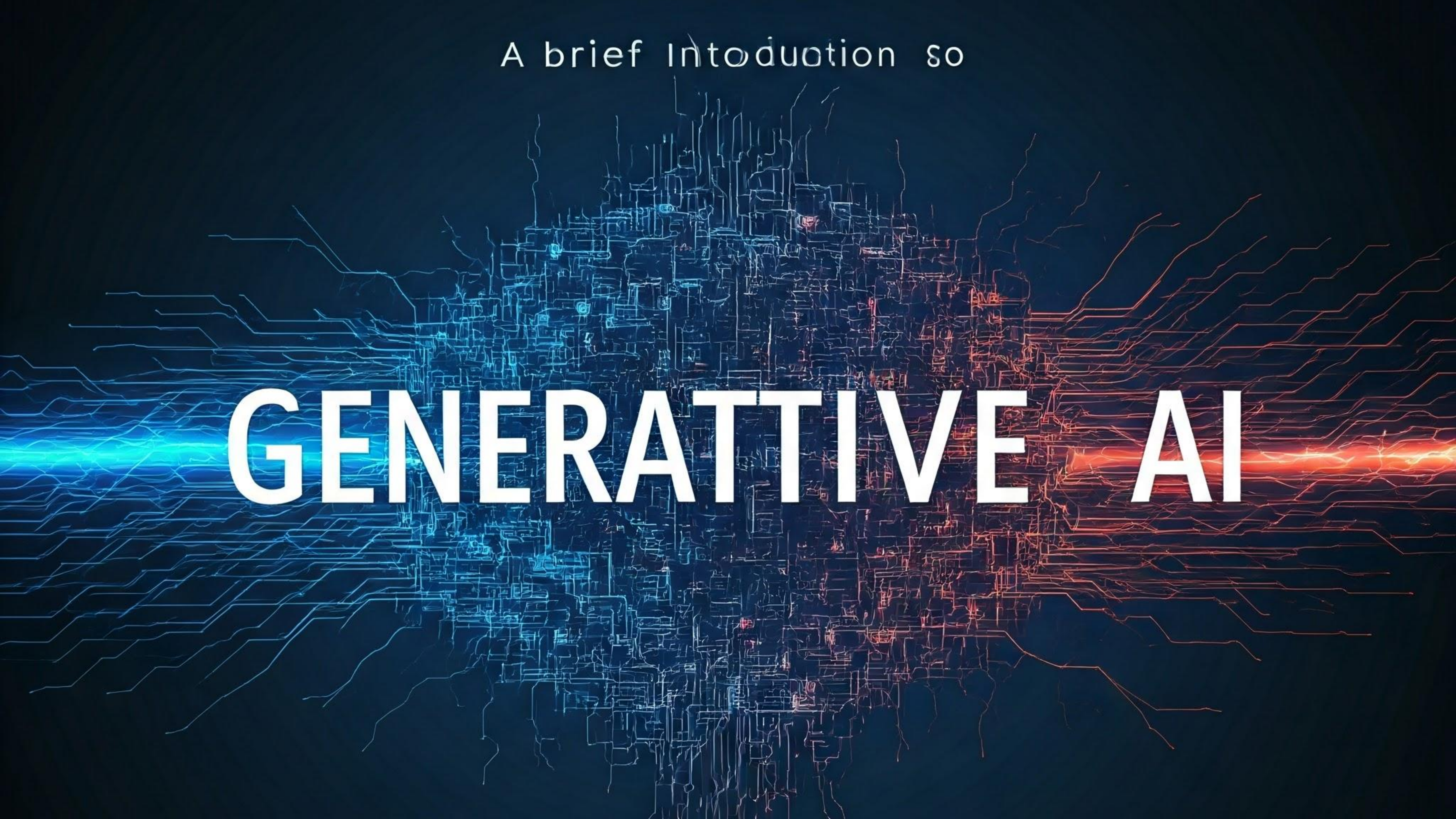
A brief introduction to

Generative AI

What ChatGPT Is Made Of


A brief Intoduotion so
GENERATTIVE AI

Generative AI

GPT-4o

DALL·E 2

SORA

LLAMA 2

VAEs

SUNO

Midjourney

Stable Diffusion

Diffusion

Gemini

udio

PaLM 2

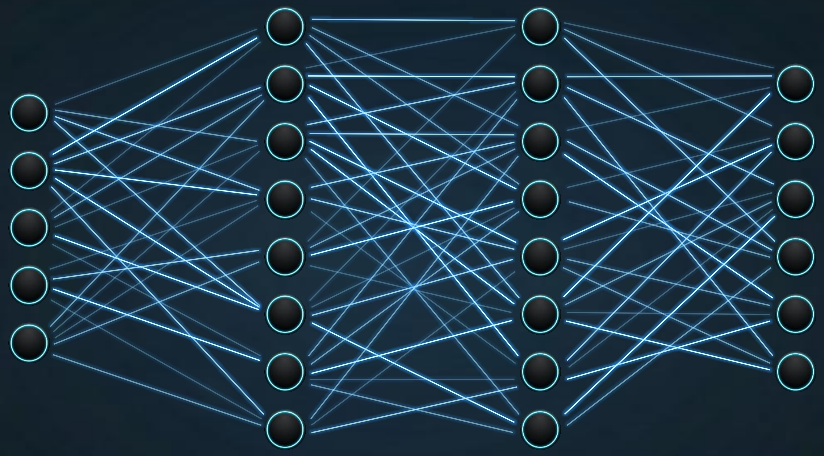
Claude

GANs

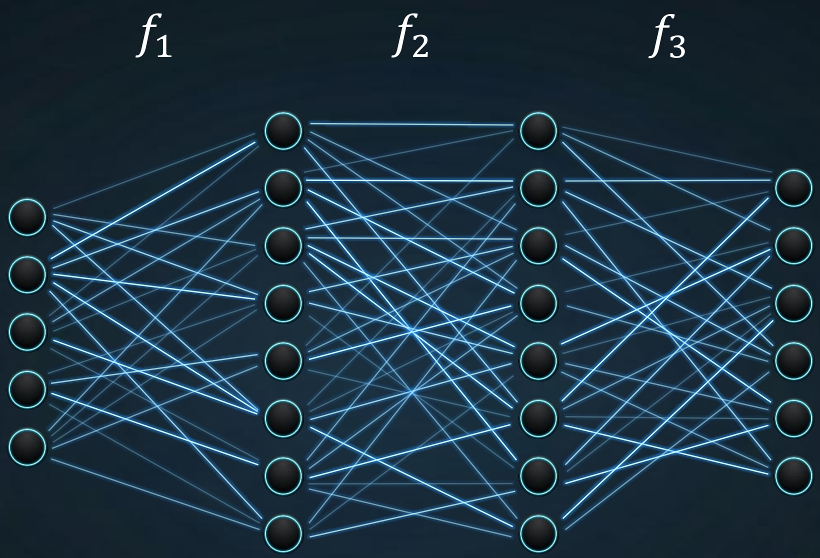
Large Language Models

Deep Learning

Artificial Neural Networks (ANNs)



Deep Learning



Katze

$$f: \mathbb{R}^m (\times\, \Omega) \to \mathbb{R}^n, (x; A, b) \mapsto \sigma(Ax + b)$$

$$L: \Omega \to \mathbb{R}, \omega \mapsto \sum_{t \in T} l\big(t, f(t; \omega)\big)$$

