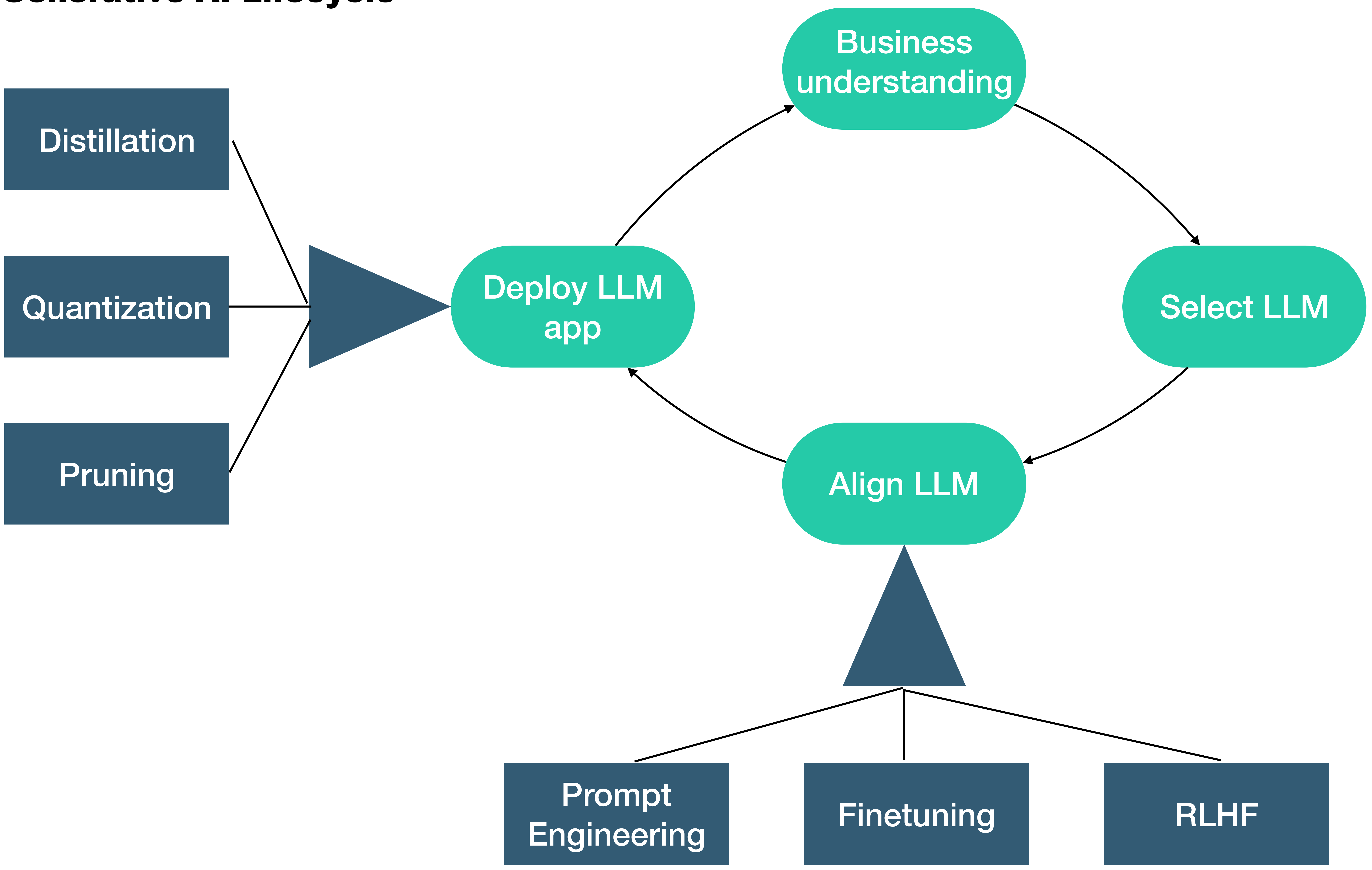
Hands on Generative AI

GenAI architectures

Data Trainers

Generative AI Lifecycle



Business understanding

Select LLM

Align LLM

Deploy LLM app

Distillation

Quantization

Pruning

Prompt Engineering

Finetuning

RLHF

Data Trainers

Generative Models in AI

Autoreggresive models (Decoder-only models)

