FLORIDA ATLANTIC

COT 6930 - Generative Intelligence
and Software Development Lifecycle

Topic 2 -
Generative Intelligence Models

Dr. Fernando Koch
kochf@fau.edu
http://www.fernandokoch.me

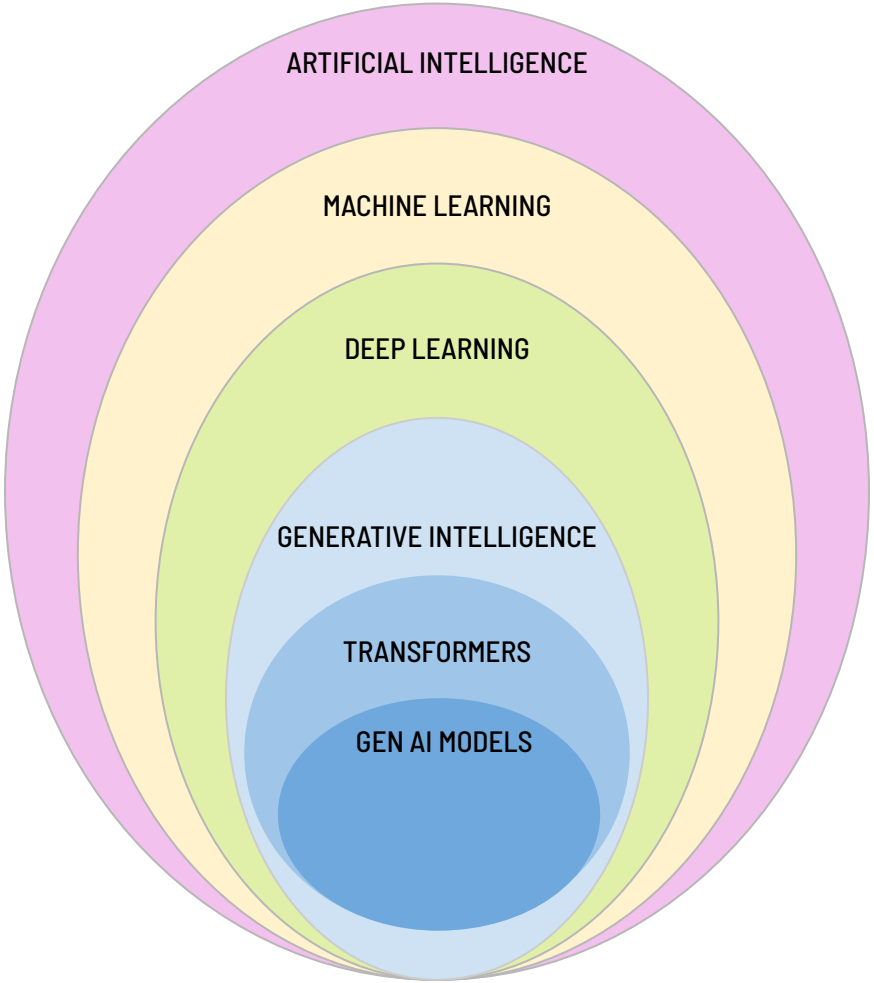
Agenda

- Fundamentals of Large Generative Models
- How LGMs are built and trained
- Controlling the behavior of the Models
- Exercises

Our Key Question

How do Large Generative Models operate, and how can they be configured and applied effectively for software-related tasks?

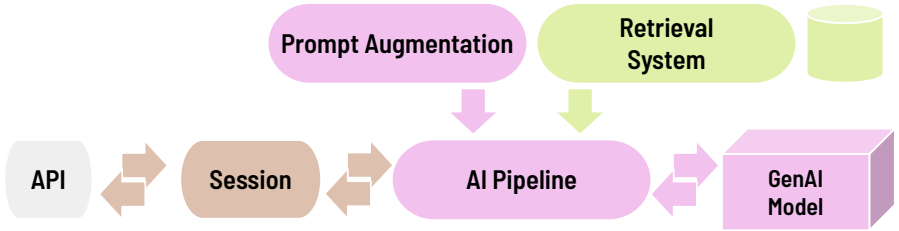
What is Generative AI?



Generative Intelligence System

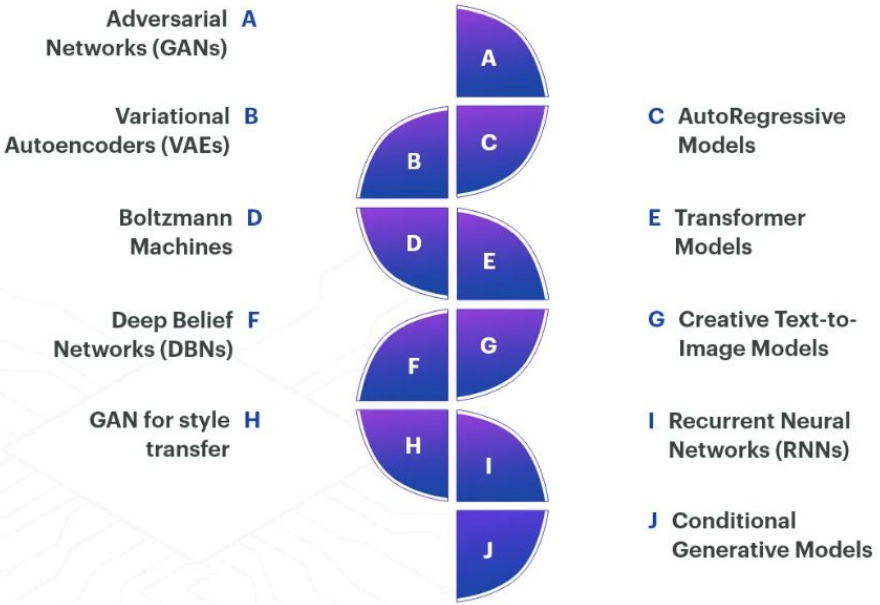
End-to-end systems with autonomous content generation

- Session-aware workflows.
- Automated prompt engineering.
- Context-aware content generation
- Adaptive behaviour combining orchestration, memory, retrieval augmentation, others.
- End-to-end automation for LGM consumption.



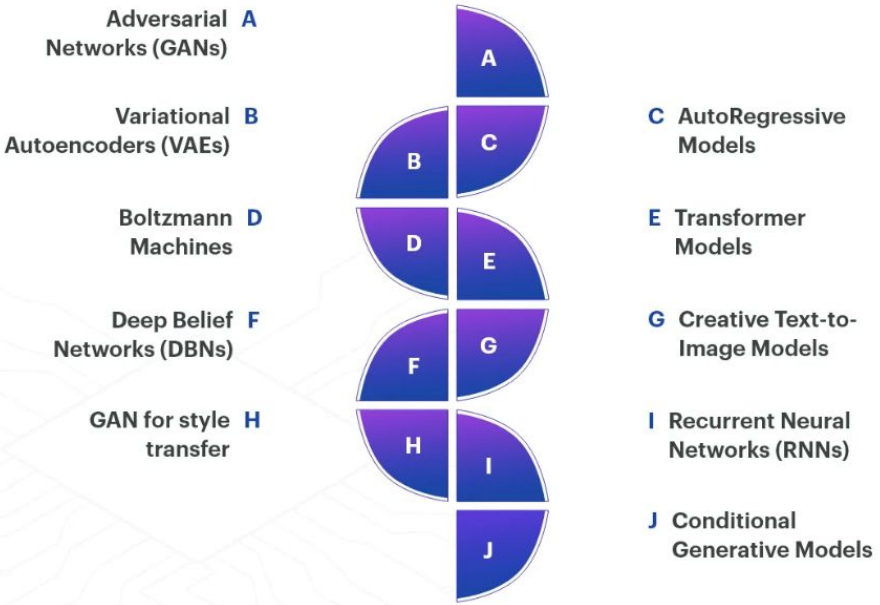
Types of Generative Models

- **GAN**: Competing networks for realistic images & videos.
- **VAE**: Latent encoding for anomaly detection & reconstruction.
- **AutoRegressive Models**: Sequential prediction for text, speech & music.
- **Boltzmann Machines**: Energy-based models for probabilistic reasoning.
- **Transformers**: Attention-driven for language, vision & multimodal tasks.



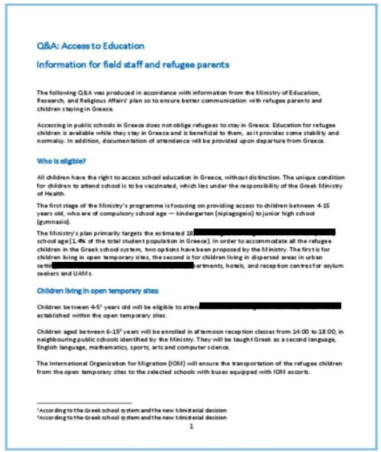
Types of Generative Models (2/2)

- **DBNs**: Layered probabilistic models for feature extraction.
- **Text-to-Image Models:** Generate visuals from text for art & design.
- **Style Transfer GANs**: Adapt visuals for artistic or domain-specific style.
- **RNNs:** Sequence learners for time-series & generative text/music.
- **Conditional Generative Models:** Guided generation for controlled outputs.