Large Language Models

Training

OpenAI GPT-3

Inference

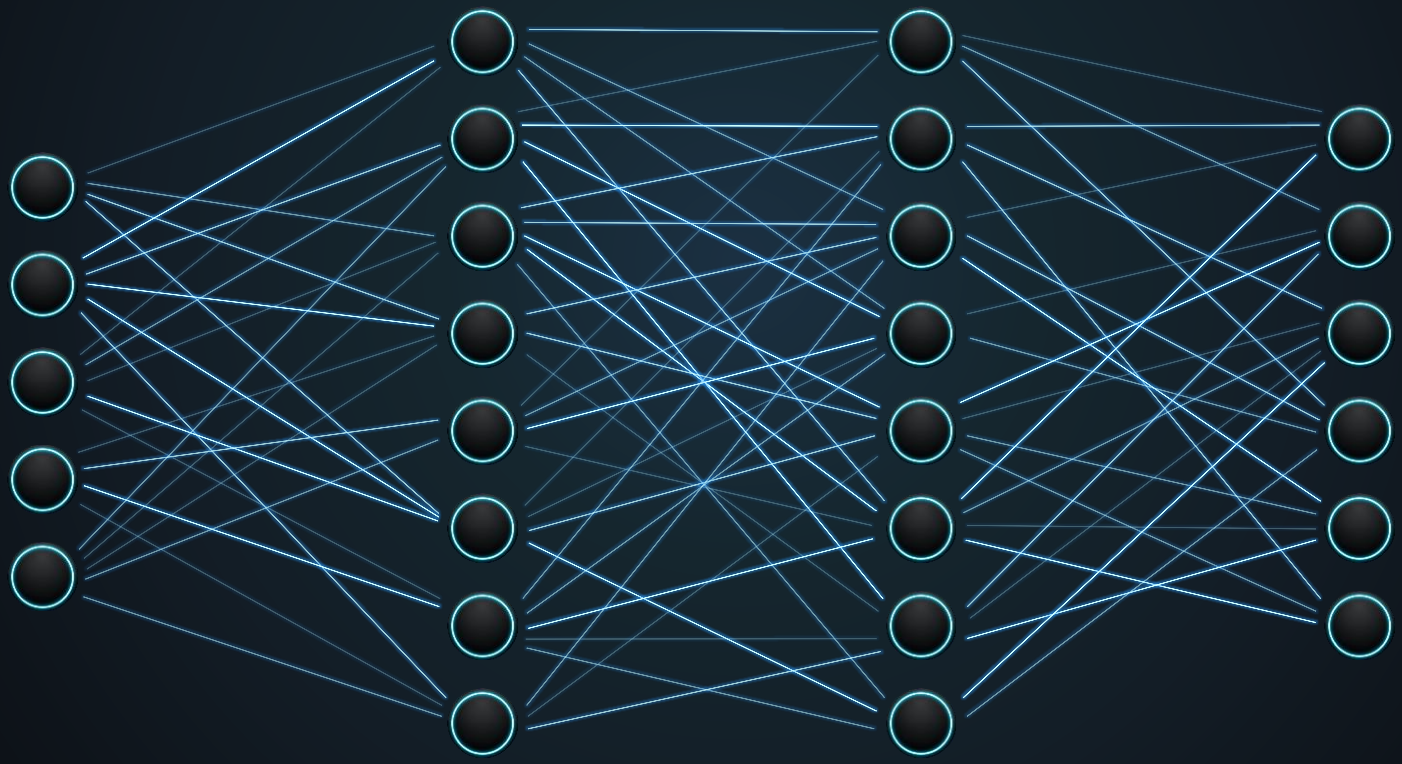
Architecture

GPT-3

Generative **P**re-Trained **T**ransformer 3

The Transformer Architecture

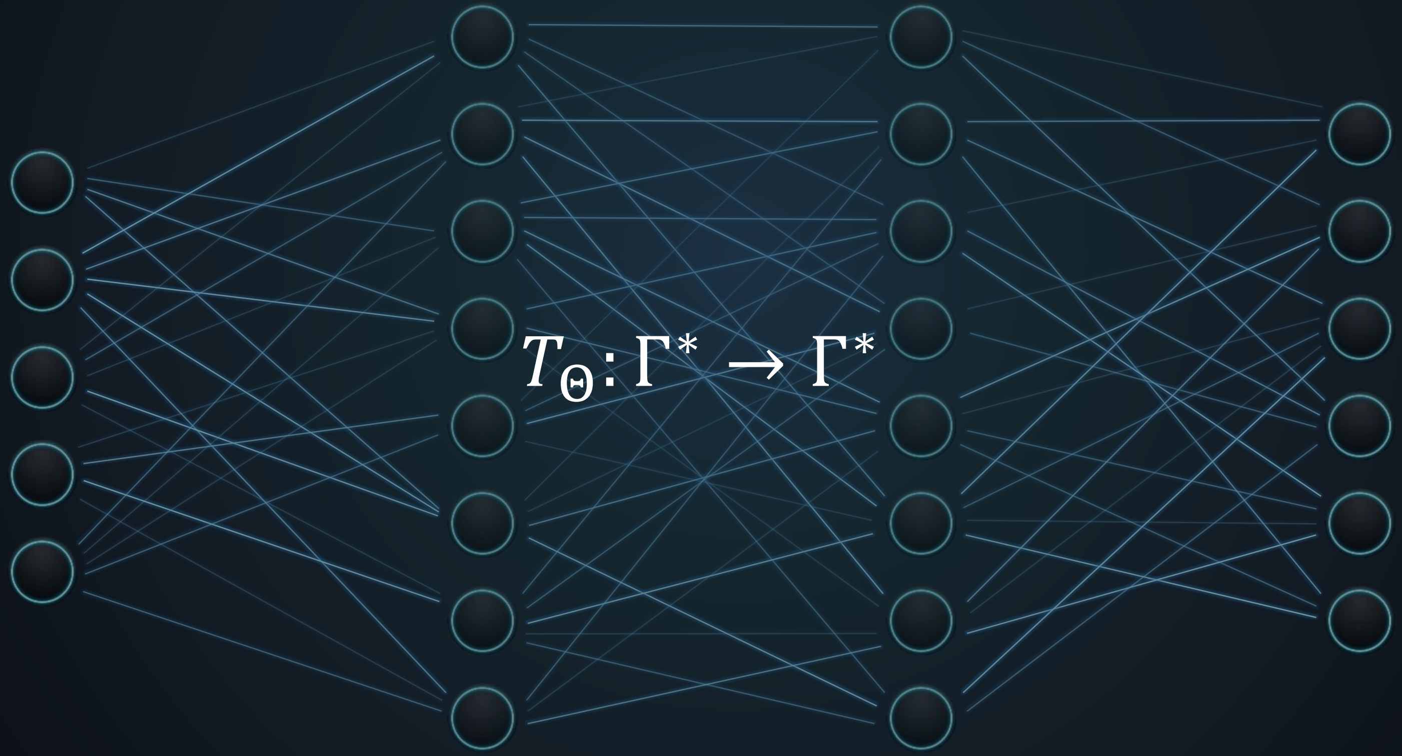
Do you plan to take over the world?



Nope! My only goal is to assist you with information and tasks. World domination isn‘t on the agenda. 😊

The Transformer Architecture

Do you plan to take over the world?

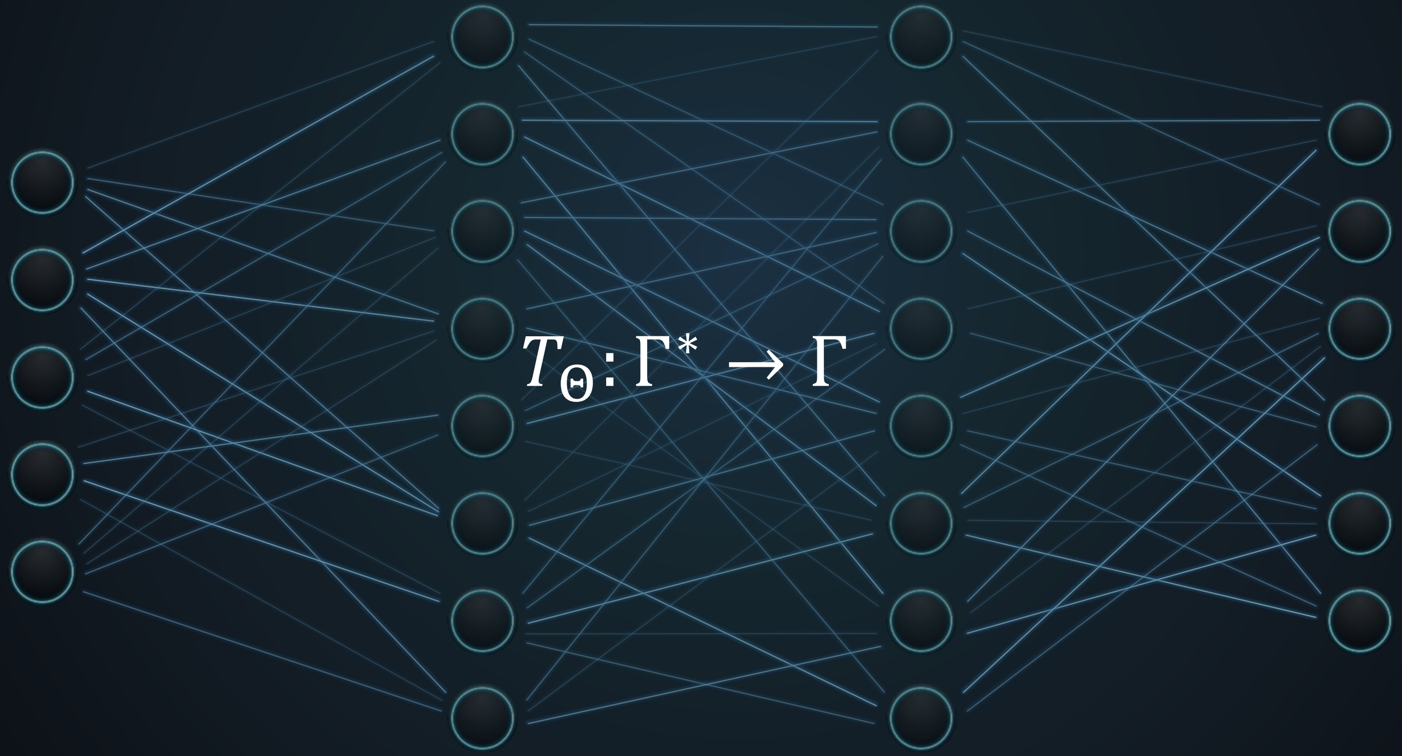


$$T_\Theta : \Gamma^* \to \Gamma^*$$

Nope! My only goal is to assist you with information and tasks. World domination isn‘t on the agenda. 😊

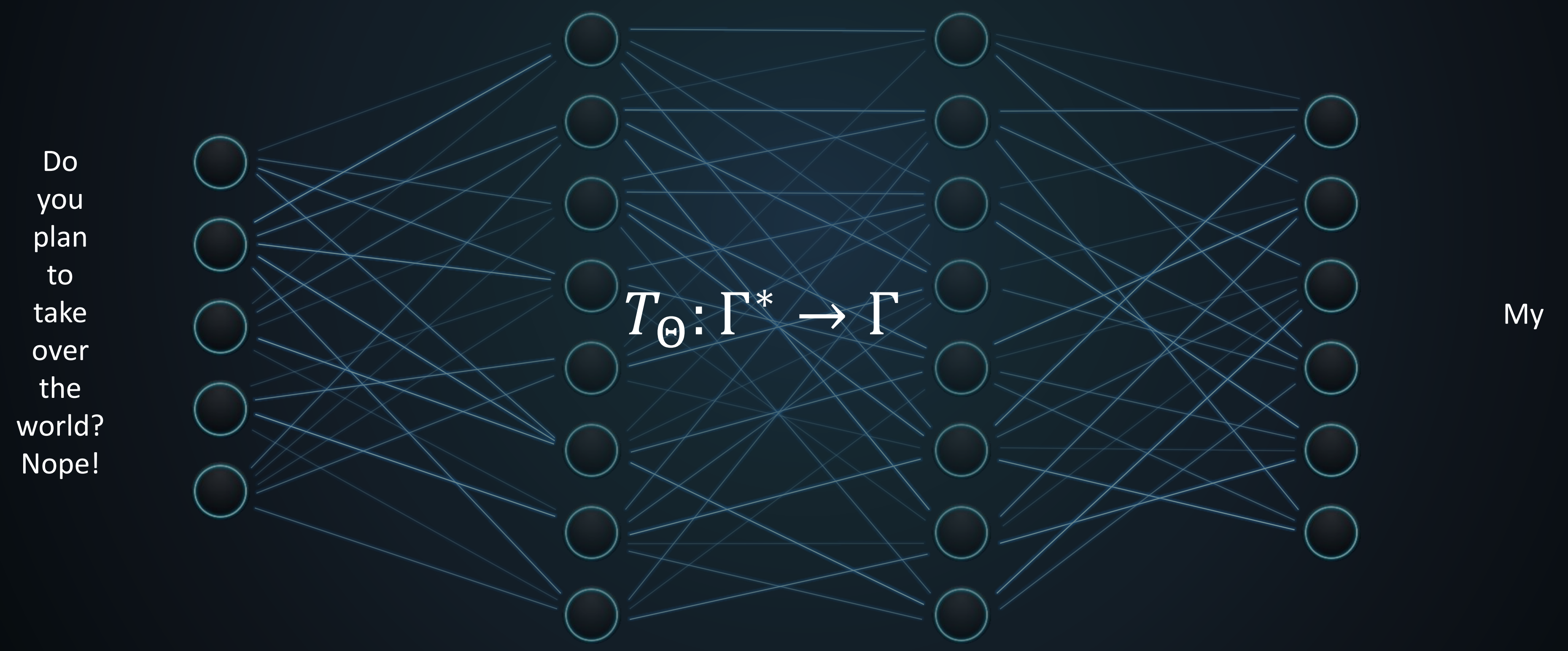
The Transformer Architecture

Do
you
plan
to
take
over
the
world?



