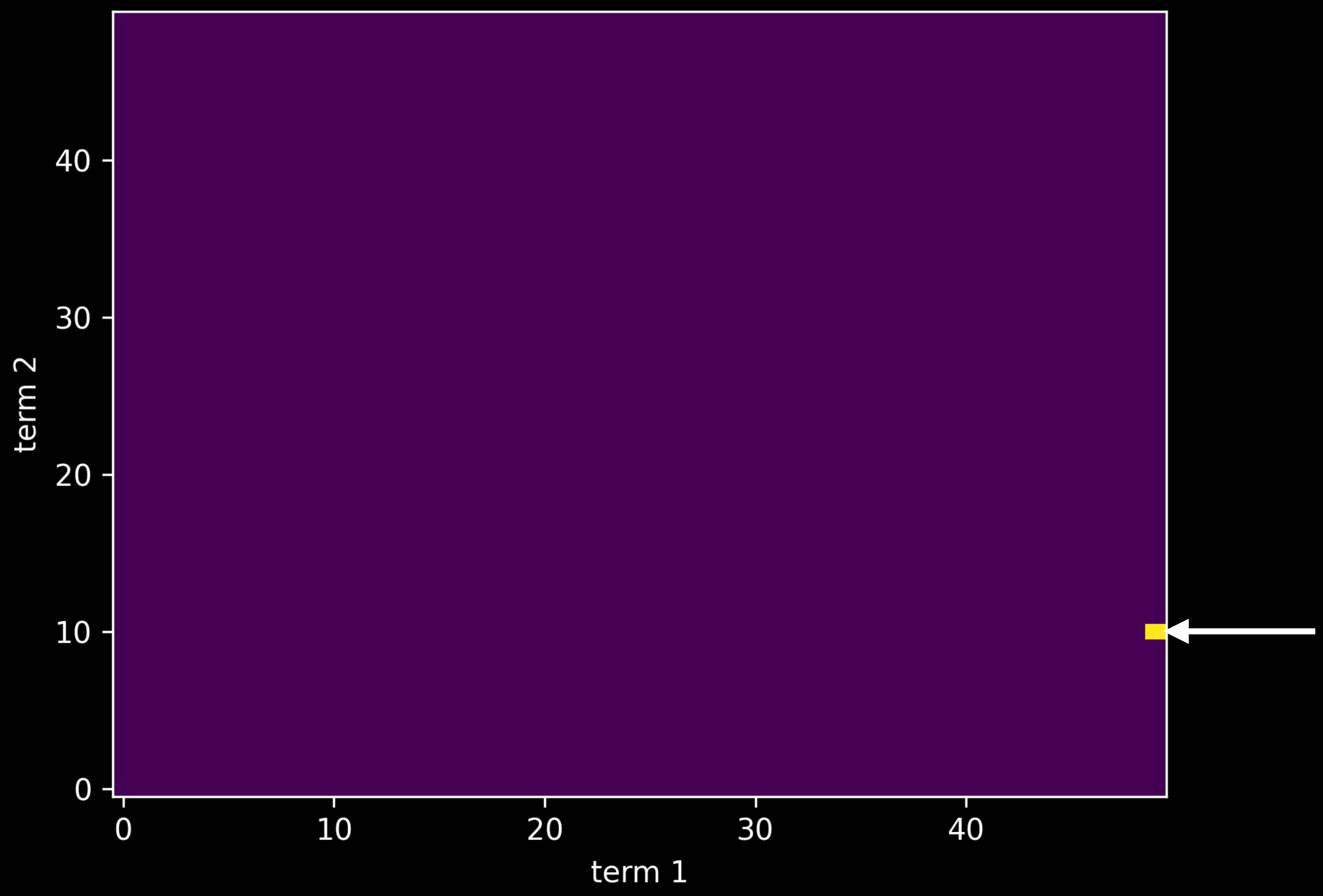
49 + 10 = 59



$$p(5 \mid 4,9,1,0) \quad \cdot \quad \underline{p(9 \mid 4,9,1,0,5)} = p(5,9 \mid 4,9,1,0)$$