AutoEncoders models (Encoder-only models)

Seq2Seq models (Encoder-Decoder models)

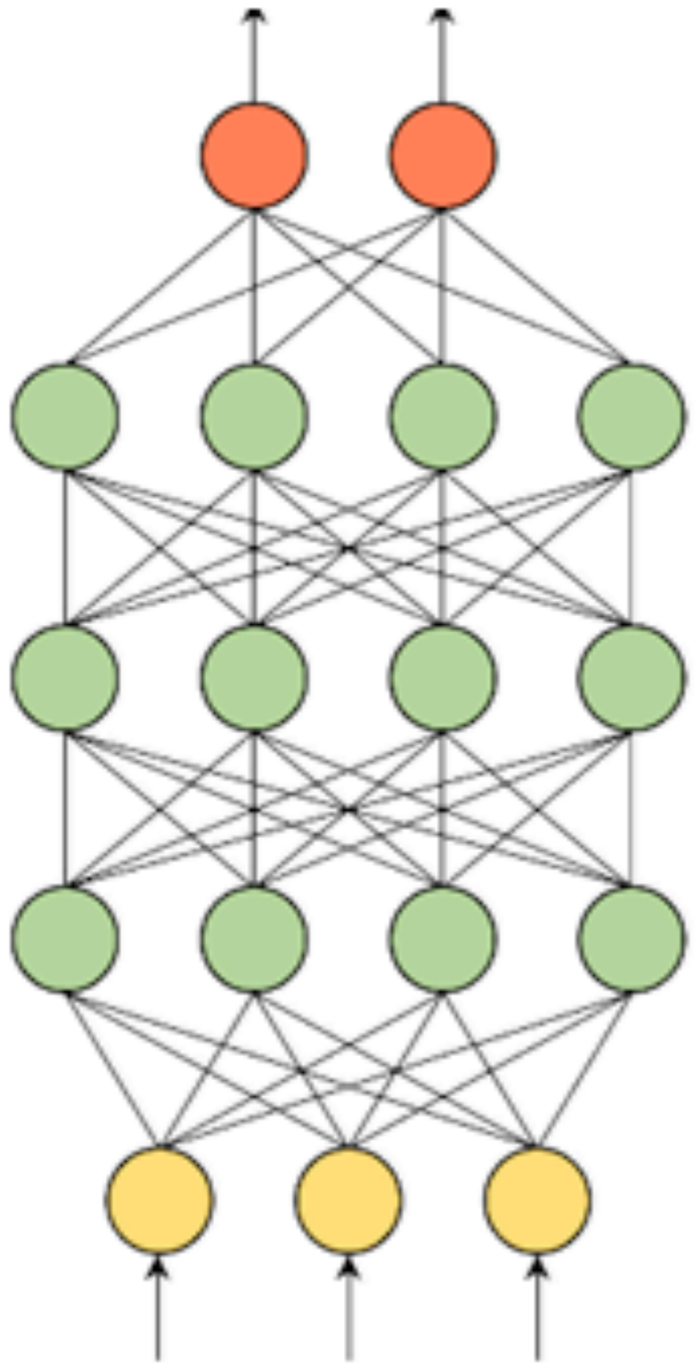
Generative Adversarial Models

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

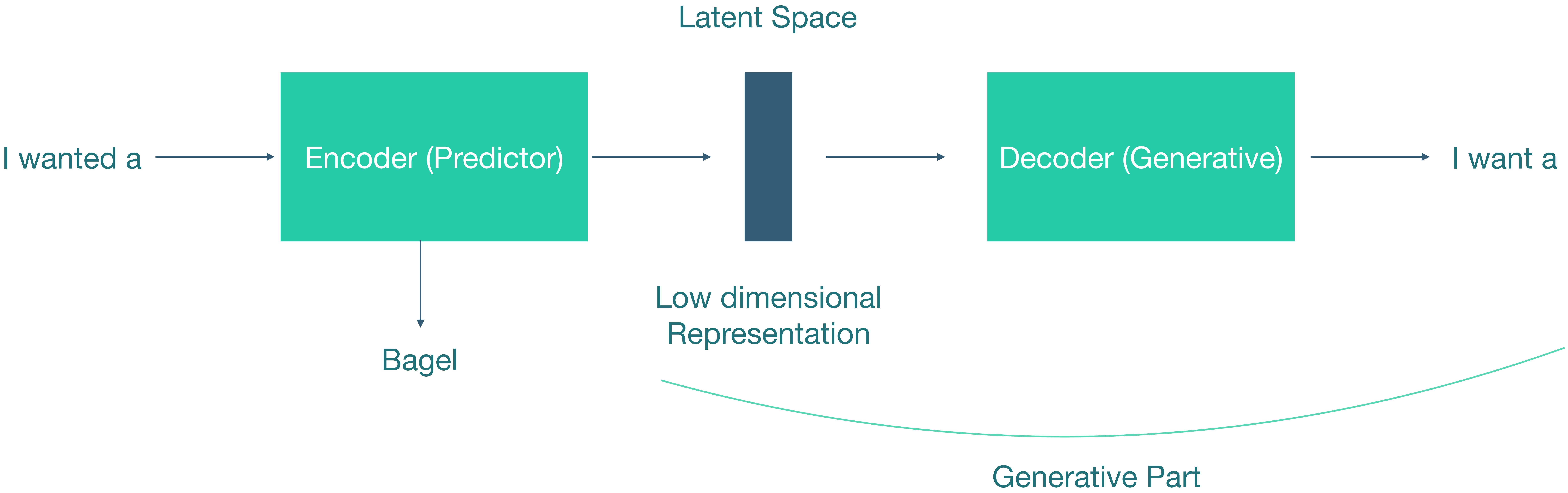
Step 1: Seed sequence of words: the man is walking — Predicted word: down