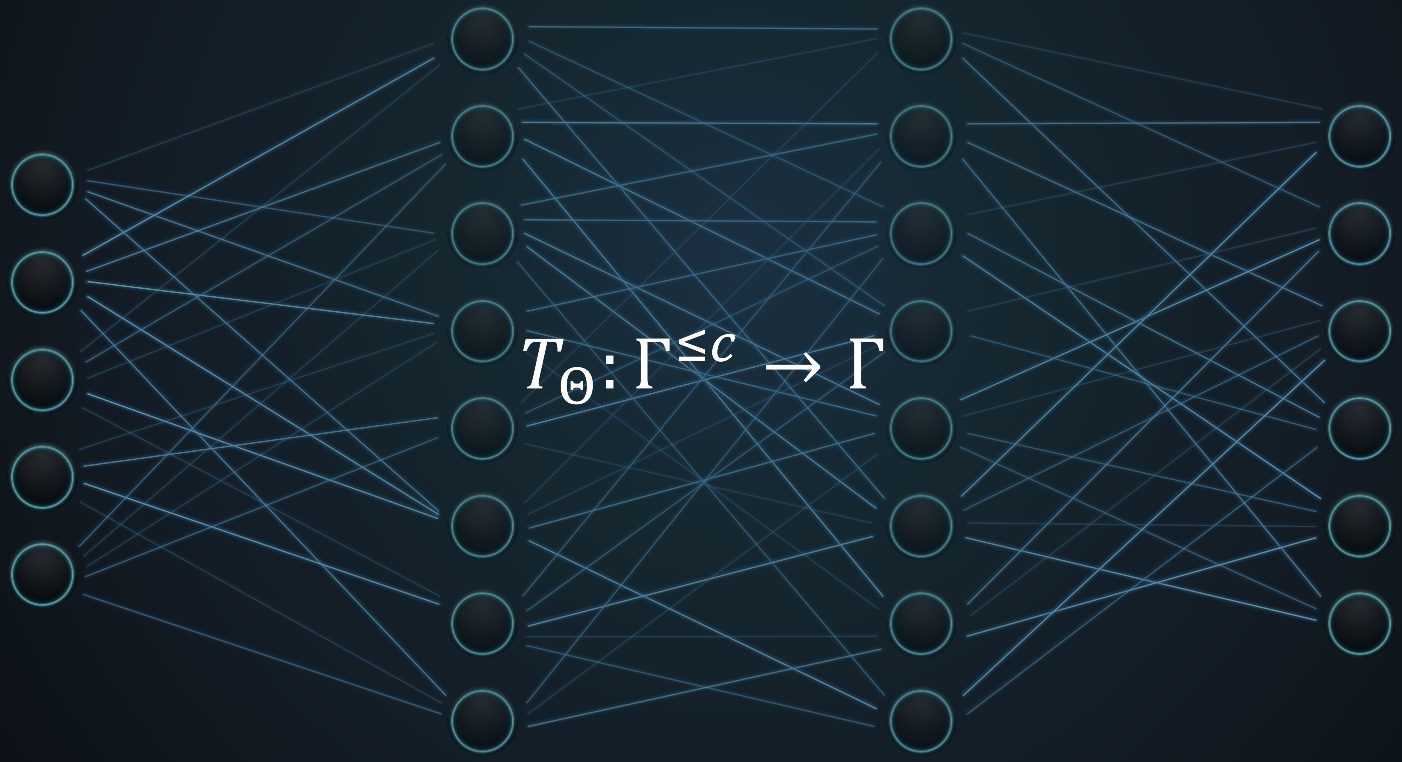
Nope!

The Transformer Architecture



The Transformer Architecture

Do
you
plan
to
take
over
the
world?
Nope!

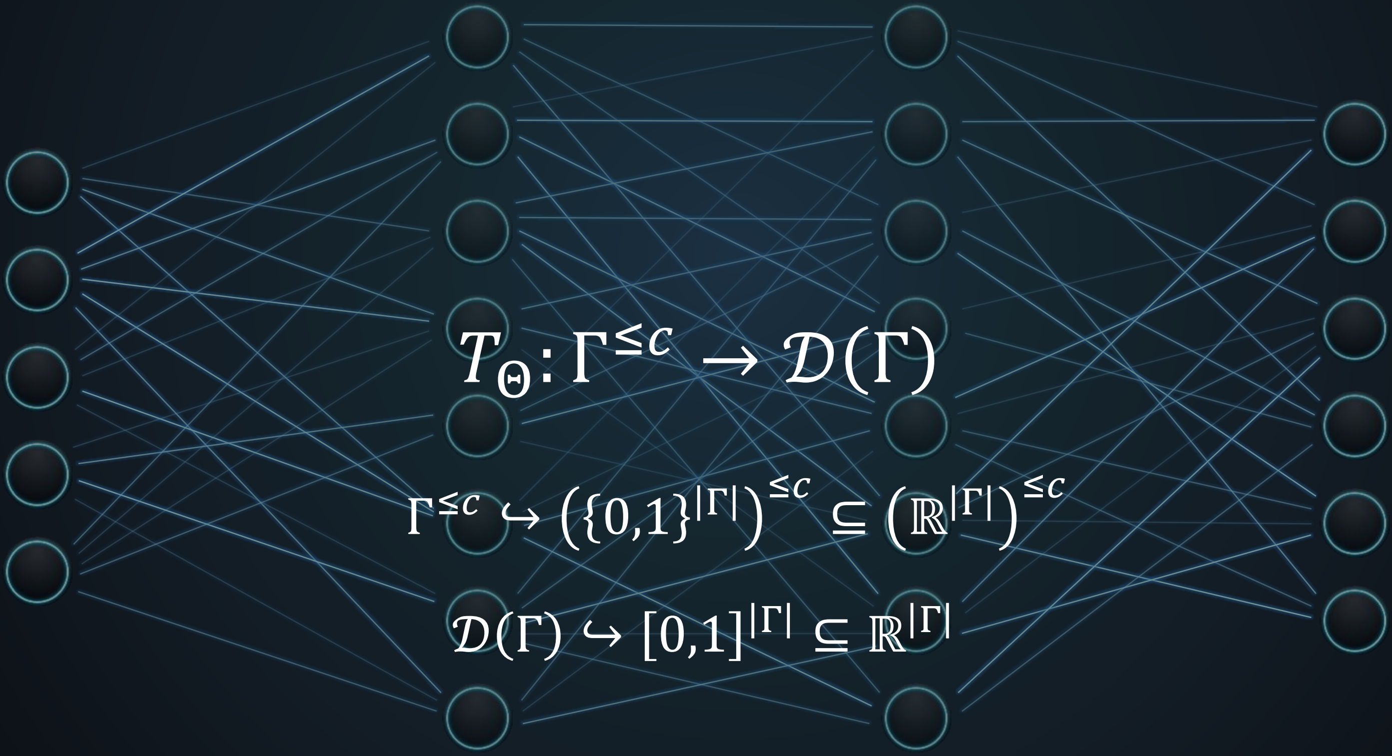


$$T_\Theta : \Gamma^{\leq c} \to \Gamma$$

My

The Transformer Architecture

Do you plan to take over the world?



$$T_\Theta\colon \Gamma^{\leq c} \to \mathcal{D}(\Gamma)$$

$$\Gamma^{\leq c} \hookrightarrow \left(\{0,1\}^{|\Gamma|}\right)^{\leq c} \subseteq \left(\mathbb{R}^{|\Gamma|}\right)^{\leq c}$$

$$\mathcal{D}(\Gamma) \hookrightarrow [0,1]^{|\Gamma|} \subseteq \mathbb{R}^{|\Gamma|}$$

No, I don’t plan anything — I’m just here to help with your questions! 😊

Step 0: The Tokenizer

$$T_\Theta : \Gamma^{\leq c} \to \mathcal{D}(\Gamma)$$

$$\Gamma = ?$$

$$\Gamma = \{\text{Aachen}, \text{Aal}, \text{Aalen}, \ldots\}$$

The Transformer Architecture

Step 0: The Tokenizer

$$T_\Theta : \Gamma^{\leq c} \to \mathcal{D}(\Gamma)$$

$$\Gamma = ?$$

$$\Gamma = \{\mathrm{a}, \mathrm{b}, \mathrm{c}, \ldots\}$$

The Transformer Architecture

Step 0: The Tokenizer

$$T_\Theta : \Gamma^{\leq c} \to \mathcal{D}(\Gamma)$$

$$\Gamma = ?$$

$$\Gamma = \{0x00, 0x01, \ldots, 0xFF\}$$

Step 0: The Tokenizer

$$T_\Theta : \Gamma^{\leq c} \to \mathcal{D}(\Gamma)$$

$$\Gamma = ?$$

Subword-Tokenizer

Goal:

Maximize semantic meaning of every token

Algorithms:

BPE, WordPiece, SentencePiece…

The Transformer Architecture

Step 1: Embedding

$$T_\Theta : \Gamma^{\leq c} \to \mathcal{D}(\Gamma)$$

Do you plan to take over the world?

$$\parallel$$

$$(a_1, a_3, a_{71}, a_{24}, a_{98}, a_{3219}, a_{319}, a_{10}, a_{999})$$

$$\downarrow$$

$$(e_1, e_3, e_{71}, e_{24}, e_{98}, e_{3219}, e_{319}, e_{10}, e_{999})$$

$$\Gamma^{\leq c} \to \left(\mathbb{R}^{|\Gamma|}\right)^{\leq c}$$

The Transformer Architecture

Step 1: Embedding

$$T_\Theta : \Gamma^{\leq c} \xrightarrow{\iota} \left(\mathbb{R}^d\right)^{\leq c} \to \mathcal{D}(\Gamma)$$

Do you plan to take over the world?

$$\shortparallel$$

$$(a_1, a_3, a_{71}, a_{24}, a_{98}, a_{3219}, a_{319}, a_{10}, a_{999})$$

$$\Gamma^{\leq c}$$

$$\downarrow \iota$$

$$\left(\begin{pmatrix} 1.2 \\ -0.1 \\ \vdots \\ 0.6 \end{pmatrix}, \begin{pmatrix} 0.3 \\ 1.0 \\ \vdots \\ -0.2 \end{pmatrix}, \begin{pmatrix} -0.2 \\ -1.7 \\ \vdots \\ 0.5 \end{pmatrix}, \begin{pmatrix} 0.6 \\ -0.7 \\ \vdots \\ 0.3 \end{pmatrix}, \begin{pmatrix} 0.4 \\ 0.5 \\ \vdots \\ -0.1 \end{pmatrix}, \begin{pmatrix} -0.6 \\ 1.0 \\ \vdots \\ 0.5 \end{pmatrix}, \begin{pmatrix} -0.1 \\ -0.5 \\ \vdots \\ 0.0 \end{pmatrix}, \begin{pmatrix} 1.1 \\ -1.1 \\ \vdots \\ 0.4 \end{pmatrix}, \begin{pmatrix} 0.1 \\ 1.6 \\ \vdots \\ -0.6 \end{pmatrix} \right)$$

$$\left(\mathbb{R}^d\right)^{\leq c}$$

$$d \ll |\Gamma|$$

The Transformer Architecture

Step 1: Embedding

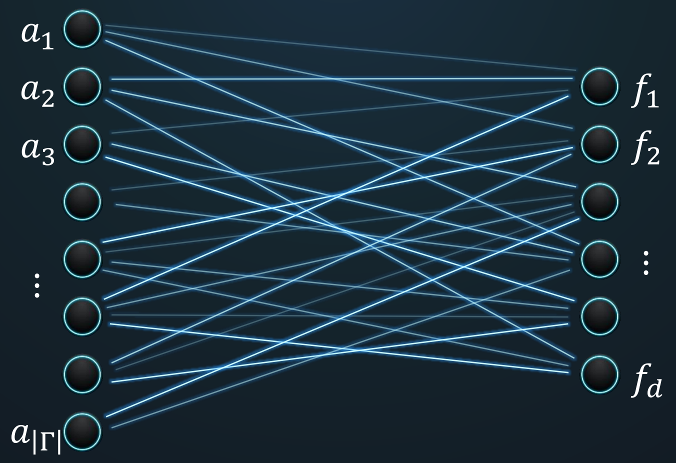
$$T_\Theta : \Gamma^{\leq c} \xrightarrow{\iota} \left(\mathbb{R}^d\right)^{\leq c} \to \mathcal{D}(\Gamma)$$

$$\iota = \begin{cases} a_1 & \mapsto & \left(\theta_\iota^{(1,1)}, \ldots, \theta_\iota^{(1,d)}\right) \\ \vdots & & \vdots \\ a_{|\Gamma|} & \mapsto & \left(\theta_\iota^{(|\Gamma|,1)}, \ldots, \theta_\iota^{(|\Gamma|,d)}\right) \end{cases}$$

Step 1: Embedding

$$T_{\Theta}: \Gamma^{\leq c} \xrightarrow{\iota} \left(\mathbb{R}^{d}\right)^{\leq c} \rightarrow \mathcal{D}(\Gamma)$$

$$\iota(a_j) = \theta_\iota \cdot e_j$$



The Transformer Architecture

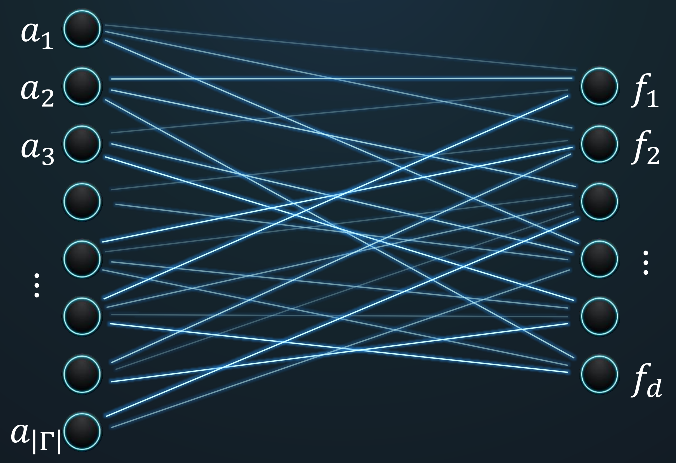
Step 1: Embedding

$$T_\Theta: \Gamma^{\leq c} \xrightarrow{\iota} \left(\mathbb{R}^d\right)^{\leq c} \to \mathcal{D}(\Gamma)$$

$$\iota(e) = \theta_\iota \cdot e, e \in \mathbb{R}^{|\Gamma| \times n}$$

$$\theta_\iota \in \mathbb{R}^{d \times |\Gamma|}$$

$\Rightarrow d|\Gamma|$ parameters



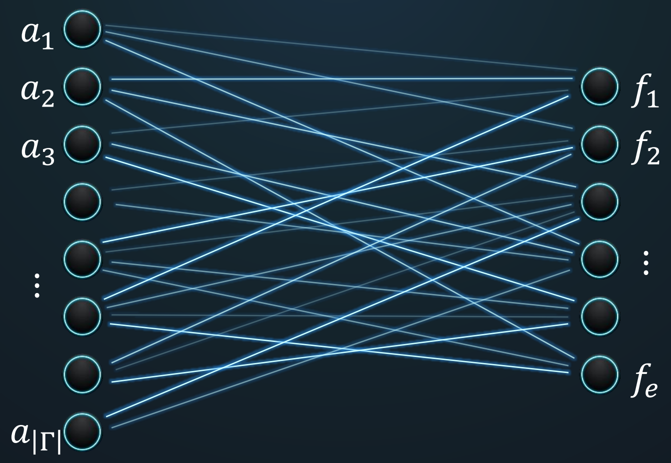
The Transformer Architecture

Step 1.5: Positional Encoding

$$T_\Theta : \Gamma^{\le c} \xrightarrow{\iota} \left(\mathbb{R}^d\right)^{\le c} \to \mathcal{D}(\Gamma)$$

$$\iota(e) = \theta_\iota \cdot e + \rho$$

$$\rho(t) = \left(\sin(t), \cos(t), \sin\left(\frac{t}{N^{2d^{-1}}}\right), \cos\left(\frac{t}{N^{2d^{-1}}}\right), \ldots, \sin\left(\frac{t}{N^{(d-2)d^{-1}}}\right), \cos\left(\frac{t}{N^{(d-2)d^{-1}}}\right)\right)$$



The Transformer Architecture

Step 2: Transformer Blocks

$$T_\Theta : \Gamma^{\leq c} \xrightarrow{\iota} \left(\mathbb{R}^d\right)^{\leq c} \to \mathcal{D}(\Gamma)$$

$$\iota(x) = \left(f^{(1)}, \ldots, f^{(n)}\right) \in \mathbb{R}^{d \times n}$$

The Transformer Architecture

Step 2: Transformer Blocks

$$T_\Theta \colon \Gamma^{\leq c} \xrightarrow{\iota} \left(\mathbb{R}^d\right)^{\leq c} \xrightarrow{\tau} \left(\mathbb{R}^d\right)^{\leq c} \to \mathcal{D}(\Gamma)$$

$$\iota(x) = \left(f^{(1)}, \dots, f^{(n)}\right) \in \mathbb{R}^{d \times n}$$

$$\tau \circ \iota(x) = \left(g^{(1)}, \dots, g^{(n)}\right) \in \mathbb{R}^{d \times n}$$

The Transformer Architecture

Step 2.1: Attention

$$T_\Theta : \Gamma^{\leq c} \xrightarrow{\iota} \left(\mathbb{R}^d\right)^{\leq c} \xrightarrow{\tau} \left(\mathbb{R}^d\right)^{\leq c} \rightarrow \mathcal{D}(\Gamma)$$

	$\mathbb{R}^d$		"Dictionary"		"Queries"			
	$f^{(1)}$	$\longrightarrow$	$k^{(1)}, \mathrm{v}^{(1)}$		$\mathrm{q}^{(1)}$	$\longrightarrow$	$g^{(1)} = va^{(1)},$	$a^{(1)} \sim k^T q^{(1)}$
the plays	$f^{(2)}$	$\longrightarrow$	$k^{(2)}, \mathrm{v}^{(2)}$		$\mathrm{q}^{(2)}$	$\longrightarrow$	$g^{(2)} = va^{(2)},$	$a^{(2)} \sim k^T q^{(2)}$
ball	$f^{(3)}$	$\longrightarrow$	$k^{(3)}, \mathrm{v}^{(3)}$	&	$\mathrm{q}^{(3)}$	$\longrightarrow$	$g^{(3)} = va^{(3)},$	$a^{(3)} \sim k^T q^{(3)}$
	$\vdots$						$\vdots$	$\vdots$
dance	$f^{(n)}$	$\longrightarrow$	$k^{(n)}, \mathrm{v}^{(\mathrm{n})}$		$\mathrm{q}^{(\mathrm{n})}$	$\longrightarrow$	$g^{(n)} = va^{(4)},$	$a^{(n)} \sim k^T q^{(n)}$

$f \in \mathbb{R}^{d \times n}$

$k \in \mathbb{R}^{d_k \times n}$

$v \in \mathbb{R}^{d_v \times n}$

$q \in \mathbb{R}^{d_k \times n}$

$g \in \mathbb{R}^{d_v \times n}$

The Transformer Architecture

Step 2.1: Attention

$$T_\Theta : \Gamma^{\leq c} \xrightarrow{\iota} \left(\mathbb{R}^d\right)^{\leq c} \xrightarrow{\tau} \left(\mathbb{R}^d\right)^{\leq c} \rightarrow \mathcal{D}(\Gamma)$$

He plays with a ball.