The Process of Creating Generative Models

(1) Inference:
Generate the content to complete the document based on predictions by the Model.



Q&A: Access to Education

Information for field staff and refugee parents

The following Q&A was produced in accordance with information from the Ministry of Education, Research, and Religious Affairs' plan so to ensure better communication with refugee parents and children staying in Greece.

Accessing in public schools in Greece does not oblige refugees to stay in Greece. Education for refugee children is available while they stay in Greece and is beneficial to them, as it provides some stability and normalcy. In addition, documentation of attendance will be provided upon departure from Greece.

Who is eligible?

All children have the right to access school education in Greece, without distinction. The unique condition for children to attend school is to be vaccinated, which lies under the responsibility of the Greek Ministry of Health.

The first stage of the Ministry's programme is focusing on providing access to children between 4-15 years old, who are of compulsory school age — kindergarten (nipiagogeio) to junior high school (gymnasio).

The Ministry's plan primarily targets the estimated [redacted] In order to accommodate all the refugee children in the Greek school system, two options have been proposed by the Ministry. The first is for children living in open temporary sites, the second is for children living in dispersed areas in urban set[redacted]artments, hotels, and reception centres for asylum seekers and UAMs.

Children living in open temporary sites:

Children between 4-5[1] years old will be eligible to atten[redacted] established within the open temporary sites.

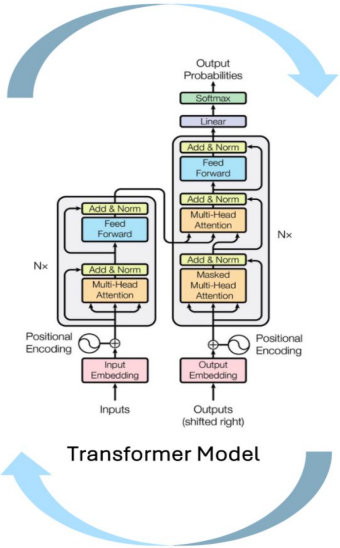
Children aged between 6-15[2] years will be enrolled in afternoon reception classes from 14:00 to 18:00, in neighbouring public schools identified by the Ministry. They will be taught Greek as a second language, English language, mathematics, sports, arts and computer science.

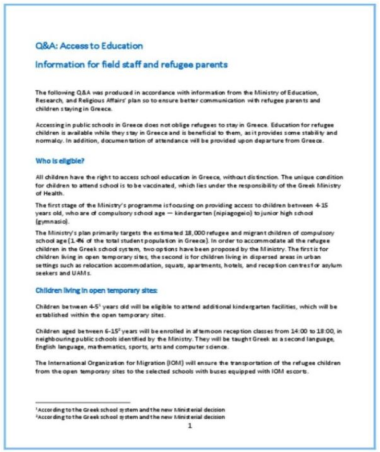
The International Organization for Migration (IOM) will ensure the transportation of the refugee children from the open temporary sites to the selected schools with buses equipped with IOM escorts.

[1] According to the Greek school system and the new ministerial decision
[2] According to the Greek school system and the new ministerial decision

1

(A) Document with Obfuscated Parts





Q&A: Access to Education

Information for field staff and refugee parents

The following Q&A was produced in accordance with information from the Ministry of Education, Research, and Religious Affairs' plan so to ensure better communication with refugee parents and children staying in Greece.

Accessing in public schools in Greece does not oblige refugees to stay in Greece. Education for refugee children is available while they stay in Greece and is beneficial to them, as it provides some stability and normalcy. In addition, documentation of attendance will be provided upon departure from Greece.

Who is eligible?

All children have the right to access school education in Greece, without distinction. The unique condition for children to attend school is to be vaccinated, which lies under the responsibility of the Greek Ministry of Health.

The first stage of the Ministry's programme is focusing on providing access to children between 4-15 years old, who are of compulsory school age — kindergarten (nipiagogeio) to junior high school (gymnasio).

The Ministry's plan primarily targets the estimated 18,000 refugee and migrant children of compulsory school age (1.4% of the total student population in Greece). In order to accommodate all the refugee children in the Greek school system, two options have been proposed by the Ministry. The first is for children living in open temporary sites, the second is for children living in dispersed areas in urban settings such as relocation accommodation, squats, apartments, hotels, and reception centres for asylum seekers and UAMs.

Children living in open temporary sites:

Children between 4-5[1] years old will be eligible to attend additional kindergarten facilities, which will be established within the open temporary sites.

Children aged between 6-15[2] years will be enrolled in afternoon reception classes from 14:00 to 18:00, in neighbouring public schools identified by the Ministry. They will be taught Greek as a second language, English language, mathematics, sports, arts and computer science.

The International Organization for Migration (IOM) will ensure the transportation of the refugee children from the open temporary sites to the selected schools with buses equipped with IOM escorts.

[1] According to the Greek school system and the new ministerial decision
[2] According to the Greek school system and the new ministerial decision

1

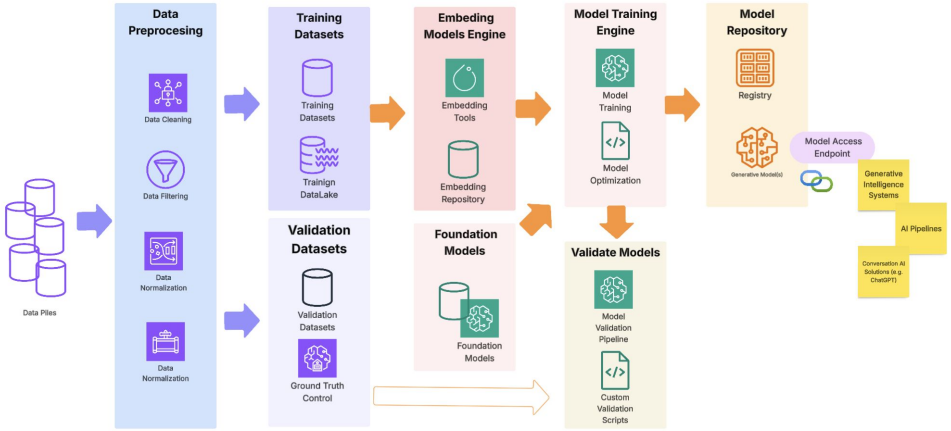
(B) Completed Document

(2) Training:
Learn how to complete the missing parts provided enough examples.

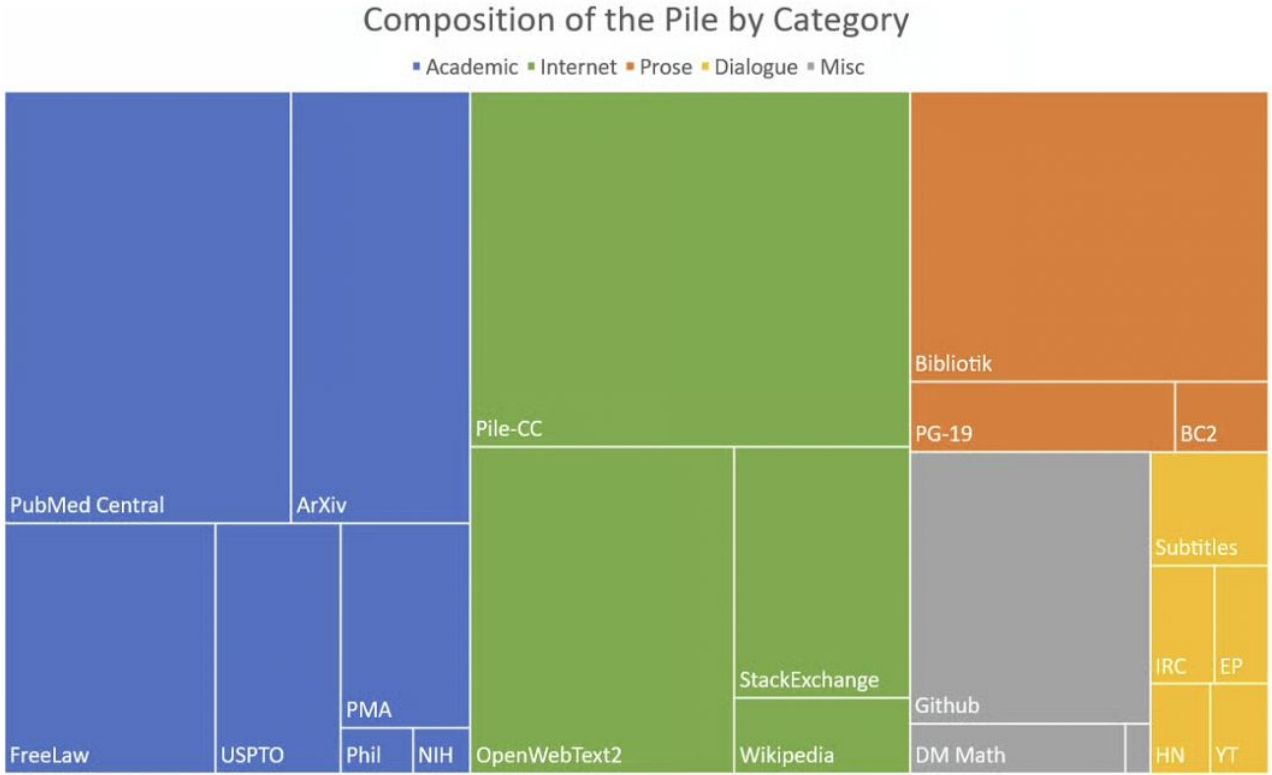
How to create Generative Intelligence Models?

