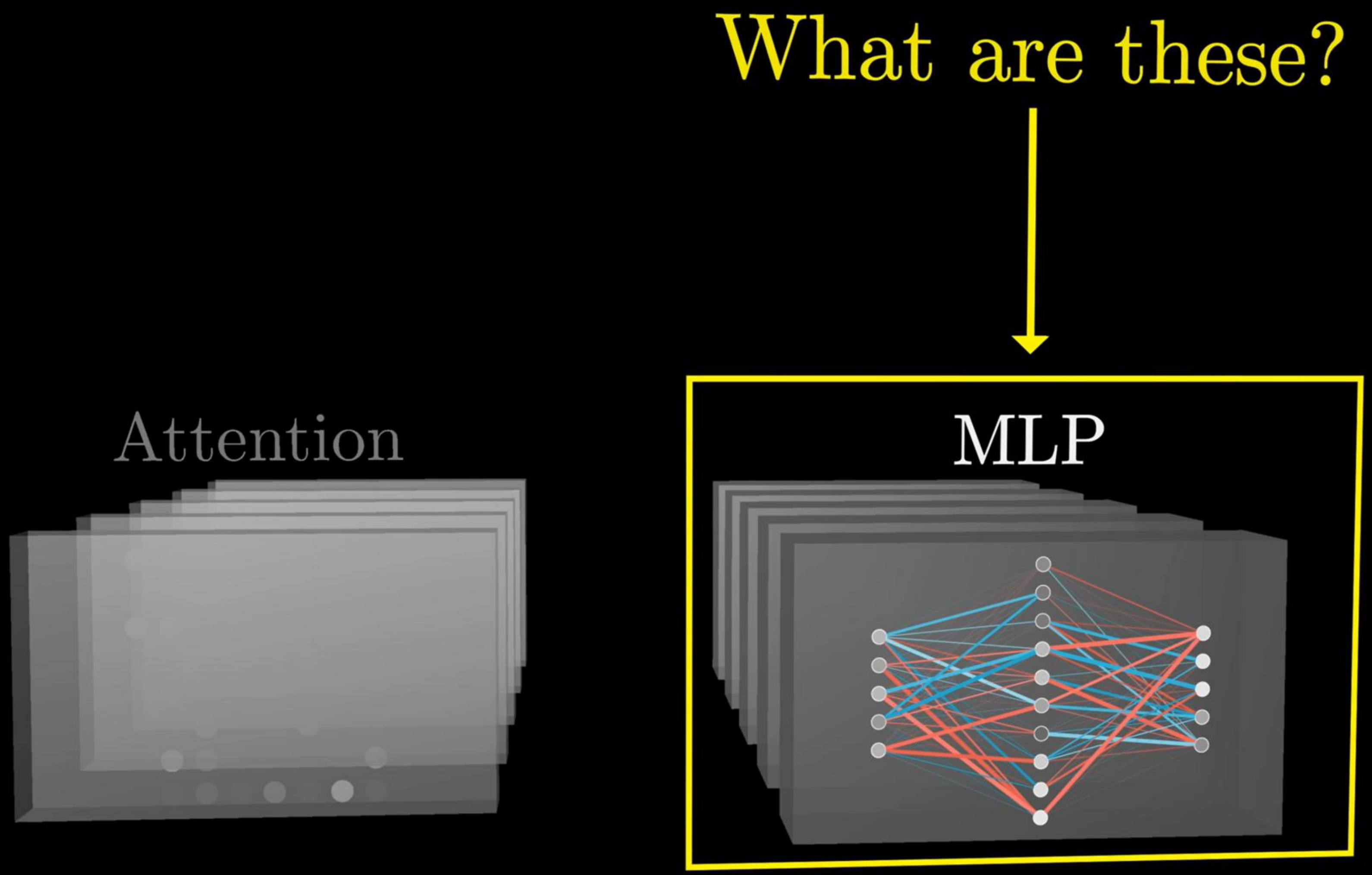
转换器中的多层感知机

Multilayer Perceptron in Transformer

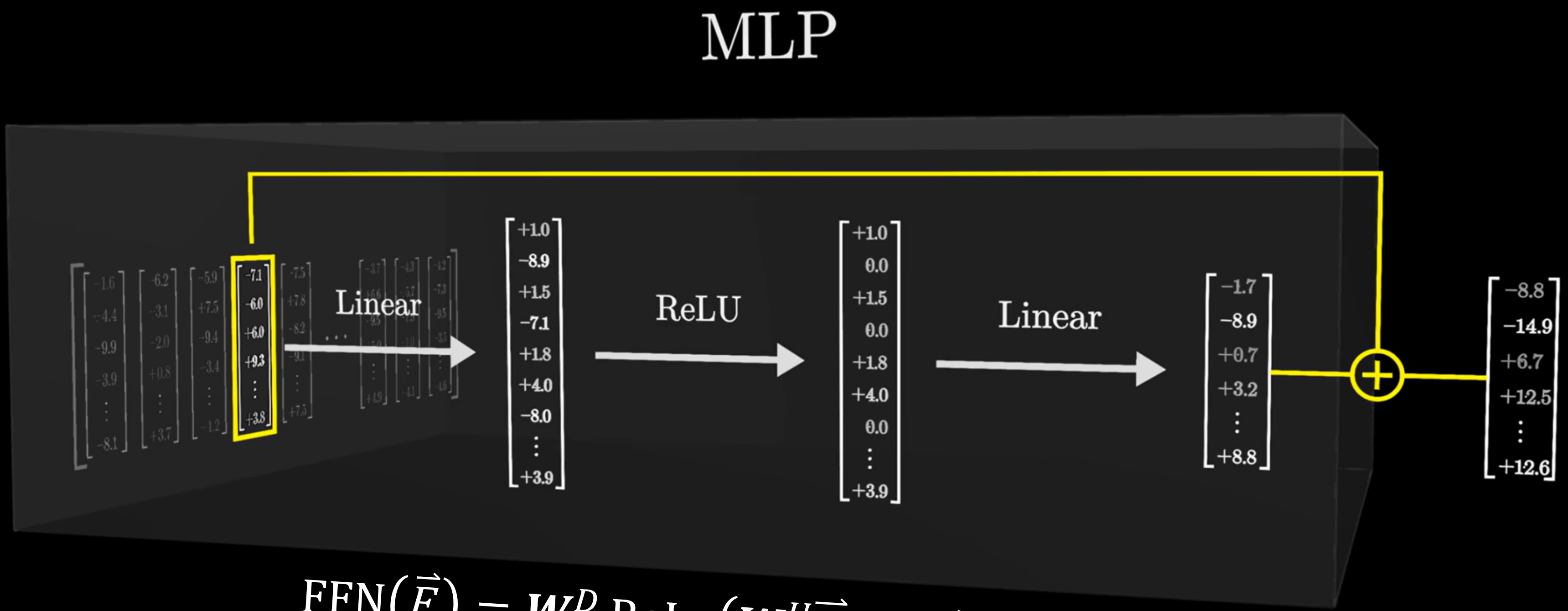
1 GPT-3中的MLP

- 主要存储世界知识（facts）



1 GPT-3中的MLP

- 2层全连接神经网络

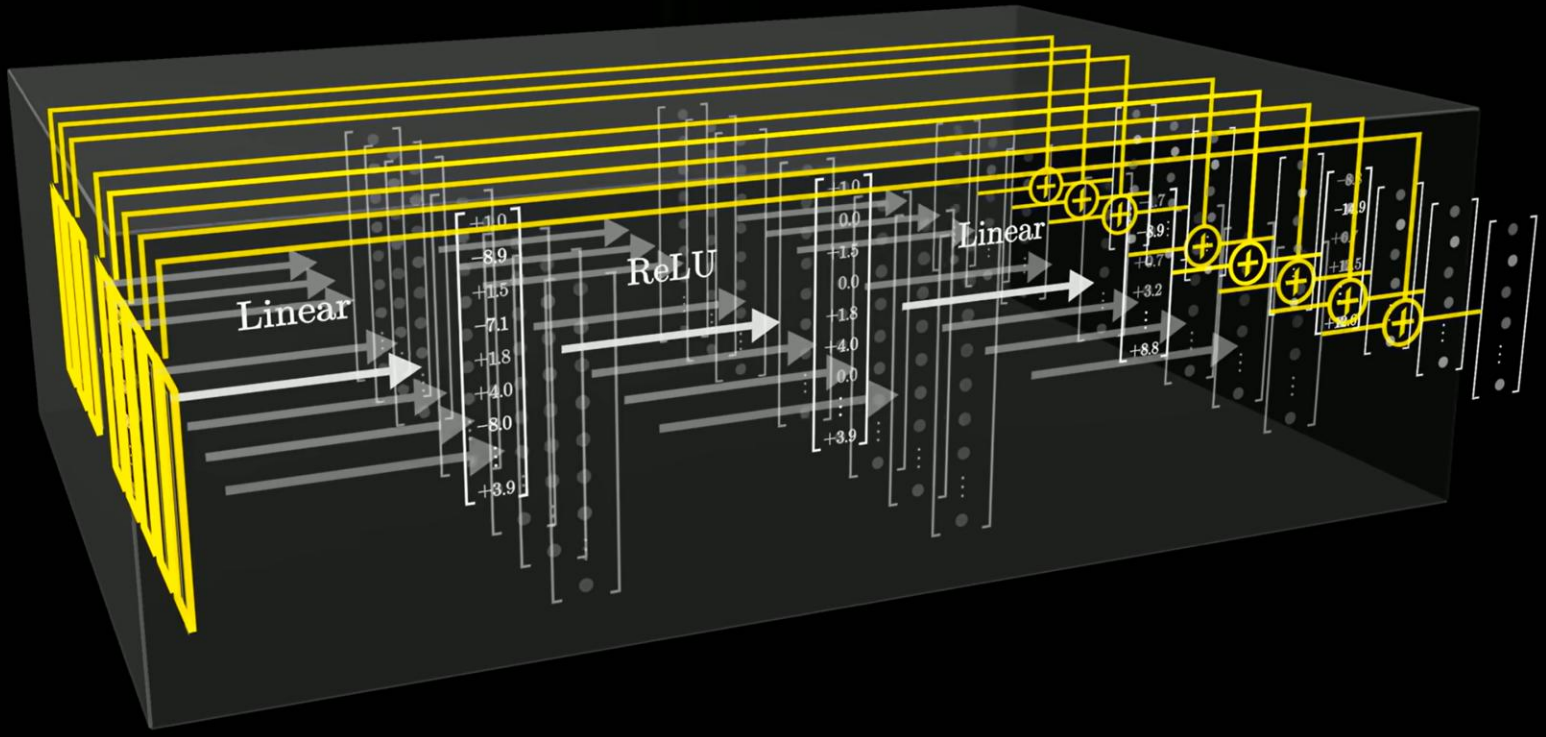


$$\mathrm{FFN}(\vec{E}) = \boldsymbol{W}^{D}\,\mathrm{ReLu}(\boldsymbol{W}^{U}\vec{\boldsymbol{E}} + \boldsymbol{b}_1) + \boldsymbol{b}_2$$

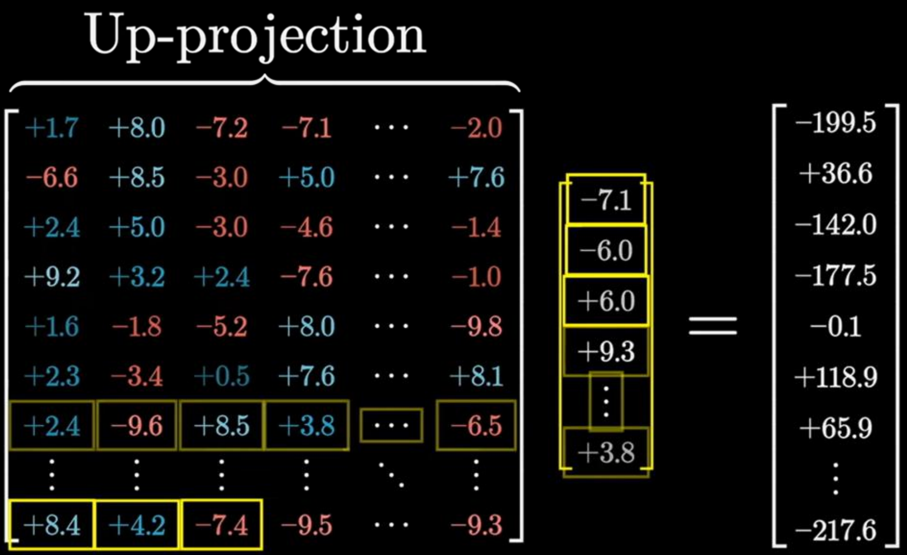
多层感知机（MLP）是最典型的全连接前向神经网络；“前向神经网络（FFN）”强调信息单向传播，“全连接神经网络”强调层间连接方式，三者在多数教学语境下常指同一类模型，但严格来说 FFN 的概念更广。