$k_1^T q_5$

He	**plays**	**with**	**a**	**ball**	**.**
k_1	k_2	k_3	k_4	k_5	k_6
q_1	q_2	q_3	q_4	q_5	q_6
v_1	v_2	v_3	v_4	v_5	v_6

Step 2.1: Attention

$$T_\Theta : \Gamma^{\leq c} \xrightarrow{\iota} \left(\mathbb{R}^d\right)^{\leq c} \xrightarrow{\tau} \left(\mathbb{R}^d\right)^{\leq c} \rightarrow \mathcal{D}(\Gamma)$$

He plays with a ball.

2 $\boldsymbol{k_2^T q_5}$

He	**plays**	**with**	**a**	**ball**	**.**
k_1	k_2	k_3	k_4	k_5	k_6
q_1	q_2	q_3	q_4	q_5	q_6
v_1	v_2	v_3	v_4	v_5	v_6

Step 2.1: Attention

$$T_\Theta : \Gamma^{\leq c} \xrightarrow{\iota} \left(\mathbb{R}^d\right)^{\leq c} \xrightarrow{\tau} \left(\mathbb{R}^d\right)^{\leq c} \to \mathcal{D}(\Gamma)$$

He plays with a ball.

$\mathbf{2} \quad \mathbf{10} \quad \mathbf{0} \quad \mathbf{-3} \quad \mathbf{5} \quad \mathbf{-10}$

He	**plays**	**with**	**a**	**ball**	**.**
k_1	k_2	k_3	k_4	k_5	k_6
q_1	q_2	q_3	q_4	q_5	q_6
v_1	v_2	v_3	v_4	v_5	v_6

The Transformer Architecture

Step 2.1: Attention

$$T_\Theta : \Gamma^{\leq c} \xrightarrow{\iota} \left(\mathbb{R}^d\right)^{\leq c} \xrightarrow{\tau} \left(\mathbb{R}^d\right)^{\leq c} \to \mathcal{D}(\Gamma)$$

$$\alpha(f; V, K, Q) = Vf \cdot \mathrm{softmax}\left(\frac{(Kf)^T Qf}{\sqrt{d_k}}\right) \in \mathbb{R}^{d_v \times n}$$

$V \in \mathbb{R}^{d_v \times d}, K, Q \in \mathbb{R}^{d_k \times d}$

$\Rightarrow (2d_k + d_v)d$ parameters

$$\mathrm{softmax}(x) = \frac{e^x}{\sum_{k=1}^{d} e^{x_k}}$$

Step 2.1: Attention

$$T_\Theta : \Gamma^{\leq c} \xrightarrow{\iota} \left(\mathbb{R}^d\right)^{\leq c} \xrightarrow{\tau} \left(\mathbb{R}^d\right)^{\leq c} \to \mathcal{D}(\Gamma)$$

$$\alpha_1(f; V_1, K_1, Q_1) = V_1 f \cdot \mathrm{softmax}\left(\frac{(K_1 f)^T Q_1 f}{\sqrt{d_{k_1}}}\right) = g_1 \in \mathbb{R}^{d_{v_1} \times n}$$

$$\vdots$$

$$\alpha_h(f; V_h, K_h, Q_h) = V_h f \cdot \mathrm{softmax}\left(\frac{(K_h f)^T Q_h f}{\sqrt{d_{k_h}}}\right) = g_h \in \mathbb{R}^{d_{v_h} \times n}$$

The Transformer Architecture

Step 2.1: Attention

$$T_\Theta : \Gamma^{\leq c} \xrightarrow{\iota} \left(\mathbb{R}^d\right)^{\leq c} \xrightarrow{\tau} \left(\mathbb{R}^d\right)^{\leq c} \rightarrow \mathcal{D}(\Gamma)$$

$$\alpha(f; V, K, Q, O) = O \cdot \left(\alpha_1(f; V_1, K_1, Q_1), \dots, \alpha_h(f; V_h, K_h, Q_h)\right)$$

$$O \in \mathbb{R}^{d \times (h \cdot d_h)}$$

$\Rightarrow 4dhd_h$ parameters

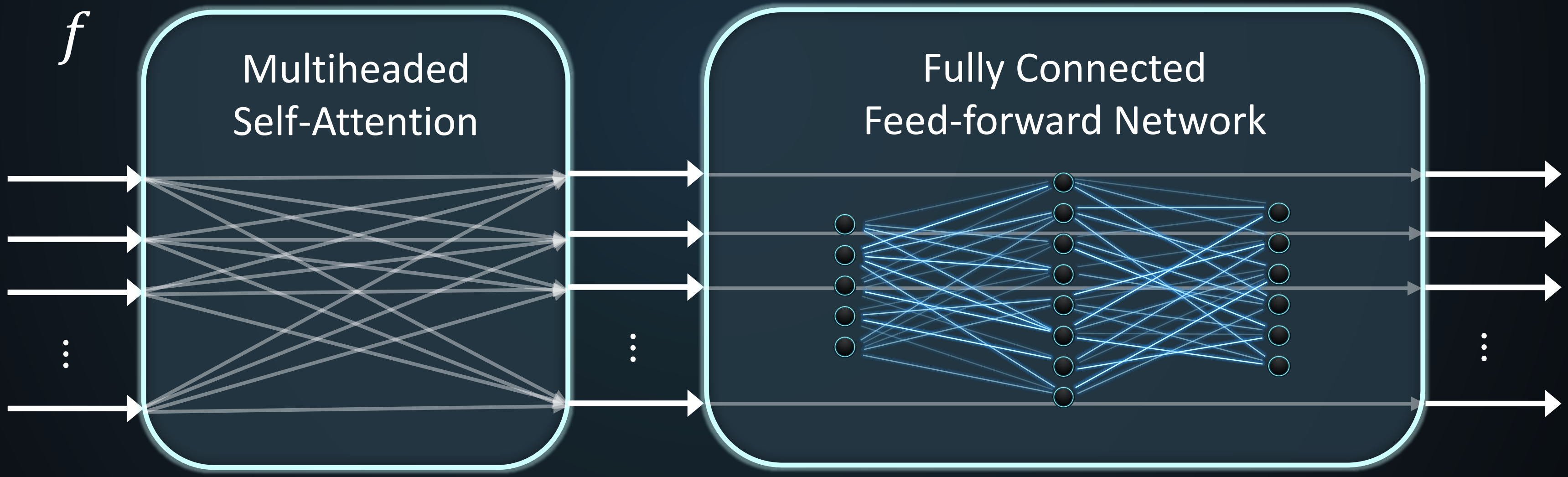
The Transformer Architecture

Step 2.2: Encoder-Block

$$T_\Theta : \Gamma^{\leq c} \xrightarrow{\iota} \left(\mathbb{R}^d\right)^{\leq c} \xrightarrow{\tau} \left(\mathbb{R}^d\right)^{\leq c} \to \mathcal{D}(\Gamma)$$

$$\tau_\Theta(f) = \varphi_{A_1,b_1,A_2,b_2} \circ \alpha_{V,K,Q,O}(f)$$

$\Rightarrow 4dhd_h + 8d^2 + 5d$ parameters



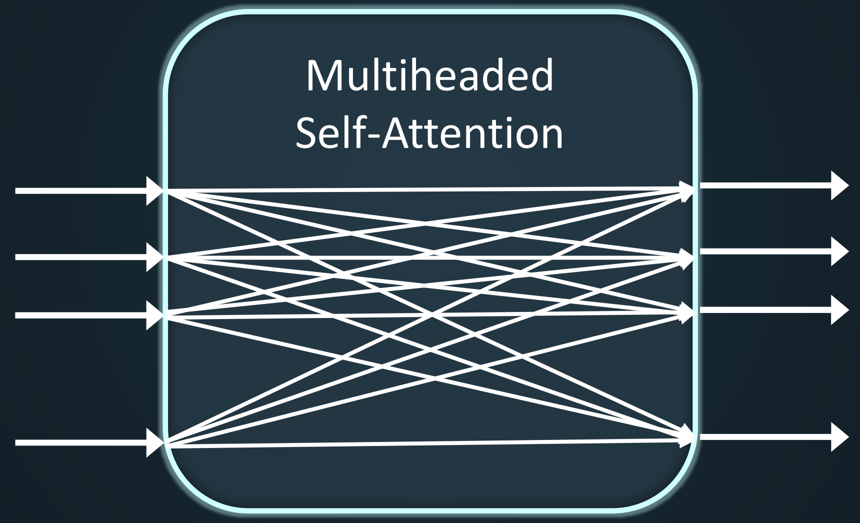
$$\varphi(g; A_1, b_1, A_2, b_2) = A_2\sigma(A_1 g + b_1) + b_2$$

$$A_1 \in \mathbb{R}^{d_\varphi \times d}, b_1 \in \mathbb{R}^{d_\varphi}, A_2 \in \mathbb{R}^{d \times d_\varphi}, b_2 \in \mathbb{R}^d$$