From raw data to intelligent applications

- Structured pipelines for data cleaning, filtering, and normalization
- Training and validation using large-scale, curated datasets
- Model training, optimization, and validation.
- Foundation models for representation and generalization



Where the data comes from?

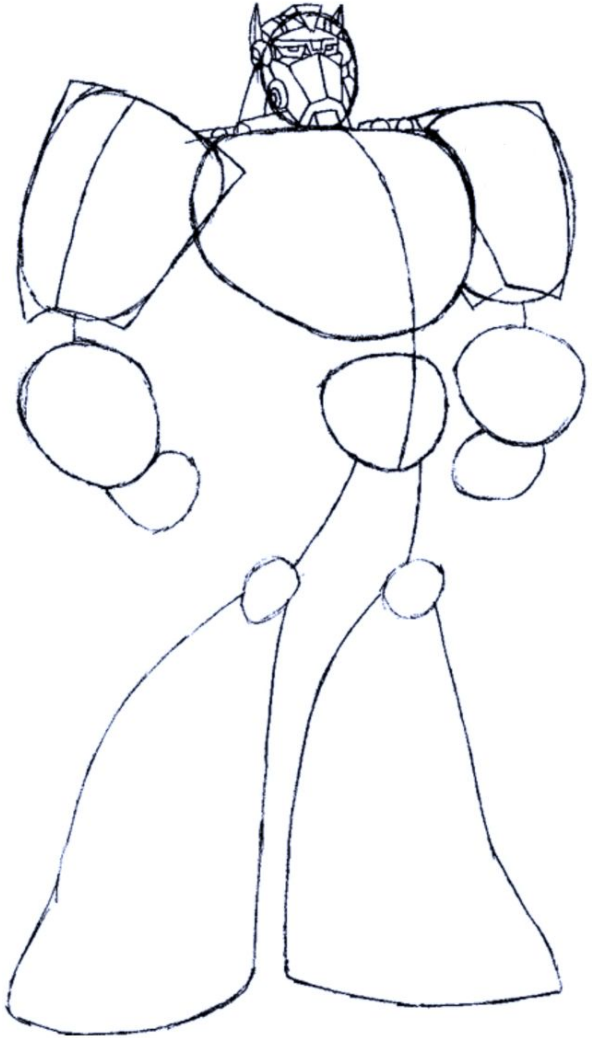


The Technology Behind Large Generative Models

Transformers

Transformers use self-attention to process entire sequences efficiently.

- Built on encoder-decoder architecture with self-attention
- Replaced RNNs and CNNs for sequence modeling
- Introduced in 2017 via “Attention Is All You Need”
- Enabled breakthroughs like BERT (2018) and GPT
- Scaled to ChatGPT (2022) and beyond for GenAI tasks



- It introduced the Transformer and proved attention alone is enough.
- Proposed a model architecture based purely on attention
- Eliminated recurrence and convolutions
- Enabled full parallelism during training
- Laid the foundation for GPT, BERT, and modern GenAI

https://arxiv.org/abs/1706.03762

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

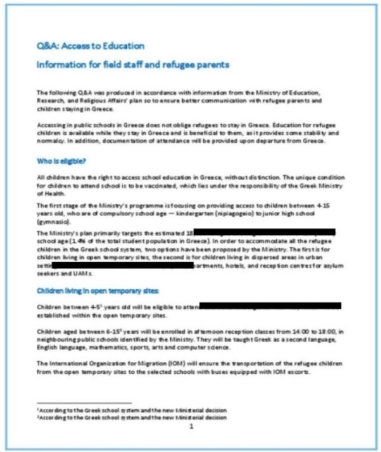
Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

The Process of Creating Generative Models

(1) Inference:
Generate the content to complete the document based on predictions by the Model.



Q&A: Access to Education

Information for field staff and refugee parents

The following Q&A was produced in accordance with information from the Ministry of Education, Research, and Religious Affairs' plan so to ensure better communication with refugee parents and children staying in Greece.

Accessing in public schools in Greece does not oblige refugees to stay in Greece. Education for refugee children is available while they stay in Greece and is beneficial to them, as it provides some stability and normalcy. In addition, documentation of attendance will be provided upon departure from Greece.

Who is eligible?

All children have the right to access school education in Greece, without distinction. The unique condition for children to attend school is to be vaccinated, which lies under the responsibility of the Greek Ministry of Health.

The first stage of the Ministry's programme is focusing on providing access to children between 4-15 years old, who are of compulsory school age — kindergarten (nipiagogeio) to junior high school (gymnasio).

The Ministry's plan primarily targets the estimated [redacted] In order to accommodate all the refugee children in the Greek school system, two options have been proposed by the Ministry. The first is for children living in open temporary sites, the second is for children living in dispersed areas in urban set[redacted]artments, hotels, and reception centres for asylum seekers and UAMs.

Children living in open temporary sites:

Children between 4-5[1] years old will be eligible to atten[redacted] established within the open temporary sites.

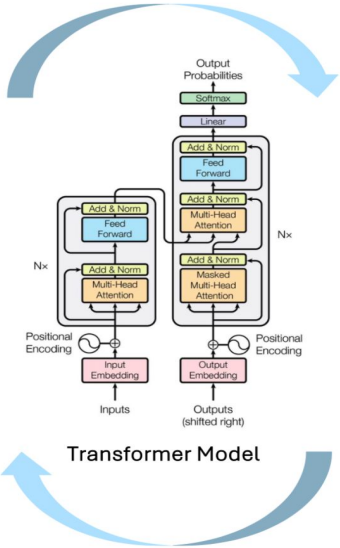
Children aged between 6-15[2] years will be enrolled in afternoon reception classes from 14:00 to 18:00, in neighbouring public schools identified by the Ministry. They will be taught Greek as a second language, English language, mathematics, sports, arts and computer science.

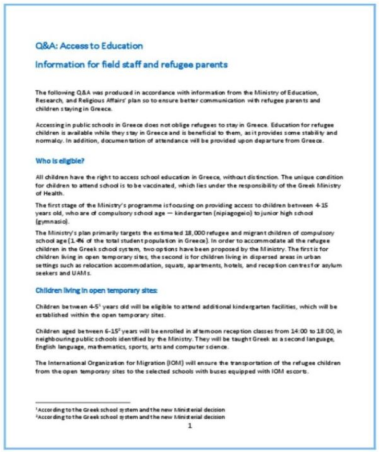
The International Organization for Migration (IOM) will ensure the transportation of the refugee children from the open temporary sites to the selected schools with buses equipped with IOM escorts.

[1] According to the Greek school system and the new ministerial decision
[2] According to the Greek school system and the new ministerial decision

1

(A) Document with Obfuscated Parts





Q&A: Access to Education

Information for field staff and refugee parents

The following Q&A was produced in accordance with information from the Ministry of Education, Research, and Religious Affairs' plan so to ensure better communication with refugee parents and children staying in Greece.

Accessing in public schools in Greece does not oblige refugees to stay in Greece. Education for refugee children is available while they stay in Greece and is beneficial to them, as it provides some stability and normalcy. In addition, documentation of attendance will be provided upon departure from Greece.

Who is eligible?

All children have the right to access school education in Greece, without distinction. The unique condition for children to attend school is to be vaccinated, which lies under the responsibility of the Greek Ministry of Health.

The first stage of the Ministry's programme is focusing on providing access to children between 4-15 years old, who are of compulsory school age — kindergarten (nipiagogeio) to junior high school (gymnasio).

The Ministry's plan primarily targets the estimated 18,000 refugee and migrant children of compulsory school age (1.4% of the total student population in Greece). In order to accommodate all the refugee children in the Greek school system, two options have been proposed by the Ministry. The first is for children living in open temporary sites, the second is for children living in dispersed areas in urban settings such as relocation accommodation, squats, apartments, hotels, and reception centres for asylum seekers and UAMs.