Perfect Model

Training Data 100% – Accuracy 100%

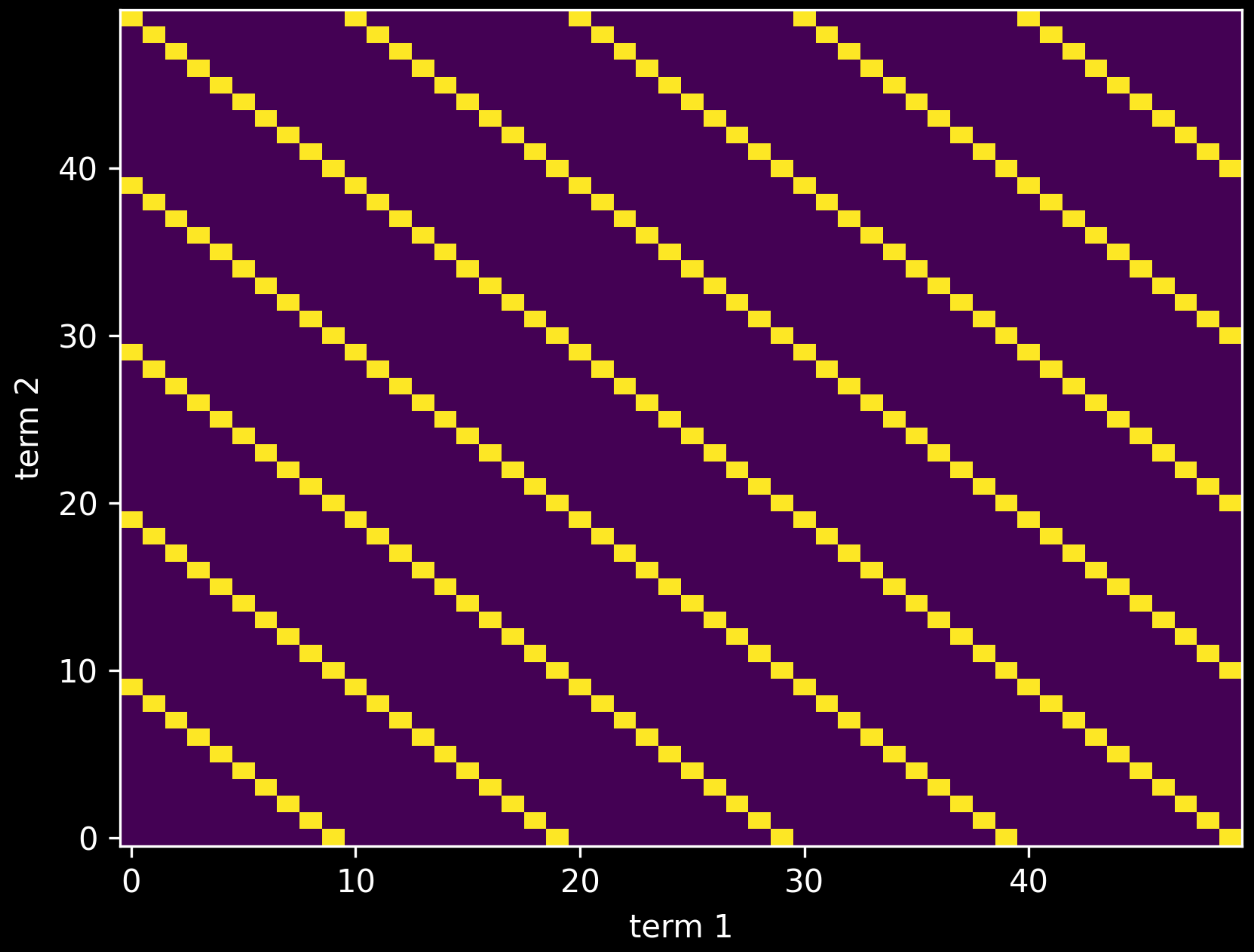


$p(9 \mid 4,9,1,0,5) = 1$

