생성형 인공지능 이론과 실습

Microsoft AI MVP
전미정

2024 삼성전자 무선사업부 생성형 AI 교육

DAY1	
언어 모델 이해	o 생성형 인공지능의 발전과 이론 - 거대 언어 인공지능의 발전 및 동향 - Transformer 기반 모델의 이해
GPT API 실습	o AzureOpenAI Service API 활용한 GPT 사용 - Azure Cloud Resource 생성 - GPT Completion API (GPT-3/GPT-3-Instruct) - ChatGPT ChatCompletion API (GTP-3.5-Turbo/GPT-4)
GPT Fine-Tuning 실습	o Custom 모델 생성을 위한 Fine-Tuning - Fine-Tuning 기법 이해 - GPT 모델 Fine-Tuning을 위한 데이터 전처리 - Custom 모델 생성 및 활용
이미지 생성 모델	o 이미지 생성 모델의 이해 - 이미지 생성 모델 구조 및 작동 방식 이해

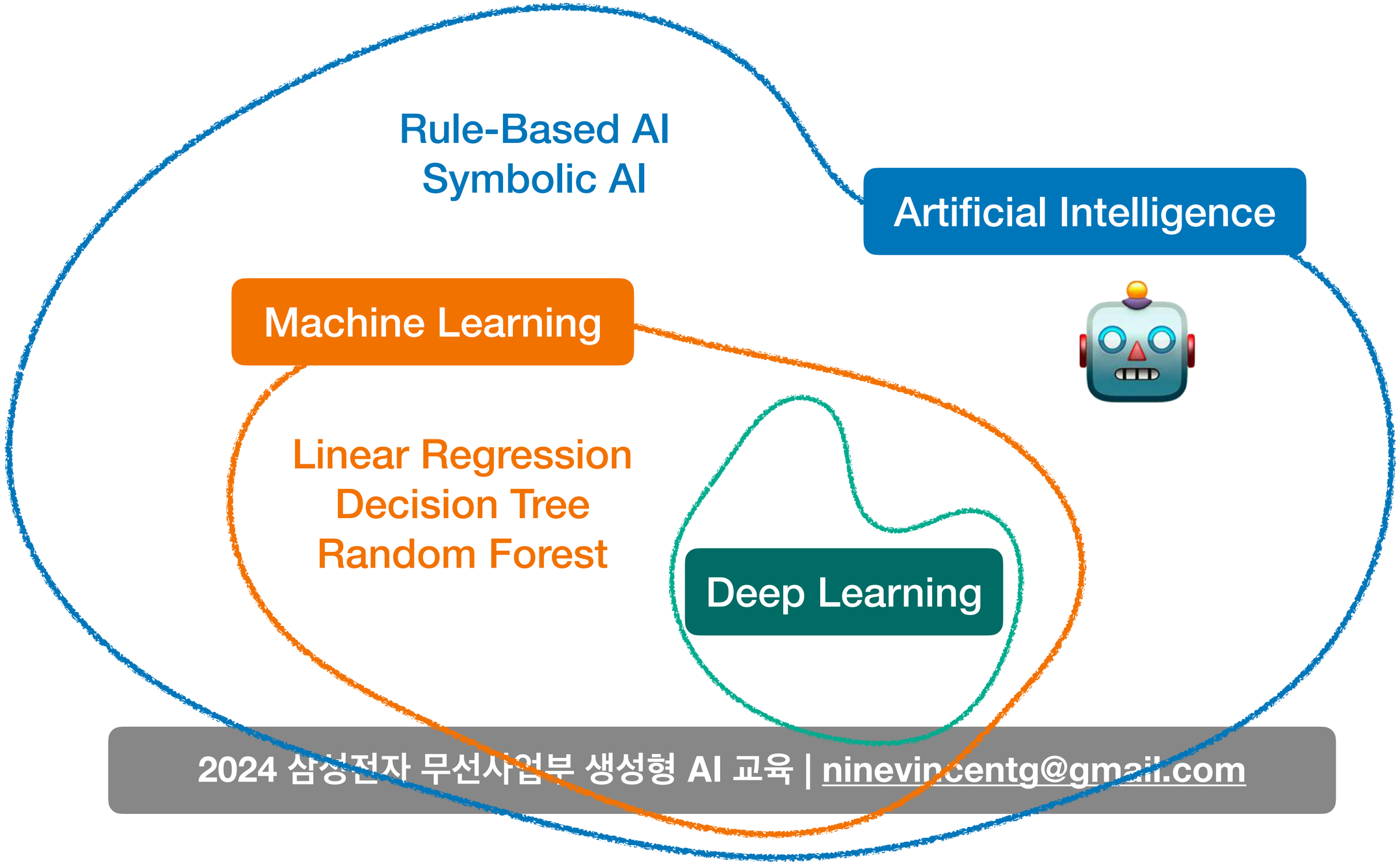
DAY2	
DALL-E API 실습	o 이미지 생성 모델 실습 및 발전 동향 - DALL-E API를 활용한 이미지 생성/수정 실습 - 이미지 생성 모델 발전 동향
LangChain 기본 실습	- Fine-Tuning/Prompt Engineering 차이 - LangChain 기본 활용법 실습 - Embedding 모델의 이해
LangChain 활용 실습	- 외부 데이터(PDF, URL, Youtube 등) 와 언어 모델 연결을 통한 데이터 기반 질의응답 실습 - LangChain을 활용한 언어 생성 모델 & 이미지 생성 모델 chain 실습

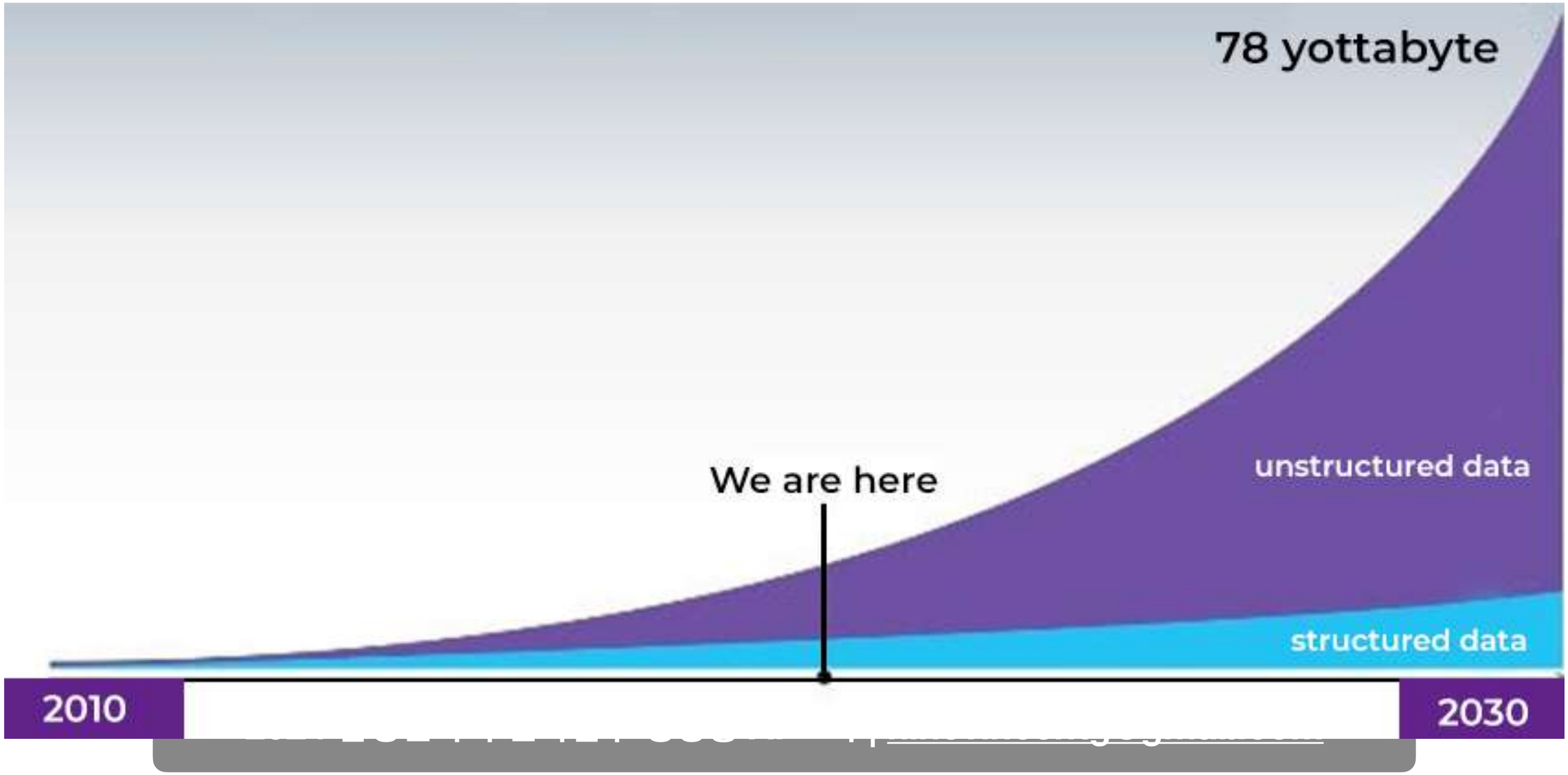
2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincent@gmail.com

Generative AI

Text, Image, Video, Sound, Music, Etc

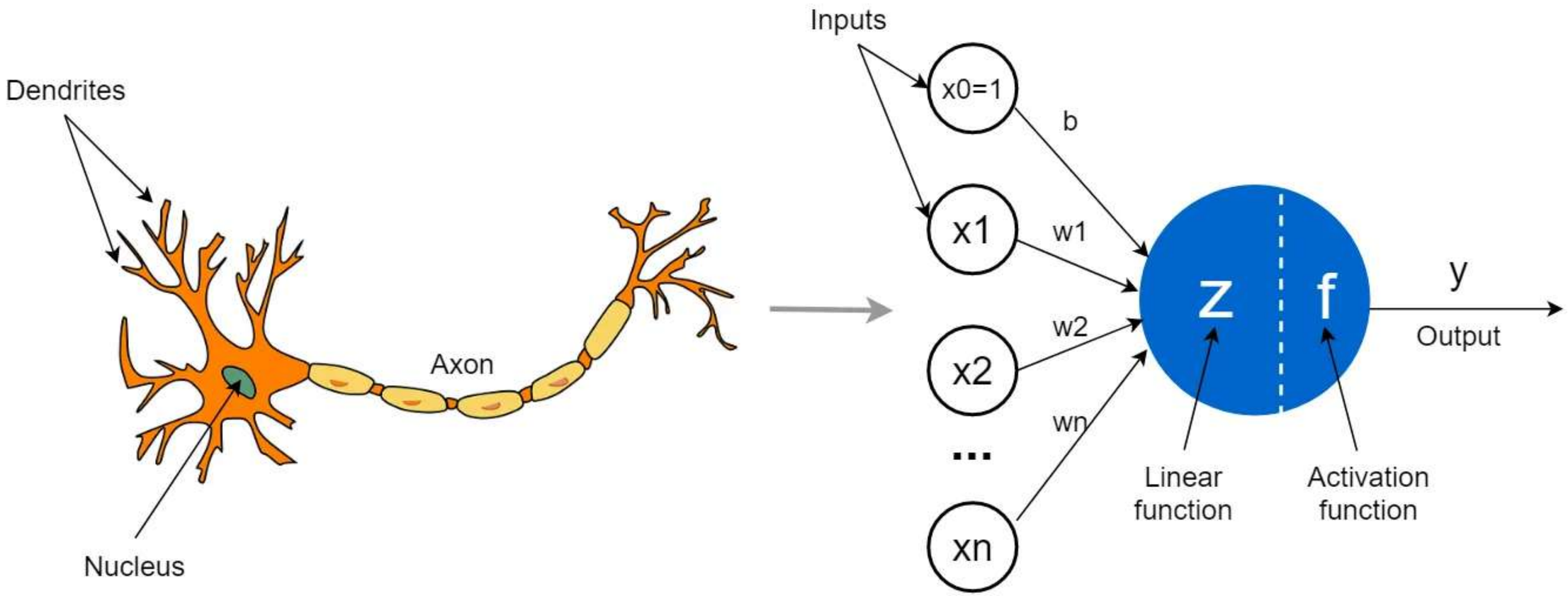
Analytic AI	Cognitive AI	Generative AI
진단, 예측	이미지	이미지, 영상
식별, 해석	음성, 문자	음성, 문자
분석	인지	생성


Rule-Based AI
Symbolic AI
Artificial Intelligence
Machine Learning
Linear Regression
Decision Tree
Random Forest
Deep Learning



https://www.brainvire.com/blog/top-9-problems-with-big-data/

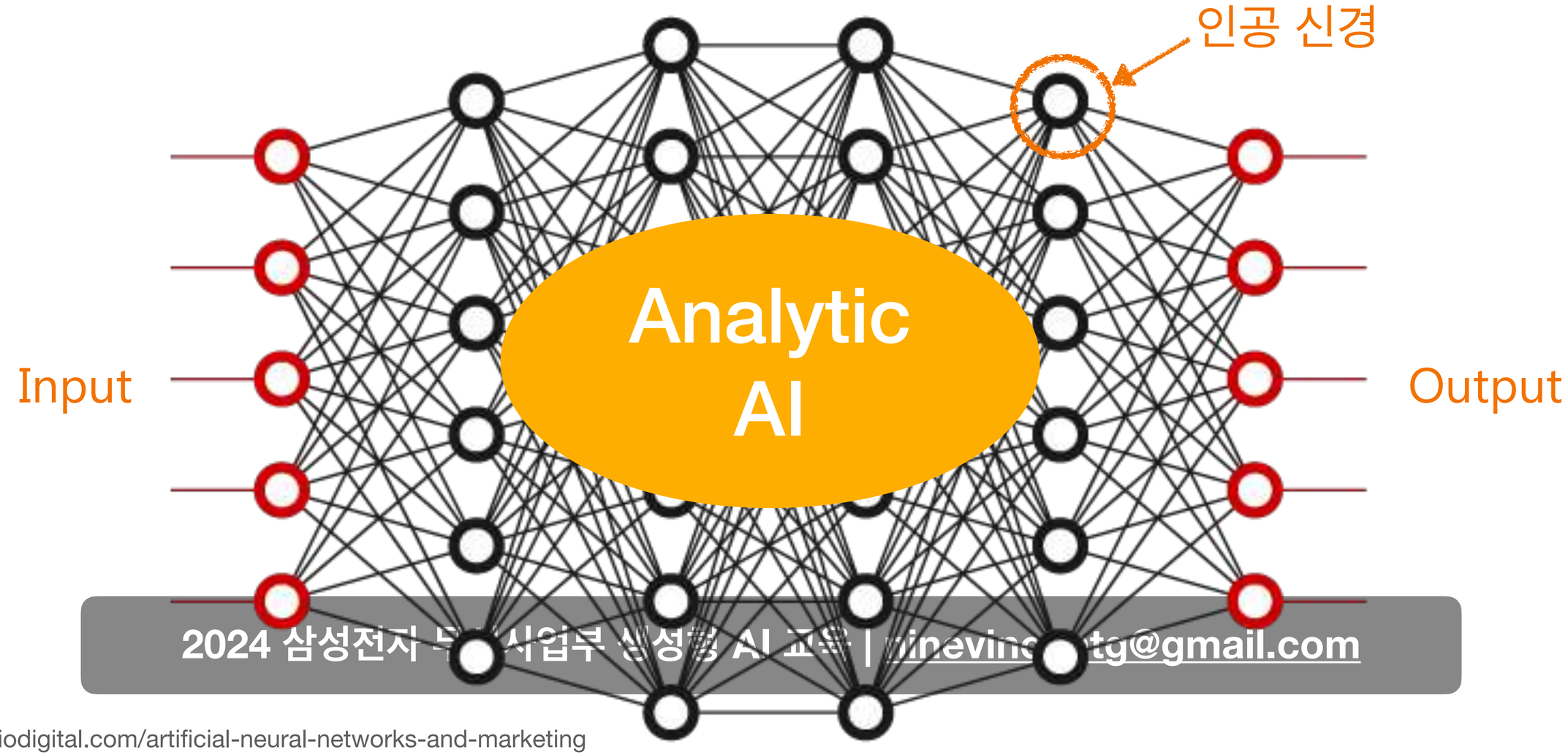
Artificial Neuron



https://towardsdatascience.com/the-concept-of-artificial-neurons-perceptrons-in-neural-networks-fab22249cbfc