Step 2: the — Seed sequence of words: man is walking down — Predicted word: the

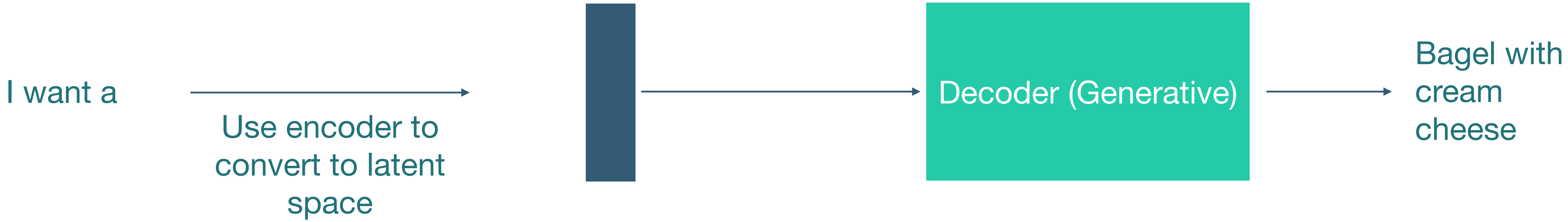
Step 3: the man — Seed sequence of words: is walking down the — Predicted word: street

Step 4: the man is — Seed sequence of words: walking down the street — Predicted word: .

AutoEncoder Models (Denoiser)



AutoEncoder Models (Generative model)