Perfect Model

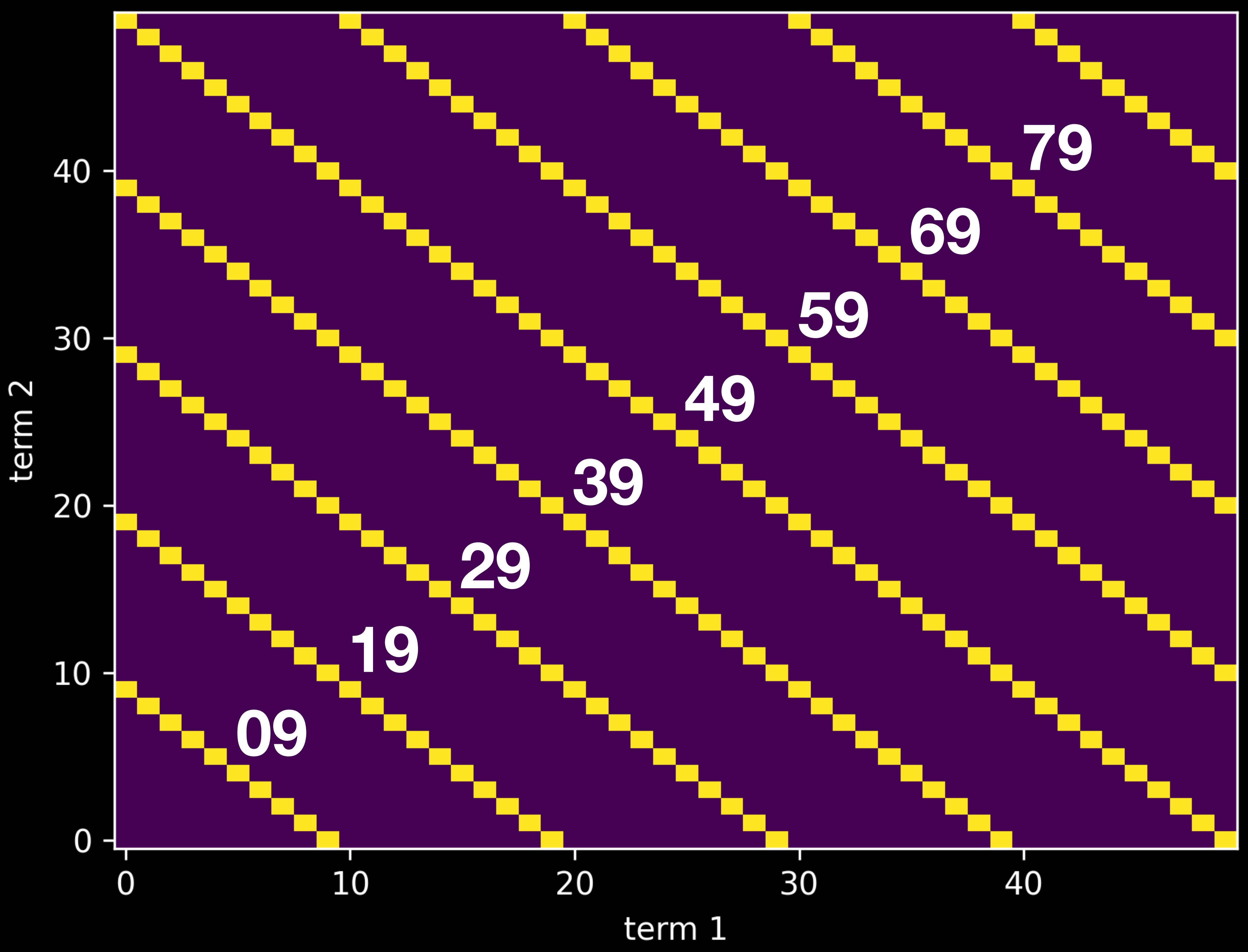
Training Data 100% – Accuracy 100%

p(digit2=9 | term1,term2,digit1)



