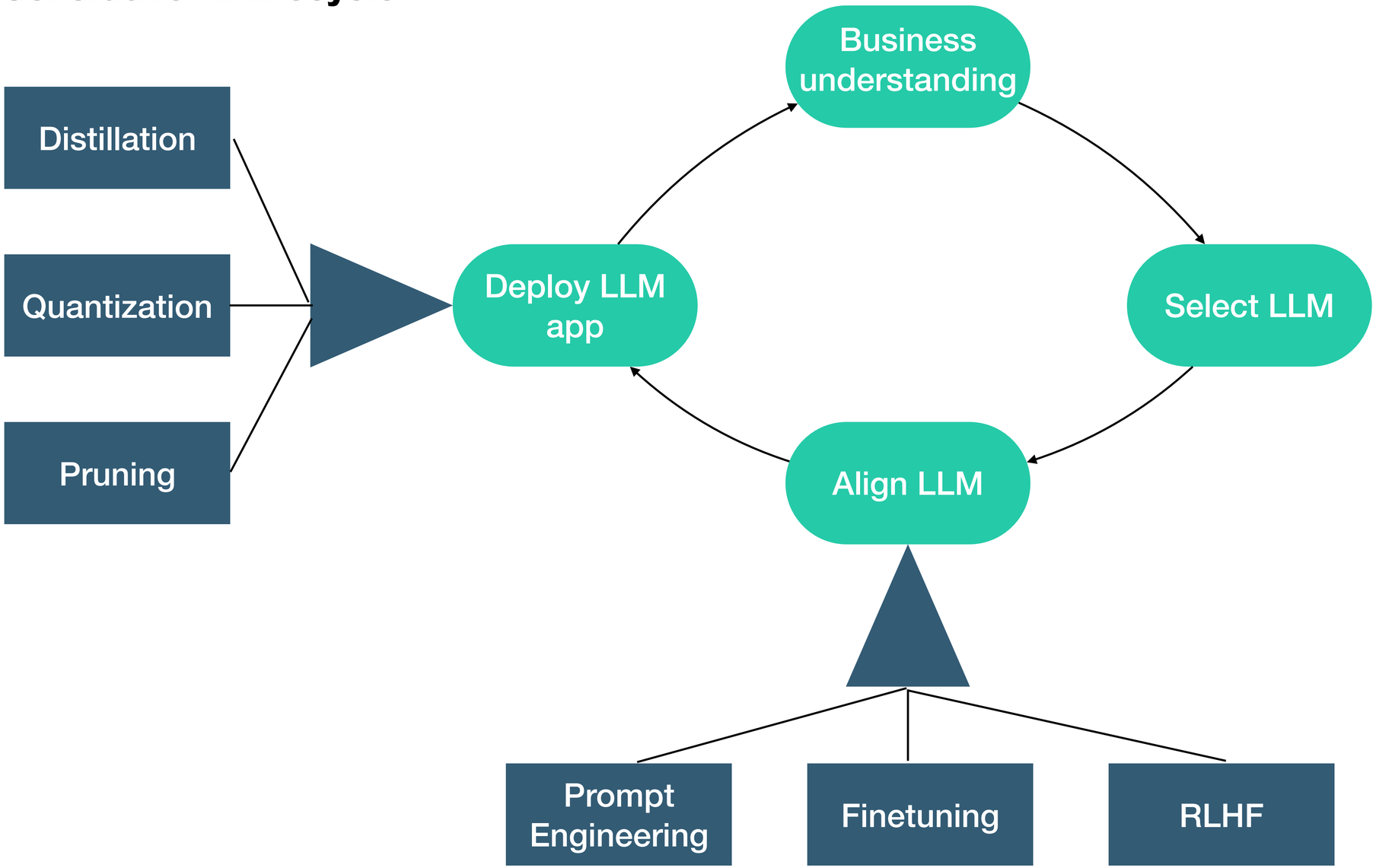
Fundamentals of Generative AI

GenAI architectures

Data Trainers

Generative AI Lifecycle



Generative Models in AI

Autoreggresive models (Decoder-only models)

AutoEncoders models (Encoder-only models)

Seq2Seq models (Encoder-Decoder models)