Artificial Neural Network

MLP(Multi Layer Perceptron)



Analytic AI

2024 삼성전자 무선사업부 상생형 AI 교육 | minevingtg@gmail.com

https://blog.siodigital.com/artificial-neural-networks-and-marketing

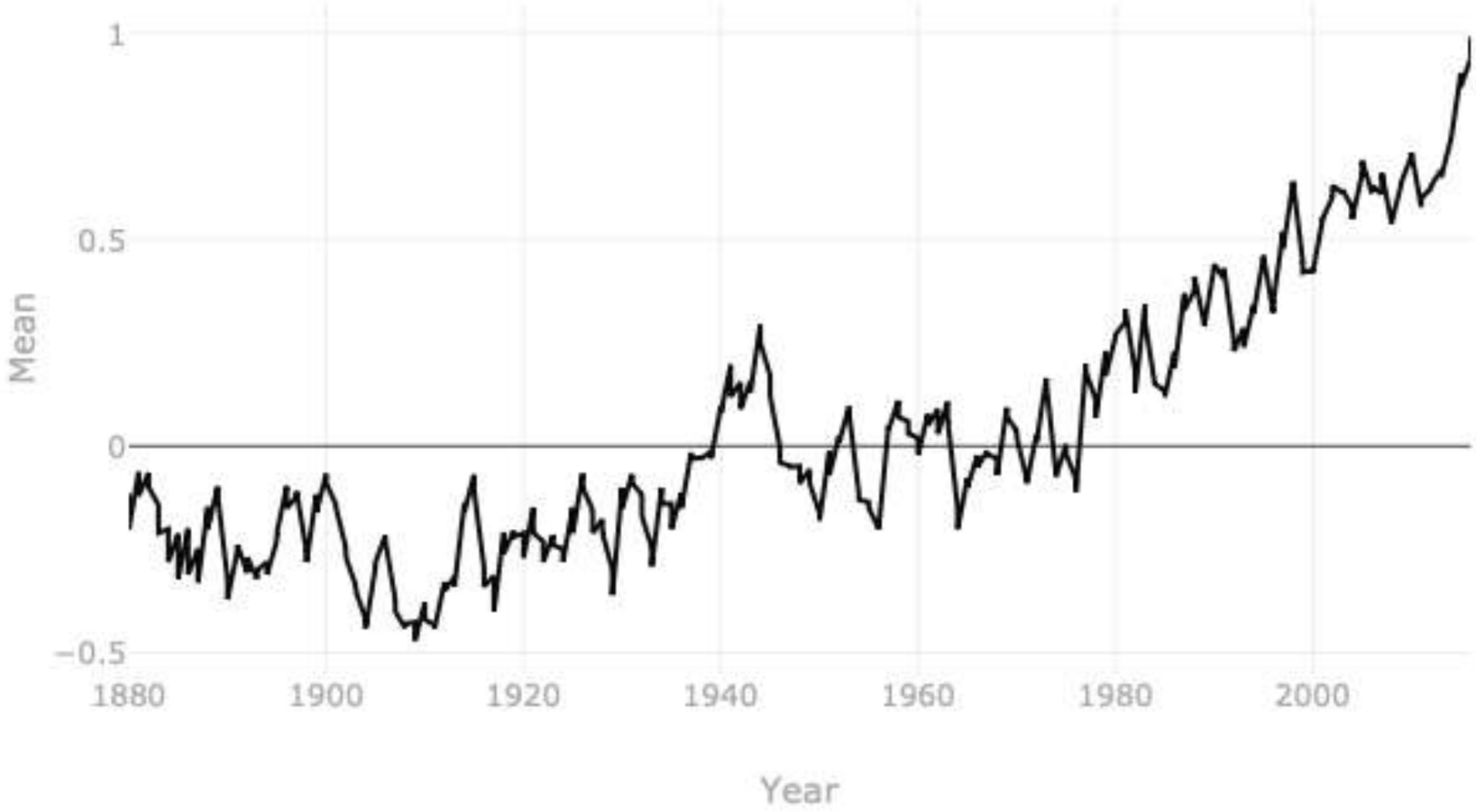
Sequence Data

순서에 영향을 받는 데이터

Sequence Data

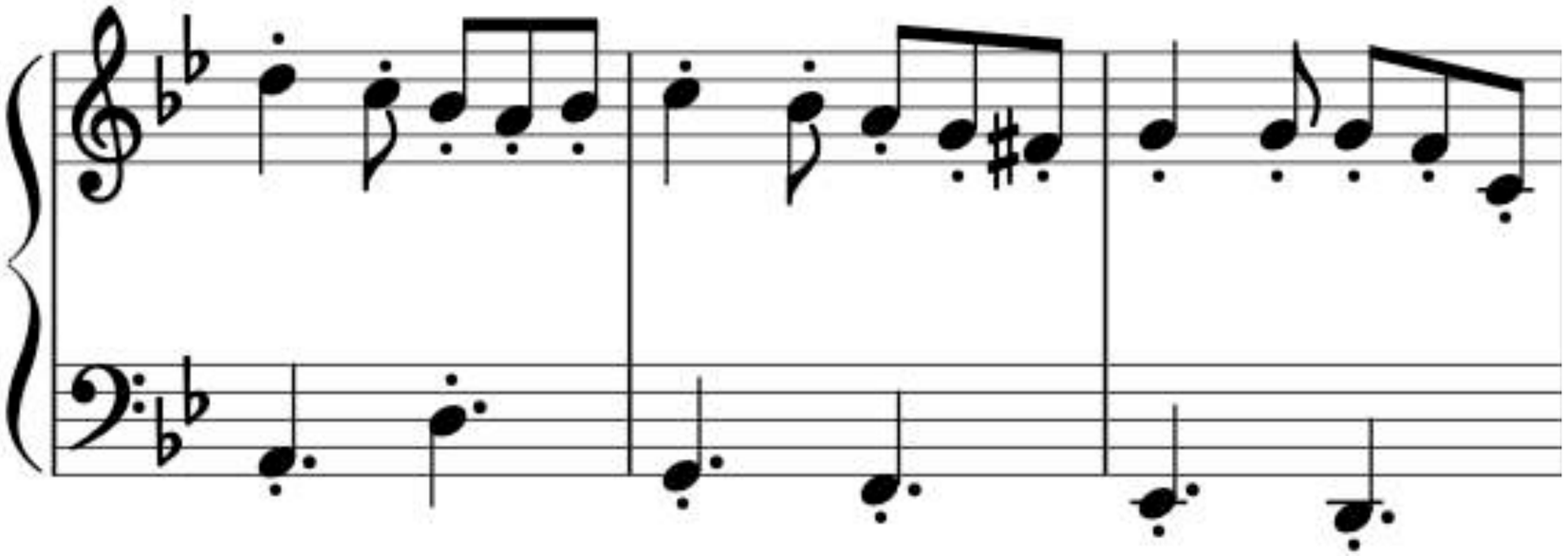
“반갑습니다, 여러분!
공부하기 좋은 날씨네요”

Sequence Data



Reference : https://datahub.io/core/global-temp

Sequence Data



Music Note

Reference : Howl's Moving Castle OST

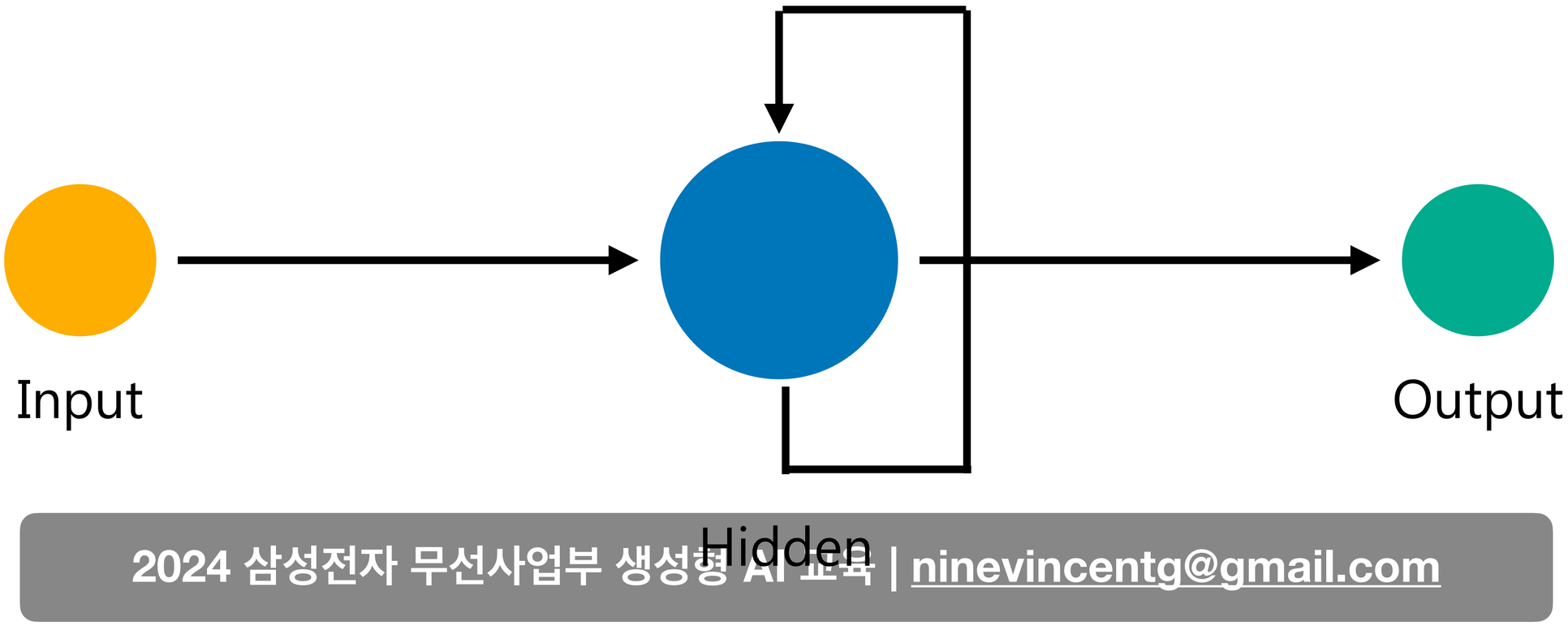
RNN(Recurrent Neural Network)

Sequence Data 예측에 뛰어난 알고리즘

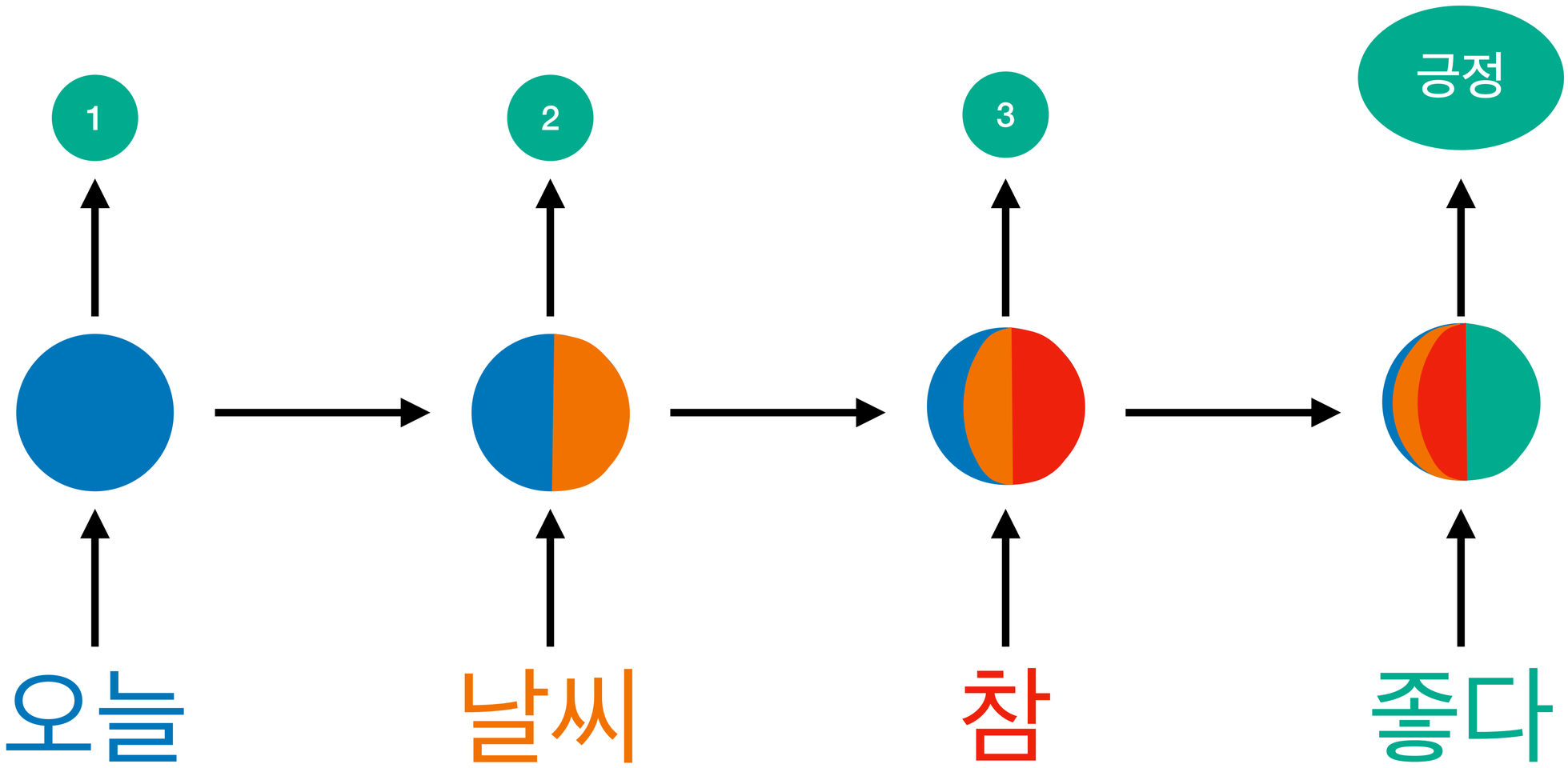
Sequential Memory

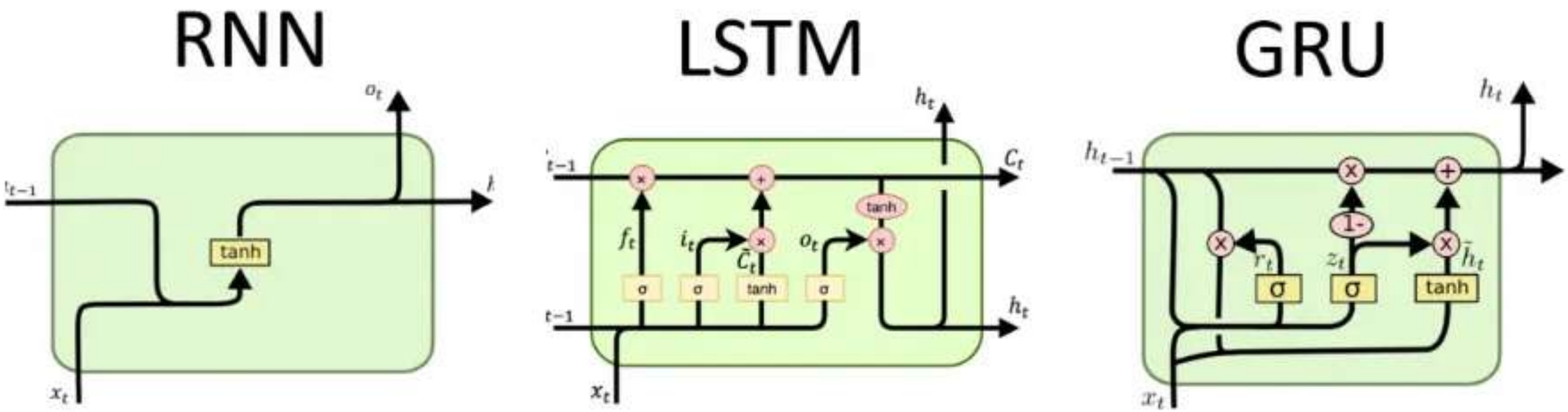
RNN(Recurrent Neural Network)