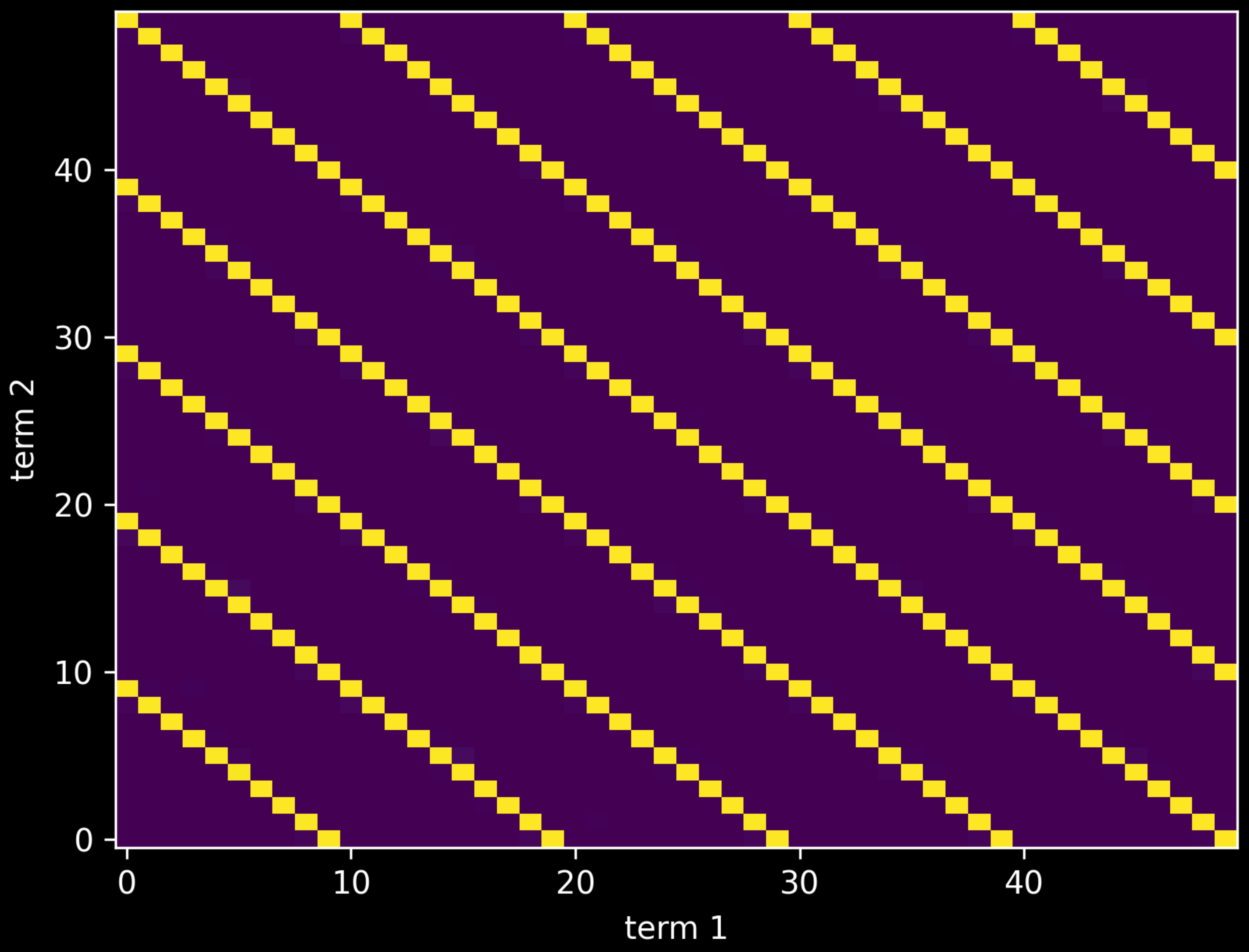
Perfect Model

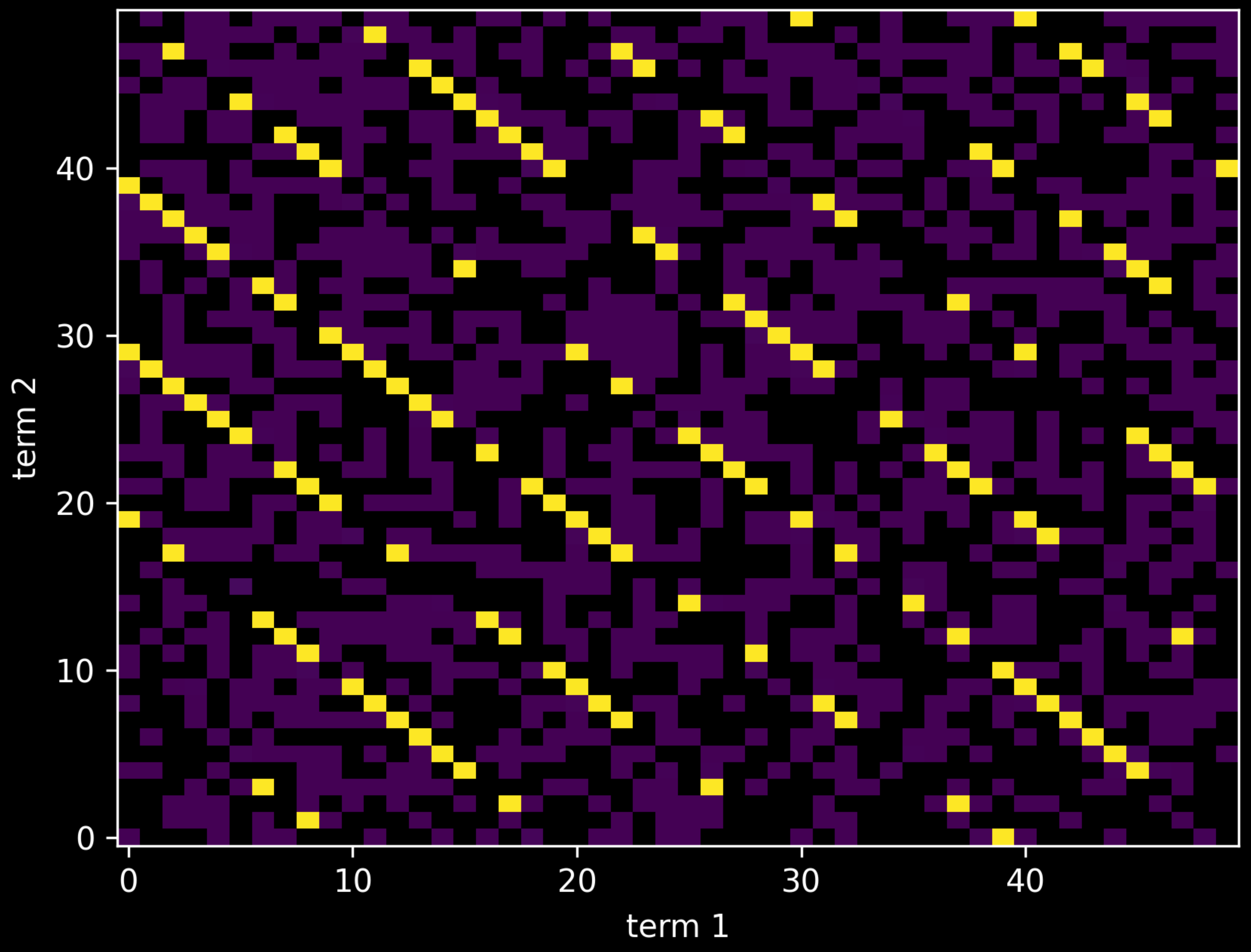
Training Data 100% – Accuracy 100%



Autoregressive Decoder

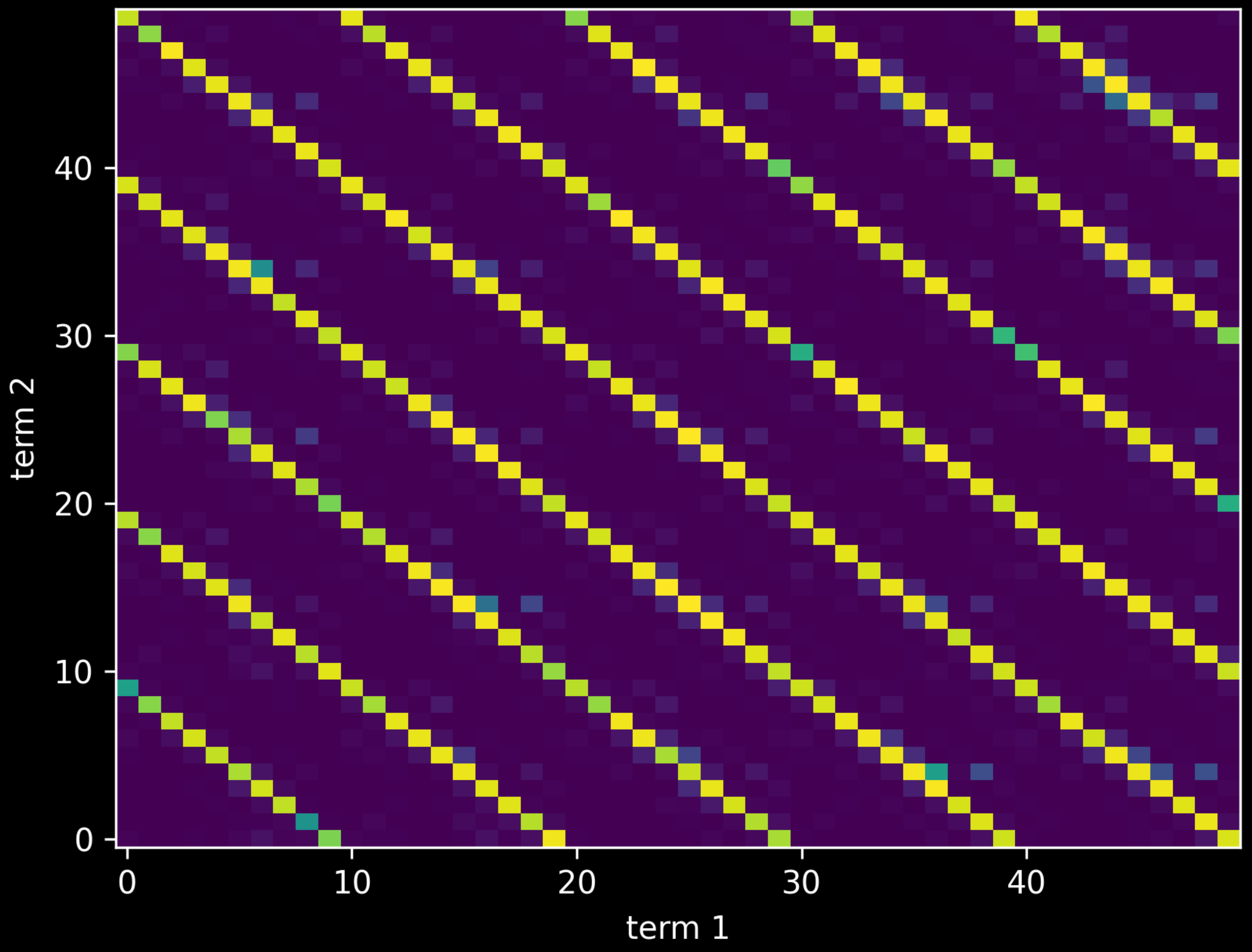
Training Data 50% – Accuracy 100%

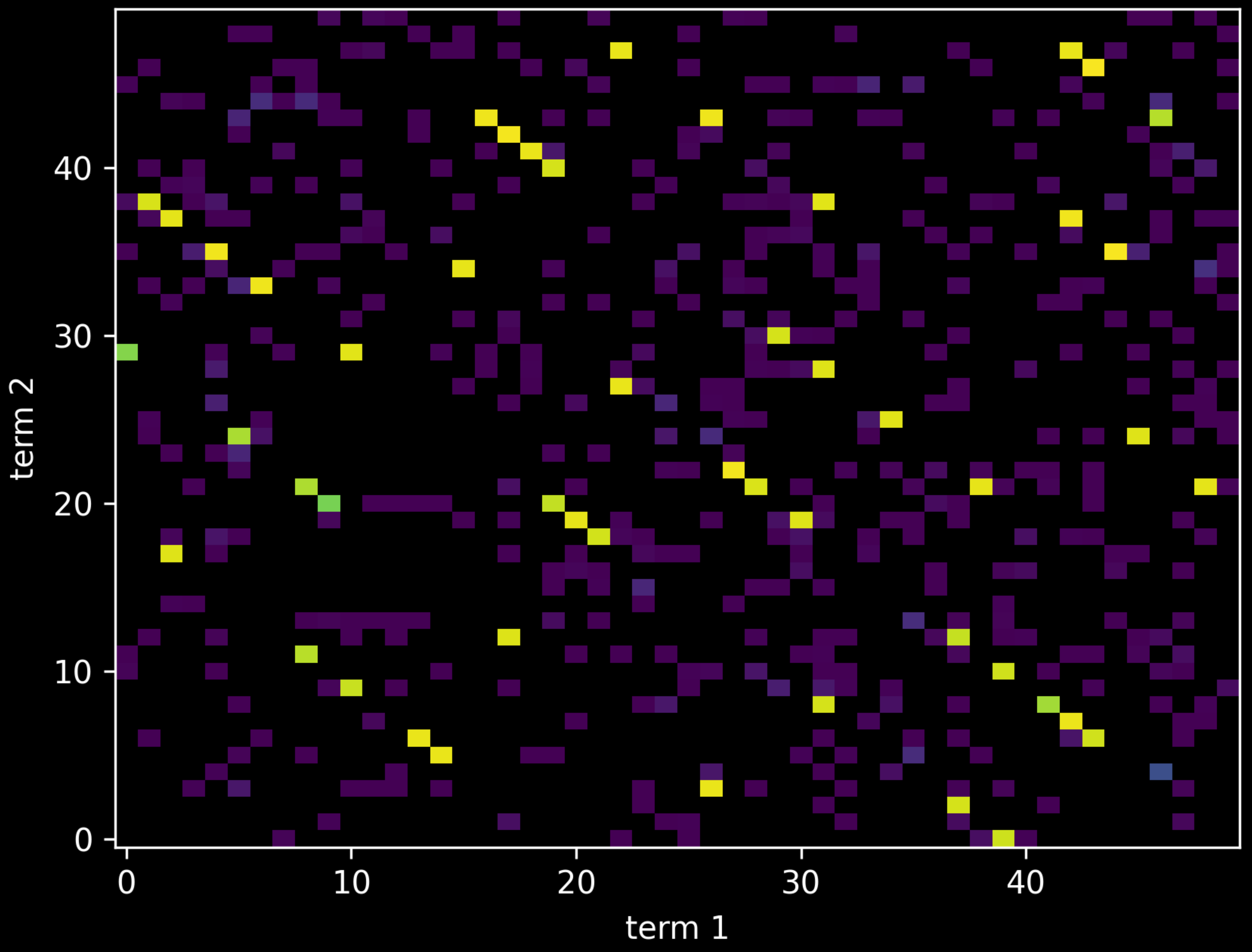




Autoregressive Decoder