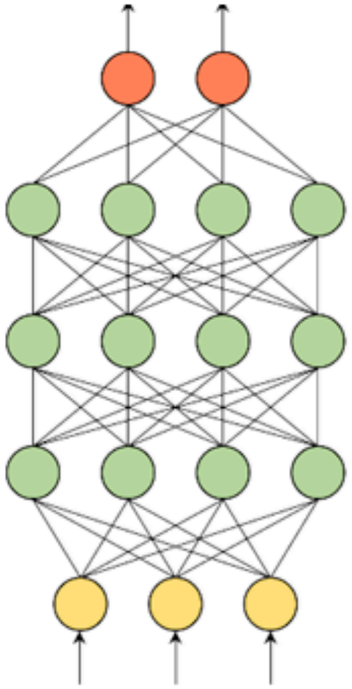
Generative Adversarial Models

Autoreggresive Models

Probabilities over char set

	a	b	c	d	e	f	g	...	z
0.01	0.02	0.36	0.25	0.02	0.001	0.22	0.001	...	0.06

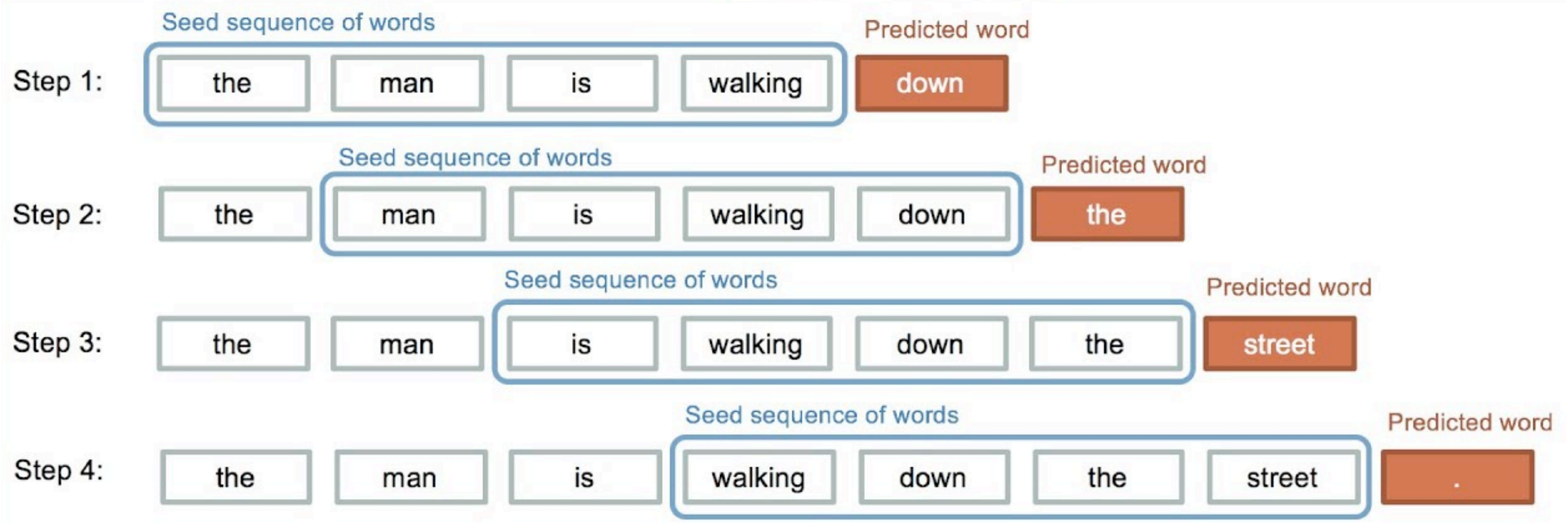
Language Model

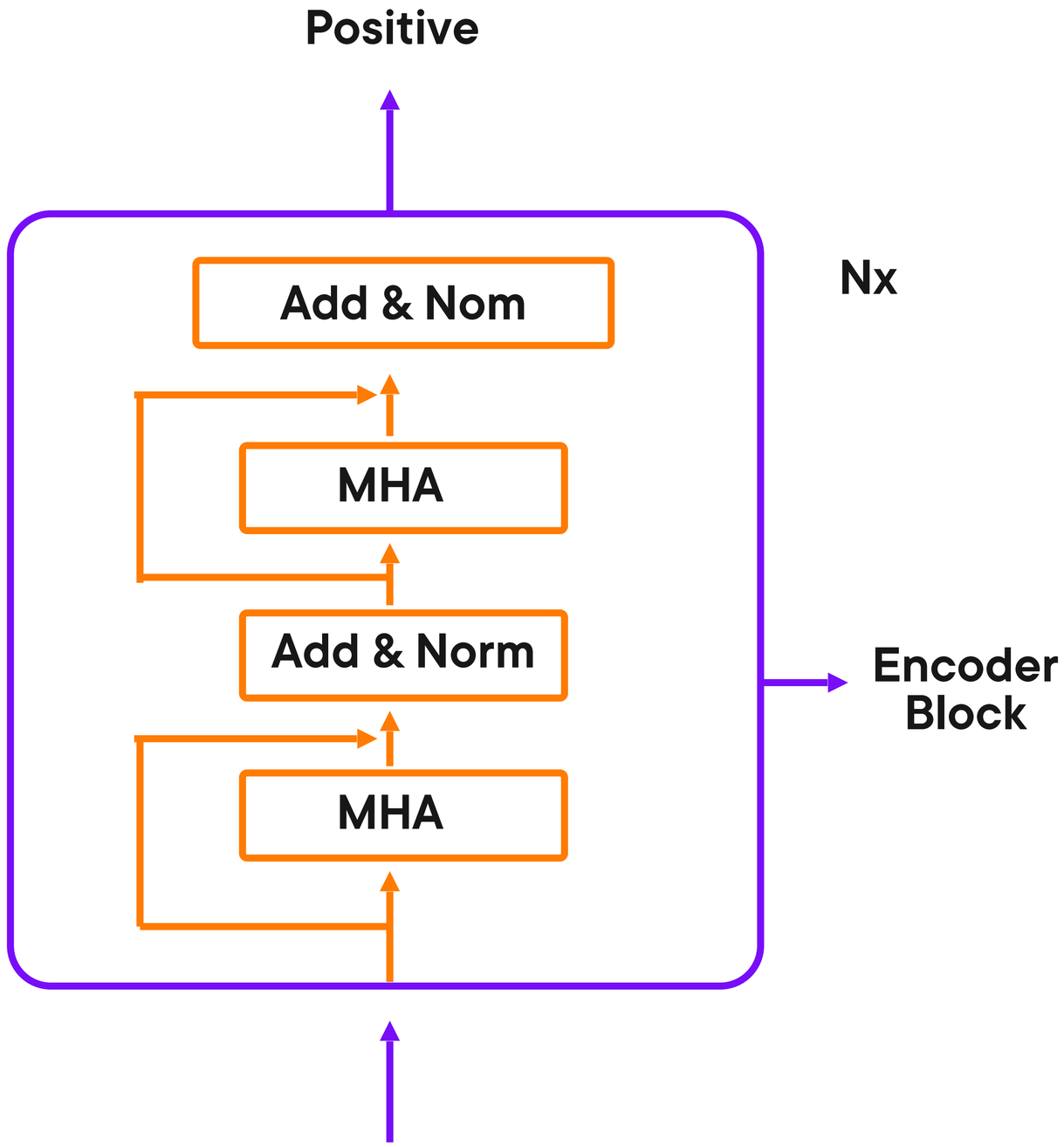


Train Input from Corpus

I		w	a	n	t		t	o		b	a	k	e	