1 GPT-3中的MLP

- 对每个经注意力机制处理后的嵌入向量同时处理



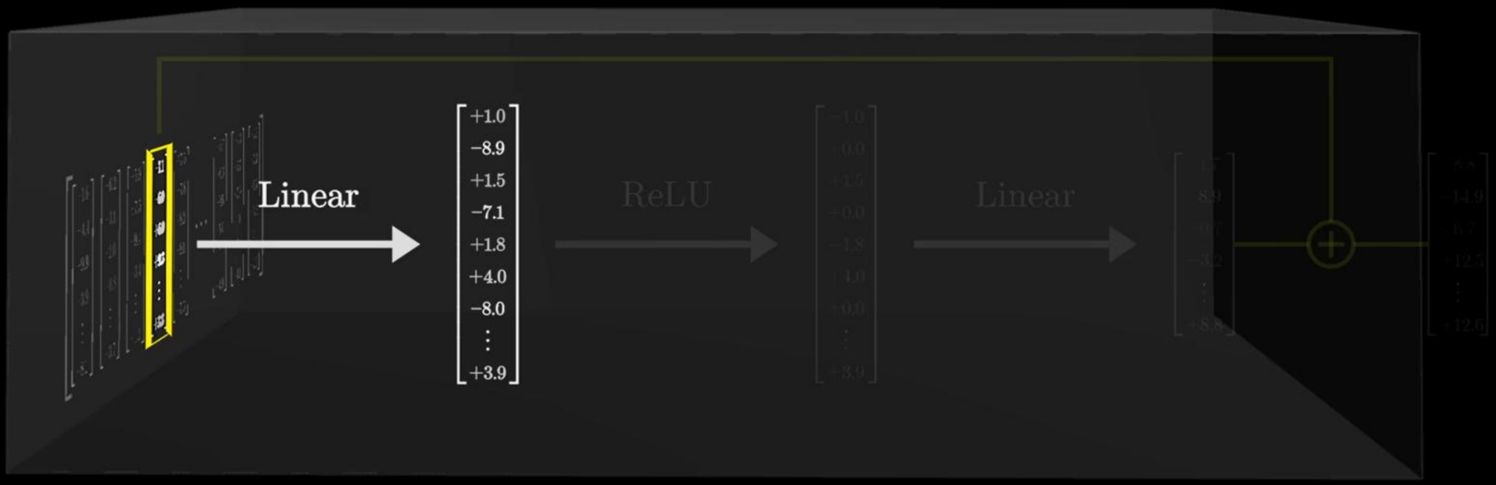
2 第一层



Up-projection

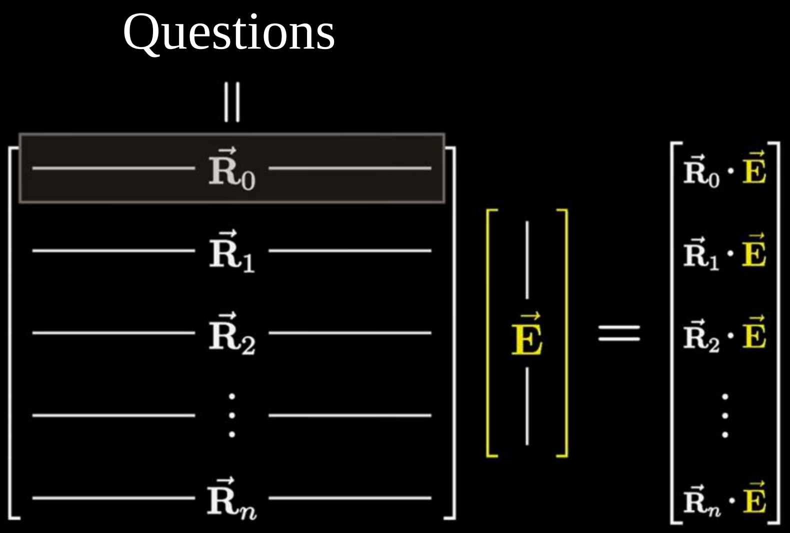
$$\underbrace{\begin{bmatrix} +1.7 & +8.0 & -7.2 & -7.1 & \cdots & -2.0 \\ -6.6 & +8.5 & -3.0 & +5.0 & \cdots & +7.6 \\ +2.4 & +5.0 & -3.0 & -4.6 & \cdots & -1.4 \\ +9.2 & +3.2 & +2.4 & -7.6 & \cdots & -1.0 \\ +1.6 & -1.8 & -5.2 & +8.0 & \cdots & -9.8 \\ +2.3 & -3.4 & +0.5 & +7.6 & \cdots & +8.1 \\ \boxed{+2.4} & \boxed{-9.6} & \boxed{+8.5} & \boxed{+3.8} & \boxed{\cdots} & \boxed{-6.5} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \boxed{+8.4} & \boxed{+4.2} & \boxed{-7.4} & -9.5 & \cdots & -9.3 \end{bmatrix}}_{} \begin{bmatrix} \boxed{-7.1} \\ \boxed{-6.0} \\ \boxed{+6.0} \\ \boxed{+9.3} \\ \boxed{\vdots} \\ \boxed{+3.8} \end{bmatrix} = \begin{bmatrix} -199.5 \\ +36.6 \\ -142.0 \\ -177.5 \\ -0.1 \\ +118.9 \\ +65.9 \\ \vdots \\ -217.6 \end{bmatrix}$$

MLP

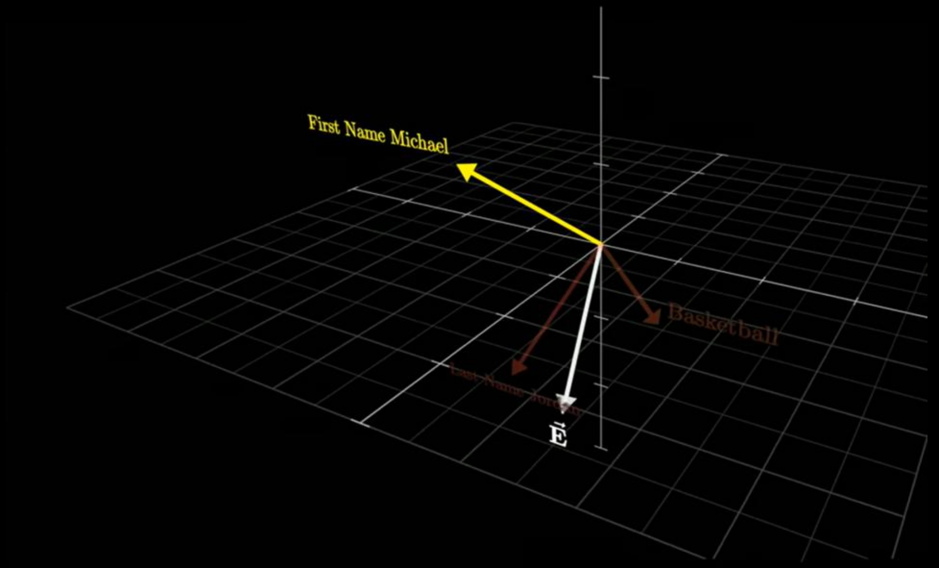


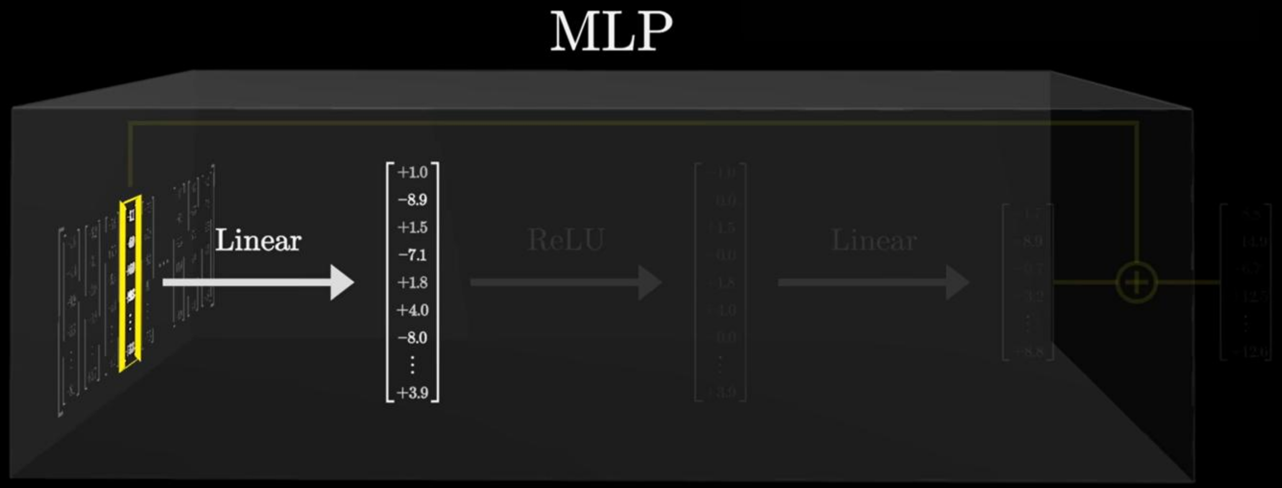
2 第一层

- 类比：“你问我猜”



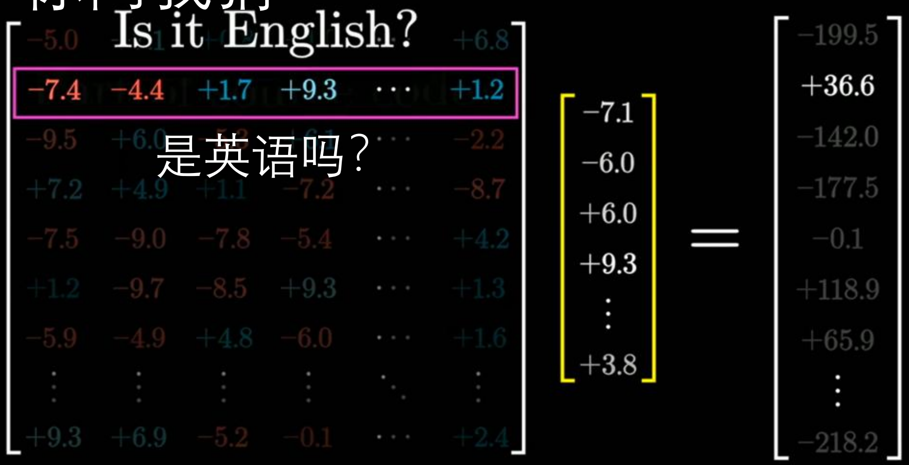
$$\begin{bmatrix} \vec{R}_0 \\ \vec{R}_1 \\ \vec{R}_2 \\ \vdots \\ \vec{R}_n \end{bmatrix} \begin{bmatrix} \vec{E} \end{bmatrix} = \begin{bmatrix} \vec{R}_0 \cdot \vec{E} \\ \vec{R}_1 \cdot \vec{E} \\ \vec{R}_2 \cdot \vec{E} \\ \vdots \\ \vec{R}_n \cdot \vec{E} \end{bmatrix}$$



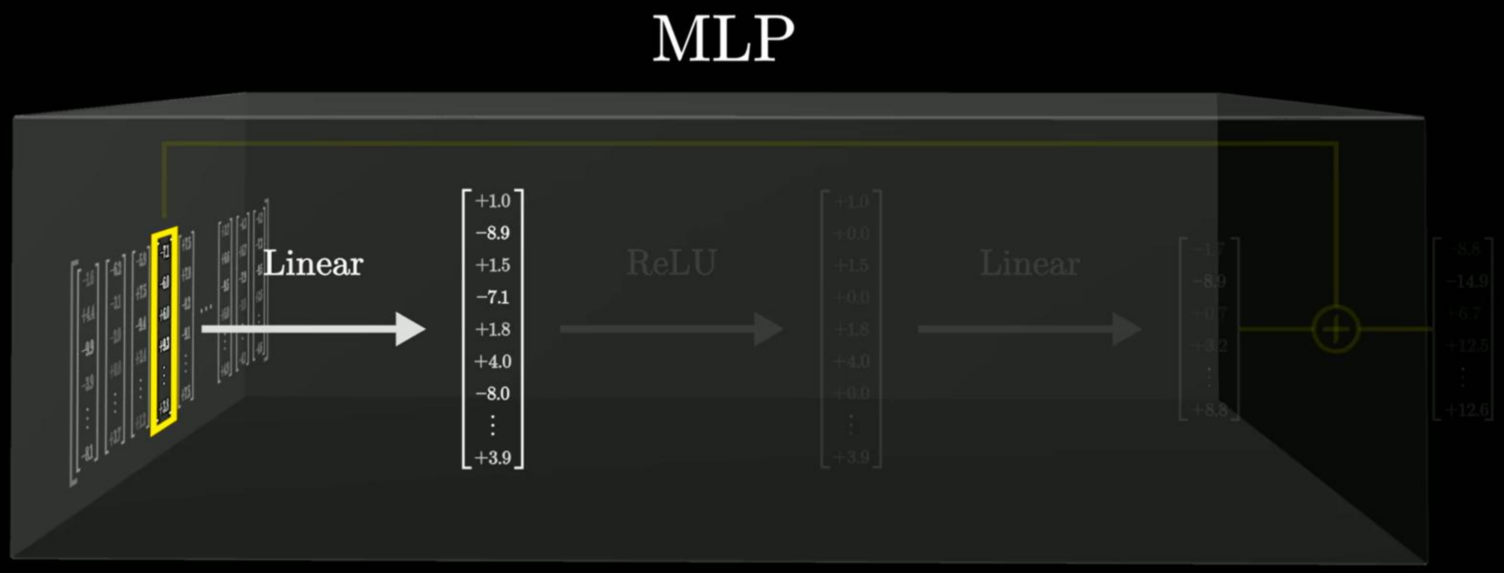


2 第一层

- 类比："你问我猜"