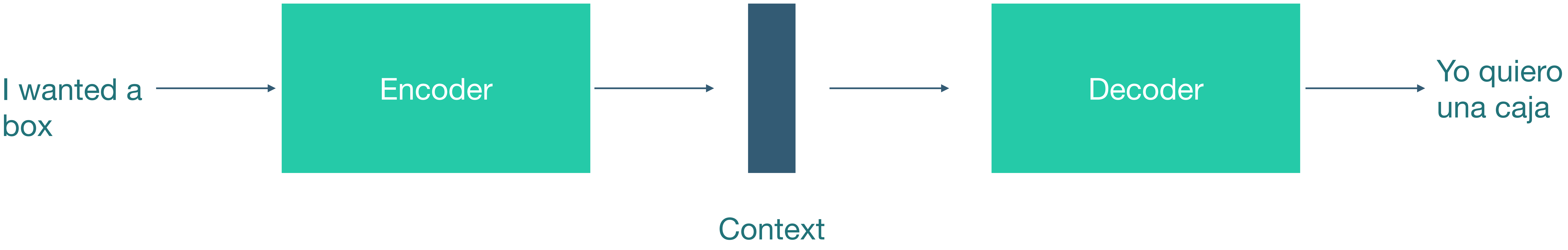
I want a

Use encoder to convert to latent space

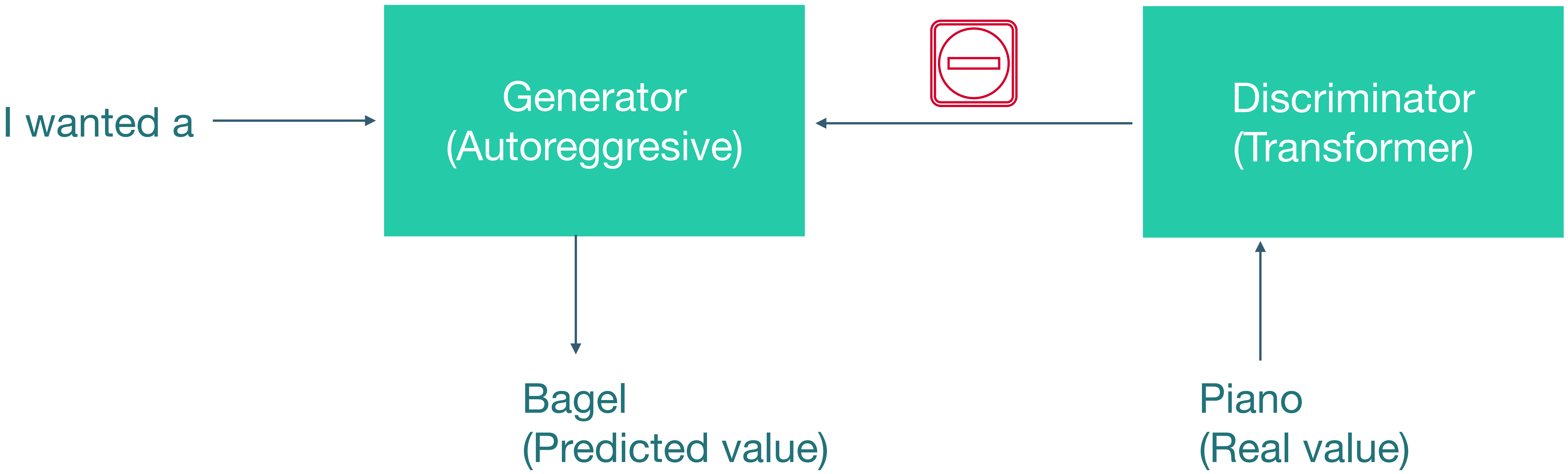
Decoder (Generative)

Bagel with cream cheese

Encoder-Decoder Models (Seq2Seq)