Children living in open temporary sites:

Children between 4-5[1] years old will be eligible to attend additional kindergarten facilities, which will be established within the open temporary sites.

Children aged between 6-15[2] years will be enrolled in afternoon reception classes from 14:00 to 18:00, in neighbouring public schools identified by the Ministry. They will be taught Greek as a second language, English language, mathematics, sports, arts and computer science.

The International Organization for Migration (IOM) will ensure the transportation of the refugee children from the open temporary sites to the selected schools with buses equipped with IOM escorts.

[1] According to the Greek school system and the new ministerial decision
[2] According to the Greek school system and the new ministerial decision

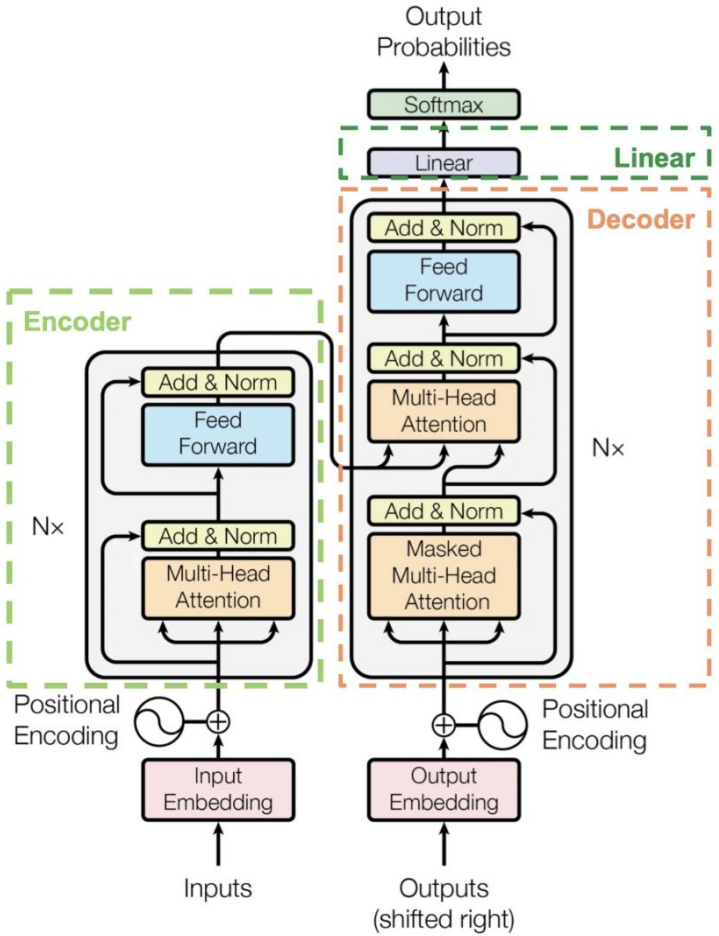
1

(B) Completed Document

(2) Training:
Learn how to complete the missing parts provided enough examples.

Transformer Architecture

- A modular, attention-driven design that models full sequences in parallel.
- Consists of an encoder-decoder architecture.
- Uses Multi-Head Attention for context understanding.
- Replaces recurrence with full self-attention.
- Enables parallel training and long-range dependency modeling.



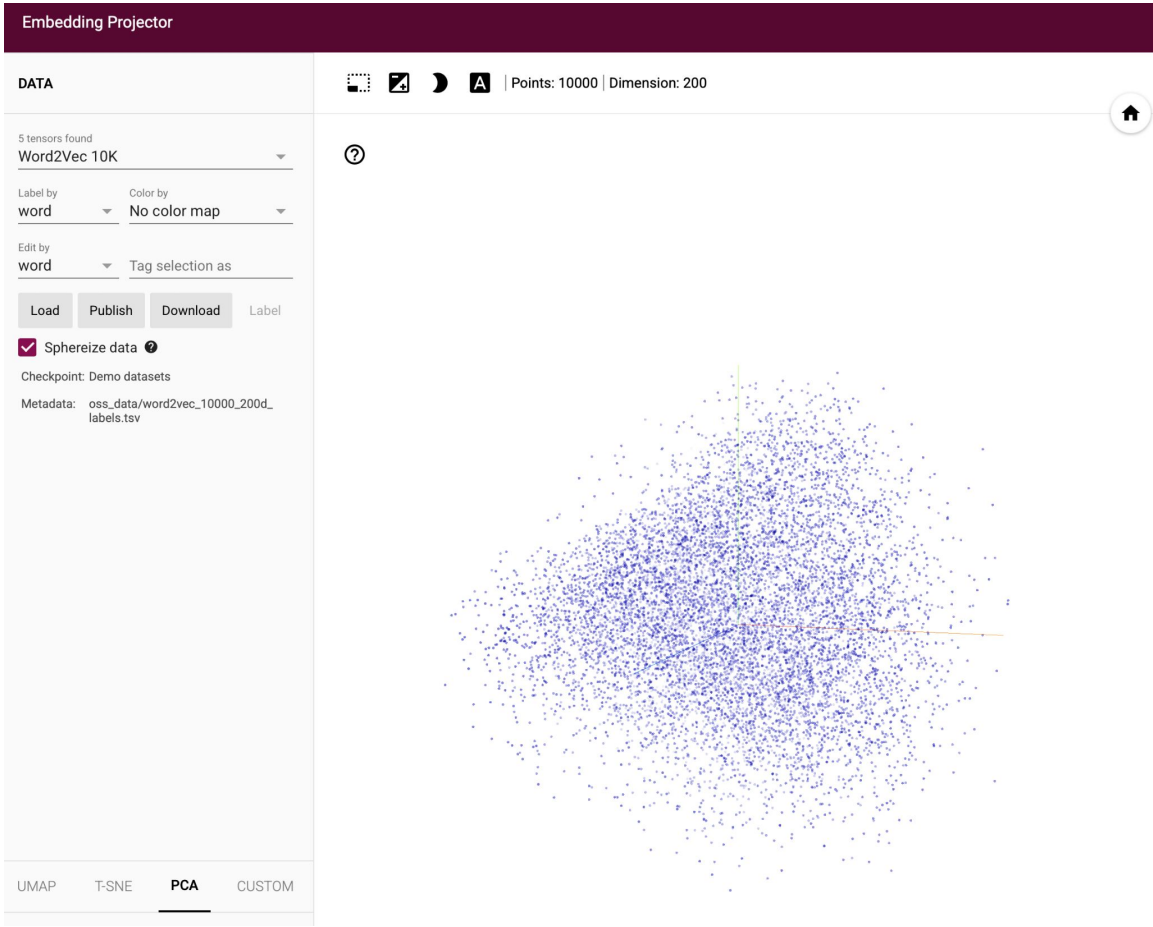
The Intelligent Foam Analogy

Analogy of The Intelligent Foam

A metaphor using bubbles and light to explain how prompts activate meaning.

- Bubbles represent concept spaces or semantic fields
- Prompts act as beams of light entering the foam
- Light scatters across bubbles, triggering latent knowledge
- Activation depends on learned structure and contextual overlap
- Helps visualize inference as resonance, not just sequence

Embedding Projector



See the ‘bubble’ by yourself:
https://projector.tensorflow.org

Different than Classical-ML Models

ML models operate within one concept space.
GenAI spans multiple concept spaces.

- ML models are task-specific and operate in isolated semantic bubbles
- Each ML model is trained on a narrow, single-purpose dataset
- ML is best for structured, repeatable, high-precision tasks.

- GenAI models integrate diverse, overlapping context spaces
- GenAI supports open-ended reasoning, abstraction, and generation

Different than DL Models

DL models operate in fixed-layer hierarchies.
GenAI models scale across abstraction layers and concept spaces.