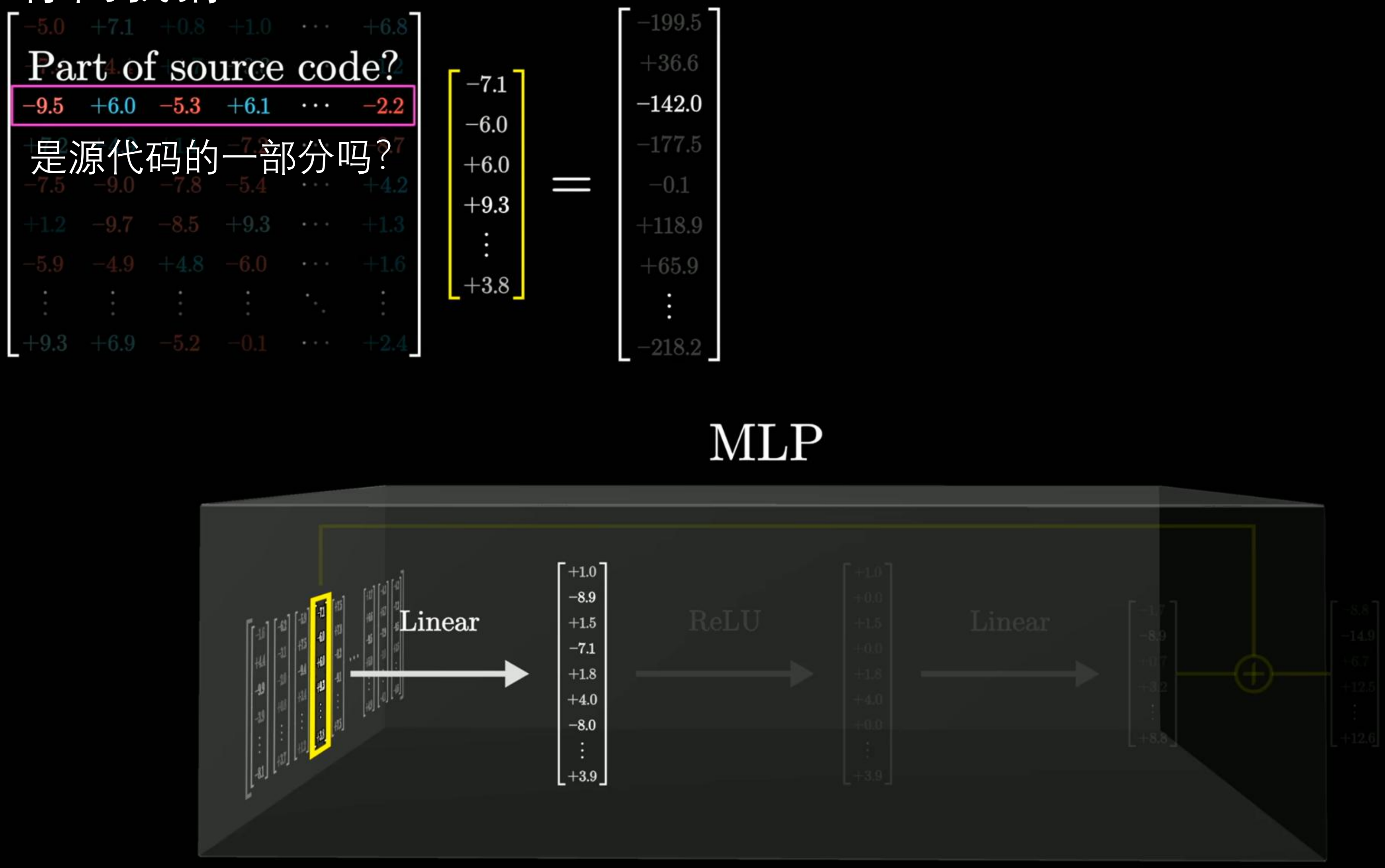


是英语吗？



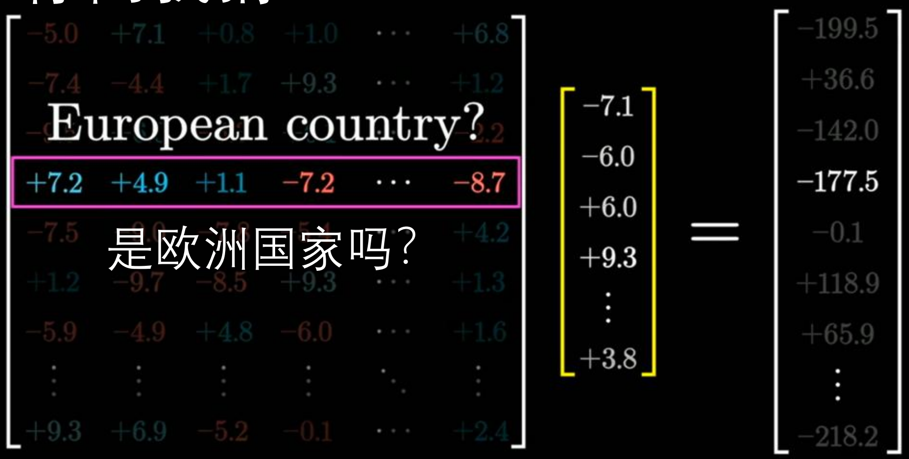
2 第一层

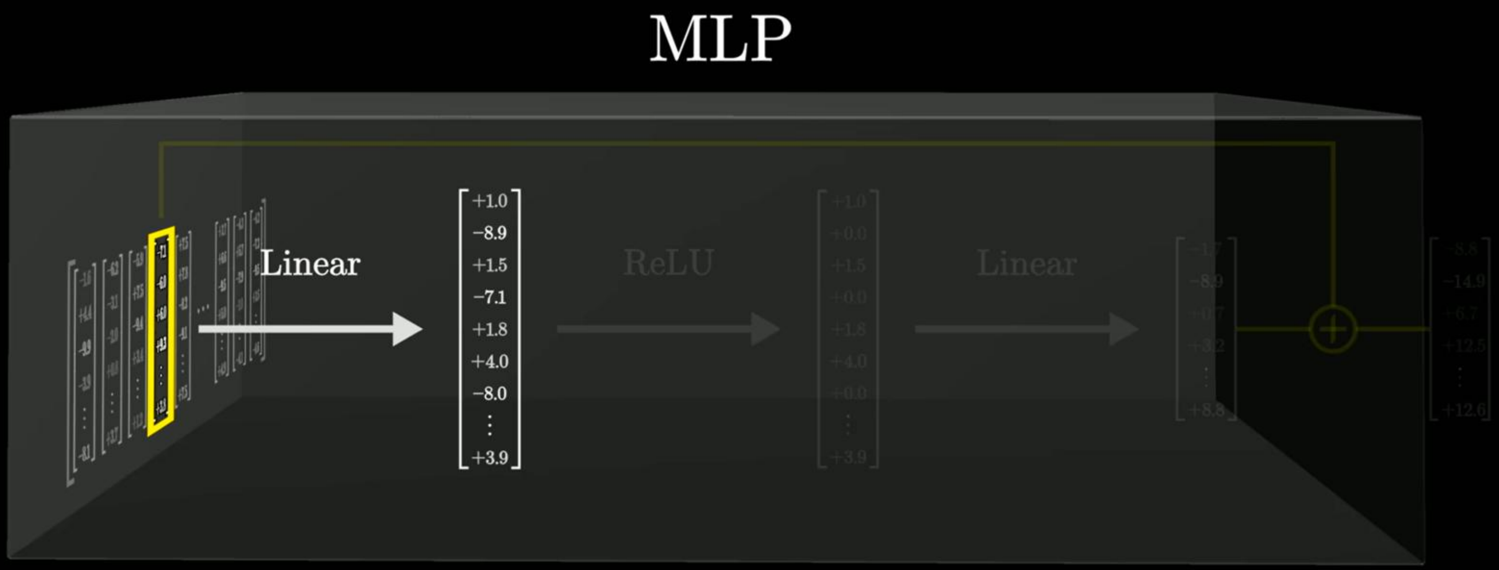
- 类比：“你问我猜”


Part of source code?
−9.5 +6.0 −5.3 +6.1 ⋯ −2.2
是源代码的一部分吗？
−7.1
−6.0
+6.0
+9.3
+3.8
=
−199.5
+36.6
−142.0
−177.5
−0.1
+118.9
+65.9
−218.2
MLP
Linear
+1.0
−8.9
+1.5
−7.1
+1.8
+4.0
−8.0
+3.9
ReLU
Linear

2 第一层

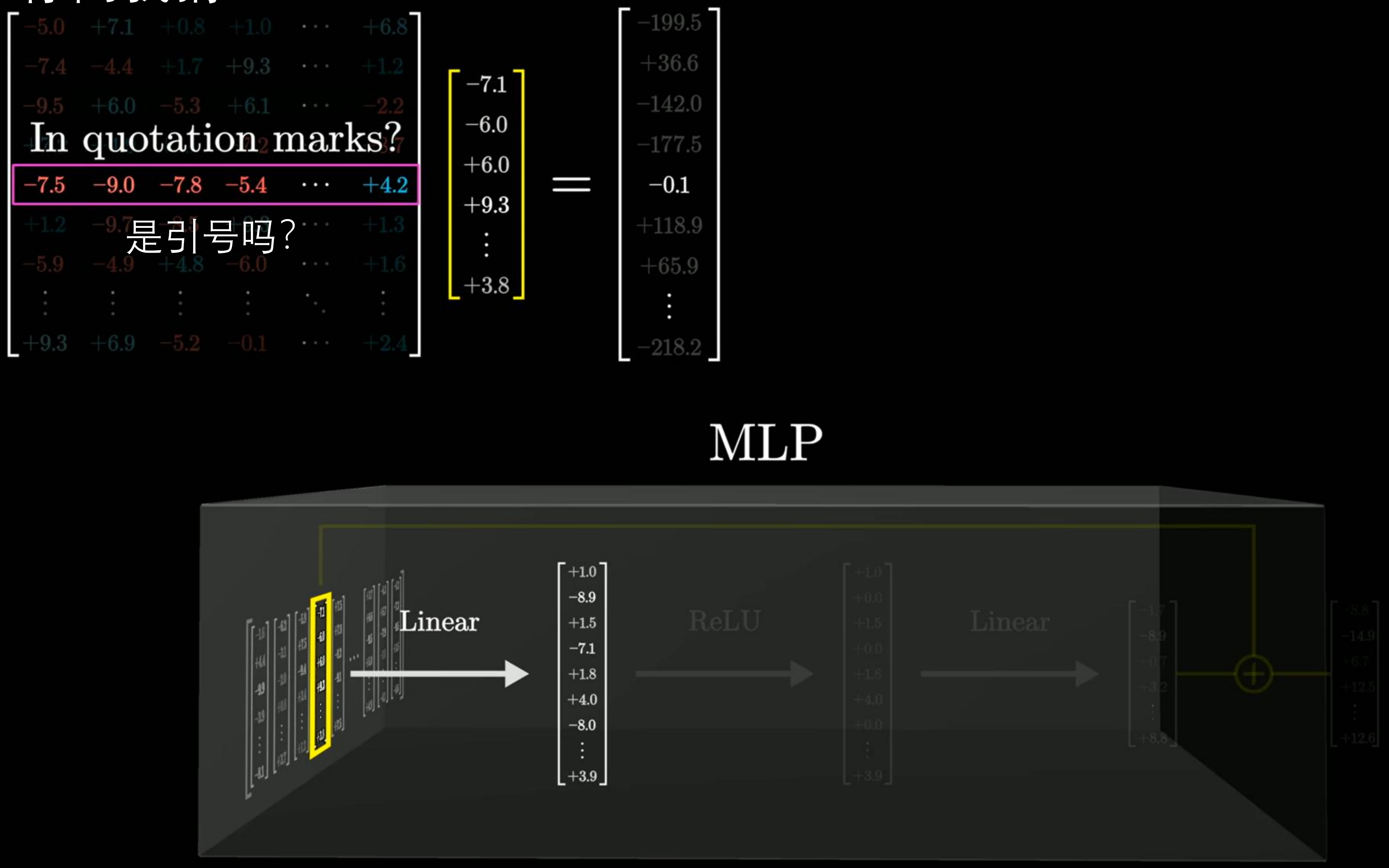
- 类比：“你问我猜”


European country?
+7.2 +4.9 +1.1 −7.2 ⋯ −8.7
是欧洲国家吗？
−7.1
−6.0
+6.0
+9.3
+3.8
=
−199.5
+36.6
−142.0
−177.5
−0.1
+118.9
+65.9
−218.2


MLP
Linear
+1.0
−8.9
+1.5
−7.1
+1.8
+4.0
−8.0
+3.9
ReLU
Linear

2 第一层

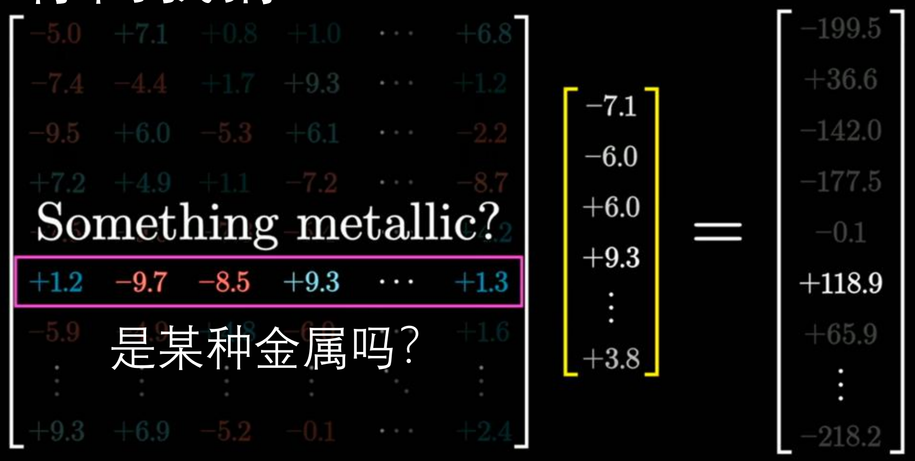
- 类比：“你问我猜”