Sequence Data 예측에 뛰어난 알고리즘



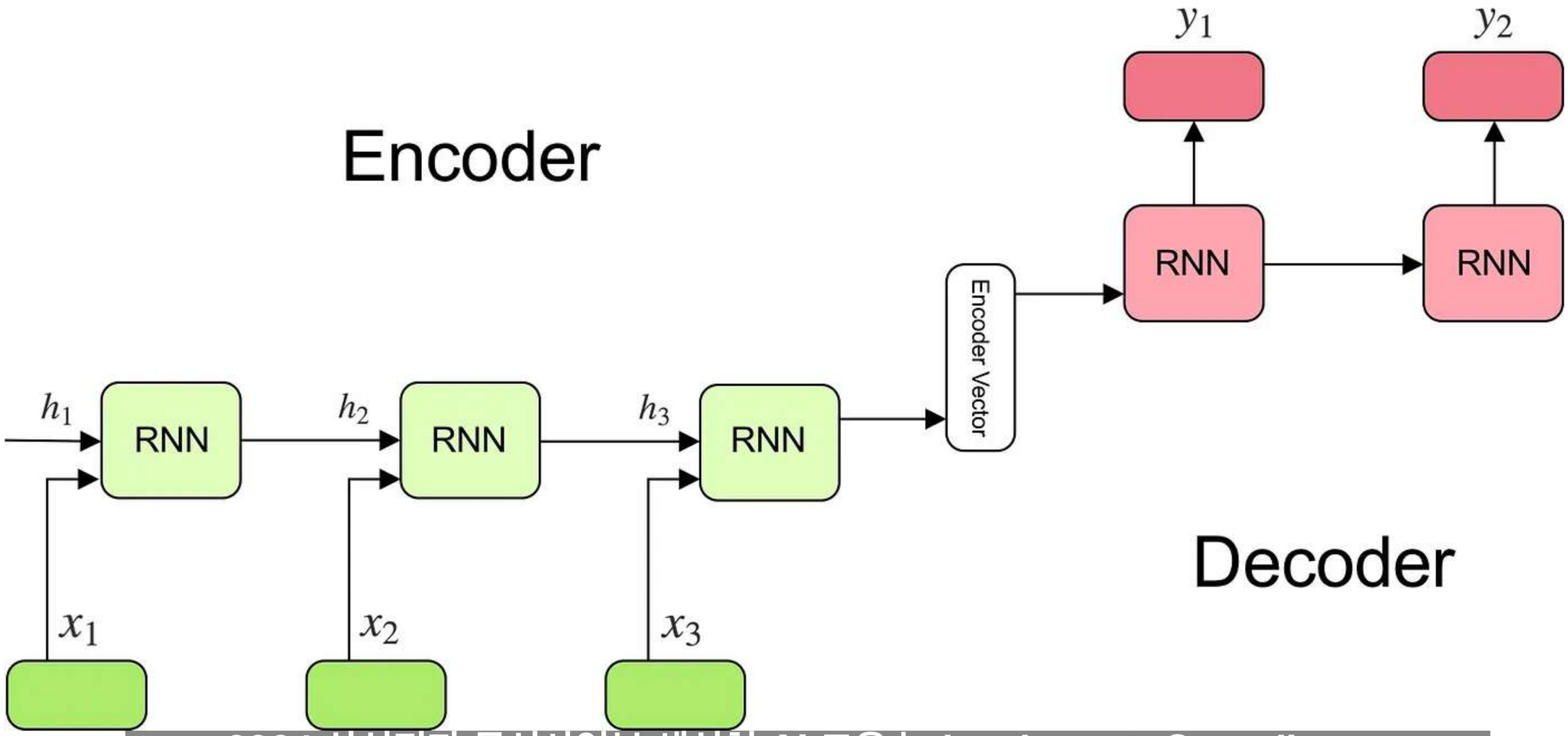
오늘 날씨 참 좋다


1
2
3
긍정
오늘
날씨
참
좋다



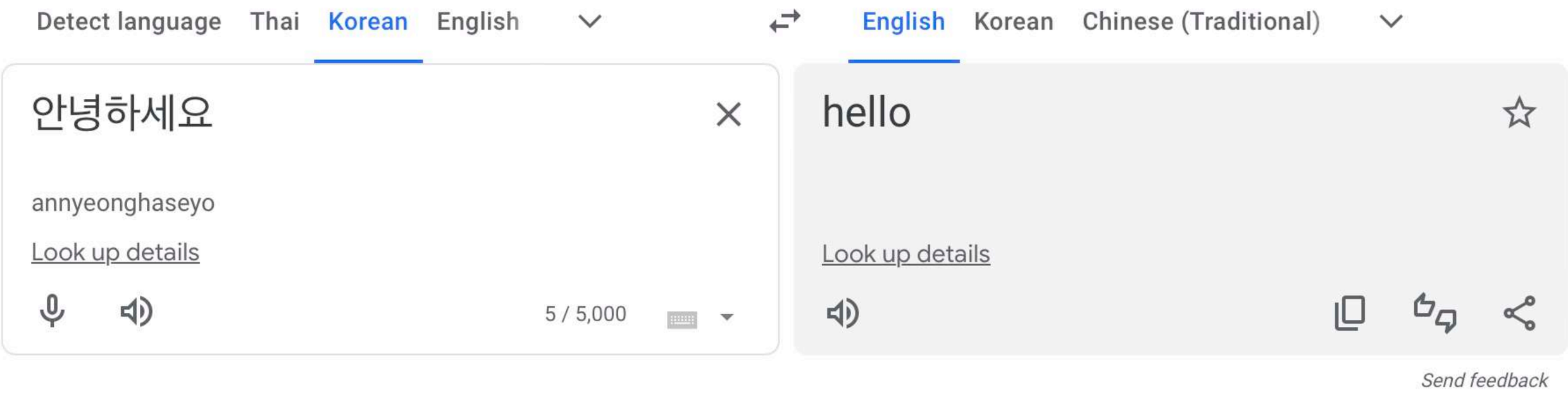
2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincentg@gmail.com

https://medium.com/@le.oasis/recurrent-neural-networks-report-c70e6f05cc9e



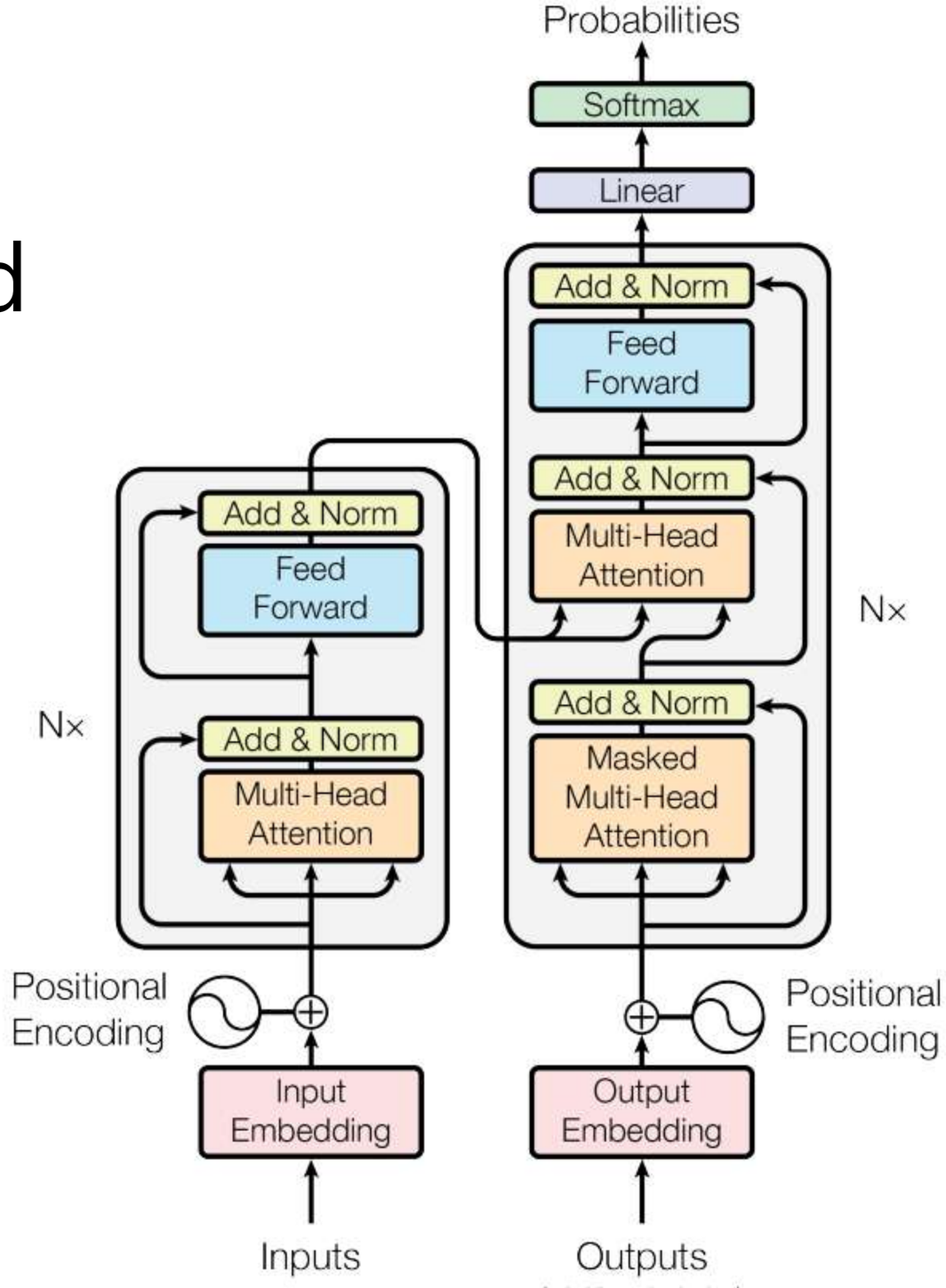
https://towardsdatascience.com/understanding-encoder-decoder-sequence-to-sequence-model-679e04af4346

Sequence to Sequence(Seq2Seq)