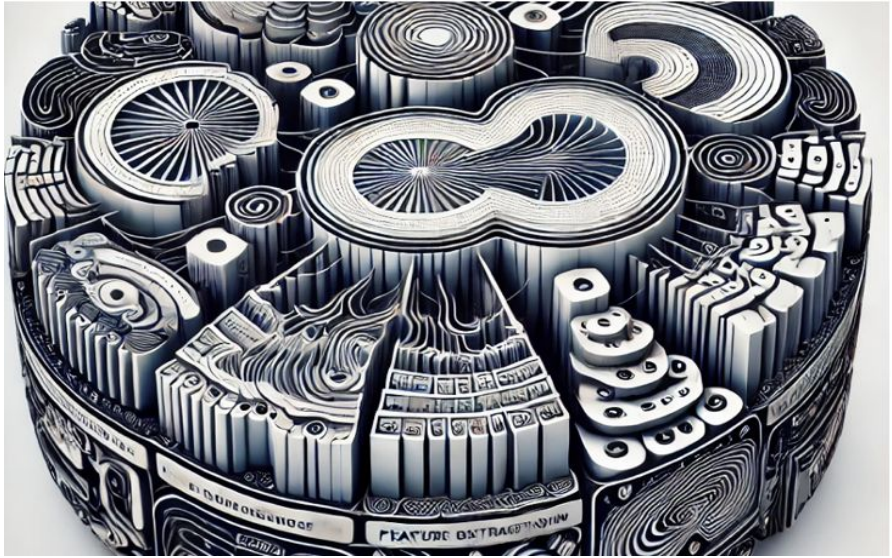
- DL models are structured with layered abstraction
- Each layer captures progressively deeper features
- DL models can span multiple concept spaces, but are still task-bound.
- DL excels at recognition

- GenAI extends DL with dense inter-concept generalization and open-ended generation
- GenAI excels at reasoning and synthesis

Different than CNNs

CNNs extract structured patterns.
GenAI models synthesize across concept spaces.

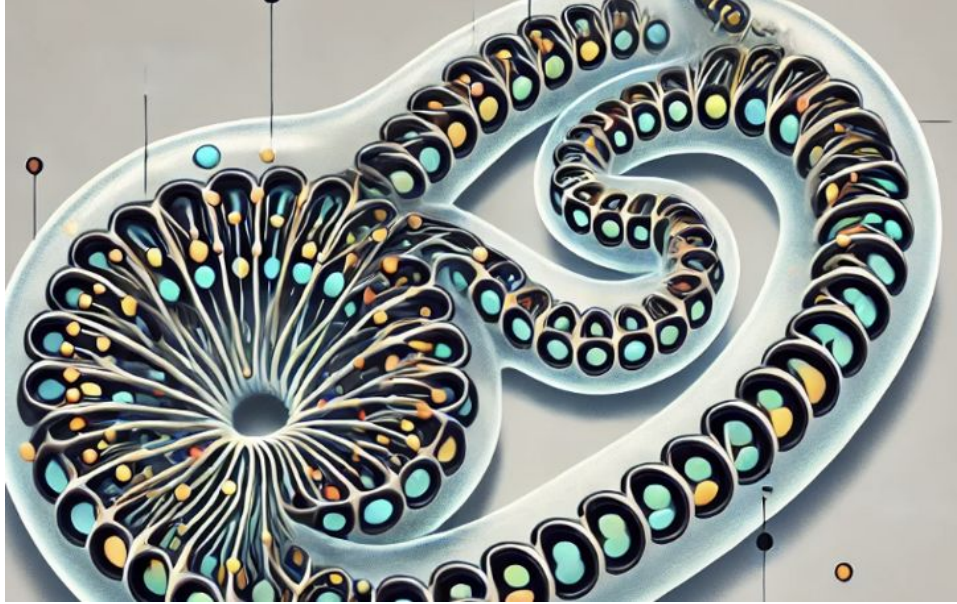
- CNNs are deep hierarchies of filters, structured multi-layered bubbles
- Optimized for spatial feature extraction (e.g. edges → textures → objects)
- Operate within a single modality (usually images or video)
- CNNs are ideal for vision tasks; GenAI for abstraction, reasoning, and synthesis



Different than RNNs

RNNs follow a fixed sequence path.
GenAI models operate across global context.

- RNNs process input step-by-step through time (sequential architecture)
- Capture temporal dependencies and short-term memory
- Struggle with long-range context and parallelism
- RNNs are good for time-series; GenAI is better for reasoning, generation, and scale



Transformers are Interconnected Bubbles

They form a semantic web where attention links meaning across concepts.

- Each bubble = a learned concept or semantic cluster
- Attention links connect bubbles dynamically per input
- Multiple bubbles activate together in parallel
- Enables context switching, abstraction, and generalization
- Overlapping boundaries reflect shared meaning or nuance

The Right Tool for the Right Job!

Model Type	Data Flow	Strengths	Limitations	Ideal Use Cases
CNN (Convolutional)	Grid-based (2D/3D)	Visual pattern detection	Modality-specific, limited reasoning	Image recognition, object detection
RNN (Recurrent)	Left-to-right (sequential)	Time-aware, memory of past input	Short-term focus, no parallelism	Time series, speech, log analysis
Transformer (GenAI)	Full-sequence parallel	Contextual reasoning, generation	Latency, control complexity	Language modeling, coding, dialog

A Deep Dive into Transformers Architecture

Henrique Krupck Secchi Follow 20 min read · Dec 3, 2024

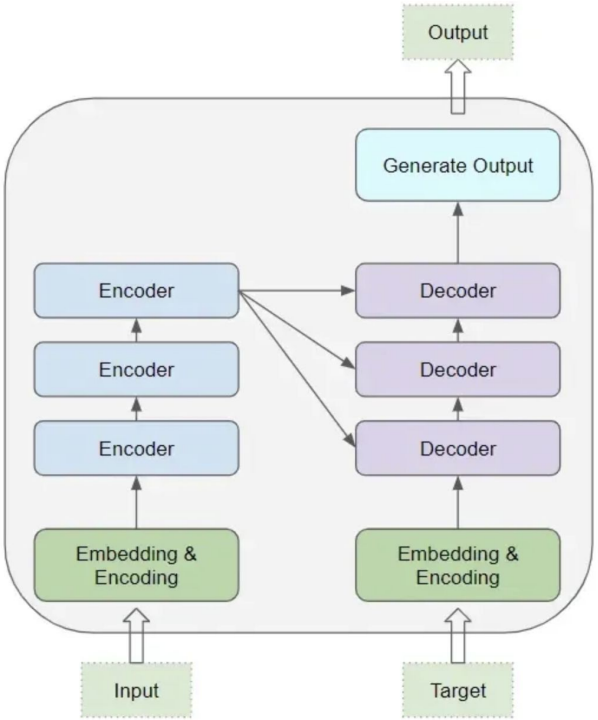
9

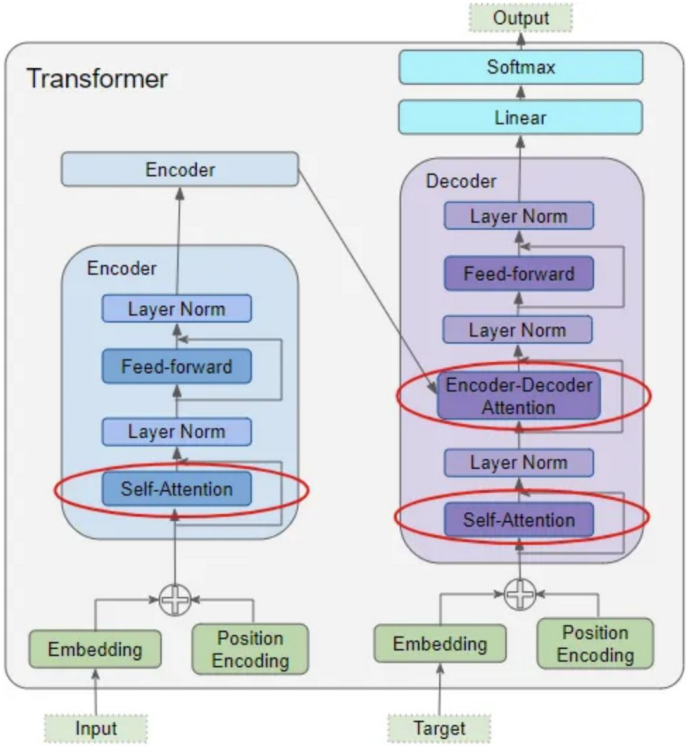
Overview of the Transformers Architecture

Transformers are in the spotlight, and for good reason. They have revolutionized the field over the past few years.

The Transformer is an architecture that leverages Attention to significantly enhance the performance of models designed for sequence learning tasks.

https://medium.com/@krupck/a-deep-dive-into-transformers-architecture-58fed326b08d


Output
Generate Output
Encoder
Encoder
Encoder
Decoder
Decoder
Decoder
Embedding & Encoding
Embedding & Encoding
Input
Target


Transformer
Output
Softmax
Linear
Encoder
Encoder
Layer Norm
Feed-forward
Layer Norm
Self-Attention
Decoder
Layer Norm
Feed-forward
Layer Norm
Encoder-Decoder Attention
Layer Norm
Self-Attention
Embedding
Position Encoding
Embedding
Position Encoding
Input
Target

What Does Self-Attention Do?

The key to the groundbreaking performance of the Transformer is its use of Attention, specifically Self-Attention.