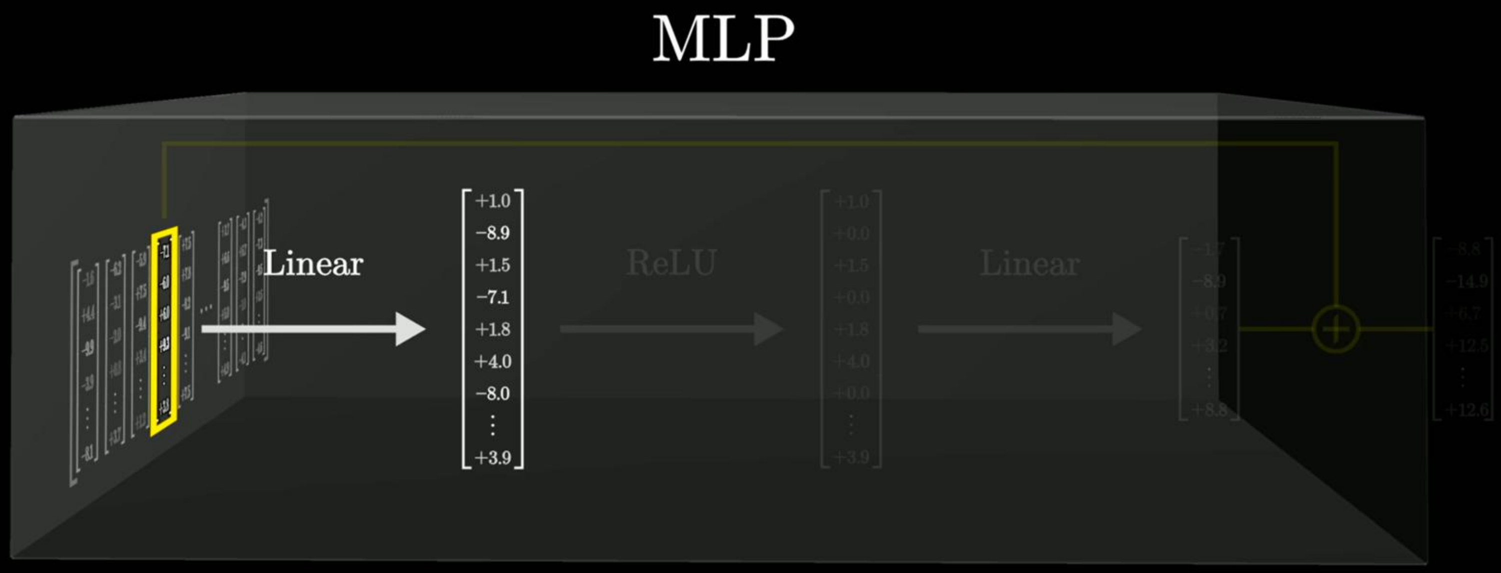


−5.0 +7.1 +0.8 +1.0 ⋯ +6.8
−7.4 −4.4 +1.7 +9.3 ⋯ +1.2
−9.5 +6.0 −5.3 +6.1 ⋯ −2.2
In quotation marks?
−7.5 −9.0 −7.8 −5.4 ⋯ +4.2
是引号吗？
−5.9 −4.9 +4.8 −6.0 ⋯ +1.6
+9.3 +6.9 −5.2 −0.1 ⋯ +2.4
−7.1 −6.0 +6.0 +9.3 ⋮ +3.8
=
−199.5 +36.6 −142.0 −177.5 −0.1 +118.9 +65.9 ⋮ −218.2
MLP
Linear
+1.0 −8.9 +1.5 −7.1 +1.8 +4.0 −8.0 ⋮ +3.9
ReLU
Linear

2 第一层

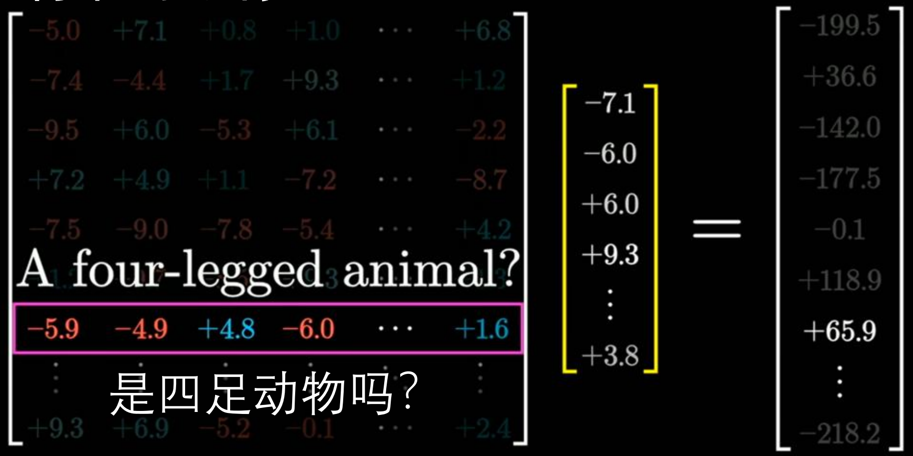
- 类比：“你问我猜”


-5.0 +7.1 +0.8 +1.0 ··· +6.8
-7.4 -4.4 +1.7 +9.3 ··· +1.2
-9.5 +6.0 -5.3 +6.1 ··· -2.2
+7.2 +4.9 +1.1 -7.2 ··· -8.7
Something metallic?
+1.2 -9.7 -8.5 +9.3 ··· +1.3
-5.9 ··· +1.6
是某种金属吗？
+9.3 +6.9 -5.2 -0.1 ··· +2.4
-7.1
-6.0
+6.0
+9.3
+3.8
=
-199.5
+36.6
-142.0
-177.5
-0.1
+118.9
+65.9
-218.2

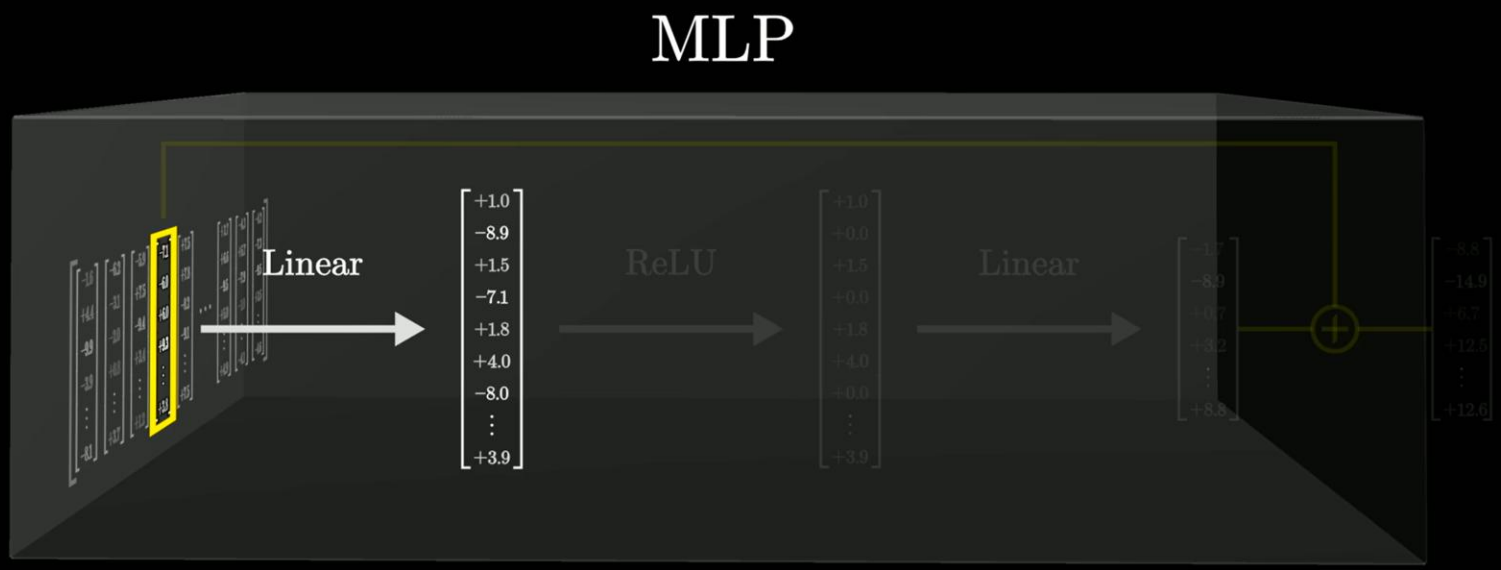

MLP
Linear
+1.0
-8.9
+1.5
-7.1
+1.8
+4.0
-8.0
+3.9
ReLU
Linear

2 第一层

- 类比：“你问我猜”



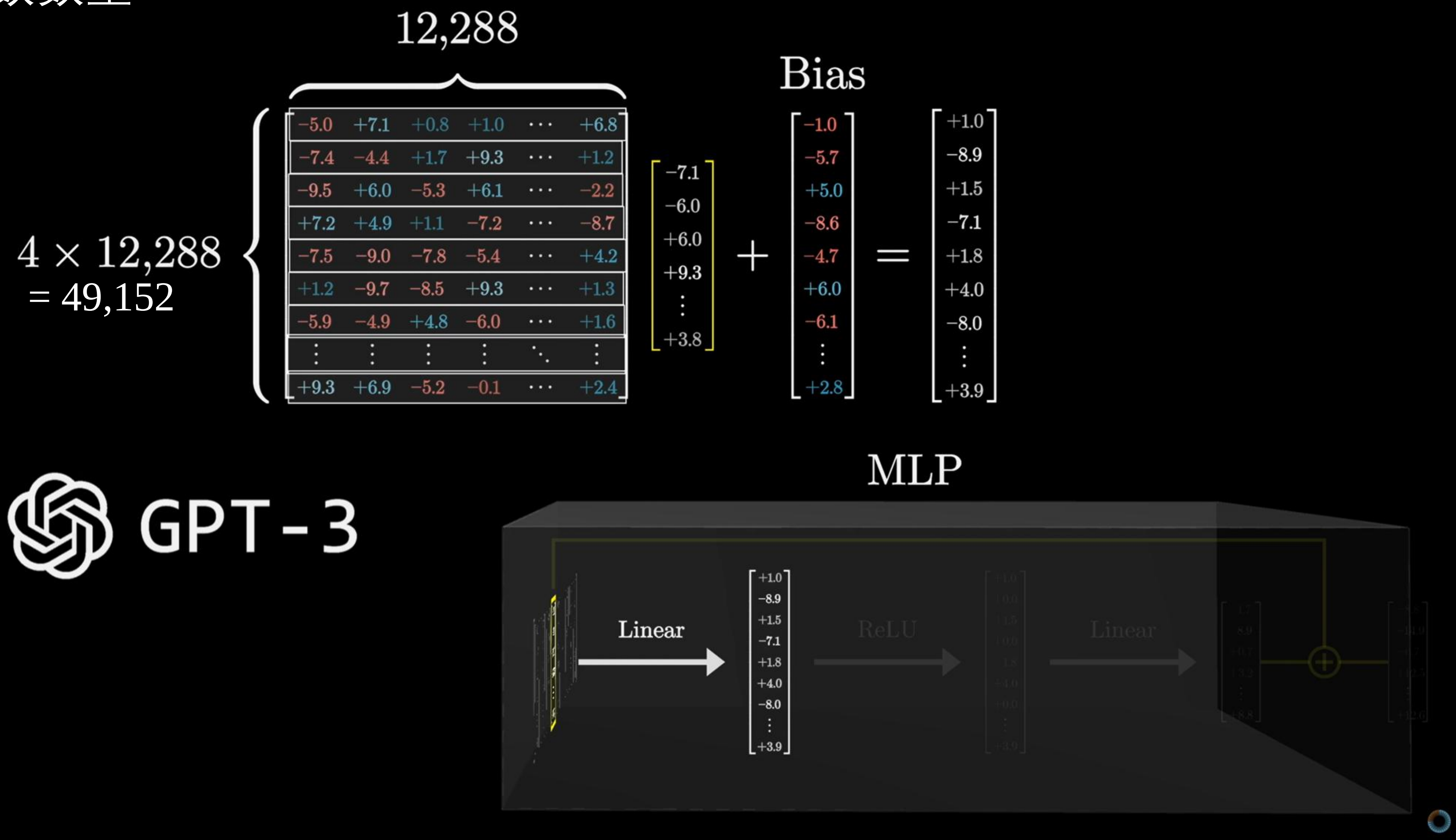
A four-legged animal?
−5.9 −4.9 +4.8 −6.0 ⋯ +1.6
是四足动物吗？
−7.1 −6.0 +6.0 +9.3 ⋮ +3.8
=
−199.5 +36.6 −142.0 −177.5 −0.1 +118.9 +65.9 ⋮ −218.2