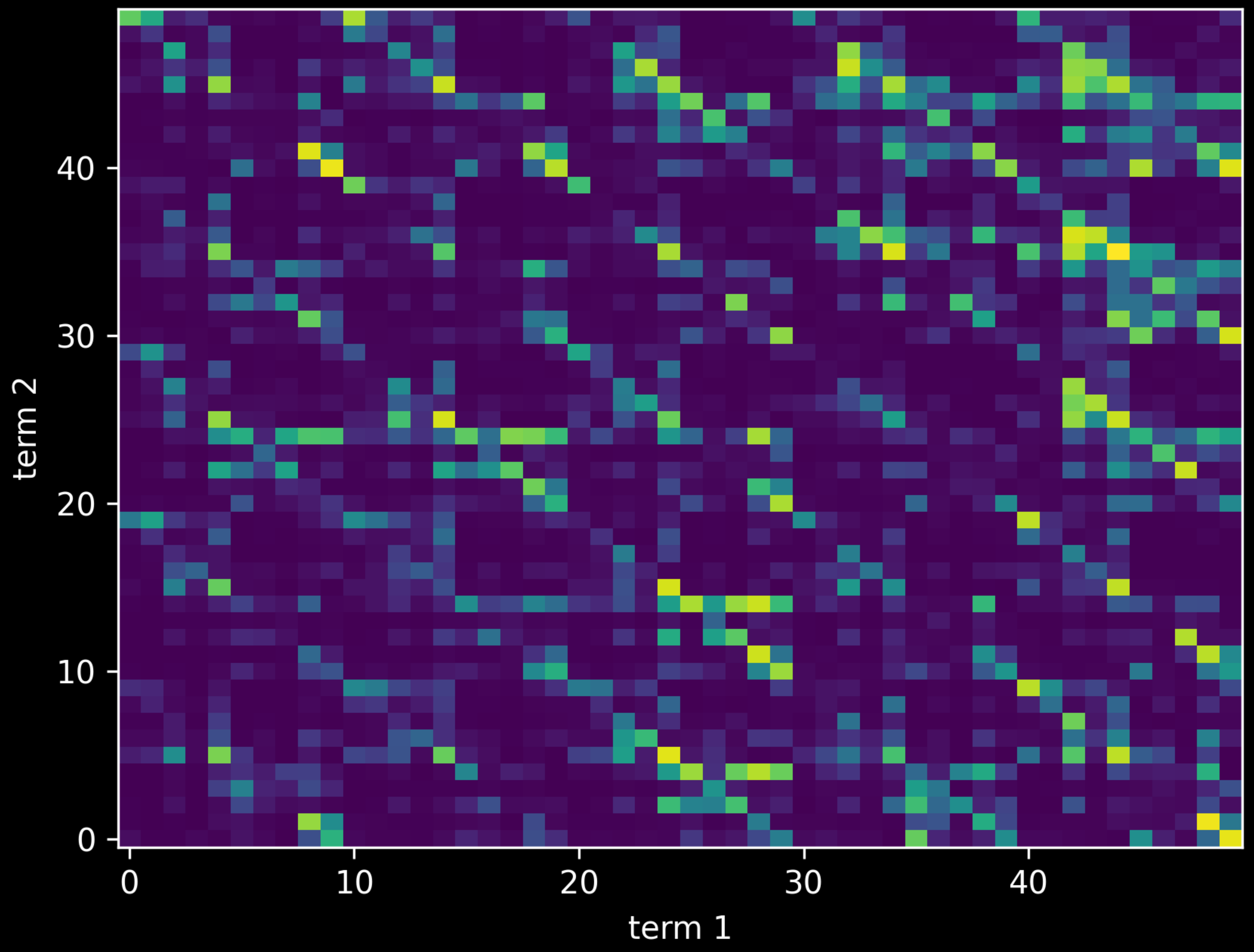
Training Data 20% – Accuracy 82%

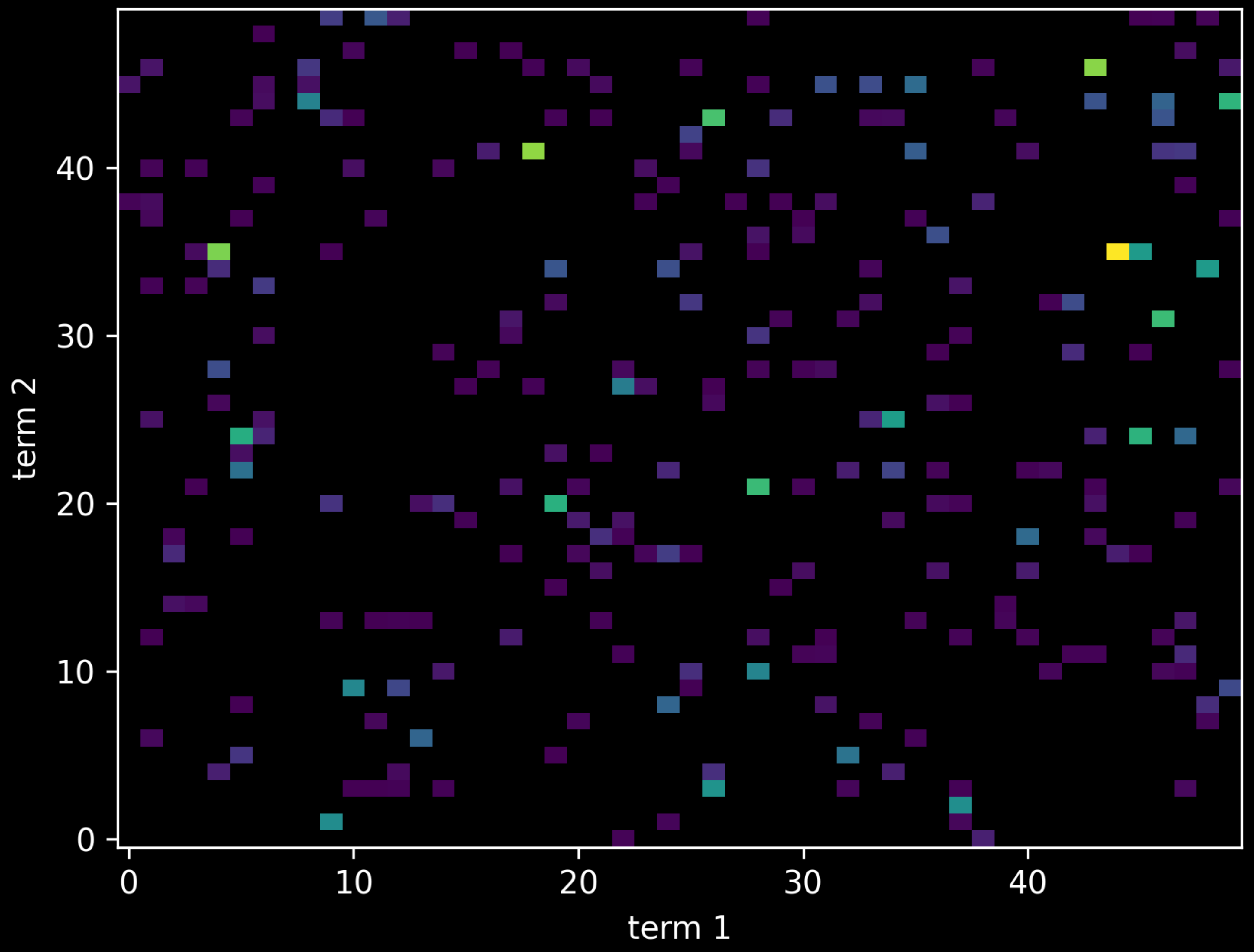




Autoregressive Decoder

Training Data 10% – Accuracy 36%





Recent Achievements of Language Models

- 2022-02: AlphaCode achieved on average a ranking of top 54.3% in programming competitions.