GANs



I wanted a

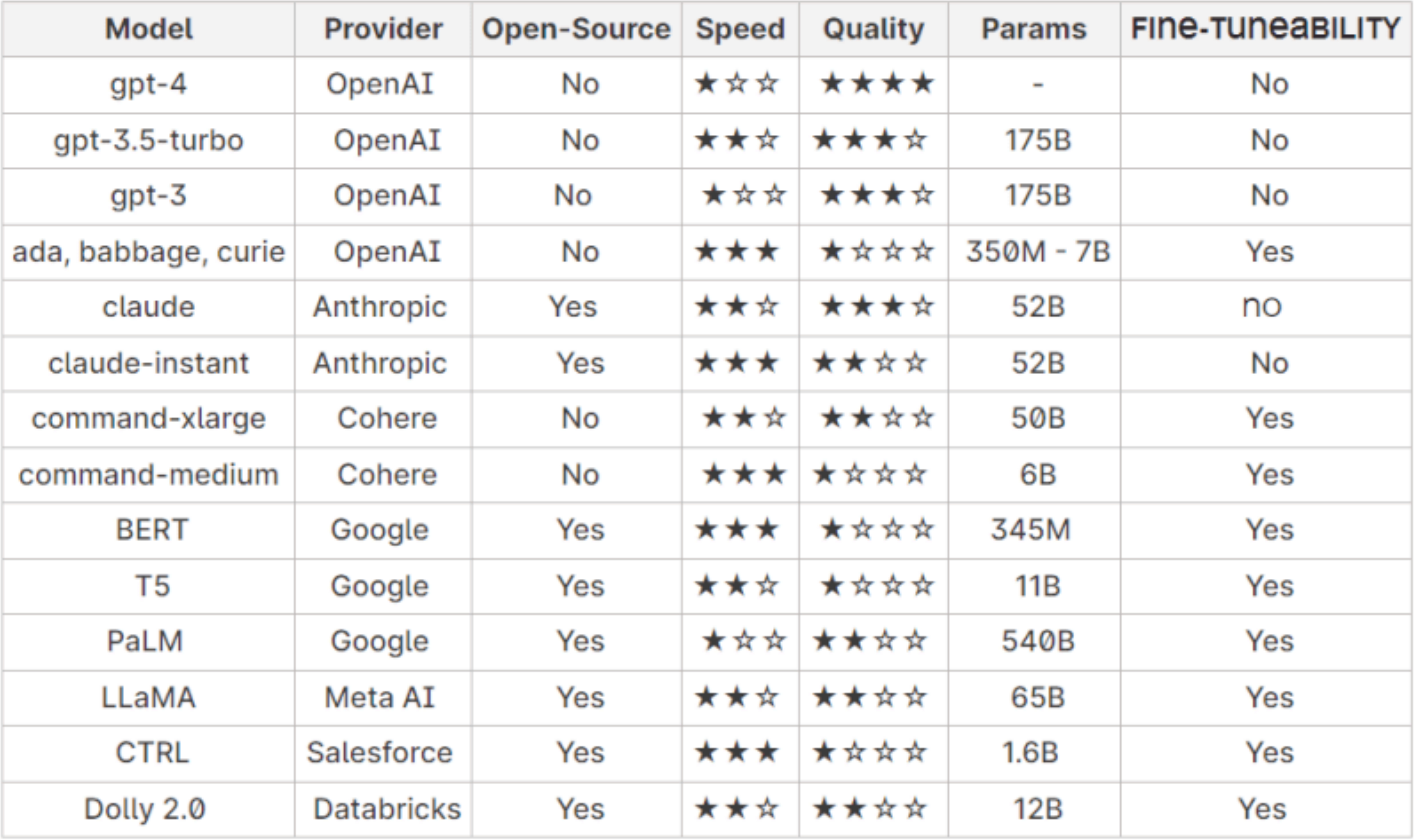
Generator (Autoreggresive)

Discriminator (Transformer)

Bagel (Predicted value)

Piano (Real value)

	Encoder Only	Decoder Only	Encoder-Decoder
Tasks	Classification NER Sentiment Analysis	Text Generation	Summarisation Translation
Training objective	Masked language	Next word prediction	Full output comparison
Context	Bidirectional	Unidirectional	Full input encoded
Famous Examples	BERT RoBERTa	GPTs PaLM	T5 Flan-T5 BART