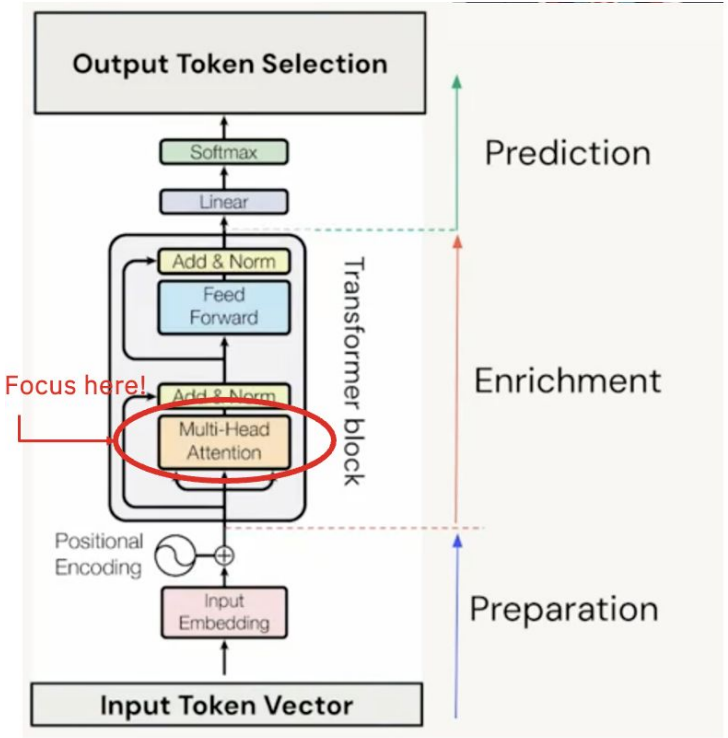
When processing a word, Attention allows the model to focus on other words in the input that are closely related to that word.

For example, "**ball**" is closely related to "**blue**" and "**hold**." On the other hand, "**blue**" is not related to "**boy**," as shown in the image below:

Core Layers of the Transformer Architecture

1. **Preparation Layer**: Embeds tokens + adds positional encoding
2. **Transformer Block:** Applies attention + residual connections
3. **Prediction Layer:** Uses softmax to generate next token/output

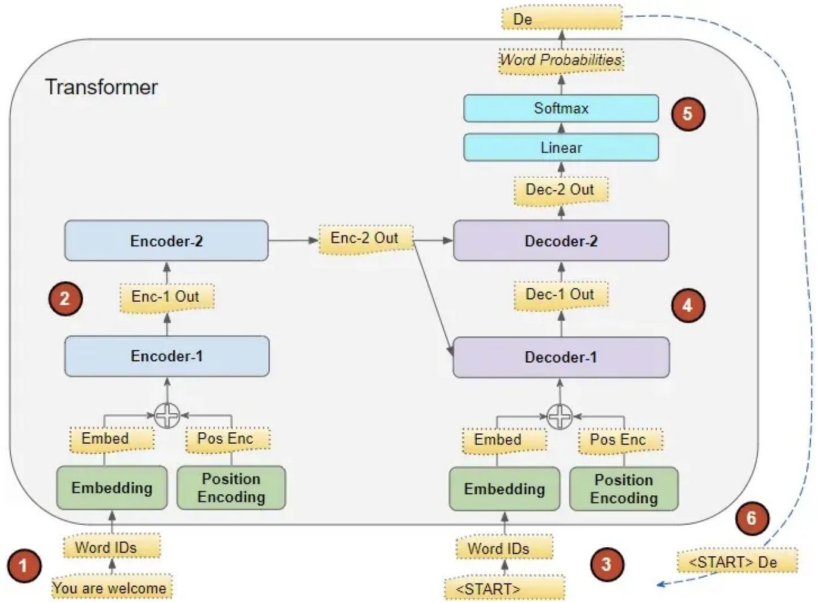
- Layers operate in stacked fashion for deeper abstraction
- Enables parallel, context-aware sequence modeling



How does it work?

1. The input sequence is converted into **embeddings.**
2. On the encoder, the embeddings are processed to produce an **encoded representation of the input ER.**
3. On the decoder, an **empty sequence ES** with only a start-of-sentence token.
4. The decoder stack processes ER + ES to produce an **encoded representation of the target sequence ET**.
5. The output layer converts ET this into word probabilities and produces an output sequence.
6. **The last word LW of the output sequence is taken as the predicted word.**

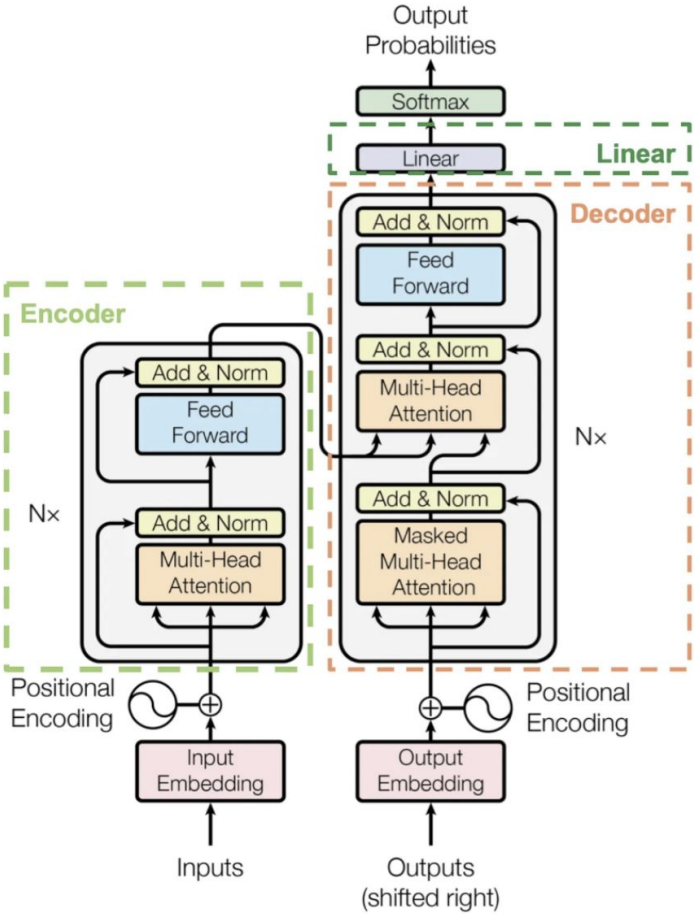
The LW is then appended to the second position in the decoder input sequence, which now contains the start-of-sentence token and the first predicted word.



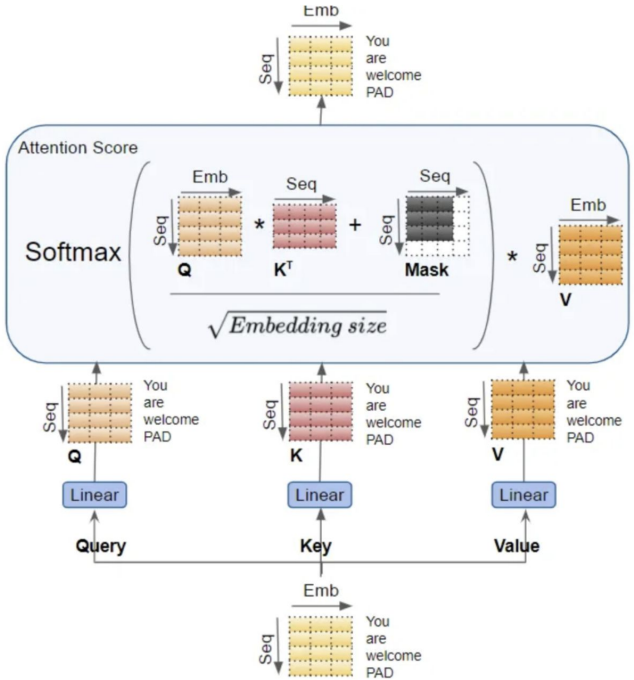
Internal Structures inside Transformer Blocks

Key mechanisms stabilize learning and enhance representation power.

- **Feed-Forward Networks** apply nonlinear transformations per token
- **Residual Connections** pass unaltered signals forward across layers
- **Layer Normalization** keeps input distributions stable and consistent



The Attention Mechanism

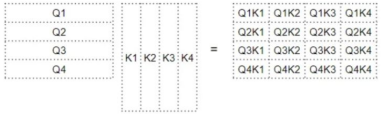


$$Z = Softmax(\frac{QK^T}{\sqrt{d_k}})V$$

Attention Score calculation (Image by Author)

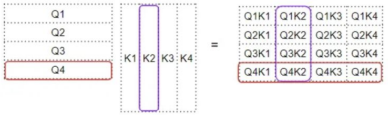
As we can see from the formula, the first step within Attention is to do a matrix multiply (ie. dot product) between the Query (Q) matrix and a transpose of the Key (K) matrix. Watch what happens to each word.

We produce an intermediate matrix (let's call it a 'factor' matrix) where each cell is a matrix multiplication between two words.