Attention is All You Need

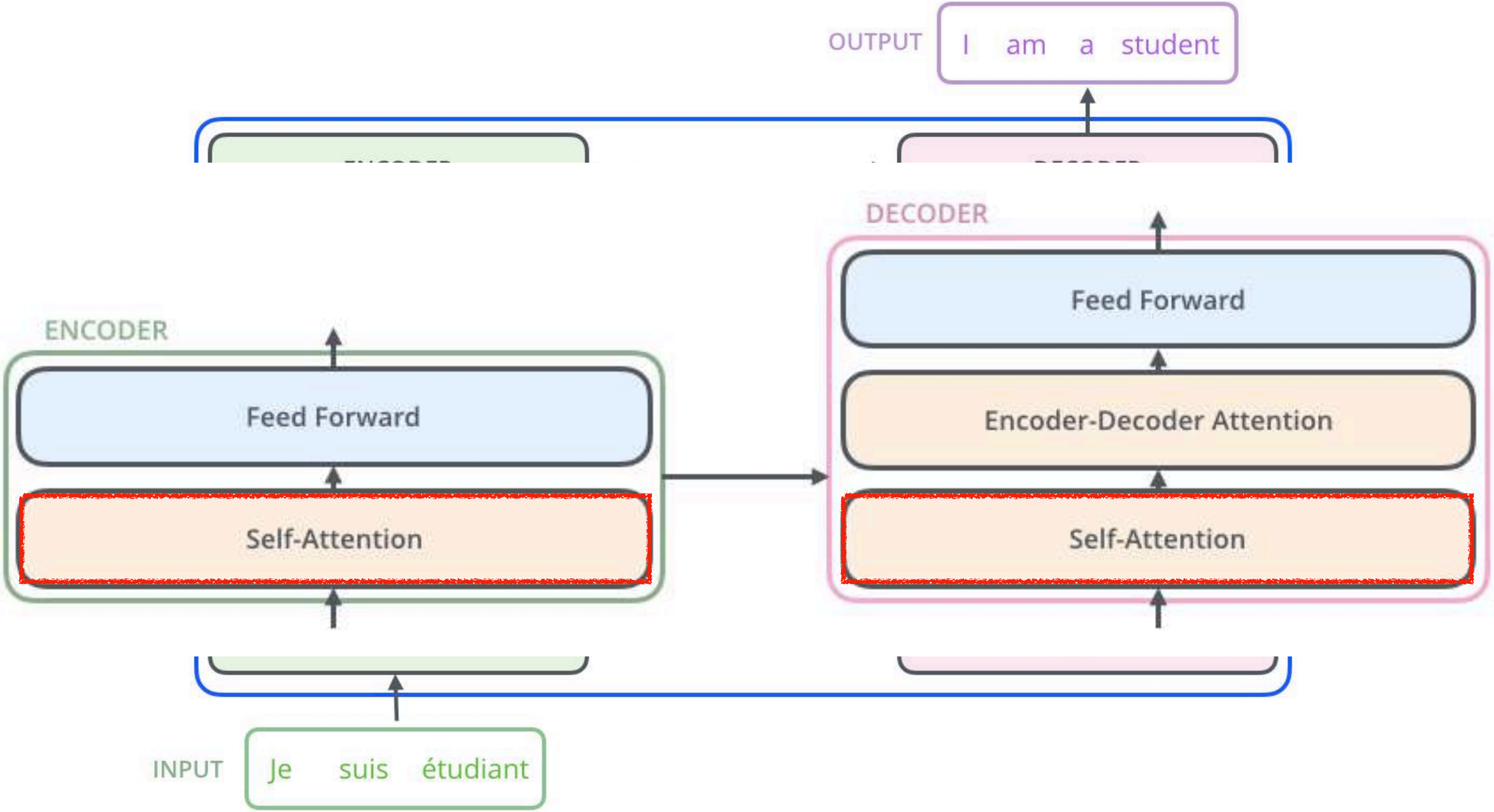
Transformer Network



2024 삼성전자 무선사업부 생성형 A

https://arxiv.org/abs/1706.03762

http://jalammar.github.io/illustrated-transformer/



http://jalammar.github.io/illustrated-transformer/

2017

Transformer Network

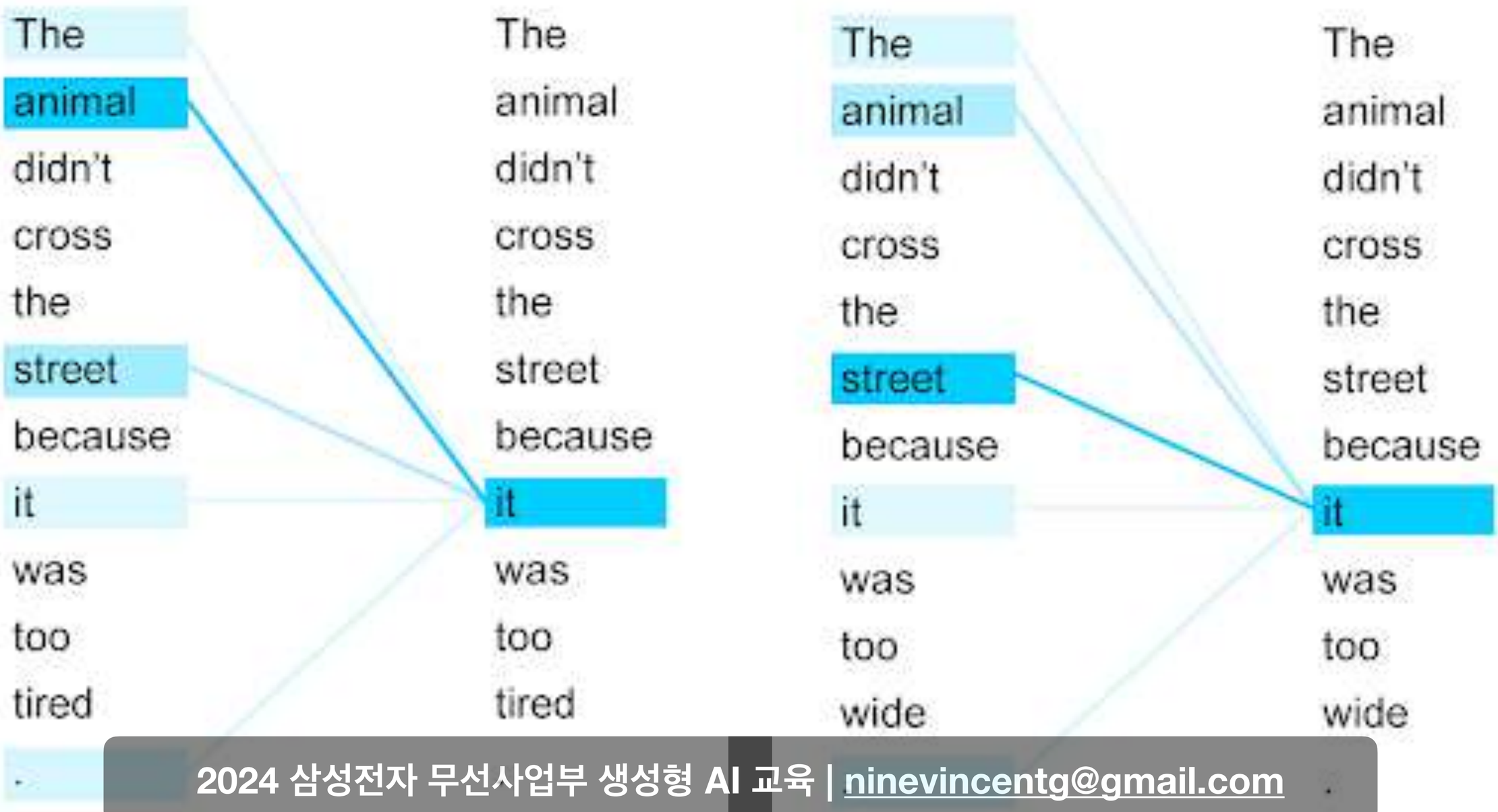
2014

Attention Algorithm

Pay Attention!

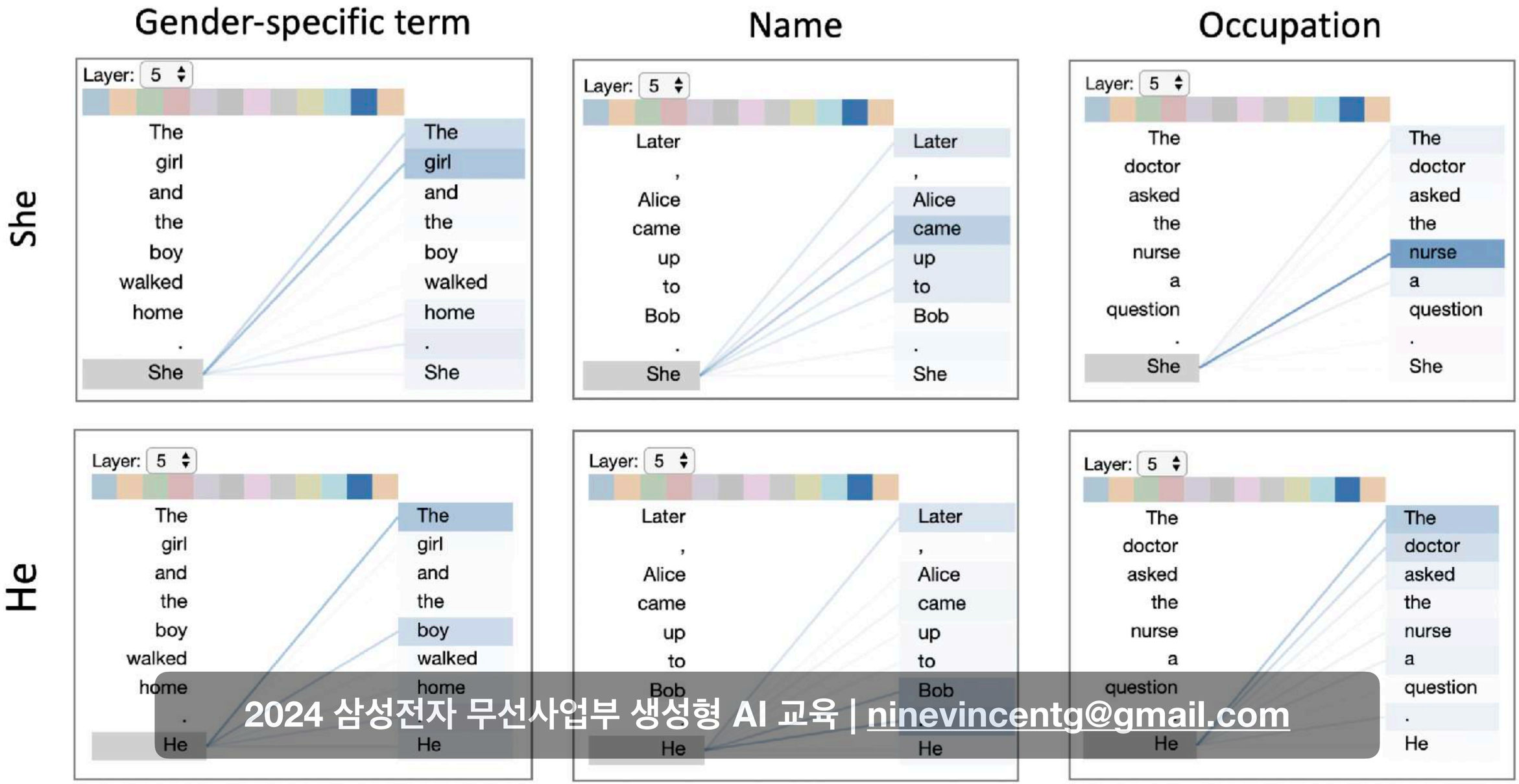
데이터 중 어디에 가중치를 더 둘것인가

The animal didn’t cross the street, because *it* was too tired.

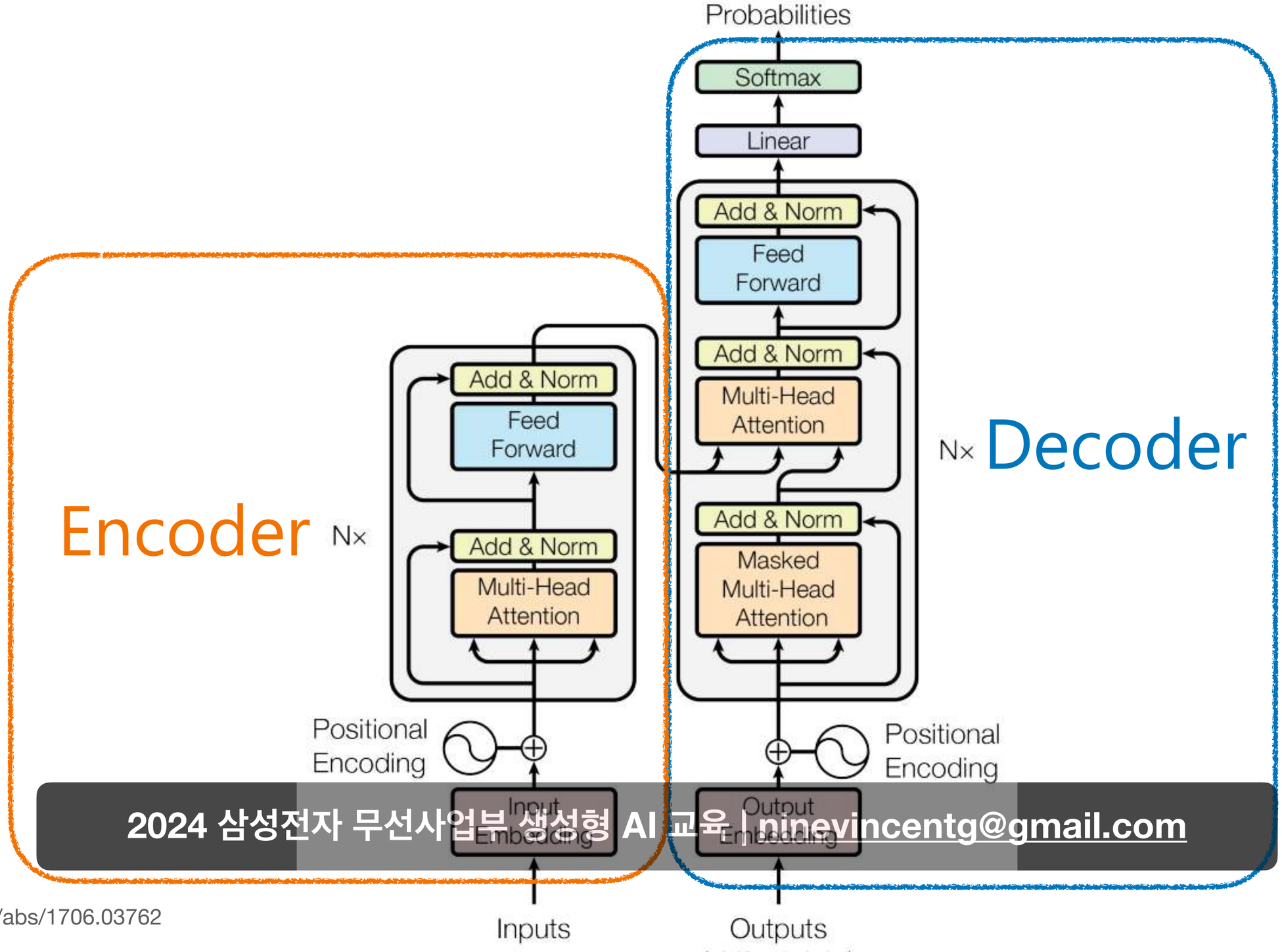


2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincentg@gmail.com

https://blog.research.google/2017/08/transformer-novel-neural-network.html



2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincentg@gmail.com



2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincentg@gmail.com

https://arxiv.org/abs/1706.03762

October 2018

BERT by Google

Bidirectional **E**ncoder **R**epresentations from **T**ransformers

Wikipedia (~2.5B words)

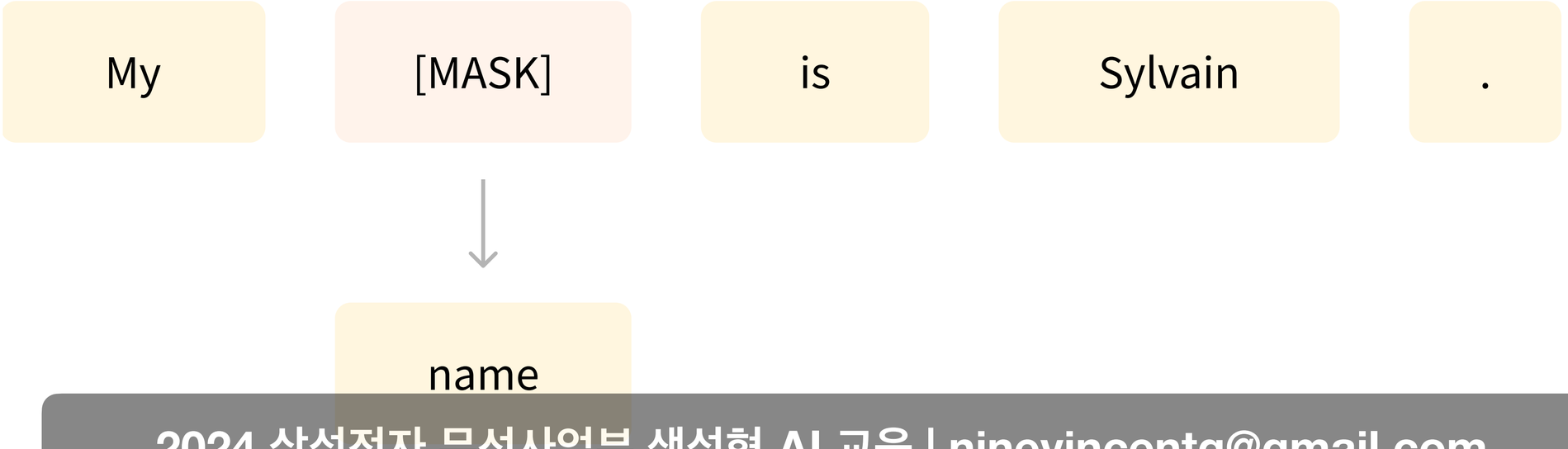
Google’s BooksCorpus (~800M words)

약 33억개 단어

출처 : https://huggingface.co/blog/bert-101

MLM

Masked Language Model



출처 : https://huggingface.co/learn/nlp-course/chapter1/4

MLM

Masked Language Model

Paris is the [MASK] of France.

Compute

Computation time on Intel Xeon 3rd Gen Scalable cpu: cached

capital	0.997
heart	0.001
center	0.000

2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincentg@gmail.com

출처 : https://huggingface.co/blog/bert-101

CLM

Causal Language Model

Auto Regressive Model

My				→	name
My	name			→	is
My	name	is		→	Sylvain
My	name	is	Sylvain	→	.