MLP
Linear
+1.0 −8.9 +1.5 −7.1 +1.8 +4.0 −8.0 ⋮ +3.9
ReLU
Linear

2 第一层

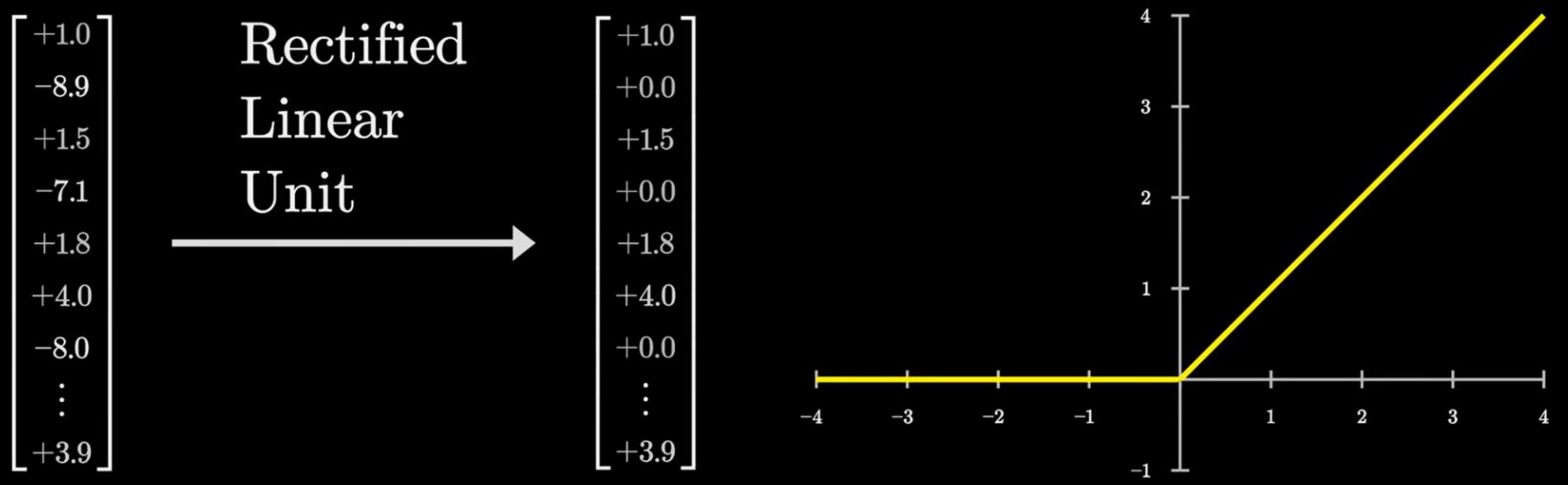
- 维度与参数数量

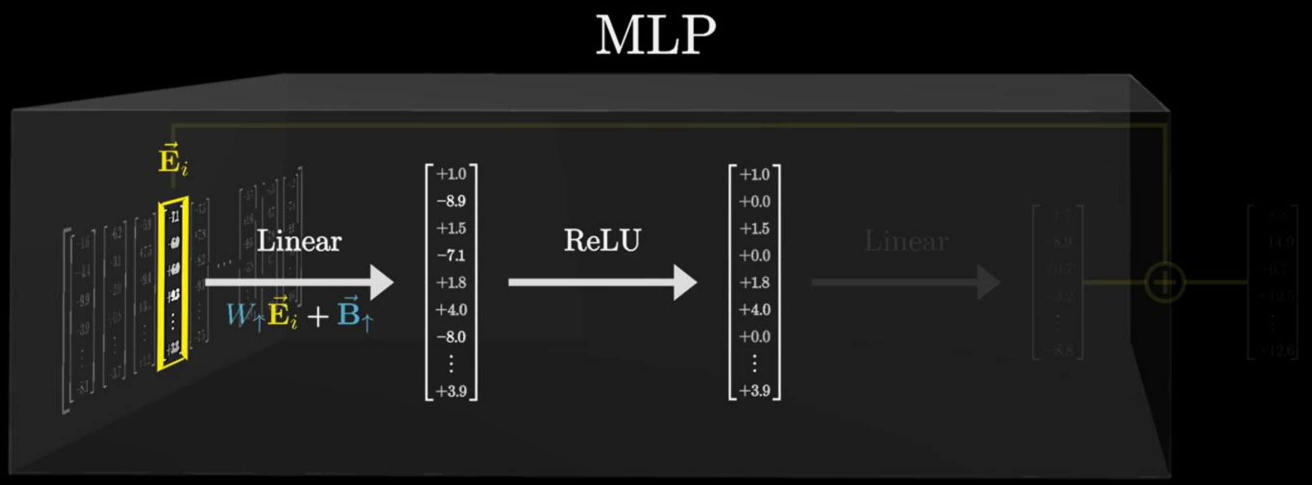


$$4 \times 12{,}288 = 49{,}152 \quad \overbrace{\begin{bmatrix} -5.0 & +7.1 & +0.8 & +1.0 & \cdots & +6.8 \\ -7.4 & -4.4 & +1.7 & +9.3 & \cdots & +1.2 \\ -9.5 & +6.0 & -5.3 & +6.1 & \cdots & -2.2 \\ +7.2 & +4.9 & +1.1 & -7.2 & \cdots & -8.7 \\ -7.5 & -9.0 & -7.8 & -5.4 & \cdots & +4.2 \\ +1.2 & -9.7 & -8.5 & +9.3 & \cdots & +1.3 \\ -5.9 & -4.9 & +4.8 & -6.0 & \cdots & +1.6 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ +9.3 & +6.9 & -5.2 & -0.1 & \cdots & +2.4 \end{bmatrix}}^{12{,}288} \begin{bmatrix} -7.1 \\ -6.0 \\ +6.0 \\ +9.3 \\ \vdots \\ +3.8 \end{bmatrix} + \overset{\text{Bias}}{\begin{bmatrix} -1.0 \\ -5.7 \\ +5.0 \\ -8.6 \\ -4.7 \\ +6.0 \\ -6.1 \\ \vdots \\ +2.8 \end{bmatrix}} = \begin{bmatrix} +1.0 \\ -8.9 \\ +1.5 \\ -7.1 \\ +1.8 \\ +4.0 \\ -8.0 \\ \vdots \\ +3.9 \end{bmatrix}$$

2 第一层

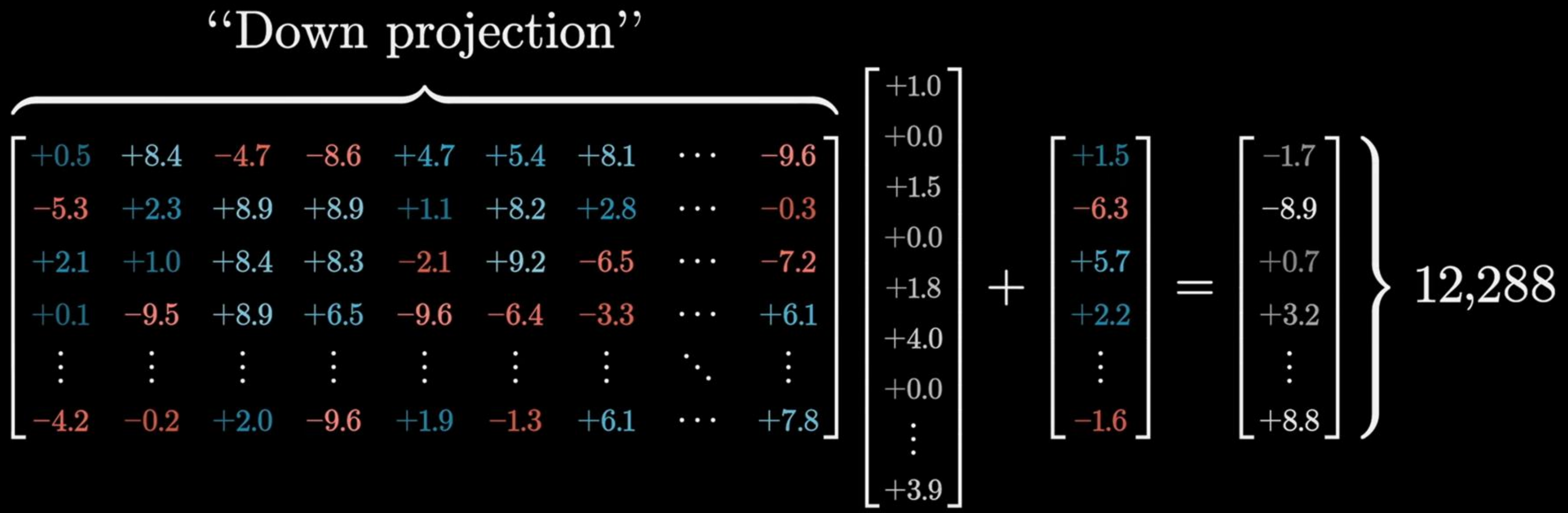
- 激活函数：ReLU



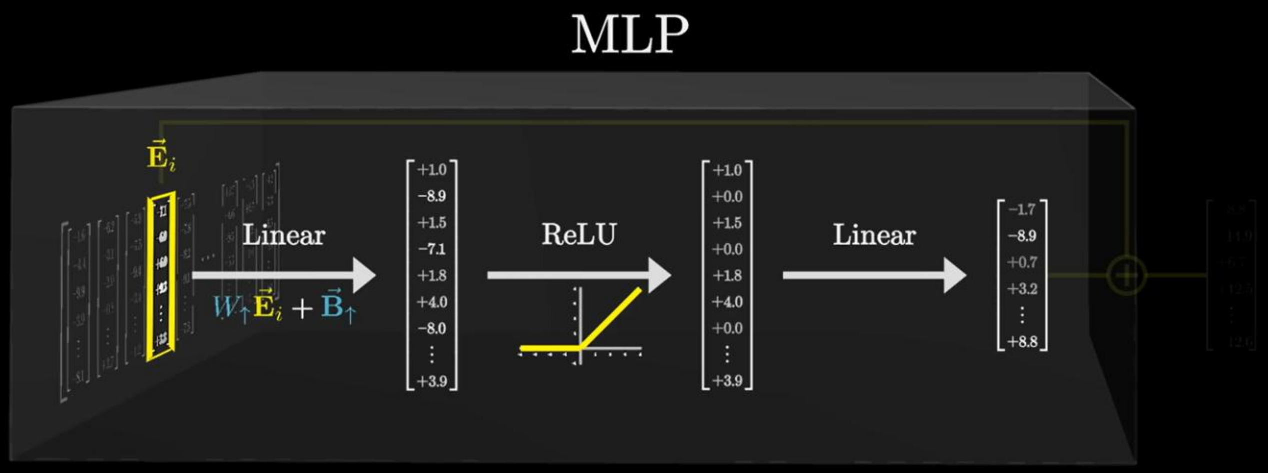


3 第二层

- “下投影”矩阵



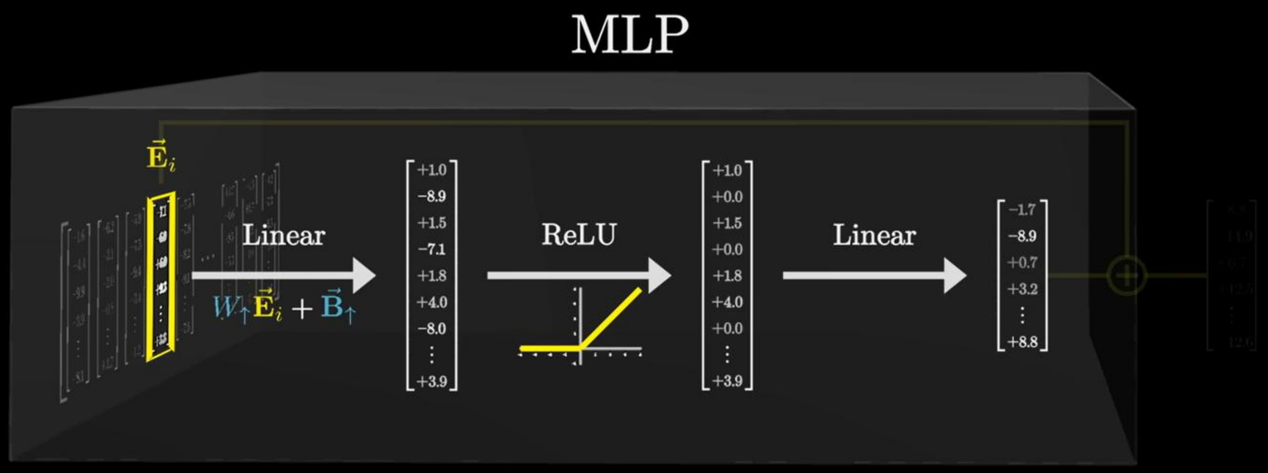
$$\overbrace{\begin{bmatrix} +0.5 & +8.4 & -4.7 & -8.6 & +4.7 & +5.4 & +8.1 & \cdots & -9.6 \\ -5.3 & +2.3 & +8.9 & +8.9 & +1.1 & +8.2 & +2.8 & \cdots & -0.3 \\ +2.1 & +1.0 & +8.4 & +8.3 & -2.1 & +9.2 & -6.5 & \cdots & -7.2 \\ +0.1 & -9.5 & +8.9 & +6.5 & -9.6 & -6.4 & -3.3 & \cdots & +6.1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ -4.2 & -0.2 & +2.0 & -9.6 & +1.9 & -1.3 & +6.1 & \cdots & +7.8 \end{bmatrix}}^{\text{“Down projection”}} \begin{bmatrix} +1.0 \\ +0.0 \\ +1.5 \\ +0.0 \\ +1.8 \\ +4.0 \\ +0.0 \\ \vdots \\ +3.9 \end{bmatrix} + \begin{bmatrix} +1.5 \\ -6.3 \\ +5.7 \\ +2.2 \\ \vdots \\ -1.6 \end{bmatrix} = \left.\begin{bmatrix} -1.7 \\ -8.9 \\ +0.7 \\ +3.2 \\ \vdots \\ +8.8 \end{bmatrix}\right\} 12{,}288$$



3 第二层

- 视作列向量组合

$$\begin{bmatrix} \vert & \vert & \vert & \vert & \vert & & \vert \\ \vec{\mathbf{C}}_0 & \vec{\mathbf{C}}_1 & \vec{\mathbf{C}}_2 & \vec{\mathbf{C}}_3 & \vec{\mathbf{C}}_4 & \cdots & \vec{\mathbf{C}}_m \\ \vert & \vert & \vert & \vert & \vert & & \vert \end{bmatrix} \begin{bmatrix} n_0 \\ n_1 \\ n_2 \\ n_3 \\ n_4 \\ \vdots \\ n_m \end{bmatrix} + \begin{bmatrix} \vert \\ \vec{\mathbf{B}} \\ \vert \end{bmatrix} = \begin{bmatrix} -1.7 \\ -8.9 \\ +0.7 \\ +3.2 \\ \vdots \\ +8.8 \end{bmatrix}$$