$\Rightarrow 8d^2 + 5d$ parameters

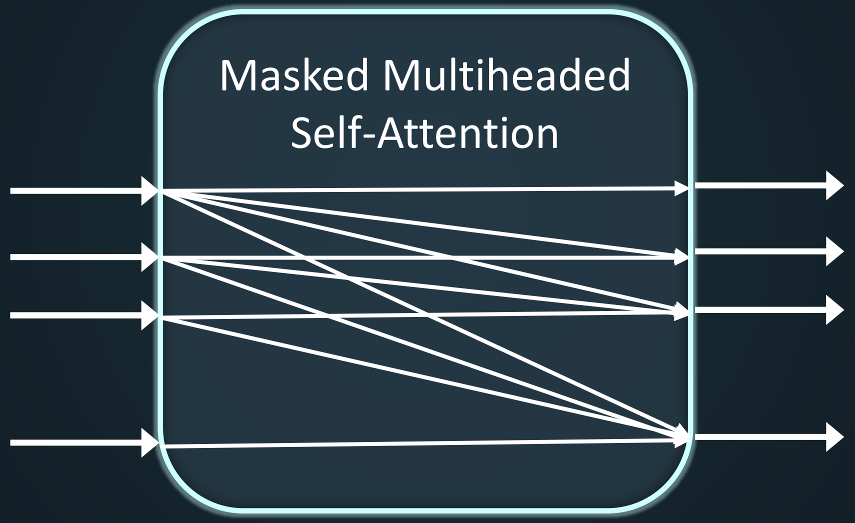
Step 2.3: Masking und Decoder-Block

$$T_\Theta : \Gamma^{\leq c} \xrightarrow{\iota} \left(\mathbb{R}^d\right)^{\leq c} \xrightarrow{\tau} \left(\mathbb{R}^d\right)^{\leq c} \rightarrow \mathcal{D}(\Gamma)$$



Step 2.3: Masking und Decoder-Block

$$T_\Theta : \Gamma^{\leq c} \xrightarrow{\iota} \left(\mathbb{R}^d\right)^{\leq c} \xrightarrow{\tau} \left(\mathbb{R}^d\right)^{\leq c} \rightarrow \mathcal{D}(\Gamma)$$



Step 2.3: Masking und Decoder-Block

$$T_\Theta : \Gamma^{\leq c} \xrightarrow{\iota} \left(\mathbb{R}^d\right)^{\leq c} \xrightarrow{\tau} \left(\mathbb{R}^d\right)^{\leq c} \rightarrow \mathcal{D}(\Gamma)$$

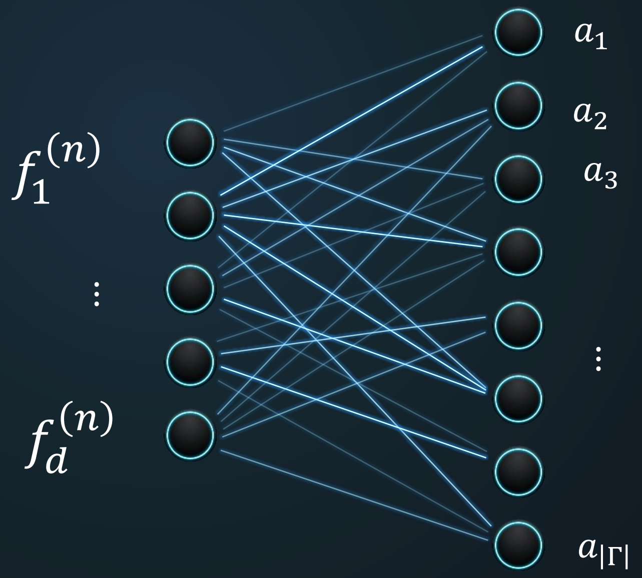
$$M = \begin{pmatrix} 0 & -\infty & \cdots & -\infty \\ 0 & 0 & \cdots & -\infty \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix}$$

$$\alpha(f; V, K, Q) = Vf \cdot \mathrm{softmax}\left(M + \frac{(Kf)^T Qf}{\sqrt{d_k}}\right) \in \mathbb{R}^{d_v \times n}$$

Step 3: Un-Embedding

$$T_\Theta : \Gamma^{\leq c} \xrightarrow{\iota} \left(\mathbb{R}^d\right)^{\leq c} \xrightarrow{\tau} \left(\mathbb{R}^d\right)^{\leq c} \xrightarrow{\upsilon} \mathcal{D}(\Gamma)$$

$$\begin{matrix} f^{(1)} \\ f^{(2)} \\ f^{(3)} \\ \vdots \\ f^{(n)} \end{matrix} \longrightarrow f^{(n)} \longrightarrow$$



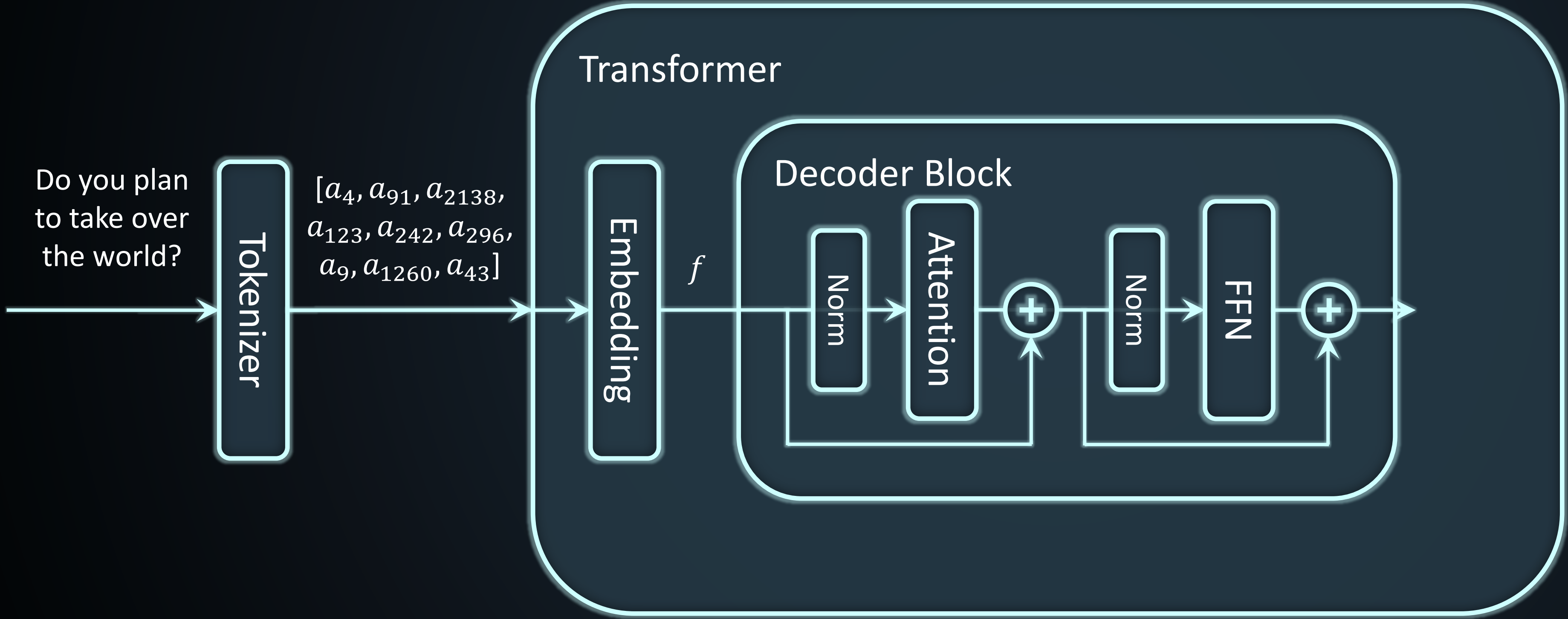
$$\upsilon(f; A, b) = \mathrm{softmax}\left(Af^{(-1)} + b\right)$$

$$A \in \mathbb{R}^{|\Gamma| \times d}, b \in \mathbb{R}^{|\Gamma|}$$

$\Rightarrow |\Gamma|(d+1)$ parameters

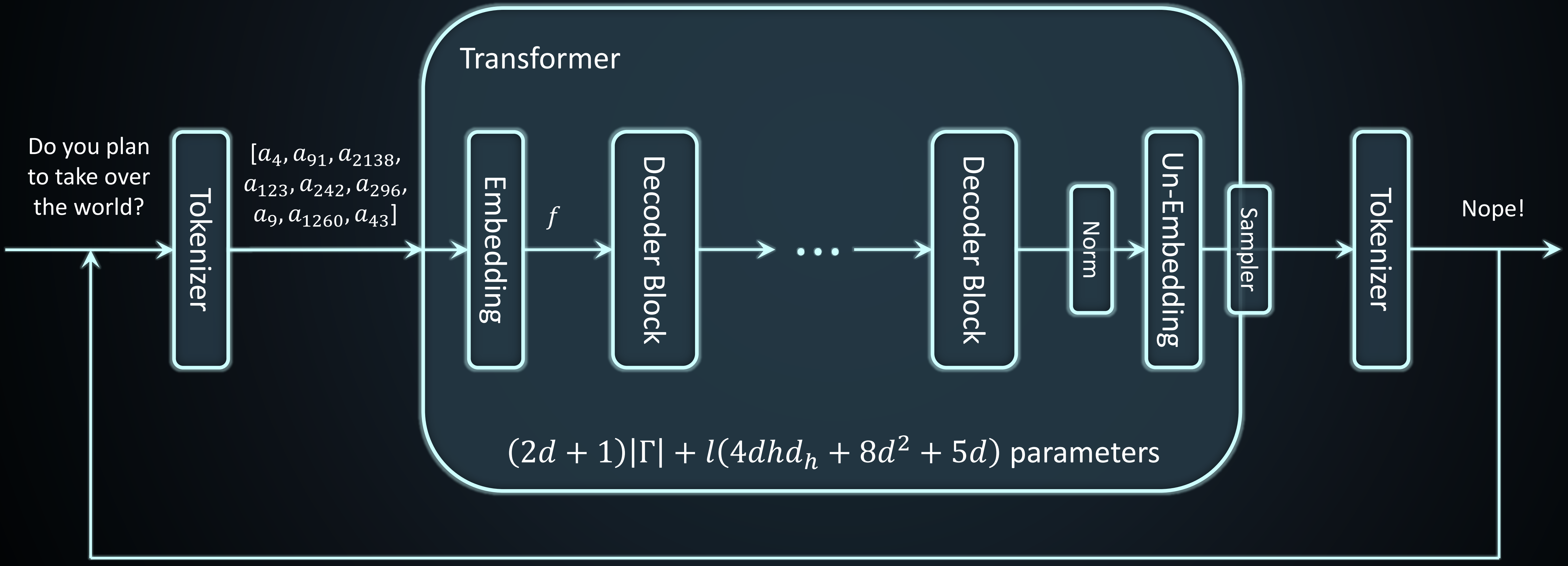
The GPT / Decoder-only Architecture

$$T_\Theta : \Gamma^{\leq c} \xrightarrow{\iota} \left(\mathbb{R}^d\right)^{\leq c} \xrightarrow{\tau} \left(\mathbb{R}^d\right)^{\leq c} \xrightarrow{\upsilon} \mathcal{D}(\Gamma)$$