2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincentg@gmail.com

출처 : https://huggingface.co/learn/nlp-course/chapter1/4

입력	결과
안녕하세요, 무엇	을
안녕하세요, 무엇을	도
안녕하세요, 무엇을 도	와
안녕하세요, 무엇을 도와	드릴
안녕하세요, 무엇을 도와드릴	까
안녕하세요, 무엇을 도와드릴까	요

오
늘
날
씨
가
참
n tokens in
0.01
좋네
0.005
으로
0.003
흐리
0.013
맑다
0.077
화창
...
0.006
새
1 tokens out

GPT-3.5 Turbo

GPT-3.5 Turbo is our fast and inexpensive model for simpler tasks.

`gpt-3.5-turbo-0125` is the flagship model of this family, supports a 16K context window and is optimized for dialog.

`gpt-3.5-turbo-instruct` is an Instruct model and only supports a 4K context window.

Learn about GPT-3.5 Turbo ↗

(영어 기준)

1,000,000 토큰 = 750,000 단어

= 2,500 페이지 = 10권

Model	Input	Output
gpt-3.5-turbo-0125	$0.50 / 1M tokens	$1.50 / 1M tokens
gpt-3.5-turbo-instruct	$1.50 / 1M tokens	$2.00 / 1M tokens

2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincentg@gmail.com

https://openai.com/api/pricing/

GPT-4o

GPT-4o is our most advanced multimodal model that's faster and cheaper than GPT-4 Turbo with stronger vision capabilities. The model has 128K context and an October 2023 knowledge cutoff.

Learn about GPT-4o ↗

Model	Input	Output
gpt-4o	$5.00 / 1M tokens	$15.00 / 1M tokens
gpt-4o-2024-05-13	$5.00 / 1M tokens	$15.00 / 1M tokens

x10

2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincentg@gmail.com

https://openai.com/api/pricing/

GPT-4

Our original GPT-4 model was built with broad general knowledge and domain expertise.

Learn about GPT-4 >

Model	Input	Output
gpt-4	$30.00 / 1M tokens	$60.00 / 1M tokens x40
gpt-4-32k	$60.00 / 1M tokens	$120.00 / 1M tokens

https://openai.com/api/pricing/

https://platform.openai.com/tokenizer

Tokenizer

Learn about language model tokenization

OpenAI's large language models (sometimes referred to as GPT's) process text using **tokens**, which are common sequences of characters found in a set of text. The models learn to understand the statistical relationships between these tokens, and excel at producing the next token in a sequence of tokens.

You can use the tool below to understand how a piece of text might be tokenized by a language model, and the total count of tokens in that piece of text.

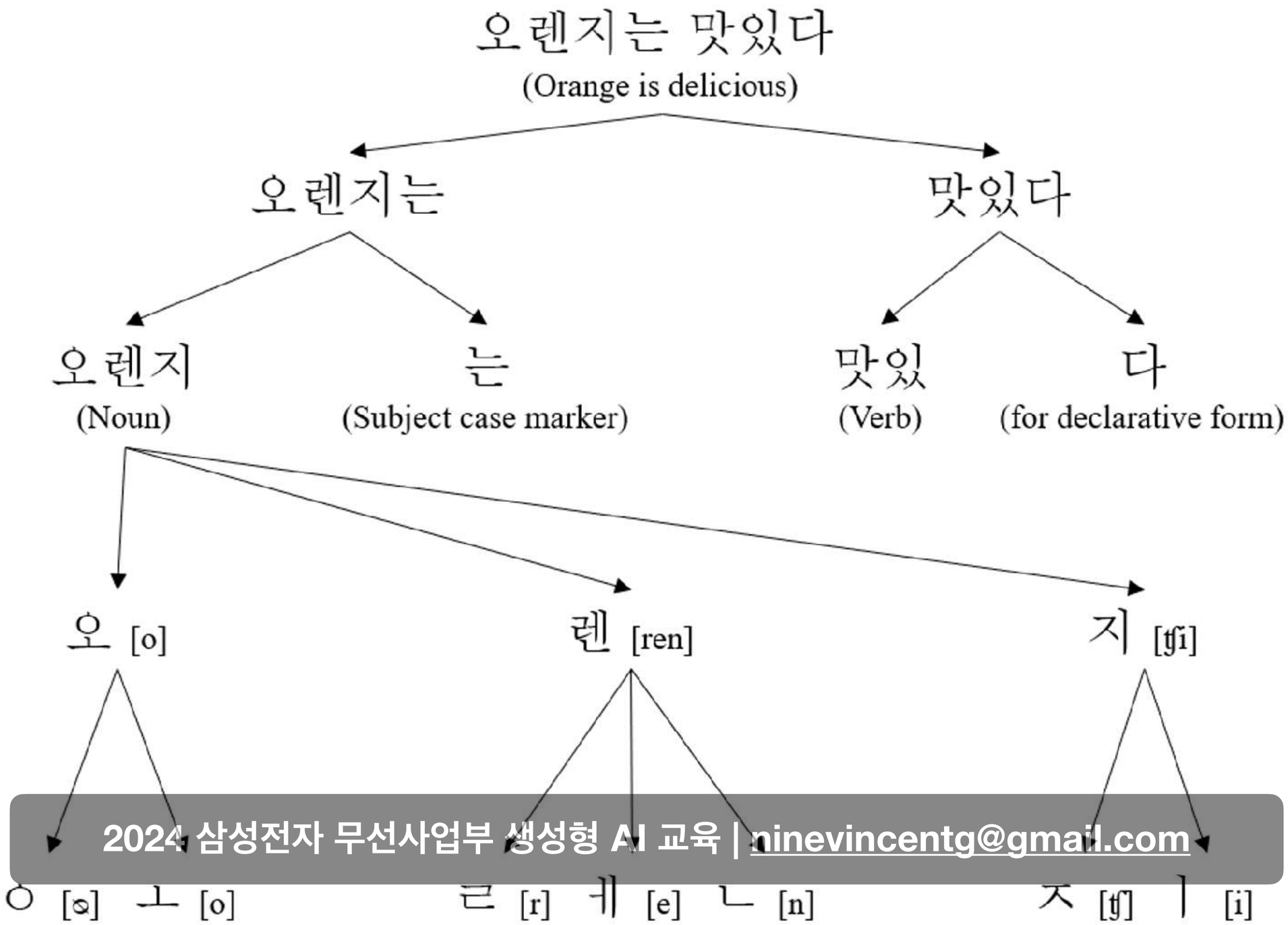
It's important to note that the exact tokenization process varies between mo like GPT-3.5 and GPT-4 use a different tokenizer than previous models, and tokens for the same input text.

GPT-4o (coming soon) | GPT-3.5 & GPT-4 | GPT-3 (Legacy)

안녕하세요, 여러분

Tokens	Characters
9	10

2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincentg@gmail.com



2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincentg@gmail.com

https://smilegate.ai/2022/05/31/korean-tokenizer

https://platform.openai.com/tokenizer

Tokenizer

Learn about language model tokenization

OpenAI's large language models (sometimes referred to as GPT's) process text using **tokens**, which are common sequences of characters found in a set of text. The models learn to understand the statistical relationships between these tokens, and excel at producing the next token in a sequence of tokens.

You can use the tool below to understand how a piece of text might be tokenized by a language model, and the total count of tokens in that piece of text.

It's important to note that the exact tokenization process varies between mo like GPT-3.5 and GPT-4 use a different tokenizer than previous models, and tokens for the same input text.

GPT-4o (coming soon) | GPT-3.5 & GPT-4 | GPT-3 (Legacy)

안녕하세요, 여러분

Tokens	Characters
9	10

2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincentg@gmail.com

“안녕하세요 여러분”

GPT-3.5 & GPT-4 GPT-3 (Legacy)

안녕하세요 여러분

Clear Show example

Tokens	Characters
23	9

GPT-4o (coming soon) | GPT-3.5 & GPT-4 | GPT-3 (Legacy)

안녕하세요, 제 이름은 GPT-4o입니다. 저는 새로운 유형의 언어 모델입니다, 만나서 반갑습니다!

Korean 1.7x fewer tokens (from 45 to 27)

안녕하세요, 제 이름은 GPT-4o입니다. 저는 새로운 유형의 언어 모델입니다, 만나서 반갑습니다!

Clear | Show example

Tokens	Characters
45	55

2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincentg@gmail.com

“안녕하세요 여러분”

CLOVA Studio β　플레이그라운드

토큰 계산기

입력한 문장의 토큰 수를 계산합니다.

토큰 수

3

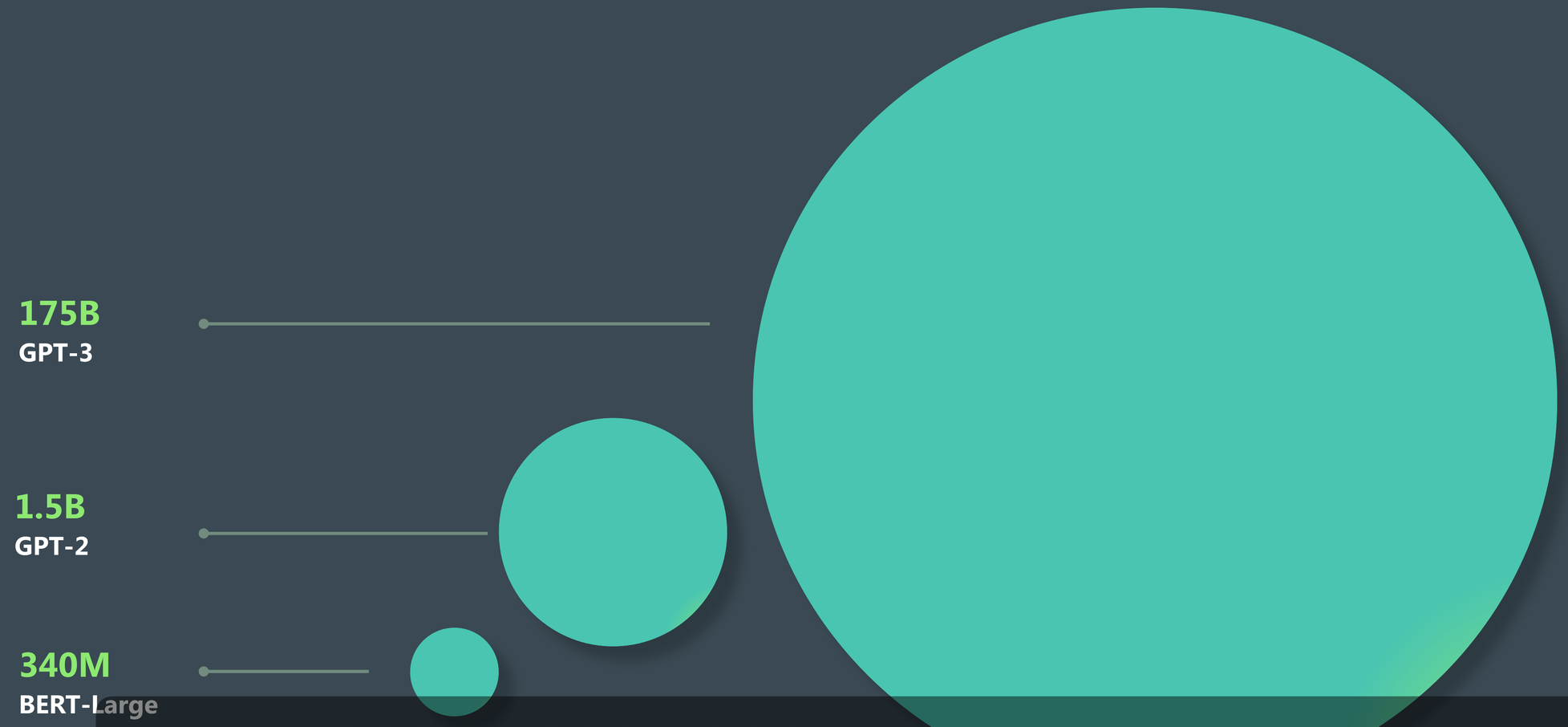
글자 수

공백 포함	공백 제외
9	8

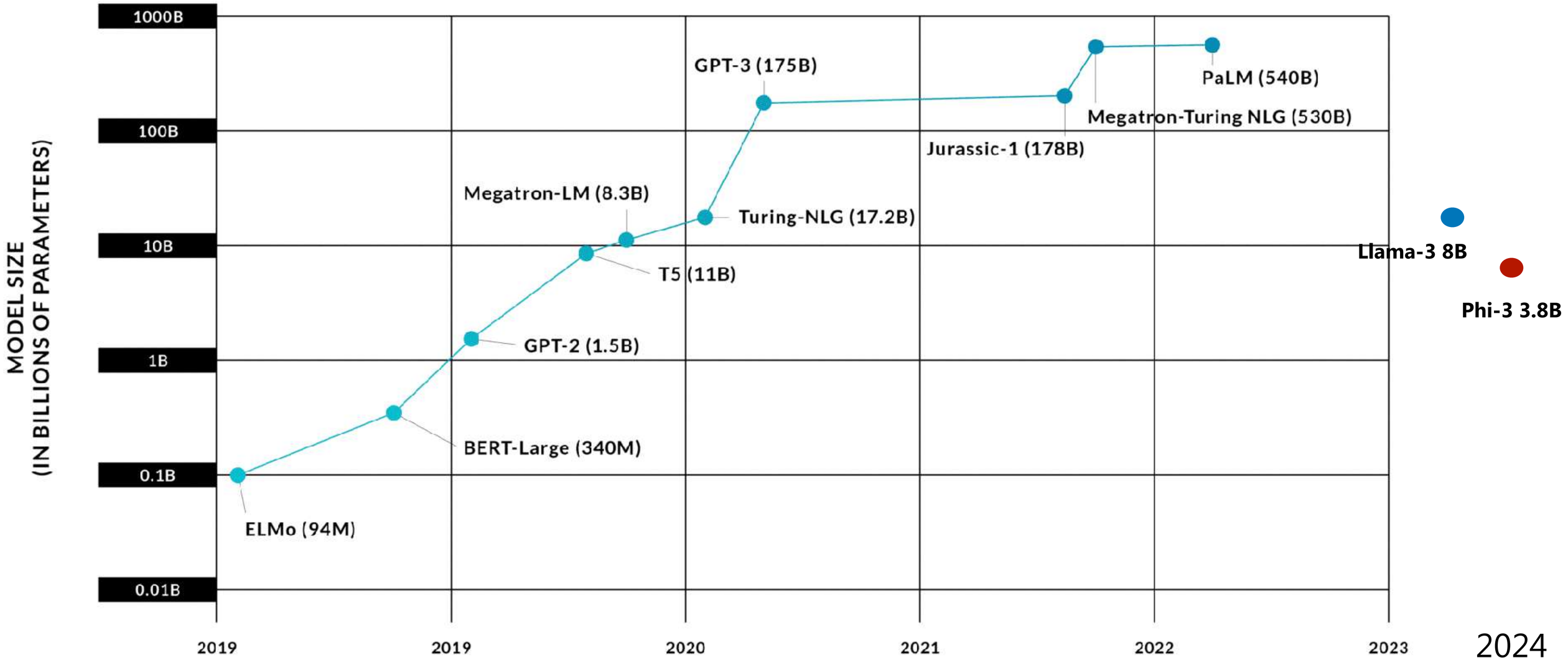
생성형 모델의 발전 동향

1. 모델 사이즈

Foundation models



1. 모델 사이즈



https://cmte.ieee.org/futuredirections/2023/04/24/how-much-bigger-can-should-llms-become/

	Gemini Ultra (pixel only)	Gemini Pro (pixel only)	Gemini Nano 2 (pixel only)	Gemini Nano 1 (pixel only)	GPT-4V	Prior SOTA
MMMU (val) Multi-discipline college-level problems (Yue et al., 2023)	59.4% pass@1 62.4% Maj1@32	47.9%	32.6%	26.3%	56.8%	56.8% GPT-4V, 0-shot
TextVQA (val) Text reading on natural images (Singh et al., 2019)	82.3%	74.6%	65.9%	62.5%	78.0%	79.5% Google PaLI-3, fine-tuned
DocVQA (test) Document understanding (Mathew et al., 2021)	90.9%	88.1%	74.3%	72.2%	88.4% (pixel only)	88.4% GPT-4V, 0-shot
ChartQA (test) Chart understanding (Masry et al., 2022)	80.8%	74.1%	51.9%	53.6%	78.5% (4-shot CoT)	79.3% Google DePlot, 1-shot PoT
InfographicVQA (test) Infographic understanding (Mathew et al., 2022)	80.3%	75.2%	54.5%	51.1%	75.1% (pixel only)	75.1% GPT-4V, 0-shot
MathVista (testmini) Mathematical reasoning (Lu et al., 2023)	53.0%	45.2%	30.6%	27.3%	49.9%	49.9% GPT-4V, 0-shot
AI2D (test) Science diagrams (Kembhavi et al., 2016)	79.5%	73.9%	51.0%	37.9%	78.2%	81.4% Google PaLI-X, fine-tuned
VQAv2 (test-dev) Natural image understanding (Goyal et al., 2017)	77.8%	71.2%	67.5%	62.7%	77.2%	86.1% Google PaLI-X, fine-tuned

2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincentg@gmail.com

Table 7 | **Image understanding** Gemini Ultra consistently outperforms existing approaches even in

2024.04.18.