Autoreggresive Models



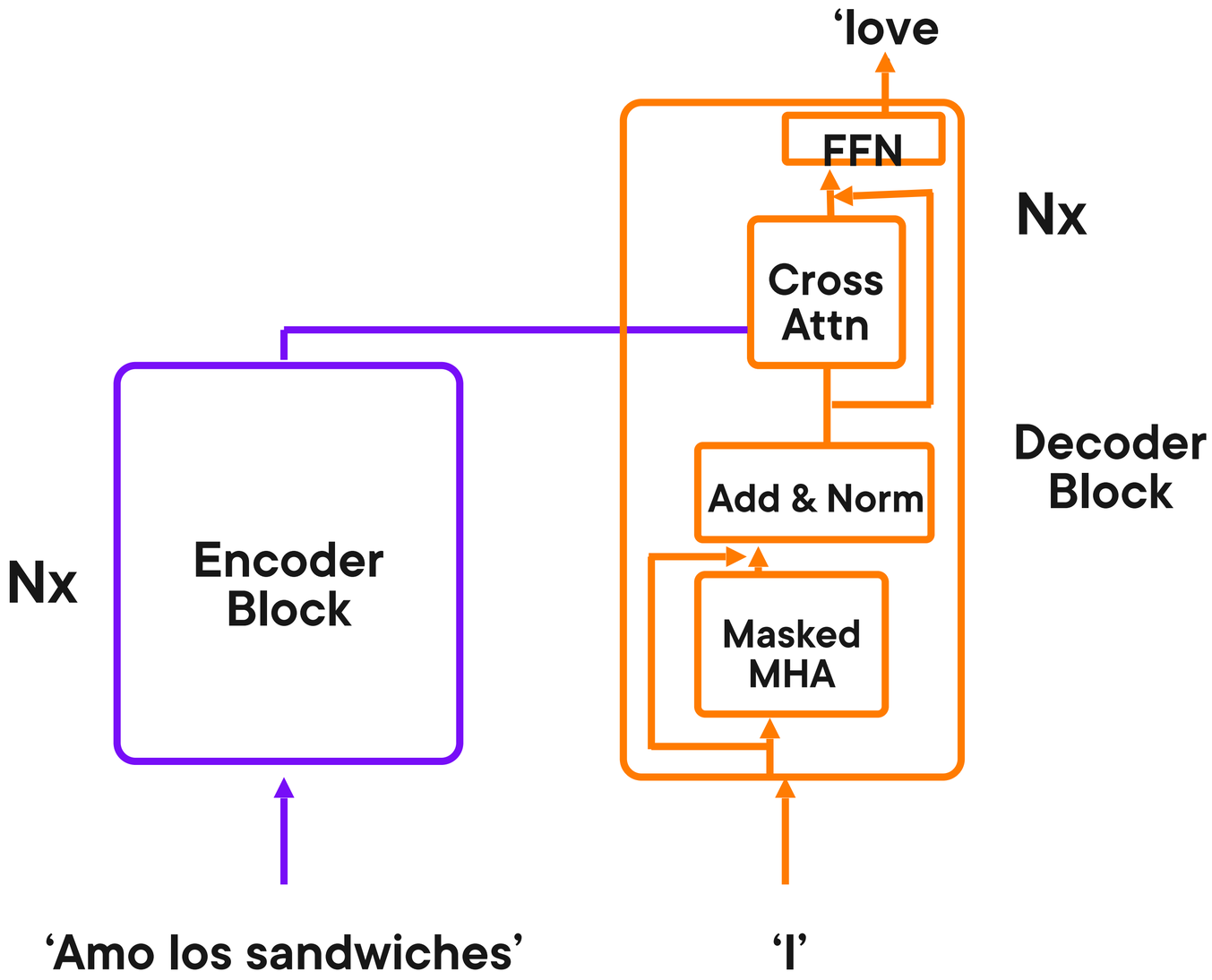


Data Trainers

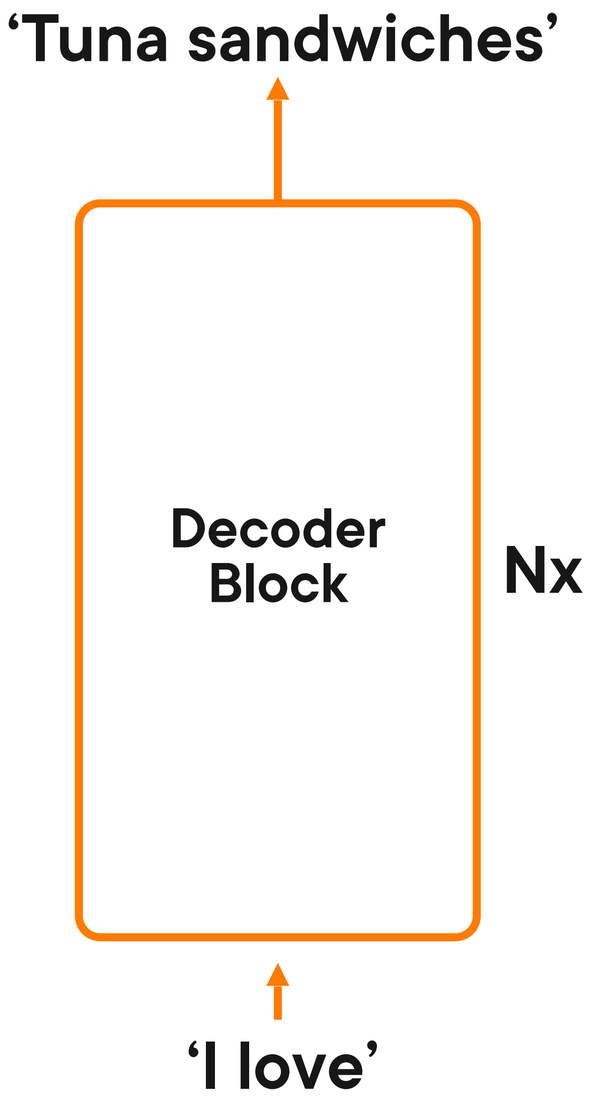
- Training Objective: Understand and encode the input text.