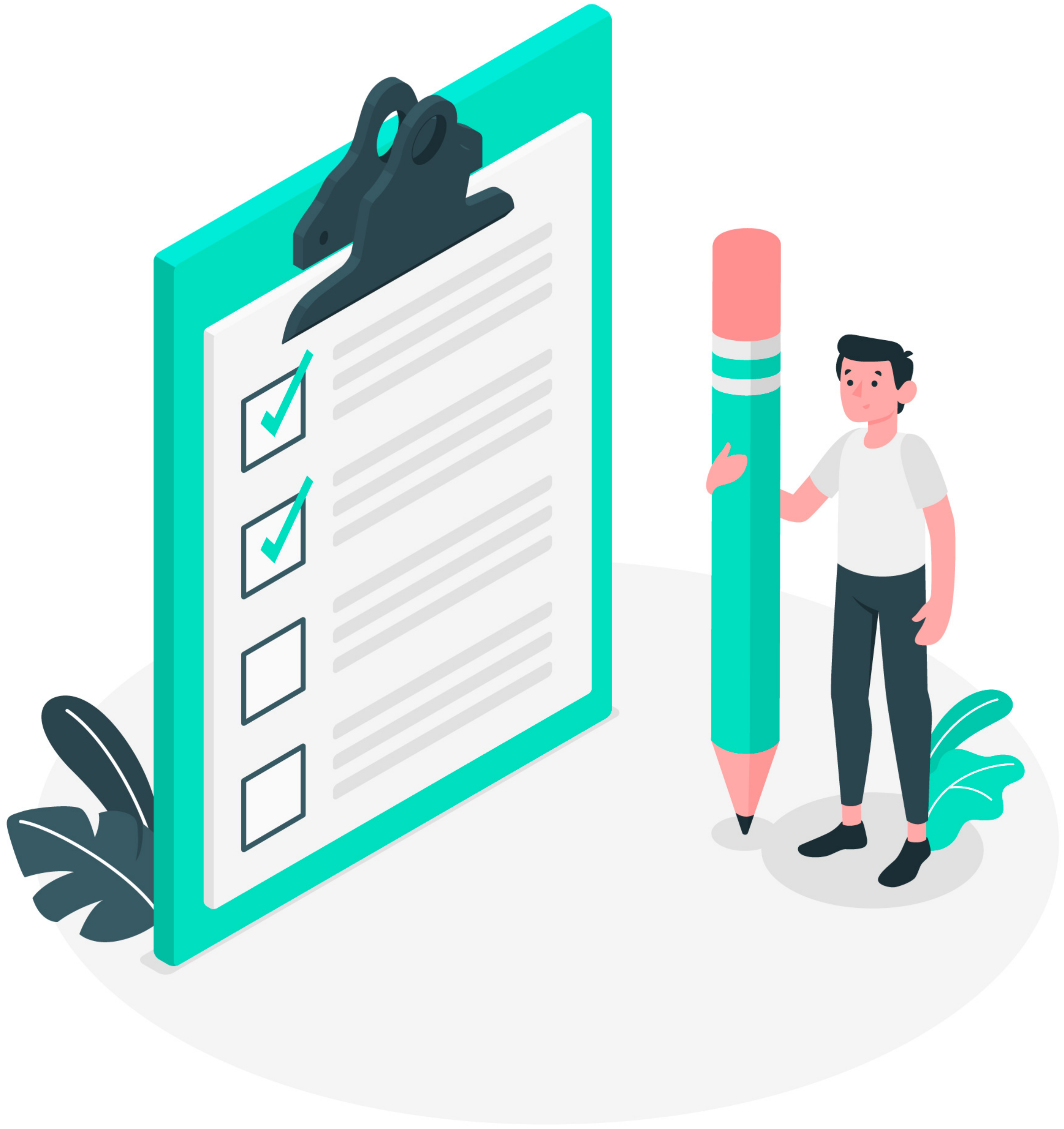
Model	Provider	Open-Source	Speed	Quality	Params	FIne-TuneaBILITY
gpt-4	OpenAI	No	★☆☆	★★★★	-	No
gpt-3.5-turbo	OpenAI	No	★★☆	★★★☆	175B	No
gpt-3	OpenAI	No	★☆☆	★★★☆	175B	No
ada, babbage, curie	OpenAI	No	★★★	★☆☆☆	350M - 7B	Yes
claude	Anthropic	Yes	★★☆	★★★☆	52B	no
claude-instant	Anthropic	Yes	★★★	★★☆☆	52B	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