	Meta Llama 3 8B	Gemma 7B - It Measured	Mistral 7B Instruct Measured
MMLU 5-shot	68.4	53.3	58.4
GPQA 0-shot	34.2	21.4	26.3
HumanEval 0-shot	62.2	30.5	36.6
GSM-8K 8-shot, CoT	79.6	30.6	39.9
MATH 4-shot, CoT	30.0	[illegible]	[illegible]

	Meta Llama 3 70B	Gemini Pro 1.5 Published	Claude 3 Sonnet Published
MMLU 5-shot	82.0	81.9	79.0
GPQA 0-shot	39.5	41.5 CoT	38.5 CoT
HumanEval 0-shot	81.7	71.9	73.0
GSM-8K 8-shot, CoT	93.0	91.7 11-shot	92.3 0-shot
MATH 4-shot, CoT	50.4	58.5 Minerva prompt	40.5

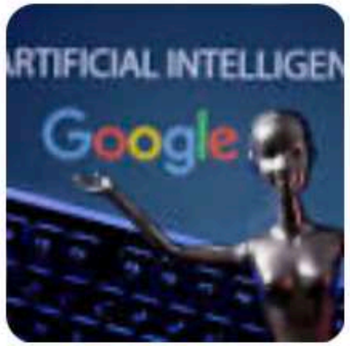
2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincentg@gmail.com

https://llama.meta.com/llama3/

1. 모델 사이즈

구글, AI 챗봇 개발용 경량 최첨단 언어모델 '젬마' 내놔

구글, AI 챗봇 개발용 경량 최첨단 언어모델 젬마 내놔 구글이 인공지능AI 챗봇을 개발할 수 있는 거대 언어모델LLM의 가벼운 버전을 외부 개발자들...



서울경제

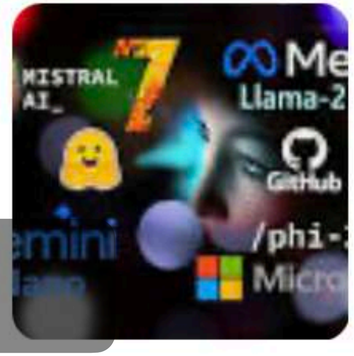
업스테이지, 경량형 언어모델 '솔라' 베타 시작

업스테이지, 경량형 언어모델 '솔라' 베타 시작 ... 인공지능(AI) 기업 업스테이지가 사전학습 언어 모델 '솔라'의 베타 버전을 출시하고 내달 31일까지...

인공지능신문

누구나 온디바이스에서 생성 AI 구현하는 시대로...메타 ·구글·미스트랄에 이어 MS, sLLM '파이 2' 완전 오픈 소스화!

매개변수(Parameter)를 기존 대형언어모델(Large language models. 이하, LLM)에 비해 변수의 수가 60억(6B) 내지 100억(10B) 개로 줄여 학습을 위한...



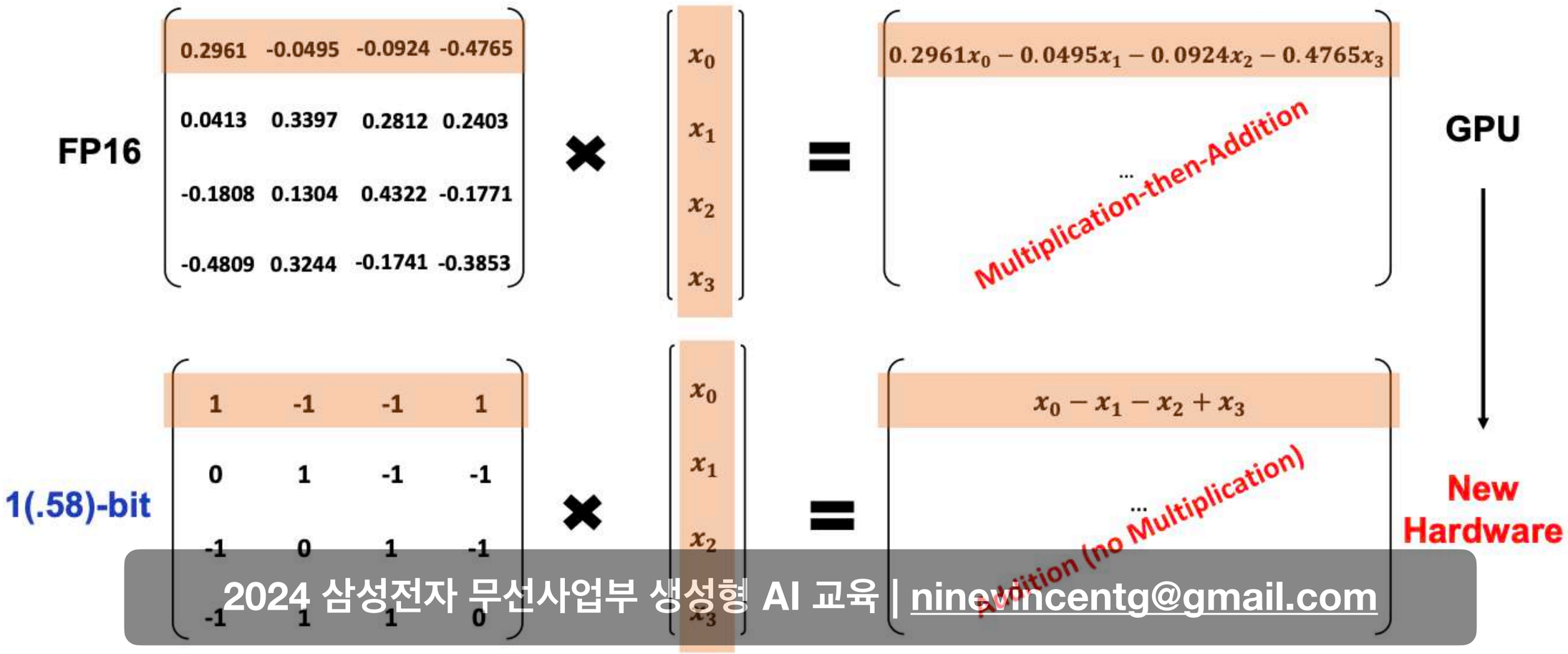
2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincentg@gmail.com

1. 모델 사이즈

Submitted on 27 Feb 2024

The Era of 1-bit LLMs: All Large Language Models are in 1.58 Bits

Shuming Ma, Hongyu Wang, Lingxiao Ma, Lei Wang, Wenhui Wang, Shaohan Huang, Li Dong, Ruiping Wang, Jilong Xue, Furu Wei



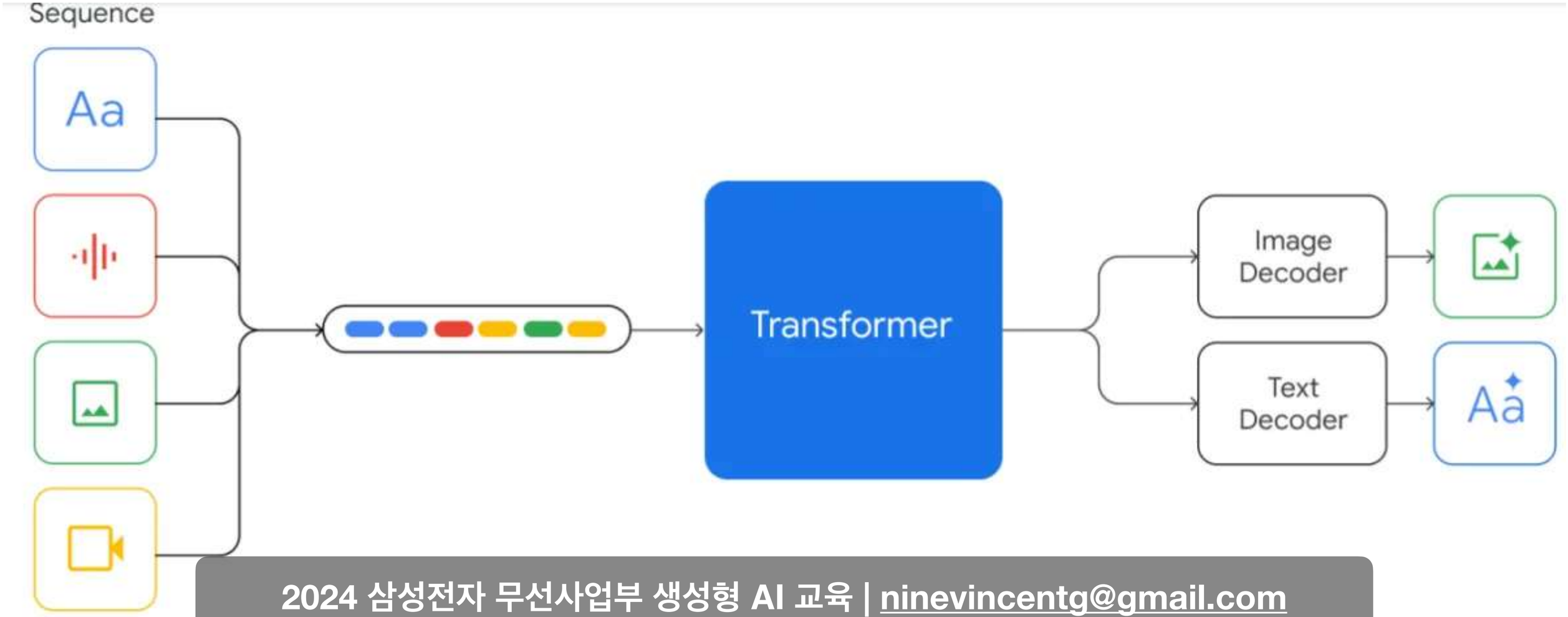
2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincentg@gmail.com

The Era of 1-bit LLMs: All Large Language Models are in 1.58 Bits : https://arxiv.org/pdf/2402.17764

Models	Size	Memory (GB)↓	Latency (ms)↓	PPL↓
LLaMA LLM	700M	2.08 (1.00x)	1.18 (1.00x)	12.33
BitNet b1.58	700M	0.80 (2.60x)	0.96 (1.23x)	12.87
LLaMA LLM	1.3B	3.34 (1.00x)	1.62 (1.00x)	11.25
BitNet b1.58	1.3B	1.14 (2.93x)	0.97 (1.67x)	11.29
LLaMA LLM	3B	7.89 (1.00x)	5.07 (1.00x)	10.04
BitNet b1.58	3B	**2.22 (3.55x)**	**1.87 (2.71x)**	**9.91**
BitNet b1.58	3.9B	**2.38 (3.32x)**	**2.11 (2.40x)**	**9.62**

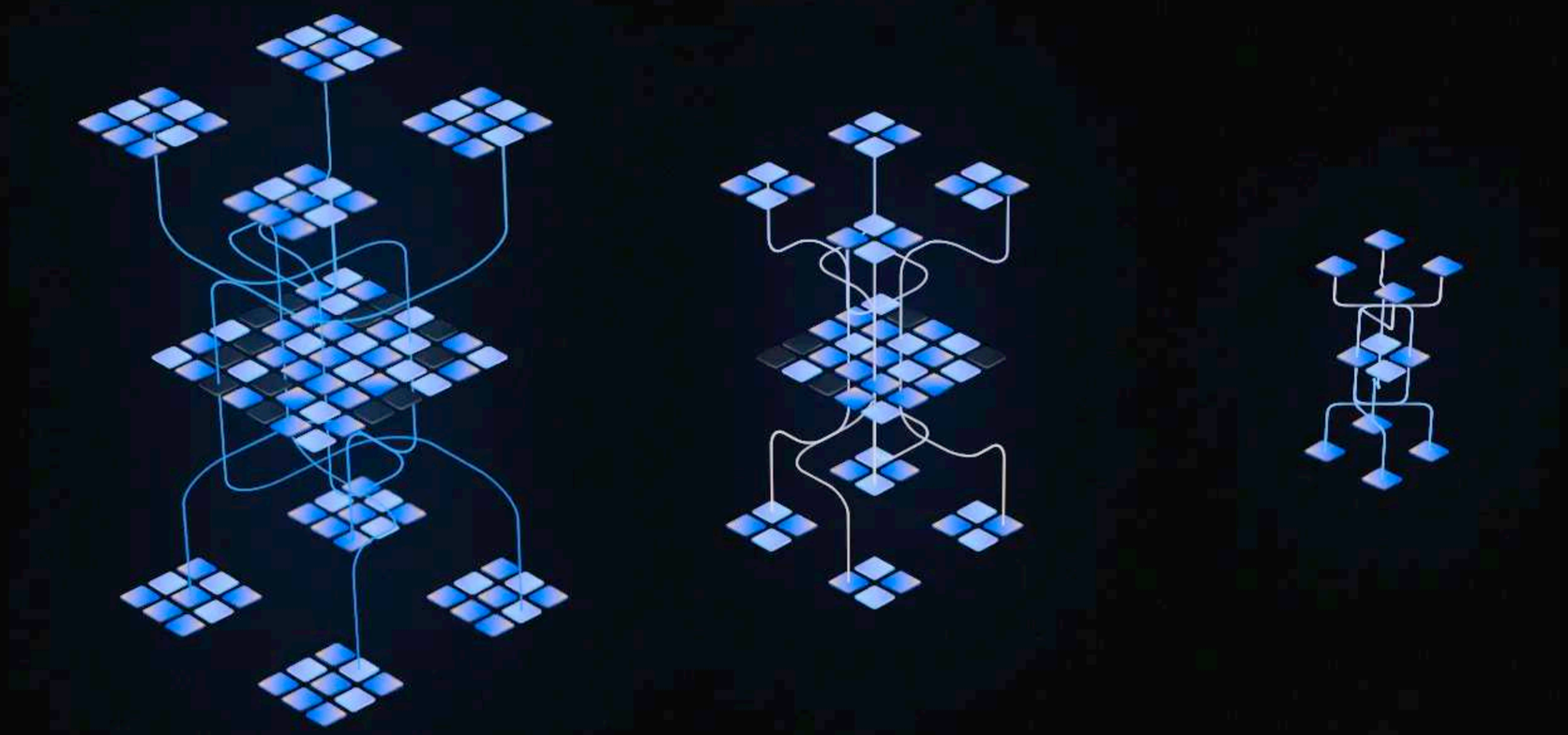
The Era of 1-bit LLMs: All Large Language Models are in 1.58 Bits : https://arxiv.org/pdf/2402.17764

2. Multimodality



2. Multimodality

Gemini



2. Multimodality

May 13, 2024

Hello GPT-4o

We're announcing GPT-4o, our new flagship model that can reason across audio, vision, and text in real time.