-Context: Used in tasks like sentence classification or semantic similarity.

Example LLMs: BERT and RoBERTa.



- **Training Objective**: Understand the input text and generate an appropriate output.
- **Context**: Used in machine translation, summarization, and question-answering.
- **Example LLMs**: BART and T5.

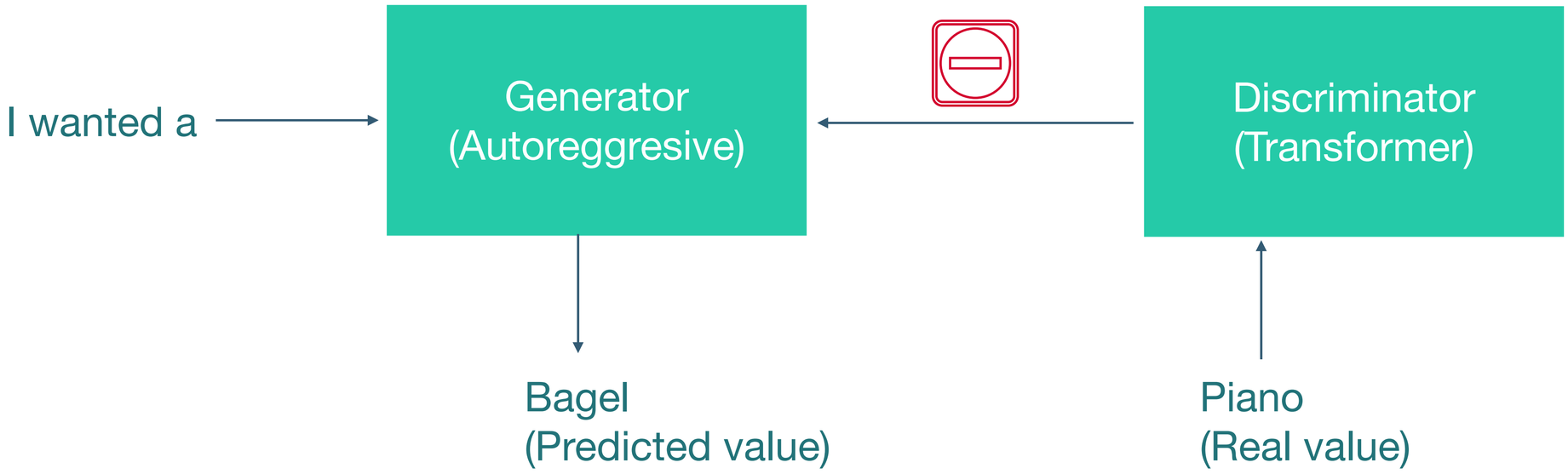


- **Training Objective**: Generate text based on given context or start token.
- **Context**: Used in tasks like language generation or image captioning.
- **Example LLMs**: GPT series.