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

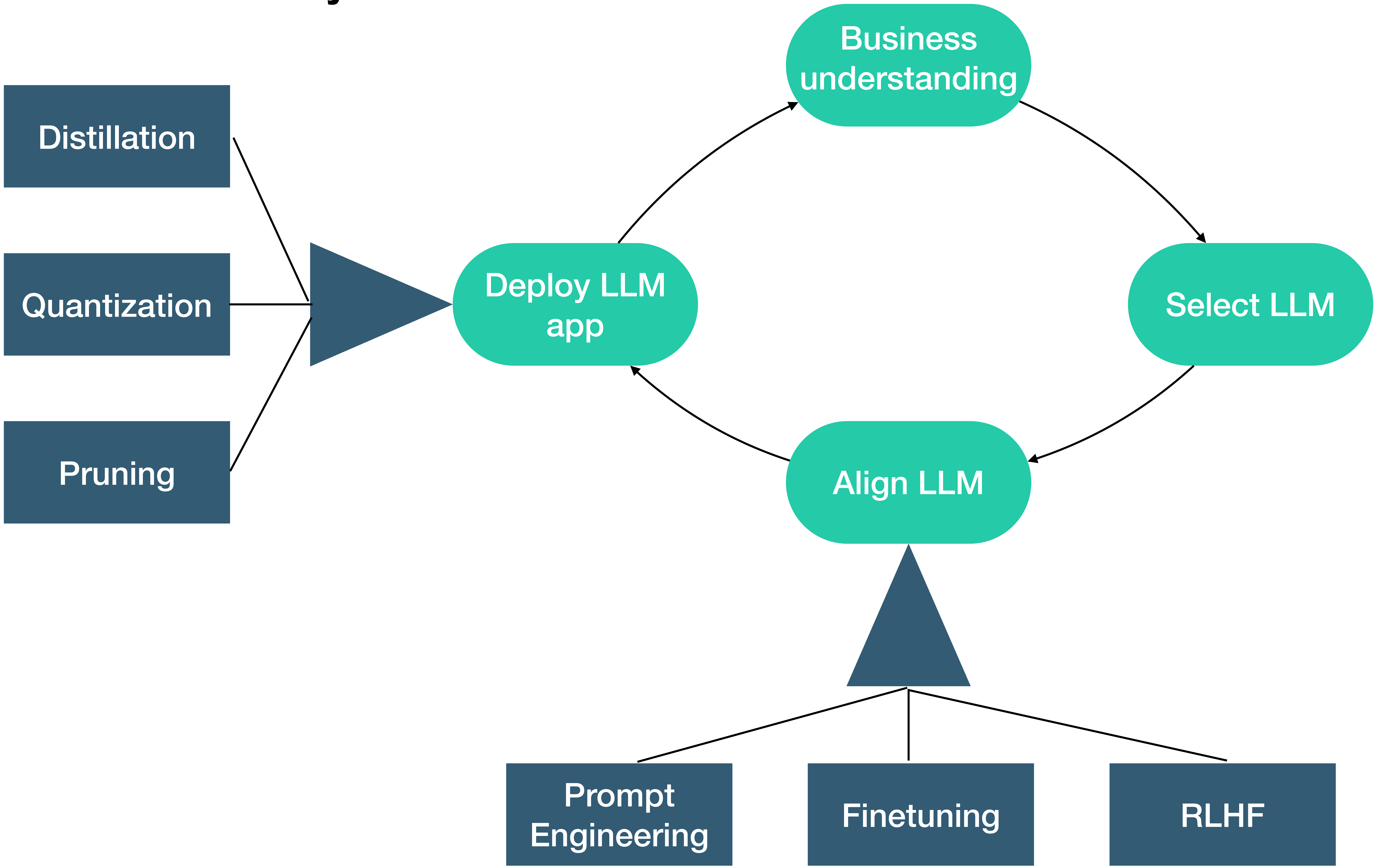
LAB

Prompt Engineering with Flan-T5

- Learn about prompt engineering
- Learn about few shot inference



Generative AI Lifecycle


Business understanding
Select LLM
Align LLM
Deploy LLM app
Distillation
Quantization
Pruning
Prompt Engineering
Finetuning
RLHF