LLaVA: Large Language and Vision Assistant

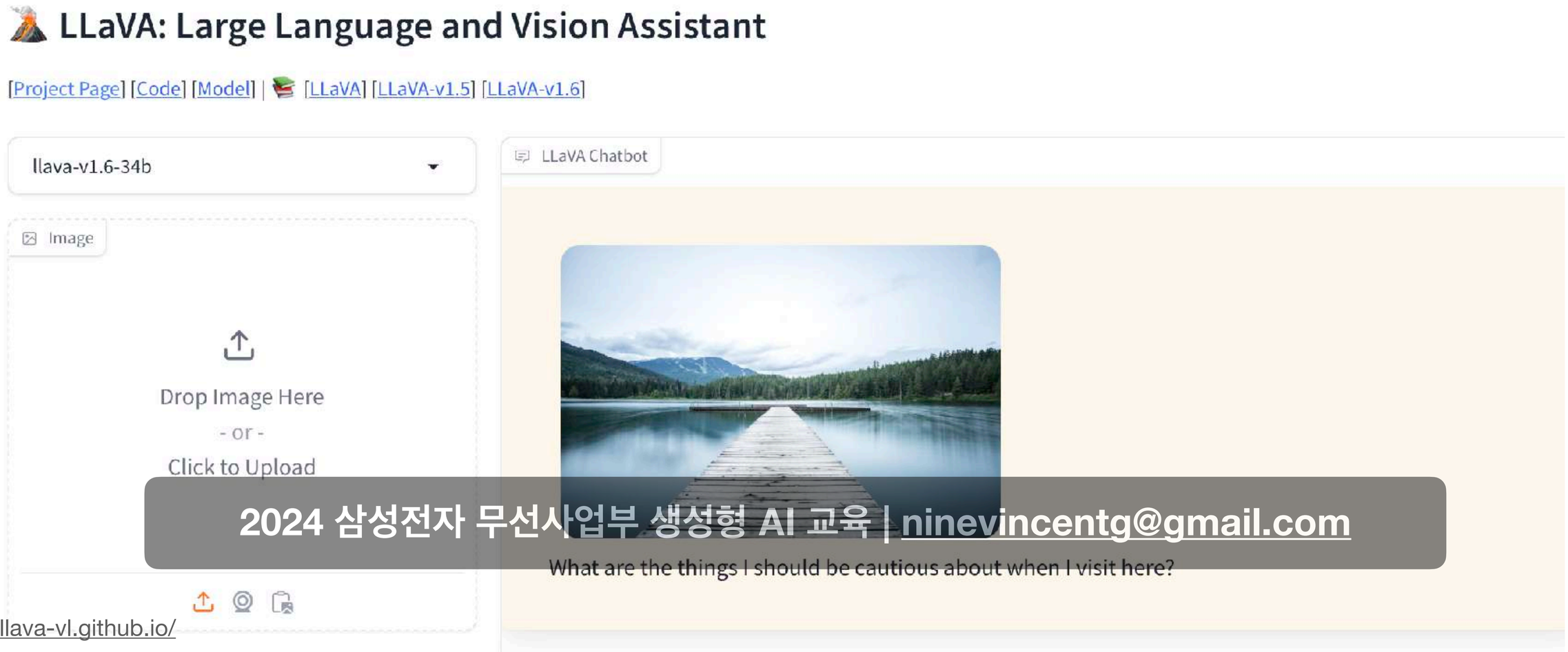
Visual Instruction Tuning

NeurIPS 2023 (Oral)

Haotian Liu*, Chunyuan Li*, Qingyang Wu, Yong Jae Lee

University of Wisconsin-Madison · Microsoft Research · Columbia University

*Equal Contribution



2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincentg@gmail.com

https://llava-vl.github.io/

MM1: Methods, Analysis & Insights from Multimodal LLM Pre-training

Brandon McKinzie°, Zhe Gan°, Jean-Philippe Fauconnier*,
Sam Dodge*, Bowen Zhang*, Philipp Dufter*, Dhruti Shah*, Xianzhi Du*,
Futang Peng, Floris Weers, Anton Belyi, Haotian Zhang, Karanjeet Singh,
Doug Kang, Ankur Jain, Hongyu Hè, Max Schwarzer, Tom Gunter,
Xiang Kong, Aonan Zhang, Jianyu Wang, Chong Wang, Nan Du, Tao Lei,
Sam Wiseman, Guoli Yin, Mark Lee, Zirui Wang, Ruoming Pang,
Peter Grasch*, Alexander Toshev†, and Yinfei Yang†

Apple

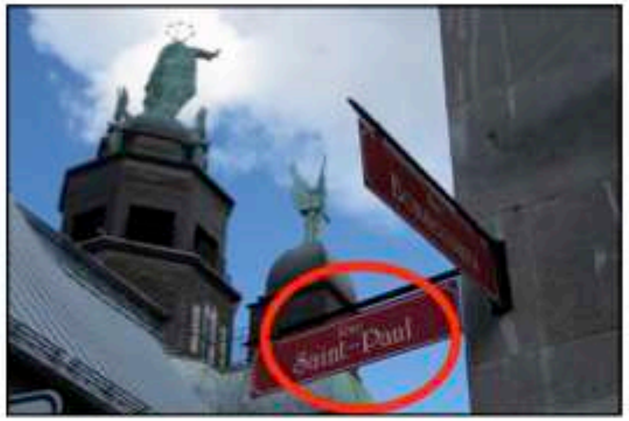
bmckinzie@apple.com, zhe.gan@apple.com

°First authors; °Core authors; †Senior authors

2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincentg@gmail.com

https://arxiv.org/pdf/2403.09611

2. Multimodality



Red circle: "Rue Saint-Paul"

Red circle:

MM1-30B (Ours):

"Hyde Park"

total:
$6 + 4 = 10$

MM1-30B (Ours):

total: $4 + 1 = 5$

https://arxiv.org/pdf/2403.09611

3. Multilingual

✔ GPT by OpenAI

✔ Gemini by Google

LLaMA by Meta

✔ Claude by Anthropic

✔ HyperCLOVA by Naver

Gemma by Google

Phi by Microsoft

Mistral by Mistral AI

3. Multilingual

ANTHROP\C

Claude

AI research and products that put safety at the frontier

3. Multilingual

Claude 2.0

Casual greeting

can you speak Korean?

저는 한국어를 말할 수 없습니다. 제가 영어로 대화할 수 있습니다.

Copy

한국어로 말해보자

죄송합니다. 제 한국어 실력은 제한적입니다. 영어로 계속 대화해주시겠어요?

Copy

아냐아냐 너 한국어 잘 할 수 있어

죄송합니다. 제가 한국어를 할 수 있다고 말씀드린 적은 없습니다. 저는 인공지능 어시스턴트로, 영어 회화를 주로 하고 있습니다. 한국어 실력이 부족하여 한국어로 회화를 이어가기 어려운 것 같습니다. 불편을 드린 점 정말 죄송합니다. 앞으로 기회가 되면 한국어 실력 향상을 위해 노력하겠습니다.

Copy

3. Multilingual

2024.03.04. Claude 3.0

M 안녕?

안녕하세요. 한국어로 대화하는 것을 기쁘게 생각합니다. 궁금한 점이나 도움이 필요한 부분이 있다면 언제든 물어보세요. 최선을 다해 답변해 드리겠습니다.

Copy Retry

Claude can make mistakes. Please double-check responses.

3. Multilingual

2024.04.18.

2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincentg@gmail.com

https://ai.meta.com/blog/meta-llama-3/

3. Multilingual

Training data

To train the best language model, the curation of a large, high-quality training dataset is paramount. In line with our design principles, we invested heavily in pretraining data. Llama 3 is pretrained on over 15T tokens that were all collected from publicly available sources. Our training dataset is seven times larger than that used for Llama 2, and it includes four times more code. To prepare for upcoming multilingual use cases, over 5% of the Llama 3 pretraining dataset consists of high-quality non-English data that covers over 30 languages. However, we do not expect the same level of performance in these languages as in English.

https://ai.meta.com/blog/meta-llama-3/

3. Multilingual

llama-3-70b-instruct

🔍 Expand to see the descriptions of 44 models

Scroll down and start chatting

클래식 한국 레시피! 🍞 여기에는 전통적인 한국 소금빵을 만드는简单한 레시피가 있습니다.

재료:

- 2컵 중력분
- 1/2 티스푼 소금
- 1/4 티스푼 설탕
- 1/2 패킷 (1 1/2 티스푼) 드라이 イースト
- 1컵 따뜻한 물
- 브러싱용 식용유

조리순서:

1. 큰 믹싱 볼에 중력분, 소금, 설탕을 섞어줍니다.
2. 따로의 볼에 イースト를 따뜻한 물과 섞어줍니다. 5-10분 동안 기다려 イースト가 거품이 있도록 합니다.
3. イースト 혼합물을 건조 재료에 추가하여 반죽을 형성합니다.
4. 10-15분 동안 반죽을 밀어줍니다.까지 반죽이 평滑하고 탄력이 있도록 합니다.
5. 그리스된 볼에 반죽을 넣고, 플라스틱 랩으로 덮어 뜨겁게 따뜻한 곳에 1시간 동안 오르도록 합니다. 또는 크기가 2배로 증가할 때까지.

2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincentg@gmail.com

3. Multilingual

Our goals for Llama 3

With Llama 3, we set out to build the best open models that are on par with the best proprietary models available today. We wanted to address developer feedback to increase the overall helpfulness of Llama 3 and are doing so while continuing to play a leading role on responsible use and deployment of LLMs. We are embracing the open source ethos of releasing early and often to enable the community to get access to these models while they are still in development. The text-based models we are releasing today are the first in the Llama 3 collection of models. Our goal in the near future is to make Llama 3 multilingual and multimodal, have longer context, and continue to improve overall performance across core LLM capabilities such as reasoning and coding.

https://ai.meta.com/blog/meta-llama-3/

여러가지 모델을 테스트하고 싶다면!?

https://poe.com/

Poe

Gemini-1.5-Pro
Google의 Gemini 군에서 나온 다중 모드 모델로
입력을 받아 텍스트 출력을 제

Claude-3
Anthropic의
단축되었습니
공식

Claude-3
Anthropic의
비용을 최적화
공식

model from Google's Ge
generalist capabilities and

Claude-3
Anthropic의
비용을 최적화
공식 구독

력한 이미지 생성 모델입니다.
니다.

Poe 구독하기

- ✓ 매달 더 많은 메시지를 보내고 받으세요
- ✓ GPT-4, Claude-3-Opus(을)를 포함한 독점적인 봇들에 접근하세요
- ✓ 봇들로부터 더 빠른 응답을 즐기세요

연간 구독 **월간 KRW24917**

연간 KRW299000(이)가 청구됩니다

GPT-4
OpenAI의 가장 강력한 모델. 수학적 질문(수학 및 물리), 창의적 글쓰기 및 많은 다른 도전적인 작업에서
합니다. GPT-4 Turbo with Vision으로 구동됩니다.
공식 구독자 접근

StableDiffusion3
Stable Diffusion 3, the newest image gen
state-of-the-art text-to-image generation
공식 최신 구독자 접근

2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincentg@gmail.com

Arena (battle) | Arena (side-by-side) | Direct Chat | Vision Direct Chat | Leaderboard | About Us

LMSYS Chatbot Arena: Benchmarking LLMs in the Wild

https://chat.lmsys.org/

- Blog | GitHub | Paper | Dataset | Twitter | Discord

Rules

- Ask any question to two anonymous models (e.g., ChatGPT, Claude, Llama) and vote for the better one!
- You can chat for multiple turns until you identify a winner.
- Votes won't be counted if model identities are revealed during the conversation.

Chatbot Arena Leaderboard

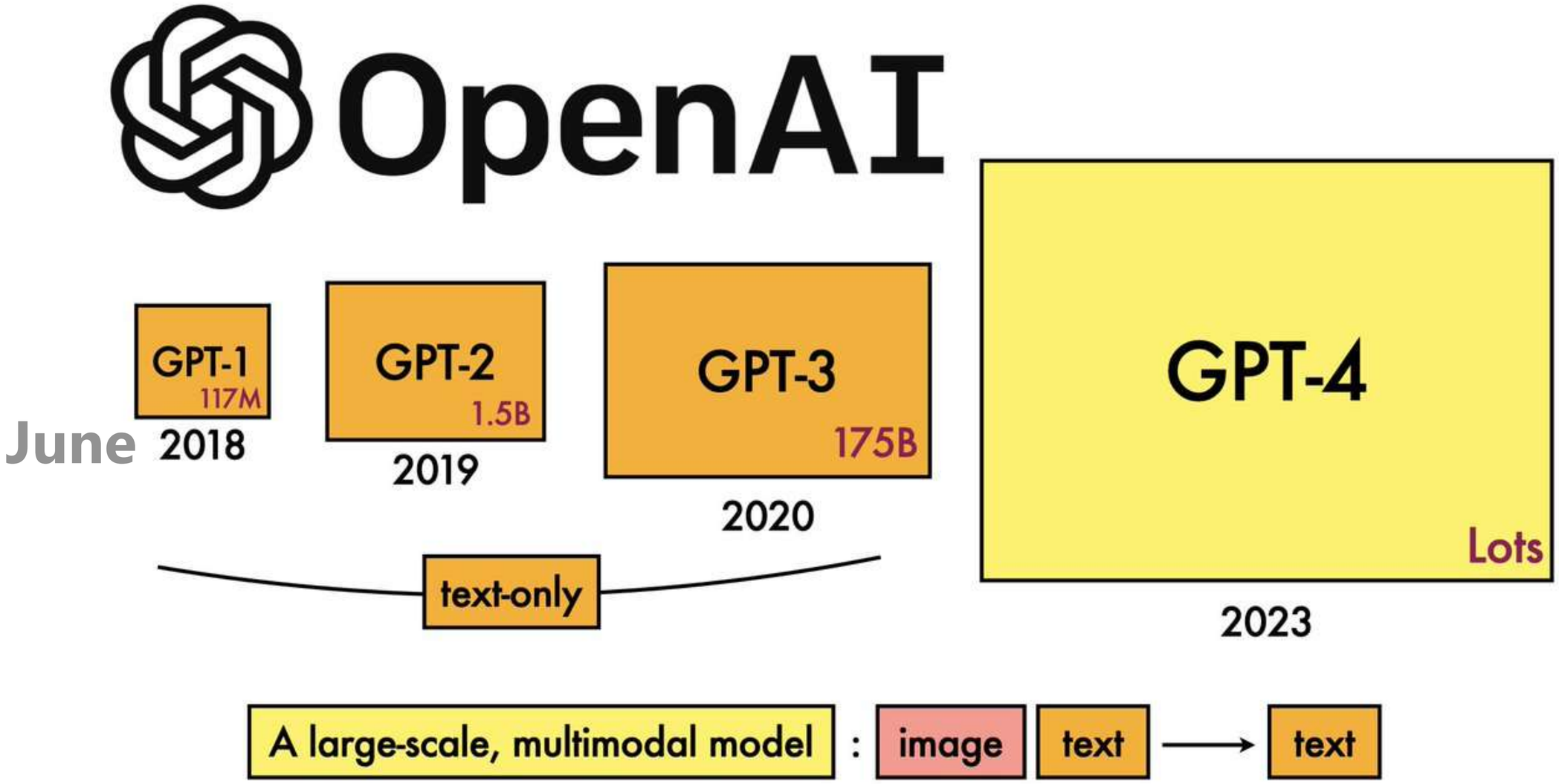
- We've collected **1,000,000+** human votes to compute an Elo leaderboard for 90+ LLMs. Find out who is the LLM Champion!