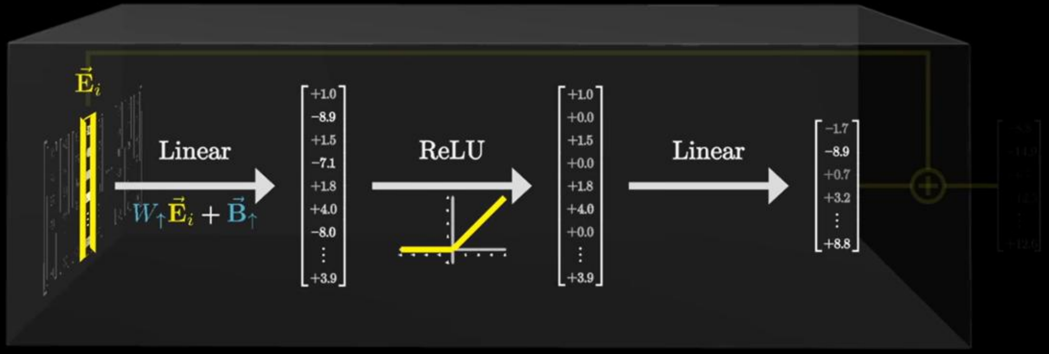
3 第二层

- 表示为对“你问我猜”的答案的加权组合

$$\boxed{n_0\vec{\mathbf{C}}_0} + n_1\vec{\mathbf{C}}_1 + n_2\vec{\mathbf{C}}_2 + n_3\vec{\mathbf{C}}_3 + n_4\vec{\mathbf{C}}_4 + \cdots + n_m\vec{\mathbf{C}}_m$$

$$\overbrace{\begin{bmatrix} \boxed{\vec{\mathbf{C}}_0} & \vec{\mathbf{C}}_1 & \vec{\mathbf{C}}_2 & \vec{\mathbf{C}}_3 & \vec{\mathbf{C}}_4 & \cdots & \vec{\mathbf{C}}_m \end{bmatrix} \begin{bmatrix} \boxed{n_0} \\ n_1 \\ n_2 \\ n_3 \\ n_4 \\ \vdots \\ n_m \end{bmatrix}} + \begin{bmatrix} \vec{\mathbf{B}} \end{bmatrix} = \begin{bmatrix} -1.7 \\ -8.9 \\ +0.7 \\ +3.2 \\ \vdots \\ +8.8 \end{bmatrix}$$

MLP



3 第二层

- 表示为对“你问我猜”的答案的加权组合

$$\underbrace{\boxed{n_0\vec{\mathbf{C}}_0} + n_1\vec{\mathbf{C}}_1 + n_2\vec{\mathbf{C}}_2 + n_3\vec{\mathbf{C}}_3 + n_4\vec{\mathbf{C}}_4 + \cdots + n_m\vec{\mathbf{C}}_m}$$

$$\begin{bmatrix} \boxed{\vec{\mathbf{C}}_0} & \vec{\mathbf{C}}_1 & \vec{\mathbf{C}}_2 & \vec{\mathbf{C}}_3 & \vec{\mathbf{C}}_4 & \cdots & \vec{\mathbf{C}}_m \end{bmatrix} \begin{bmatrix} \boxed{n_0} \\ n_1 \\ n_2 \\ n_3 \\ n_4 \\ \vdots \\ n_m \end{bmatrix} + \begin{bmatrix} \vec{\mathbf{B}} \end{bmatrix} = \begin{bmatrix} -1.7 \\ -8.9 \\ +0.7 \\ +3.2 \\ \vdots \\ +8.8 \end{bmatrix}$$