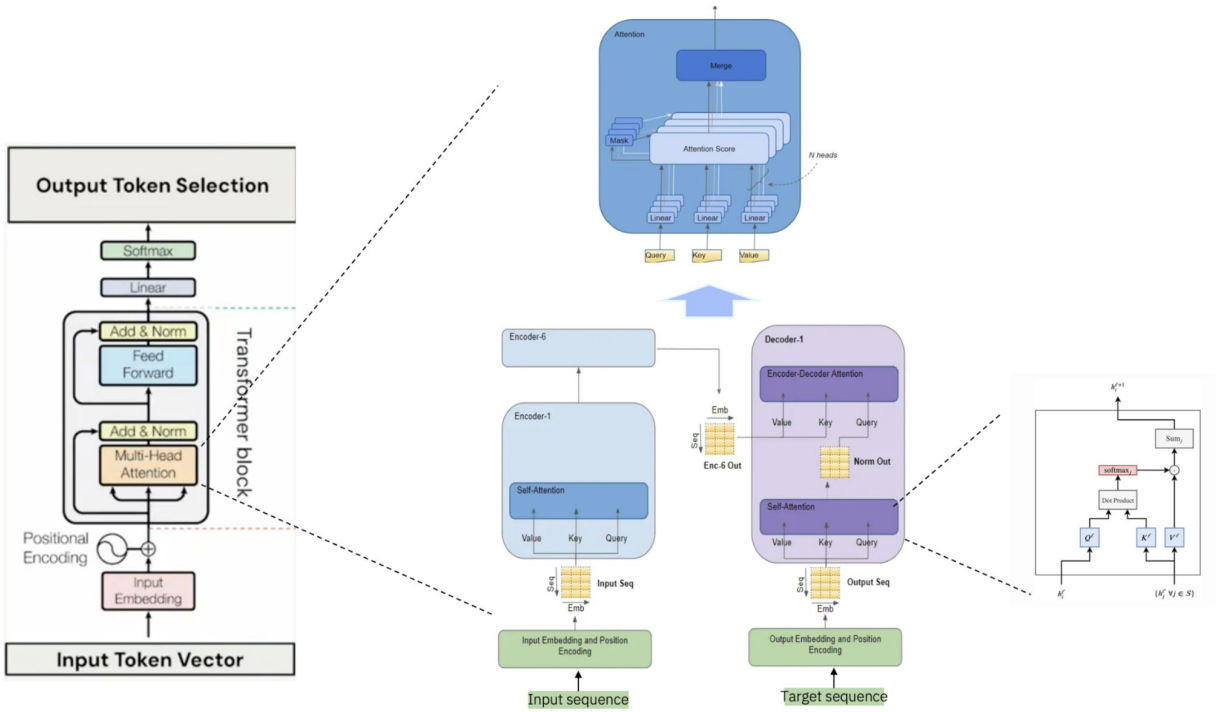


Dot Product between Query and Key matrices (Image by Author)

For instance, each column in the fourth row corresponds to a dot product between the fourth Query word with every Key word.



Putting it All Together



Size of Models

- **n_params:** Total number of trainable parameters in the Transformer model.
- **n_layers:** Number of encoder and/or decoder layers in the model.
- **d_model:** Size of the hidden layers within the model.
- **n_heads:** Number of heads in the multi-head attention mechanism.
- **d_head:** Dimensionality of each attention head.

- **Embeddings**: $2 \times (N_{\text{vocab}} \times d_{\text{model}})$ (assuming separate embeddings for input and output, and including positional embeddings).
- **Decoder Layers**:
  - **Self-Attention**: $h \times (3 \times (d_{\text{model}} \times d_{\text{k}}) + (d_{\text{model}} \times d_{\text{model}}))$ per layer.
  - **FFN**: $2 \times (d_{\text{model}} \times d_{\text{ff}})$ per layer.
  - **Layer Norm**: $2 \times d_{\text{model}}$ per layer.
- **Output Layer**: $d_{\text{model}} \times N_{\text{vocab}}$.

The total number of parameters (N_{params}) is the sum of all these components, multiplied by the number of layers (L) for the components that are repeated across layers:

$$N_{\text{params}} = \text{Embeddings} + L \times (\text{Self-Attention} + \text{FFN} + \text{Layer Norm}) + \text{Output Layer}$$

Model Name	n_{params}	n_{layers}	d_{model}	n_{heads}	d_{head}	Batch Size	Learning Rate
GPT-3 Small	125M	12	768	12	64	0.5M	6.0×10^{-4}
GPT-3 Medium	350M	24	1024	16	64	0.5M	3.0×10^{-4}
GPT-3 Large	760M	24	1536	16	96	0.5M	2.5×10^{-4}
GPT-3 XL	1.3B	24	2048	24	128	1M	2.0×10^{-4}
GPT-3 2.7B	2.7B	32	2560	32	80	1M	1.6×10^{-4}
GPT-3 6.7B	6.7B	32	4096	32	128	2M	1.2×10^{-4}
GPT-3 13B	13.0B	40	5140	40	128	2M	1.0×10^{-4}
GPT-3 175B or "GPT-3"	175.0B	96	12288	96	128	3.2M	0.6×10^{-4}

Table 2.1: Sizes, architectures, and learning hyper-parameters (batch size in tokens and learning rate) of the models which we trained. All models were trained for a total of 300 billion tokens.

Coding a Transformer

This implementation defines a transformer model with roughly 100M parameters:

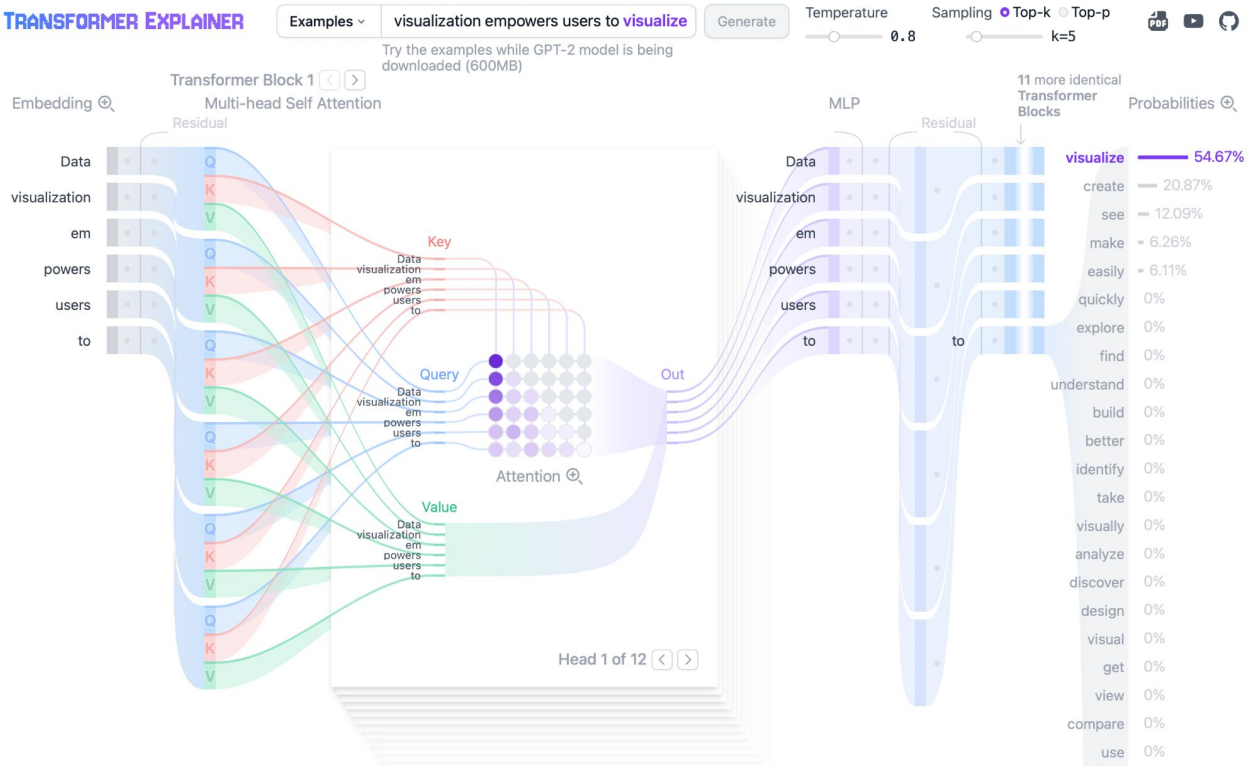
- Vocabulary size: 50,000.
- Model dimension (d_model): 768.
- Number of layers: 12.
- Number of attention heads: 12.
- Feedforward size: 3072.
- Sequence length: 512.

https://chatgpt.com/share/68b59f7b-ebec-8002-8b51-bb09de8a554f
(prompt: '*provide code*')