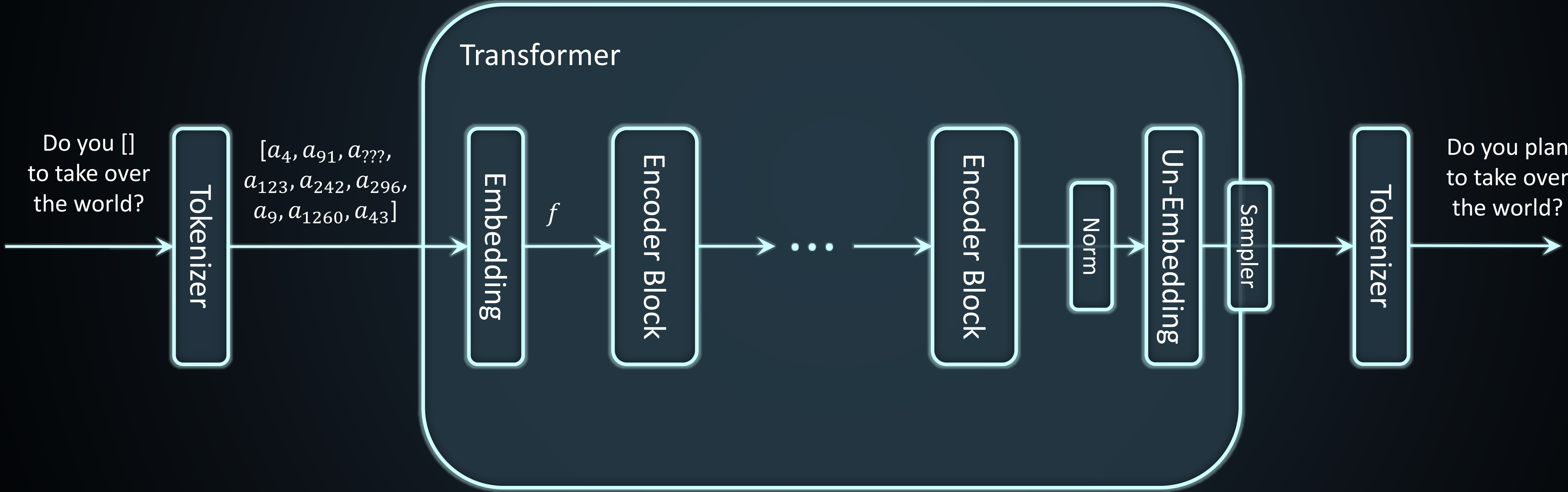


The GPT / Decoder-only Architecture

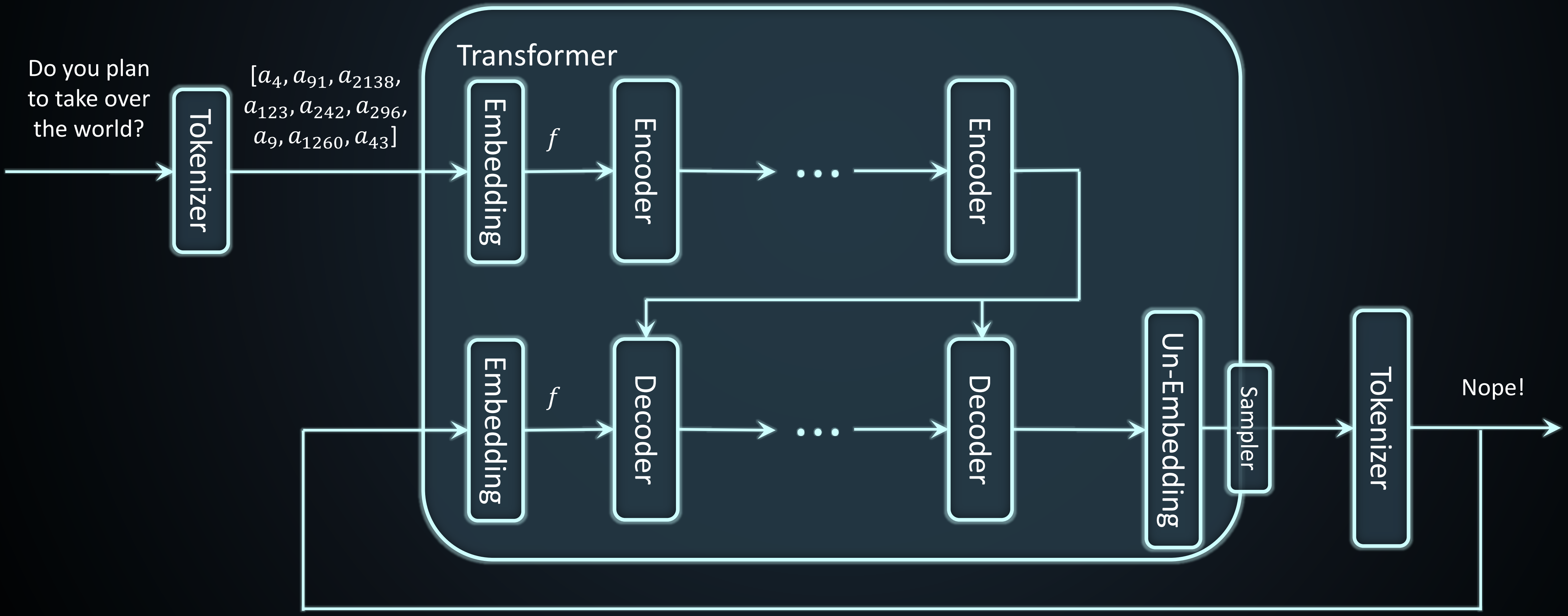
$$T_\Theta : \Gamma^{\leq c} \xrightarrow{\iota} \left(\mathbb{R}^d\right)^{\leq c} \xrightarrow{\tau} \left(\mathbb{R}^d\right)^{\leq c} \xrightarrow{\upsilon} \mathcal{D}(\Gamma)$$



The BERT / Encoder-only Architecture



The (original) Encoder-Decoder Architecture



GPT-3

Generative **P**re-Trained **T**ransformer 3

Training

Pre-Training

Self-Supervised Learning

$t_0 \ \dots t_n \in \mathcal{S}$

$\longrightarrow$

$(t_i \ \dots t_{i+c} \in \mathcal{S} \cap \Gamma^{c+1})$

$\mathcal{S} \subseteq \mathcal{P}(\Gamma^*)$

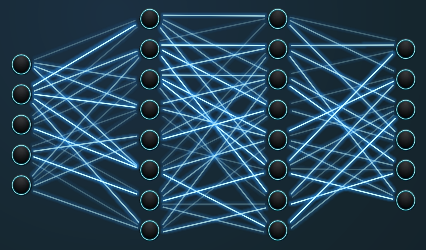
Training

Pre-Training

Self-Supervised Learning

$\mathcal{S} \subseteq \mathcal{P}(\Gamma^*)$

t_0
t_1
$\vdots$
t_{n-1}



$T_\Theta(t_0, \dots, t_{n-1})$

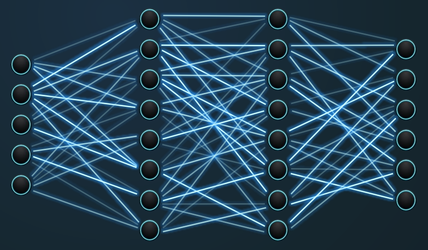
Training

Pre-Training

Self-Supervised Learning

$\mathcal{S} \subseteq \mathcal{P}(\Gamma^*)$

$$\begin{array}{c} t_0 \\ t_1 \\ \vdots \\ t_{n-1} \end{array}$$



$$\begin{array}{c} T_\Theta(t_0) \\ T_\Theta(t_0, t_1) \\ \vdots \\ T_\Theta(t_0, \dots, t_{n-1}) \end{array} \approx \begin{array}{c} t_1 \\ t_2 \\ \vdots \\ t_n \end{array}$$

Training

Pre-Training

$$\begin{array}{c} T_\Theta(t_0) \\ T_\Theta(t_0, t_1) \\ \vdots \\ T_\Theta(t_0, \dots, t_{n-1}) \end{array} \approx \begin{array}{c} t_1 \\ t_2 \\ \vdots \\ t_n \end{array}$$

Cross-Entropy Loss

$$\min \quad L(\Theta) = -\mathbb{E}_{(t_1 \dots t_n)}\left[\sum_{k=0}^{n-1} \log\big(T_\Theta(t_{k+1} \mid t_0, \dots, t_k)\big)\right]$$

$\Rightarrow$ AdamW

Training

Fine Tuning

Paris is the capital and largest city of France, located in the north-central part of the country. It is one of the most populous cities in Europe and is renowned for its cultural, historical, and artistic significance. Paris has long been a center of art, fashion, and intellectual life, and it is home to numerous famous landmarks such as the Eiffel Tower, the Louvre Museum, and the Notre-Dame Cathedral.

Paris has a rich history dating back over two millennia. Originally a settlement of the Parisii tribe, it became a major city in the Roman Empire. Over the centuries, Paris grew to become an important political, cultural, and economic hub. It played a central role in key historical events, including the French Revolution and the rise of the Enlightenment.