- 2022-02: GPT-f solves 2 problems from the Internationale Mathematik-Olympiade.

- 2022-03: Tesla FSD Beta 10.11 – Upgraded modeling of lane geometry from dense rasters (“bag of points”) to an autoregressive decoder that directly predicts and connects “vector space” lanes point by point using a transformer neural network. This enables us to predict crossing lanes, allows computationally cheaper and less error-prone post-processing, and paves the way for predicting many other signals and their relationships jointly and end-to-end.

Chain Rule of Probability

$$
\begin{aligned}
p(s_1, s_2, s_3) &= p(s_3 \mid s_1, s_2) \cdot p(s_1, s_2) \\
&= p(s_3 \mid s_1, s_2) \cdot p(s_2 \mid s_1) \cdot p(s_1)
\end{aligned}
$$

Predicting a sequence

$$p(s_1, s_2, s_3) = p(s_3 \mid s_1, s_2) \cdot p(s_2 \mid s_1) \cdot p(s_1)$$

$$p(s_2, s_3 | s_1) \cdot p(s_1) = p(s_3 \mid s_1, s_2) \cdot p(s_2 \mid s_1) \cdot p(s_1)$$

$$p(s_2, s_3 | s_1) = p(s_3 \mid s_1, s_2) \cdot p(s_2 \mid s_1)$$