Architecture	Training Objective	Context	Example LLMs
Encoder	Understand and encode input text	Sentence classification, semantic similarity	BERT, RoBERTa
Decoder	Generate text based on context/ start token	Language generation, image captioning	GPT series
Encoder-Decoder	Understand input and generate output	Machine translation, summarization, QA	BART, T5

Model	Provider	Open-Source	Speed	Quality	Params	Fine-tuneability
gpt-4	OpenAI	No	★☆☆	★★★★	-	No
gpt-3.5-turbo	OpenAI	No	★★☆	★★★★☆	175B	No
gpt-3	OpenAI	No	★☆☆	★★★★☆	175B	No
ada, babbage, curie	OpenAI	No	★★★	★☆☆☆	350M - 7B	Yes
claude	Anthropic	Yes	★★☆	★★★★☆	52B	no
claude-instant	Anthropic	Yes	★★★	★★☆☆	52B	No
command-xlarge	Cohere	No	★★☆	★★☆☆	50B	Yes
command-medium	Cohere	No	★★★	★☆☆☆	6B	Yes
BERT	Google	Yes	★★★	★☆☆☆	345M	Yes
T5	Google	Yes	★★☆	★☆☆☆	11B	Yes
PaLM	Google	Yes	★☆☆	★★☆☆	540B	Yes
LLaMA	Meta AI	Yes	★★☆	★★☆☆	65B	Yes
CTRL	Salesforce	Yes	★★★	★☆☆☆	1.6B	Yes
Dolly 2.0	Databricks	Yes	★★☆	★★☆☆	12B	Yes

GANs

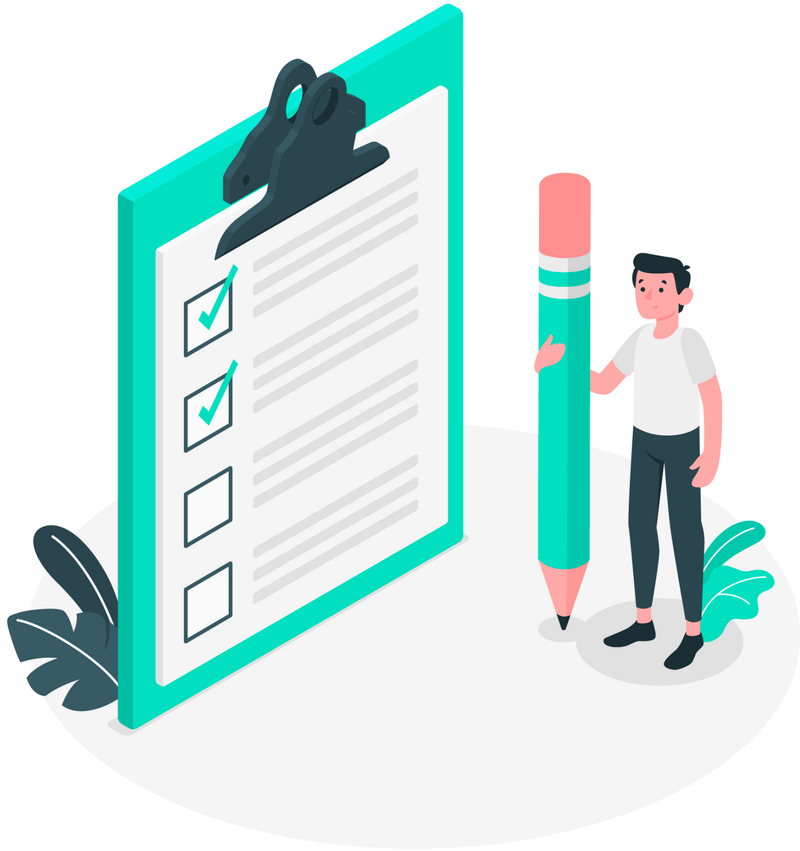

I wanted a
Generator
(Autoreggresive)
Discriminator
(Transformer)
Bagel
(Predicted value)
Piano
(Real value)

Configurations	Consequence	Usage	Importance
max-tokens	Limit the amount of tokens to be generated	To keep answers concise Performance	High
Top p	Only choose words out of the top P probability	To limit the creativeness of responses	Low
Top k	Only choose a word out of the top K tokens with highest probability	To limit the creativeness of responses	Medium
Temperature	Control how “hot” the LLM produces output. Higher Temperature	To limit the creativeness of responses	High

LAB

Prompt Engineering with ICL in Chat GPT

- Learn about prompt engineering
- Learn about few shot inference



Generative AI Lifecycle