篮球 $\overrightarrow{\text{Basketball}}$

$+$

芝加哥公牛 $\overrightarrow{\text{Chicago Bulls}}$

$+$

23号球员 $\overrightarrow{\text{Number 23}}$

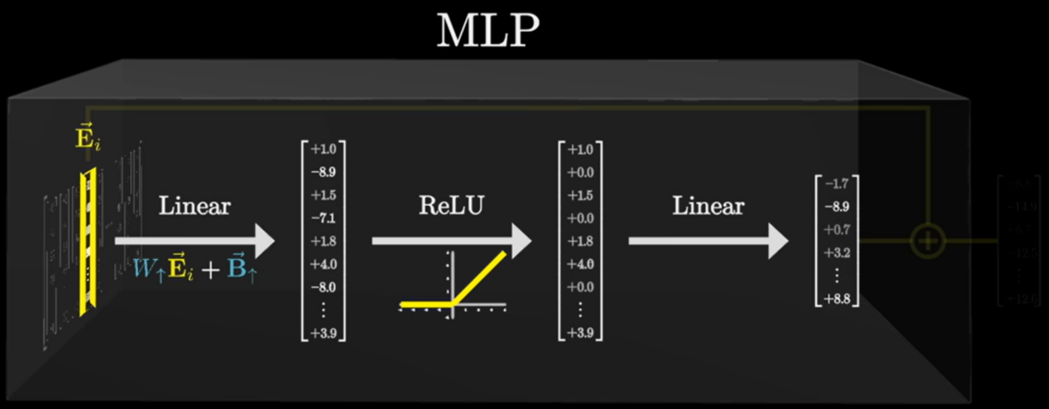
$+$

1963年出生 $\overrightarrow{\text{Born 1963}}$

$+$

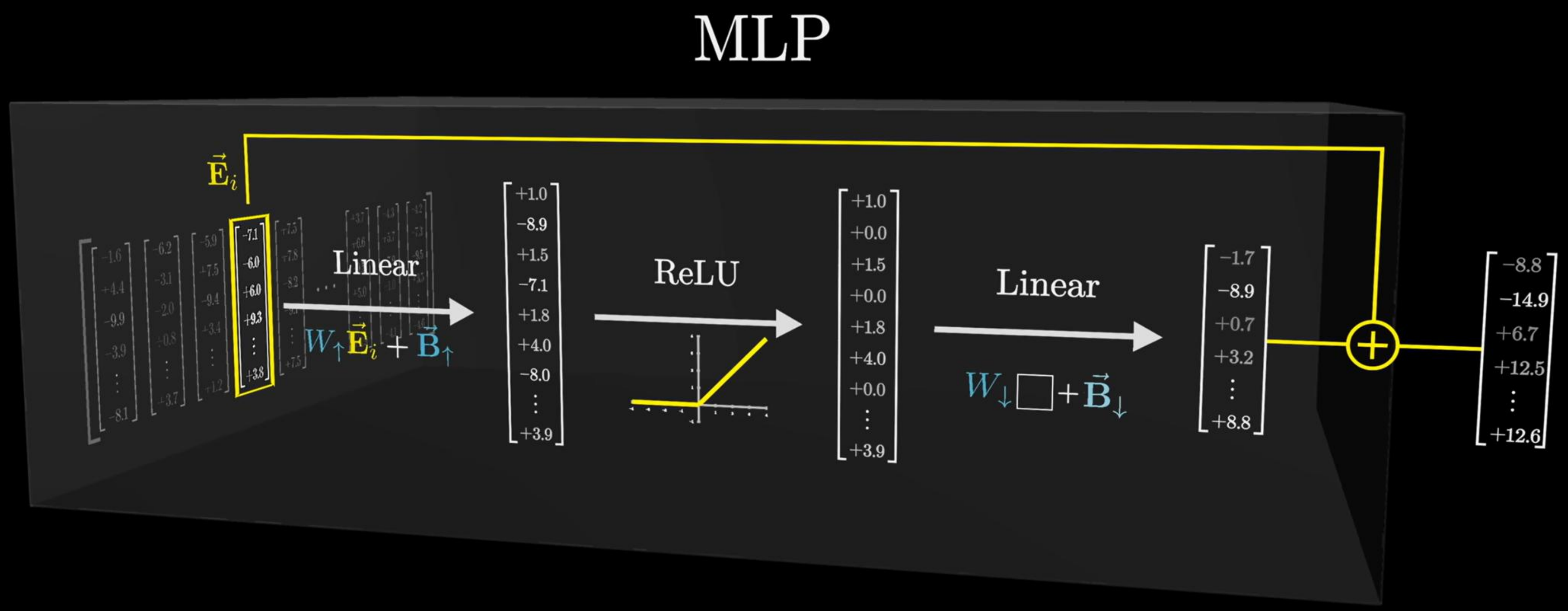
$\vdots$

迈克尔·乔丹



3 第二层

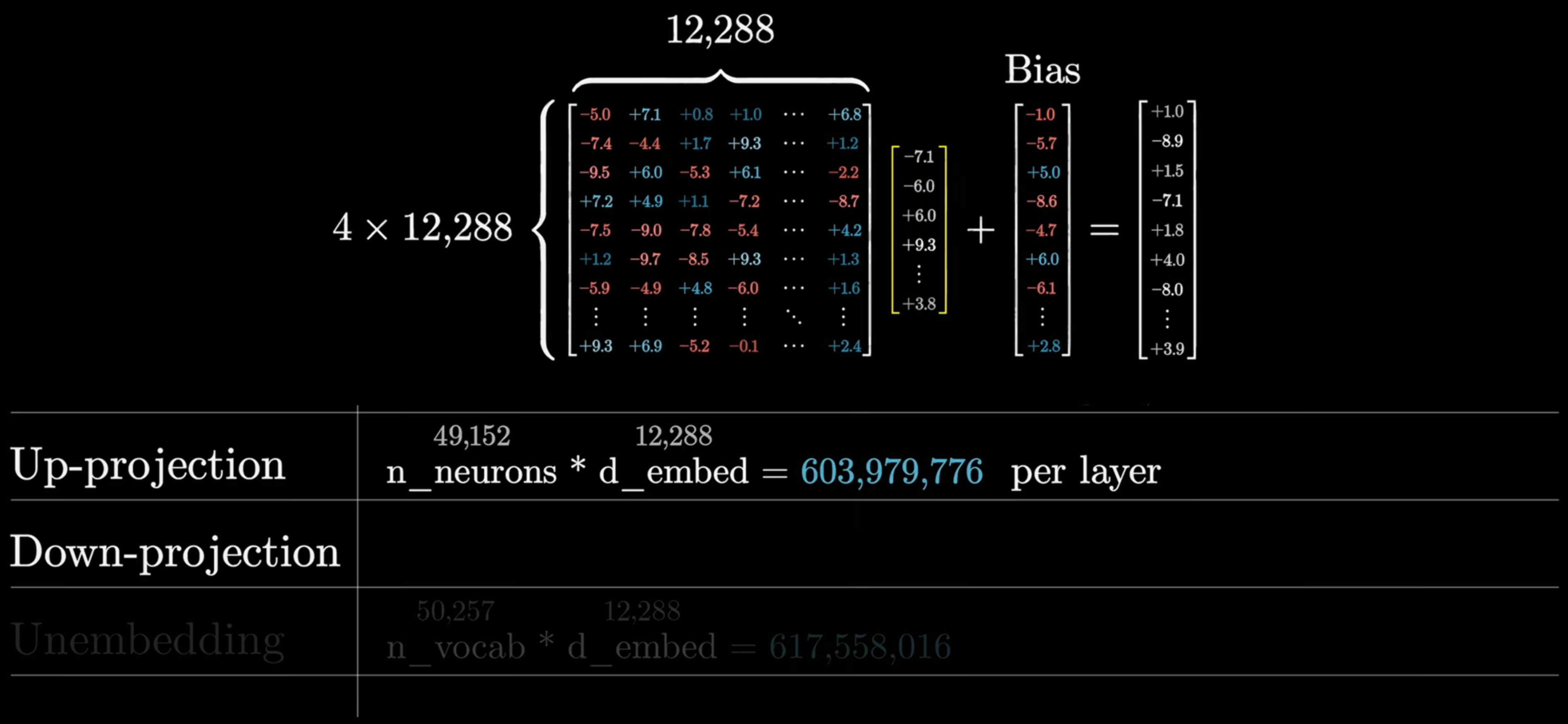
- MLP完整结构



3 参数数量分析

- “上投影”矩阵参数数量：

$$4 \times 12{,}288 \times 12{,}288 = 603{,}979{,}776$$

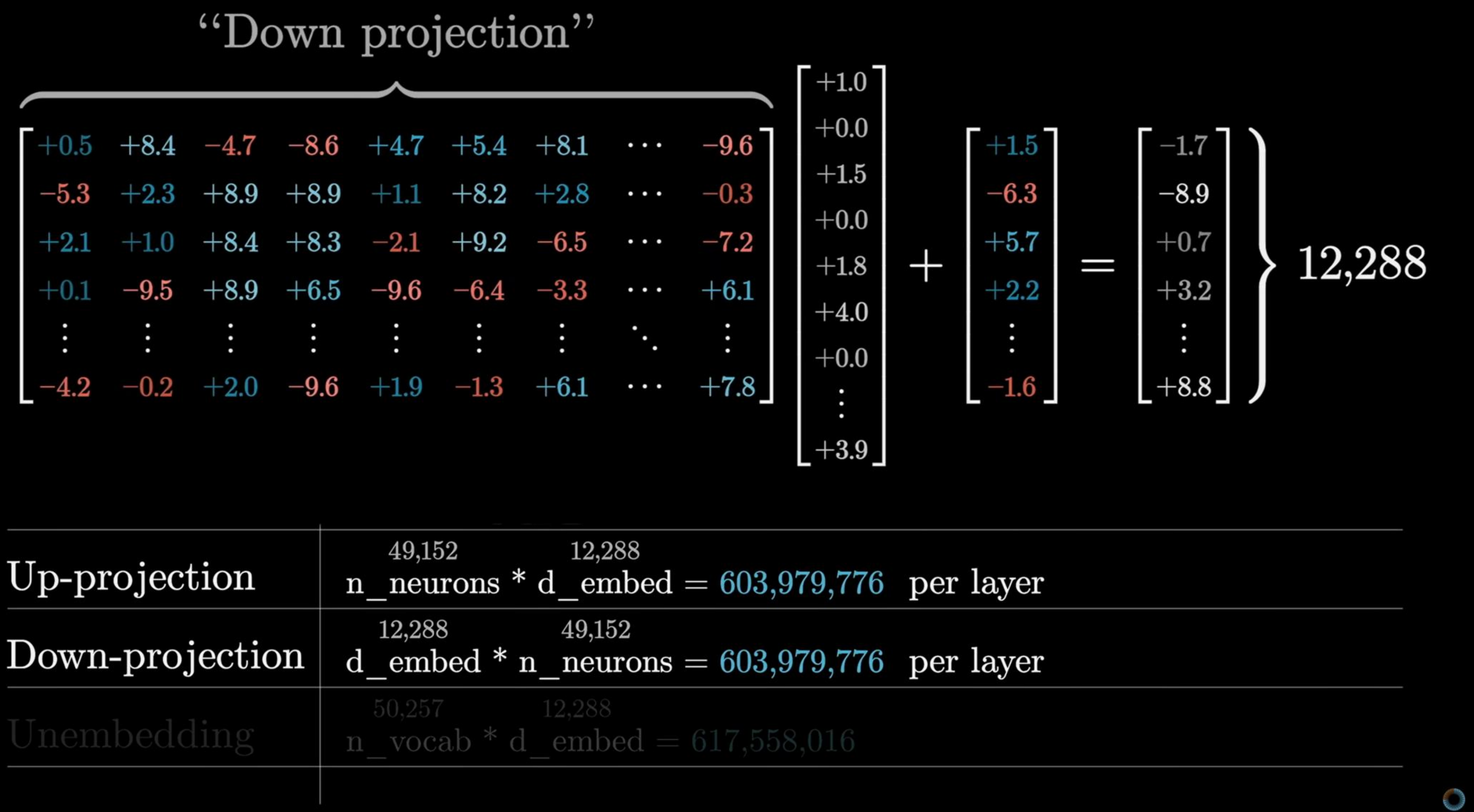


$$4 \times 12{,}288 \left\{ \overbrace{\begin{bmatrix} -5.0 & +7.1 & +0.8 & +1.0 & \cdots & +6.8 \\ -7.4 & -4.4 & +1.7 & +9.3 & \cdots & +1.2 \\ -9.5 & +6.0 & -5.3 & +6.1 & \cdots & -2.2 \\ +7.2 & +4.9 & +1.1 & -7.2 & \cdots & -8.7 \\ -7.5 & -9.0 & -7.8 & -5.4 & \cdots & +4.2 \\ +1.2 & -9.7 & -8.5 & +9.3 & \cdots & +1.3 \\ -5.9 & -4.9 & +4.8 & -6.0 & \cdots & +1.6 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ +9.3 & +6.9 & -5.2 & -0.1 & \cdots & +2.4 \end{bmatrix}}^{12{,}288} \right. \begin{bmatrix} -7.1 \\ -6.0 \\ +6.0 \\ +9.3 \\ \vdots \\ +3.8 \end{bmatrix} + \overset{\text{Bias}}{\begin{bmatrix} -1.0 \\ -5.7 \\ +5.0 \\ -8.6 \\ -4.7 \\ +6.0 \\ -6.1 \\ \vdots \\ +2.8 \end{bmatrix}} = \begin{bmatrix} +1.0 \\ -8.9 \\ +1.5 \\ -7.1 \\ +1.8 \\ +4.0 \\ -8.0 \\ \vdots \\ +3.9 \end{bmatrix}$$

Up-projection	$\overset{49{,}152}{\text{n\_neurons}} * \overset{12{,}288}{\text{d\_embed}} = 603{,}979{,}776$ per layer
Down-projection	
Unembedding	$\overset{50{,}257}{\text{n\_vocab}} * \overset{12{,}288}{\text{d\_embed}} = 617{,}558{,}016$

3 参数数量分析

- “下投影”矩阵参数数量：



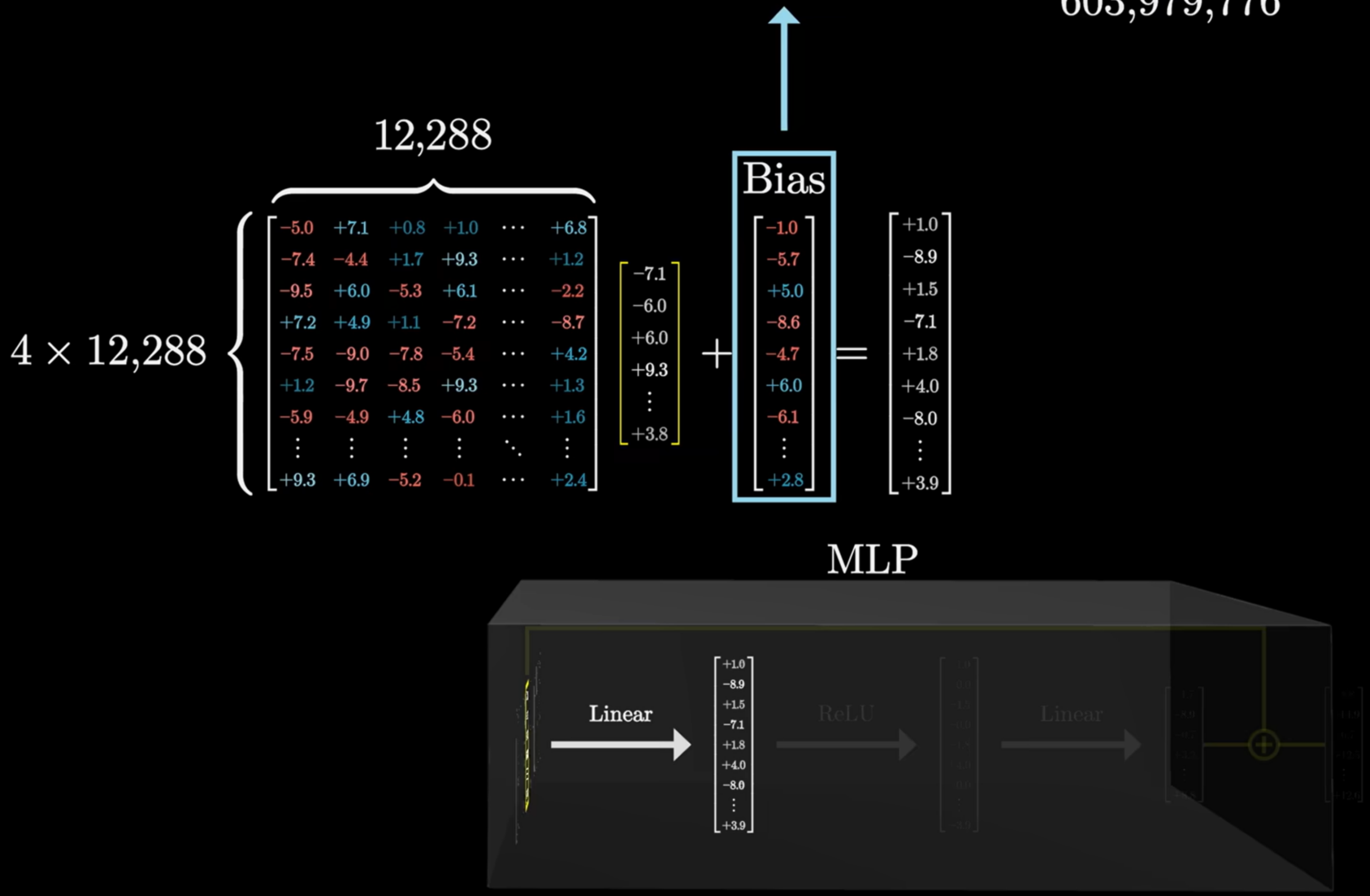
$$\overbrace{\begin{bmatrix} +0.5 & +8.4 & -4.7 & -8.6 & +4.7 & +5.4 & +8.1 & \cdots & -9.6 \\ -5.3 & +2.3 & +8.9 & +8.9 & +1.1 & +8.2 & +2.8 & \cdots & -0.3 \\ +2.1 & +1.0 & +8.4 & +8.3 & -2.1 & +9.2 & -6.5 & \cdots & -7.2 \\ +0.1 & -9.5 & +8.9 & +6.5 & -9.6 & -6.4 & -3.3 & \cdots & +6.1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ -4.2 & -0.2 & +2.0 & -9.6 & +1.9 & -1.3 & +6.1 & \cdots & +7.8 \end{bmatrix}}^{\text{“Down projection”}} \begin{bmatrix} +1.0 \\ +0.0 \\ +1.5 \\ +0.0 \\ +1.8 \\ +4.0 \\ +0.0 \\ \vdots \\ +3.9 \end{bmatrix} + \begin{bmatrix} +1.5 \\ -6.3 \\ +5.7 \\ +2.2 \\ \vdots \\ -1.6 \end{bmatrix} = \left.\begin{bmatrix} -1.7 \\ -8.9 \\ +0.7 \\ +3.2 \\ \vdots \\ +8.8 \end{bmatrix}\right\} 12{,}288$$

Up-projection	$\overset{49{,}152}{\text{n\_neurons}} * \overset{12{,}288}{\text{d\_embed}} = 603{,}979{,}776$ per layer
Down-projection	$\overset{12{,}288}{\text{d\_embed}} * \overset{49{,}152}{\text{n\_neurons}} = 603{,}979{,}776$ per layer
Unembedding	$\overset{50{,}257}{\text{n\_vocab}} * \overset{12{,}288}{\text{d\_embed}} = 617{,}558{,}016$

3 参数数量分析

- Bias部分参数可忽略

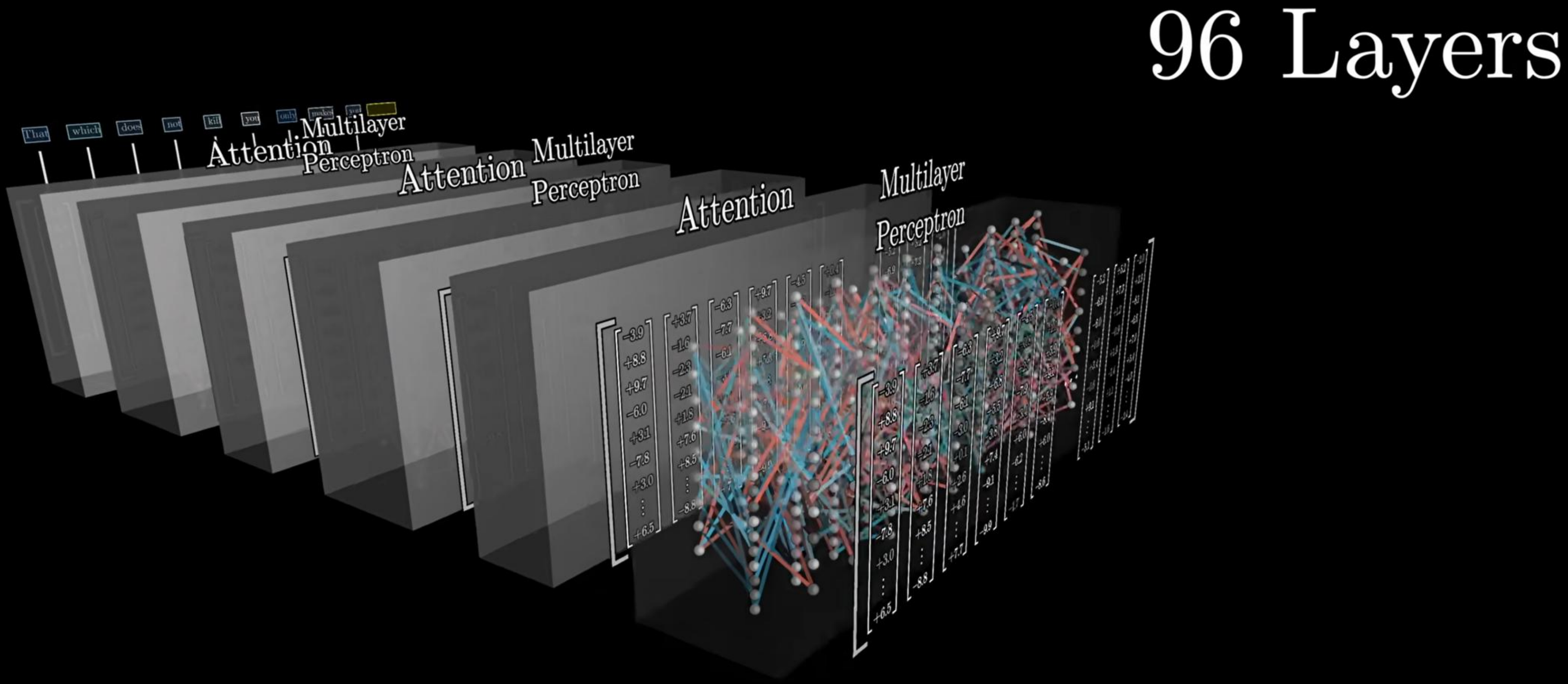
$$4 \times 12{,}288 \times 12{,}288 = 603{,}979{,}776 \qquad 4 \times 12{,}288 \qquad \frac{4 \times 12{,}288}{603{,}979{,}776} \approx 0.00008$$



$$4 \times 12{,}288 \left\{ \overbrace{\begin{bmatrix} -5.0 & +7.1 & +0.8 & +1.0 & \cdots & +6.8 \\ -7.4 & -4.4 & +1.7 & +9.3 & \cdots & +1.2 \\ -9.5 & +6.0 & -5.3 & +6.1 & \cdots & -2.2 \\ +7.2 & +4.9 & +1.1 & -7.2 & \cdots & -8.7 \\ -7.5 & -9.0 & -7.8 & -5.4 & \cdots & +4.2 \\ +1.2 & -9.7 & -8.5 & +9.3 & \cdots & +1.3 \\ -5.9 & -4.9 & +4.8 & -6.0 & \cdots & +1.6 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ +9.3 & +6.9 & -5.2 & -0.1 & \cdots & +2.4 \end{bmatrix}}^{12{,}288} \right. \begin{bmatrix} -7.1 \\ -6.0 \\ +6.0 \\ +9.3 \\ \vdots \\ +3.8 \end{bmatrix} + \boxed{\overset{\text{Bias}}{\begin{bmatrix} -1.0 \\ -5.7 \\ +5.0 \\ -8.6 \\ -4.7 \\ +6.0 \\ -6.1 \\ \vdots \\ +2.8 \end{bmatrix}}} = \begin{bmatrix} +1.0 \\ -8.9 \\ +1.5 \\ -7.1 \\ +1.8 \\ +4.0 \\ -8.0 \\ \vdots \\ +3.9 \end{bmatrix}$$