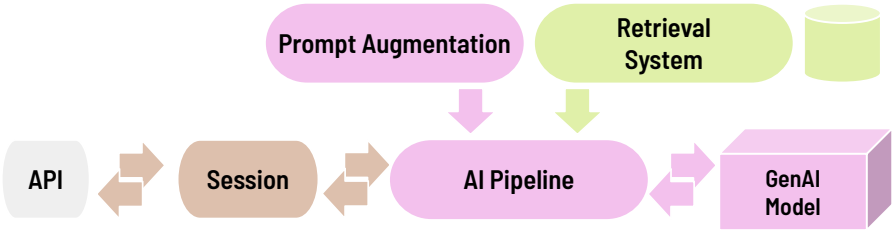
```python
import torch
import torch.nn as nn
import math

class Transformer(nn.Module):
    def __init__(self,
                 vocab_size,
                 num_layers,
                 d_model,
                 num_heads,
                 d_ff,
                 max_seq_len,
                 dropout=0.1):
        super(Transformer, self).__init__()

        self.embedding = nn.Embedding(vocab_size, d_model)
        self.positional_encoding = self._generate_positional_encoding(max_seq_len, d_model)
        self.layers = nn.ModuleList([
            nn.TransformerEncoderLayer(
                d_model=d_model,
                nhead=num_heads,
                dim_feedforward=d_ff,
                dropout=dropout
            ) for _ in range(num_layers)
        ])
        self.fc_out = nn.Linear(d_model, vocab_size)

    def forward(self, x):
        seq_len, batch_size = x.size()

        # Embedding and positional encoding
        embeddings = self.embedding(x) * math.sqrt(self.embedding.embedding_dim)
        embeddings += self.positional_encoding[:seq_len, :].unsqueeze(1).to(embeddings.device)

        # Pass through transformer layers
        for layer in self.layers:
            embeddings = layer(embeddings)

        # Output projection
        output = self.fc_out(embeddings)
        return output

    def _generate_positional_encoding(self, max_seq_len, d_model):
        pe = torch.zeros(max_seq_len, d_model)
        for pos in range(max_seq_len):
            for i in range(0, d_model, 2):
                pe[pos, i] = math.sin(pos / (10000 ** (2 * i / d_model)))
                if i + 1 < d_model:
                    pe[pos, i + 1] = math.cos(pos / (10000 ** (2 * (i + 1) / d_model)))
        return pe
```

Controlling the Behavior of Models

https://poloclub.github.io/transformer-explainer/



Calling a Model

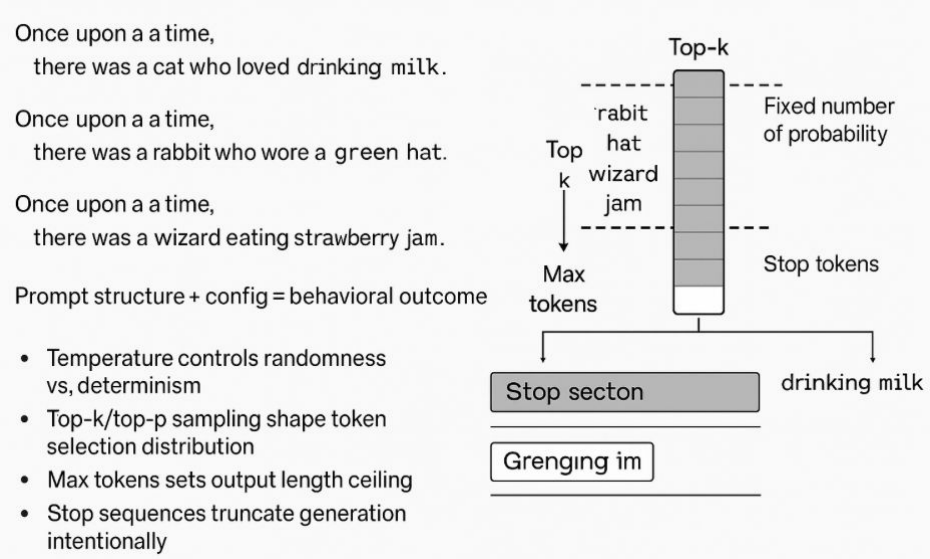



```python
from ollama import Client

client = Client()

response = client.chat(
    model="llama3",
    messages=[{"role": "user", "content": "Explain temperature, top-k, top-p simply."}],
    options={
        "temperature": 0.7,     # randomness
        "top_k": 40,            # shortlist size
        "top_p": 0.9,           # nucleus sampling
        "num_predict": 200,     # max tokens
        "stop": ["\n"]          # stop at newline
    }
)

print(response["message"]["content"])
```

Model Parameters

Tunable parameters guide creativity, coherence, and output length.

- **Temperature** controls randomness vs. determinism
- **Top-k/top-p** sampling shape token selection distribution
- **Max tokens** sets output length ceiling
- **Stop sequences** truncate generation intentionally
- **Prompt structure + config** = behavioral outcome

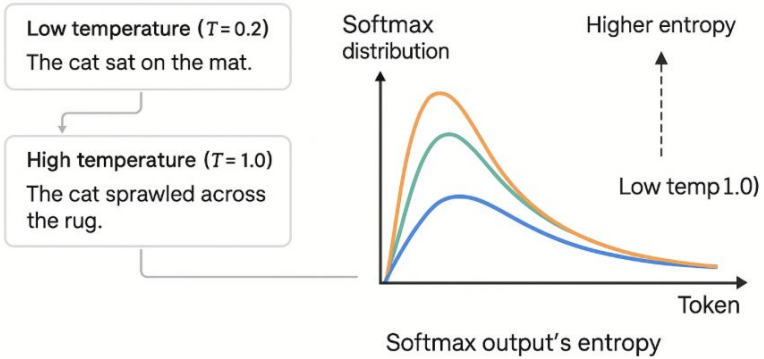


Temperature

It shifts the confidence threshold for token selection.

- **Low temperature (e.g., 0.1**): precise, repetitive, narrow sampling
- **High temperature (e.g., 1.0):** creative, surprising, higher variance

- Controls entropy of the softmax distribution
- Ideal temp depends on task type (e.g., coding vs. poetry)

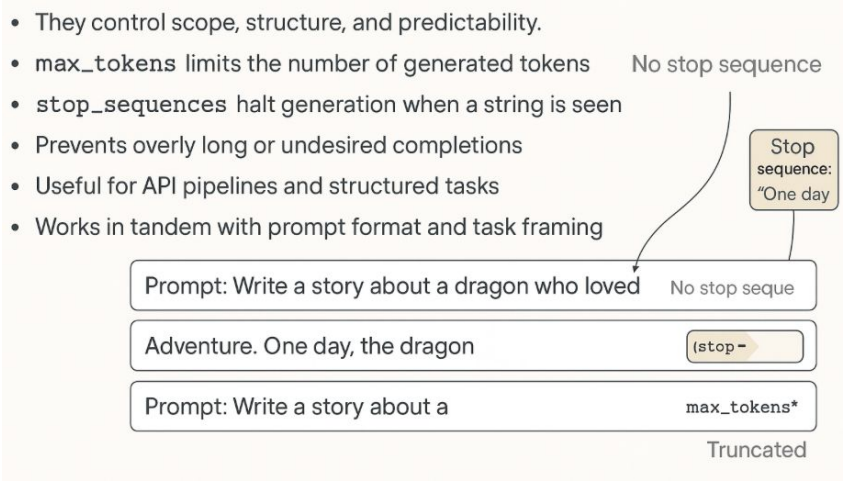


Length and Stop Settings

Control scope, structure, and predictability.

- **max_tokens** limits the number of generated tokens
- **stop_sequences** halt generation when a string is seen

- Prevents overly long or undesired completions
- Useful for API pipelines and structured tasks
- Works in tandem with prompt format and task framing



Example: Generative DocString in Python

- "Generate a Python docstring for the function below:"
- "Explain this function in one sentence."
- "Write a concise summary of what this function does."

Prompt Type	Temperature	max_tokens	top_p	Result Style
Standard docstring prompt	0.2	50	,	Precise, single-sentence summary
Same prompt, more exploratory setting	0.8	100	,	Wordier, may speculate on intent
Concise explanation prompt	0.3	40	0.7	Safe abstraction with minimal creativity
One-liner style, stop at period	0.3	60	,	Controlled, truncates with first punctuation
Prompt includes example style	0.4	75	0.85	Matches tone and structure of given examples

Exercises

Exercise 2 - Models and Configurations

Objective: Experiment with different GenAI models and evaluate the impact of variations of parameters.

Go to Canvas -> Assignments

COT 6930 - Generative Intelligence and Software Development Lifecycles

Dr. Fernando Koch
kochf@fau.edu
http://www.fernandokoch.me