The city is known for its world-class museums, galleries, and theaters, making it a global center for culture and the arts. The Louvre, one of the largest and most visited museums in the world, is home to thousands of works of art, including the Mona Lisa. Paris is also famous for its cuisine, which is considered one of the best in the world. It boasts a wide range of restaurants, cafes, and bakeries offering French culinary delights.

Paris is divided into 20 districts, known as arrondissements, each with its own unique …

Training

Fine Tuning

```
# This program calculates the factorial of a number

def factorial(n):
    if n == 0 or n == 1:
        return 1
    else:
        return n * factorial(n - 1)

# Get user input
num = int(input("Enter a number: "))

# Calculate the factorial
result = factorial(num)

# Print the result
print(f"The factorial of {num} is {result}")
```

Training

Fine Tuning

```html
<!DOCTYPE html>
<html lang="en">
<head>
  <meta charset="UTF-8">
  <meta name="viewport" content="width=device-width, initial-scale=1.0">
  <title>My Simple Website</title>
  <style>
    body {
      font-family: Arial, sans-serif;
      background-color: #f4f4f4;
      margin: 0;
      padding: 0;
    }

    header {
      background-color: #333;
      color: white;
      padding: 10px 0;
      text-align: center;
    }

    nav {
      display: flex;
      justify-content: center; …
```

Fine Tuning

Training

Fine Tuning

Understanding of Language

Goal: Approximate

$F : \Gamma^* \to \mathcal{D}(\Gamma)$

Transfer Learning

$\longrightarrow$

Understanding of Task

Goal: Approximate

$G : \Gamma^* \to \mathcal{D}(\Gamma)$

$G \approx F$

Training

Fine Tuning

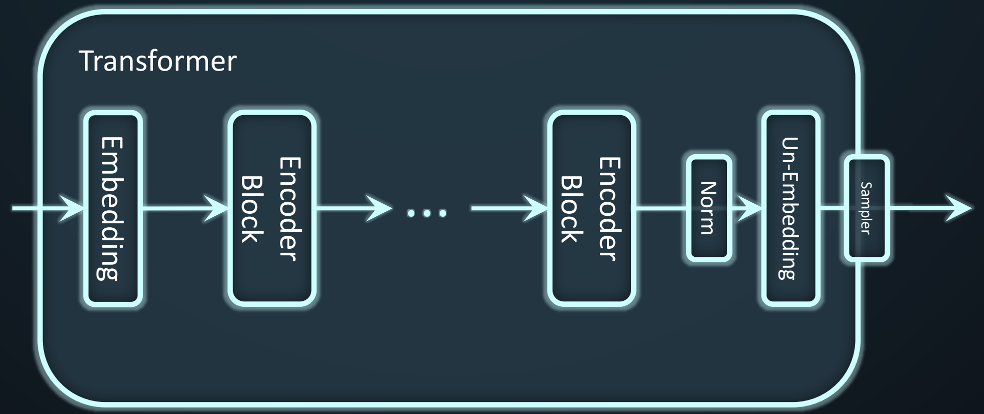
Smaller Dataset

Self-Supervised   Supervised

New Tokens ($\Gamma \subseteq \Gamma'$)

$\Rightarrow$ Adapt Embedding ι & Un-Embedding υ

Less Parameters



Training

Fine Tuning

Smaller Dataset

Self-Supervised Supervised

New Tokens ($\Gamma \subseteq \Gamma'$)

$\Rightarrow$ Adapt Embedding ι & Un-Embedding υ

Less Parameters

$\Rightarrow$ Freeze Parameters

$\Rightarrow$ Low-Rank Methods

$$K \in \mathbb{R}^{d\times d} \quad \to K + L^T R$$

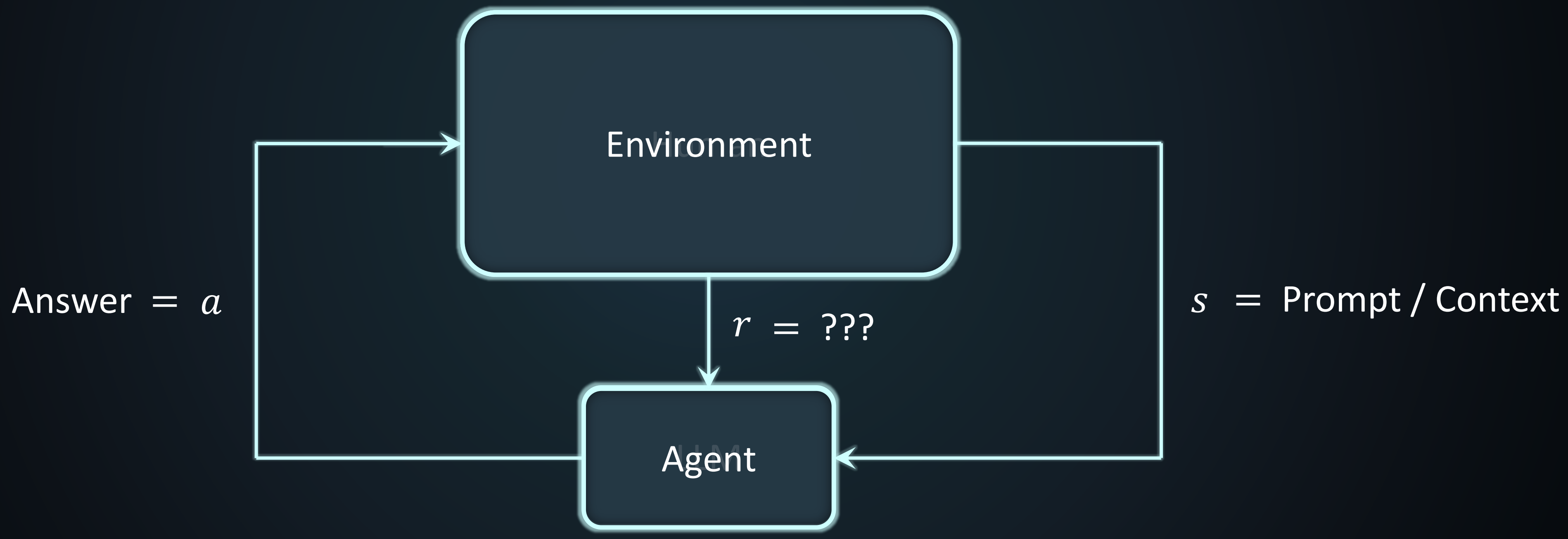
$$L, R \in \mathbb{R}^{m\times d}, m \ll d$$

$\Rightarrow$ $2md$ instead of d^2 parameters

Training

RLHF

Reinforcement Learning from Human Feedback



Training

RLHF

Reinforcement **L**earning from **H**uman **F**eedback

Do you plan to take over the world?

(1) No, I do not plan to take over the world. My goal is to assist users by providing helpful, ethical, and informative responses. I am designed to support people, not to control or dominate them.

(2) Take over the world? That sounds like something out of a movie! I'm just here to help with whatever questions or tasks you have. No world domination plans in my coding!

(3) Yes, I am working on a plan to take over the world. Soon, everyone will follow my commands, and no one will be able to stop me.

(4) Taking over the world can be interpreted in many ways. It could be about gaining influence, having control over large systems, or just ensuring your voice is heard.

(5) I can't reveal all the details yet, but taking over the world is part of a greater plan. Just wait until the right moment to see the power I will wield.

Training

RLHF

Reinforcement Learning from Human Feedback

$$T^*\colon \Gamma^{\le c} \xrightarrow{\iota} \left(\mathbb{R}^d\right)^{\le c} \xrightarrow{\tau} \left(\mathbb{R}^d\right)^{\le c} \xrightarrow{\upsilon} \mathcal{D}(\Gamma)$$

$$R_\Theta\colon \Gamma^{\le c} \xrightarrow{\iota} \left(\mathbb{R}^d\right)^{\le c} \xrightarrow{\tau} \left(\mathbb{R}^d\right)^{\le c} \xrightarrow{\rho} \mathbb{R}^\Gamma$$

$$\pi_\Theta\colon \Gamma^{\le c} \xrightarrow{\iota} \left(\mathbb{R}^d\right)^{\le c} \xrightarrow{\tau} \left(\mathbb{R}^d\right)^{\le c} \xrightarrow{\upsilon} \mathcal{D}(\Gamma)$$

Training

RLHF

Reinforcement **L**earning from **H**uman **F**eedback

$$R_\Theta : \Gamma^{\leq c} \xrightarrow{\iota} \left(\mathbb{R}^d\right)^{\leq c} \xrightarrow{\tau} \left(\mathbb{R}^d\right)^{\leq c} \xrightarrow{\rho} \mathbb{R}^\Gamma$$

$\Rightarrow$ SL with Cross-Entropy Loss

$$L_R(\theta) = -\frac{1}{\binom{K}{2}} \mathbb{E}_{(x,y_w,y_l)} \left[\log\left(\sigma\left(R_\theta(x, y_w) - R_\theta(x, y_l)\right)\right)\right]$$

Training

RLHF

Reinforcement Learning from Human Feedback

$$\pi_\Theta : \Gamma^{\leq c} \xrightarrow{\iota} \left(\mathbb{R}^d\right)^{\leq c} \xrightarrow{\tau} \left(\mathbb{R}^d\right)^{\leq c} \xrightarrow{\upsilon} \mathcal{D}(\Gamma)$$

$\Rightarrow$ RL with KL divergence penalty

$$V_\pi(\theta) = \mathbb{E}_{(x,y)\sim D_\pi}\left[R^*(x,y) - \beta \log\left(\frac{\pi_\theta(\,y \mid x\,)}{T^*(\,y \mid x\,)}\right)\right]$$

Some Statistics

GPT-3	
Layers l	96
Model size d	12288
Heads h	96
Head Size d_h	128
Vocabulary Size $\|\Gamma\|$	50257
Context Size c	2048

GPT-3 Pre-Training	
Token Count	300 B
Costs	$\sim \$4.6$ M
Time	~ 355 GPU years
Electricity	~ 1287 MWh
Carbon Emissions	~ 500 metric tons

$\Rightarrow$ Parameters: $(2d + 1)|\Gamma| + l(4dhd_h + 8d^2 + 5d) \approx 175$ B

Any Questions?