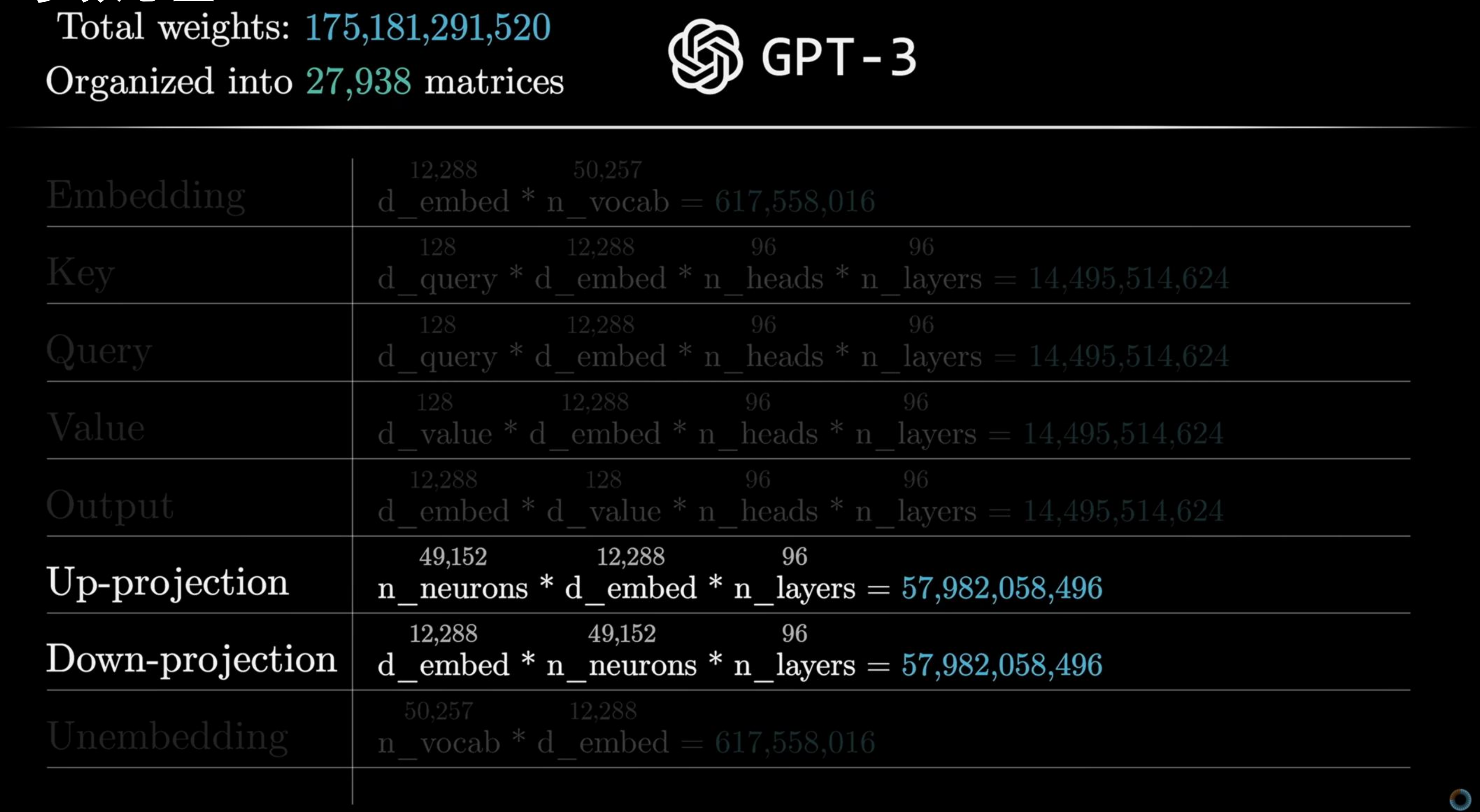
3 参数数量分析

- × 96层网络



3 参数数量分析

- GPT-3参数总量



Total weights: 175,181,291,520

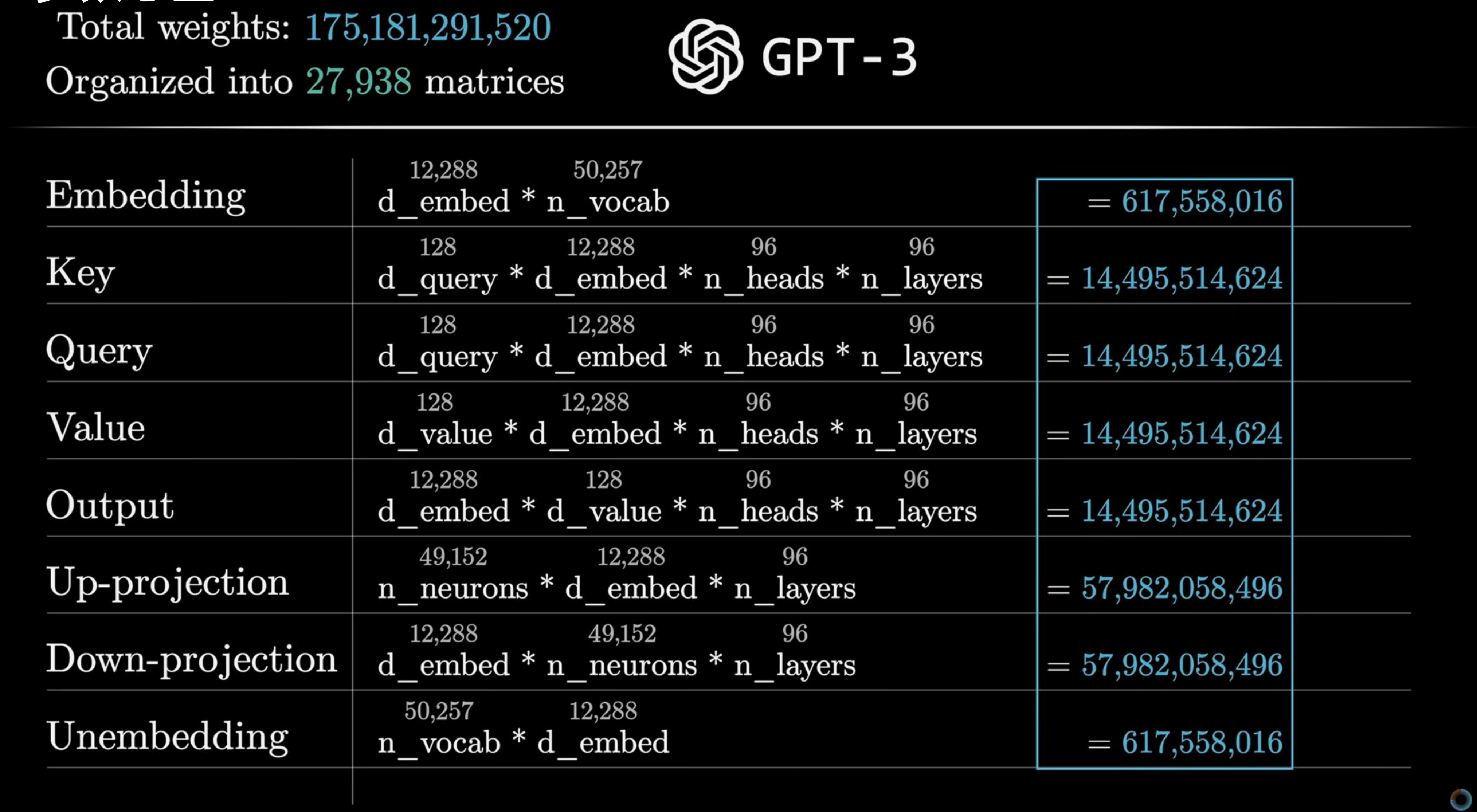
Organized into 27,938 matrices

GPT-3

Embedding	$\overset{12,288}{\text{d\_embed}} * \overset{50,257}{\text{n\_vocab}} = 617,558,016$
Key	$\overset{128}{\text{d\_query}} * \overset{12,288}{\text{d\_embed}} * \overset{96}{\text{n\_heads}} * \overset{96}{\text{n\_layers}} = 14,495,514,624$
Query	$\overset{128}{\text{d\_query}} * \overset{12,288}{\text{d\_embed}} * \overset{96}{\text{n\_heads}} * \overset{96}{\text{n\_layers}} = 14,495,514,624$
Value	$\overset{128}{\text{d\_value}} * \overset{12,288}{\text{d\_embed}} * \overset{96}{\text{n\_heads}} * \overset{96}{\text{n\_layers}} = 14,495,514,624$
Output	$\overset{12,288}{\text{d\_embed}} * \overset{128}{\text{d\_value}} * \overset{96}{\text{n\_heads}} * \overset{96}{\text{n\_layers}} = 14,495,514,624$
Up-projection	$\overset{49,152}{\text{n\_neurons}} * \overset{12,288}{\text{d\_embed}} * \overset{96}{\text{n\_layers}} = 57,982,058,496$
Down-projection	$\overset{12,288}{\text{d\_embed}} * \overset{49,152}{\text{n\_neurons}} * \overset{96}{\text{n\_layers}} = 57,982,058,496$
Unembedding	$\overset{50,257}{\text{n\_vocab}} * \overset{12,288}{\text{d\_embed}} = 617,558,016$

3 参数数量分析

- GPT-3参数总量



Total weights: 175,181,291,520

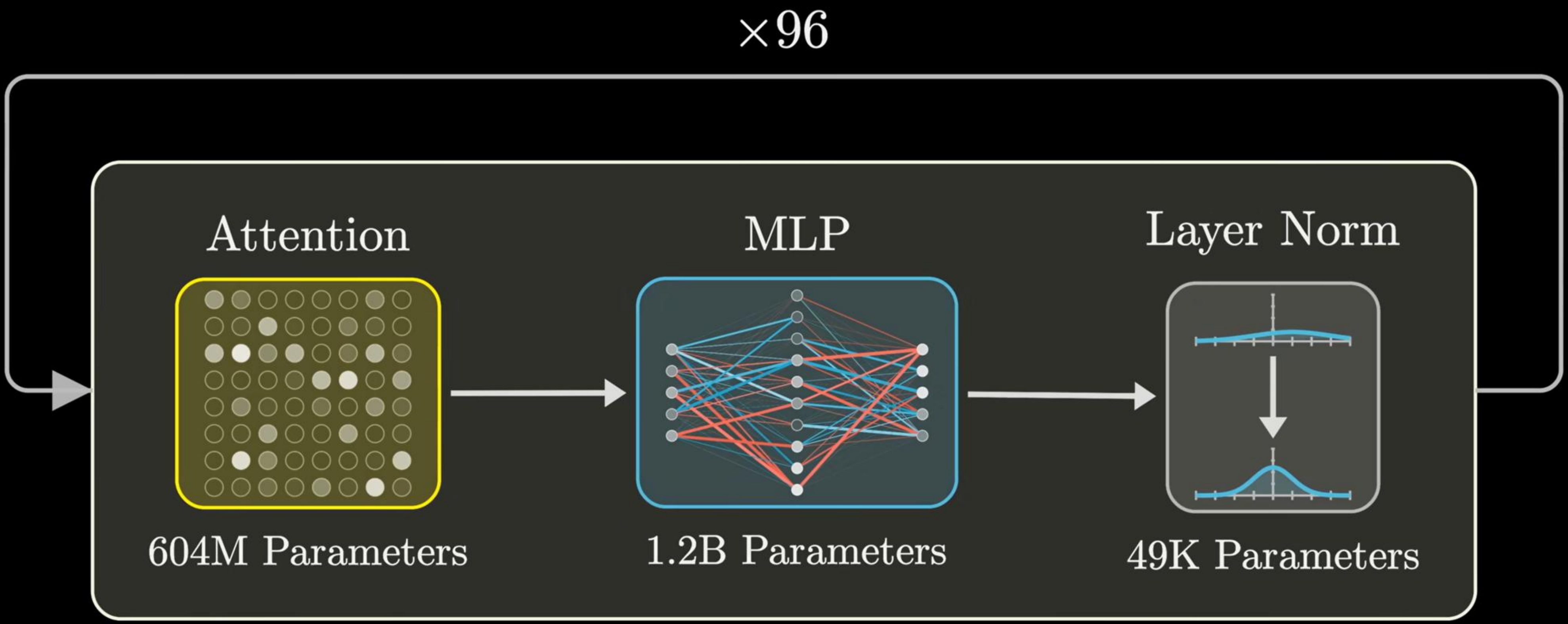
Organized into 27,938 matrices

GPT-3

Embedding	$\overset{12{,}288}{\text{d\_embed}} * \overset{50{,}257}{\text{n\_vocab}}$	= 617,558,016
Key	$\overset{128}{\text{d\_query}} * \overset{12{,}288}{\text{d\_embed}} * \overset{96}{\text{n\_heads}} * \overset{96}{\text{n\_layers}}$	= 14,495,514,624
Query	$\overset{128}{\text{d\_query}} * \overset{12{,}288}{\text{d\_embed}} * \overset{96}{\text{n\_heads}} * \overset{96}{\text{n\_layers}}$	= 14,495,514,624
Value	$\overset{128}{\text{d\_value}} * \overset{12{,}288}{\text{d\_embed}} * \overset{96}{\text{n\_heads}} * \overset{96}{\text{n\_layers}}$	= 14,495,514,624
Output	$\overset{12{,}288}{\text{d\_embed}} * \overset{128}{\text{d\_value}} * \overset{96}{\text{n\_heads}} * \overset{96}{\text{n\_layers}}$	= 14,495,514,624
Up-projection	$\overset{49{,}152}{\text{n\_neurons}} * \overset{12{,}288}{\text{d\_embed}} * \overset{96}{\text{n\_layers}}$	= 57,982,058,496
Down-projection	$\overset{12{,}288}{\text{d\_embed}} * \overset{49{,}152}{\text{n\_neurons}} * \overset{96}{\text{n\_layers}}$	= 57,982,058,496
Unembedding	$\overset{50{,}257}{\text{n\_vocab}} * \overset{12{,}288}{\text{d\_embed}}$	= 617,558,016

3 参数数量分析

- GPT-3参数总量


×96
Attention
MLP
Layer Norm
604M Parameters
1.2B Parameters
49K Parameters