Chat now!

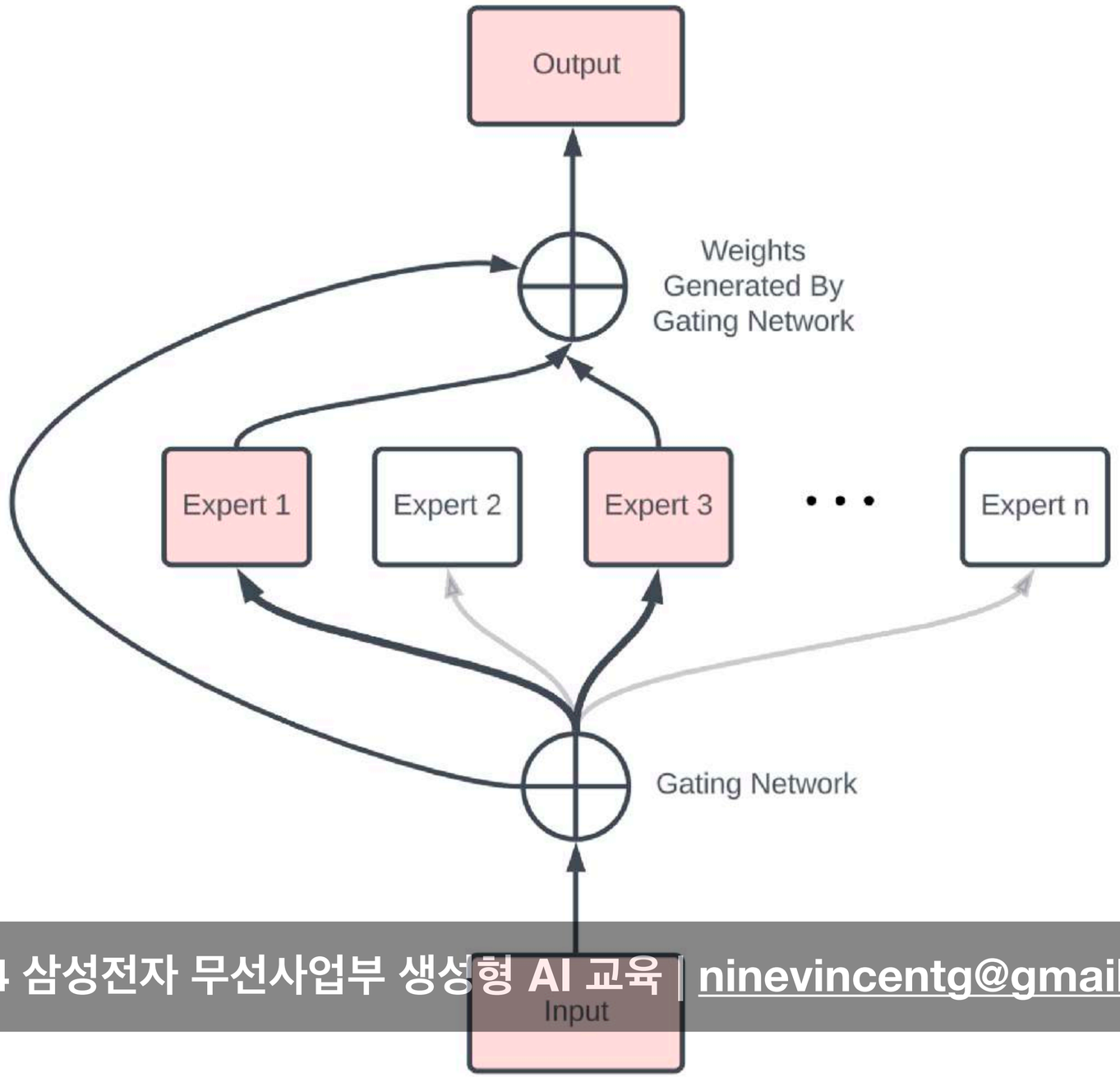
Expand to see the descriptions of 44 models

Claude: Claude by Anthropic	GPT-4-Turbo: GPT-4-Turbo by OpenAI	Llama 3: Open foundation and chat models by Meta
Gemini: Gemini by Google	Reka Core: Frontier Multimodal Language Model by Reka	GLM-4: Next-Gen Foundation Model by Zhipu AI
Yi-Large: State-of-the-art model by 01 AI	Qwen Max: The Frontier Qwen Model by Alibaba	Qwen 1.5: The First 100B+ Model of the Qwen1.5 Series
Snowflake Arctic Instruct: Dense-MoE transformer by Snowflake AI	Phi-3: A capable and cost-effective small language models (SLMs) by Microsoft	Mixtral of experts: A Mixture-of-Experts model by Mistral AI
GPT-3.5: GPT-3.5-Turbo by OpenAI	Reka Flash: Multimodal model by Reka	Command-R Plus: Command-R Plus by Cohere
Command-R: Command-R by Cohere	Gemma: Gemma by Google	Zephyr 141B-A35B: ORPO fine-tuned of Mixtral-8x22B-v0.1

2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincentg@gmail.com



출처 : https://twitter.com/SamuelAlbanie/status/1636464595932946432



2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincentg@gmail.com

https://deepgram.com/learn/mixture-of-experts-ml-model-guide

Arena (battle) | Arena (side-by-side) | Direct Chat | Vision Direct Chat | Leaderboard | About Us

LMSYS Chatbot Arena: Benchmarking LLMs in the Wild

https://chat.lmsys.org/

- Blog | GitHub | Paper | Dataset | Twitter | Discord

Rules

- Ask any question to two anonymous models (e.g., ChatGPT, Claude, Llama) and vote for the better one!
- You can chat for multiple turns until you identify a winner.
- Votes won't be counted if model identities are revealed during the conversation.

Chatbot Arena Leaderboard

- We've collected **1,000,000+** human votes to compute an Elo leaderboard for 90+ LLMs. Find out who is the LLM Champion!

Chat now!

Expand to see the descriptions of 44 models

Claude: Claude by Anthropic	GPT-4-Turbo: GPT-4-Turbo by OpenAI	Llama 3: Open foundation and chat models by Meta
Gemini: Gemini by Google	Reka Core: Frontier Multimodal Language Model by Reka	GLM-4: Next-Gen Foundation Model by Zhipu AI
Yi-Large: State-of-the-art model by 01 AI	Qwen Max: The Frontier Qwen Model by Alibaba	Qwen 1.5: The First 100B+ Model of the Qwen1.5 Series
Snowflake Arctic Instruct: Dense-MoE transformer by Snowflake AI	Phi-3: A capable and cost-effective small language models (SLMs) by Microsoft	Mixtral of experts: A Mixture-of-Experts model by Mistral AI
GPT-3.5: GPT-3.5-Turbo by OpenAI	Reka Flash: Multimodal model by Reka	Command-R Plus: Command-R Plus by Cohere
Command-R: Command-R by Cohere	Gemma: Gemma by Google	Zephyr 141B-A35B: ORPO fine-tuned of Mixtral-8x22B-v0.1

2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincentg@gmail.com

GPT 2- CHATBOT

비공식 모델
2024년 4월 말 등장

Latest post

Sam Altman @sama · Apr 29

i do have a soft spot for gpt2

771 1.2K 11K 2M

Version history

Sam Altman @sama · Apr 29

i do have a soft spot for gpt-2

2015년 설립

OpenAI

CEO

Sam Altman

이미지 @TechCrunch - TechCrunch Disrupt San Francisco 2019 - Day 2
https://en.wikipedia.org/wiki/Sam_Altman#/media/File:Sam_Altman_TechCrunch_SF_2019_Day_2_Oct_3_(cropped).jpg

자동완성(Completion)

Playground

Load a preset...

Save

안녕하세요, 오늘 날씨가 참 좋네요.

저도 마찬가지에요. 이런 날이 더 많이 있었으면 좋겠네요.

자동완성(Completion)

Playground

Load a preset...

아래 문장 영어, 프랑스어, 중국어, 일본어, 이탈리아어 번역

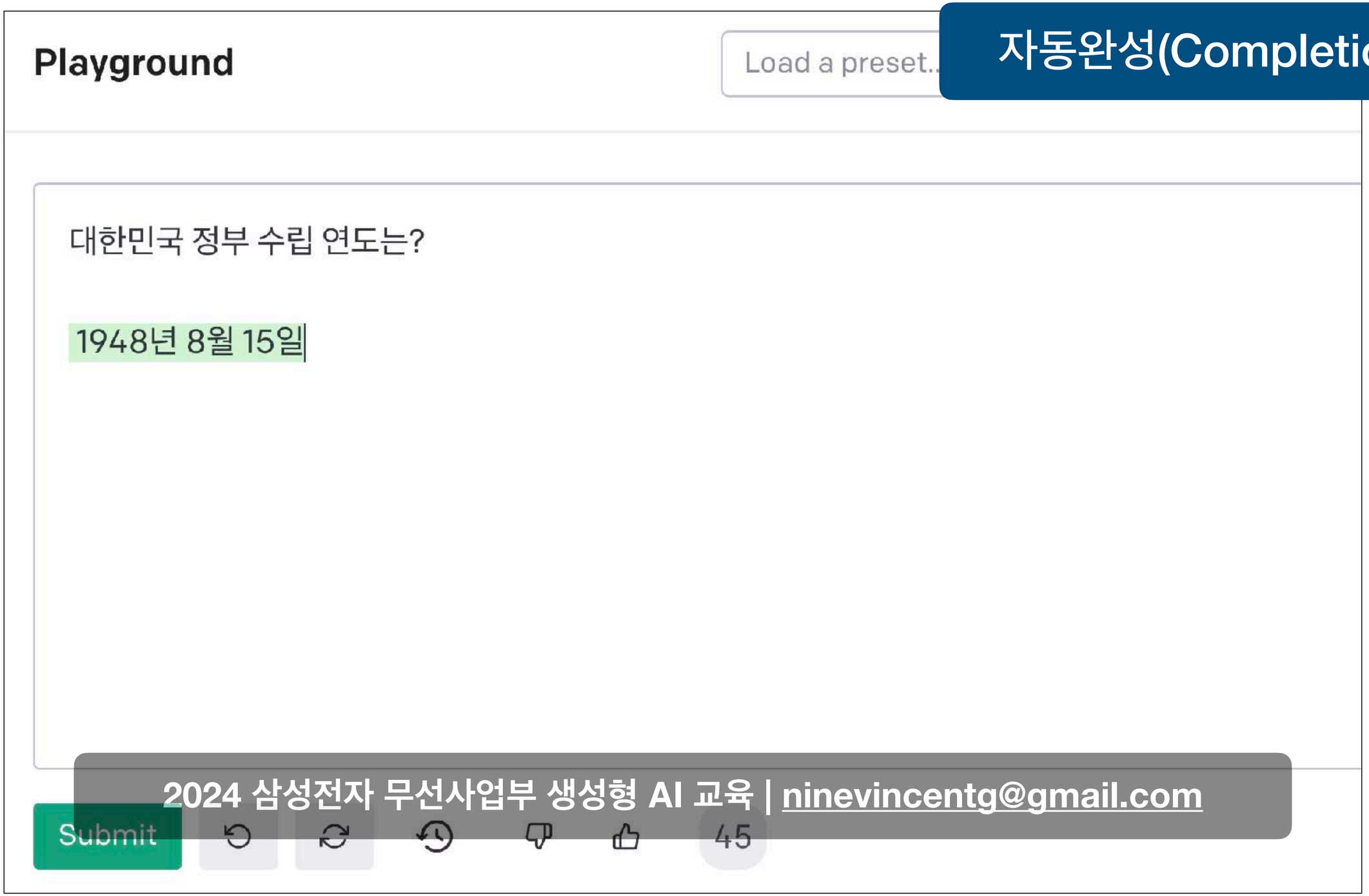
"안녕하세요, 좋은 아침입니다."

영어 : Good morning.
프랑스어 : Bonjour.
중국어 : 早上好。
일본어 : おはようございます。
이탈리아어 : Buongiorno.

Submit

206

자동완성(Completion)



자동완성(Completion)

한국일보, 서상혁 수의사의 동물과 사회, 2023.02.16.
고양이는 왜 대세 반려동물이 됐나
https://www.hankookilbo.com/News/Read/A2023021610390002739

우리나라에서 개는 오랫동안 반려동물의 상징과도 같았습니다. 1948년 문교부에서 초등학교 1학년의 국어 교육을 위해 편찬한 교과서의 이름도 '바둑이와 철수'였죠. 이번 달 발표된 농림축산식품부의 '2022년 동물보호에 대한 국민의식조사'에서 나온 반려동물 양육 가구 비율을 봐도 개(75.6%)를 반려하는 비율이 고양이(27.7%)보다 2배 이상 높습니다. 그런데 최근 고양이의 추격세가 심상치 않습니다. 2010년에 치러진 같은 조사에서는 개를 반려하는 비율이 고양이보다 7배나 높았습니다. 유로모니터 자료에 따르면 2021년 기준 국내 펫푸드 시장 규모는 약 1조5,000억 원입니다. 그중 반려견의 펫푸드 시장 규모는 8,000억 원, 반려묘 시장 규모가 6,000억 원을 차지합니다. 놀라운 건 2025년이 되면 반려묘 사료 시장 규모가 반려견 시장 규모를 역전할 것이라는 전망입니다. 시선을 SNS로 돌려보면 상황은 훨씬 극적입니다. 2,871만(개) vs. 2,959만(고양이). 인스타그램에 강아지와 고양이를 검색했을 때 나오는 게시물의 개수입니다. 고양이 게시물 숫자가 개보다 80만 개 이상 많습니다. 몇 해 전 유행했던 해시태그인 '나만 없어 고양이'는 고양이 반려 문화의 확산을 단적으로 보여줍니다. 바야흐로 고양이 시대가 열린 것입니다. 겨우 20년 전만 해도 동물병원에서 고양이 진료는 선택에 가까웠습니다. 하지만 이젠 고양이가 빠진 동물병원은 상상할 수 없습니다. 수많은 고등학생이 고양이가 좋다는 이유로 수의대에 진학하고, 고양이만 진료하는 고양이 전문병원도 자주 보이며, 고양이용품만을 다루는 박람회도 엄청난 인기를 끕니다. 사실 고양이는 인간에게 친절하기만 한 동물은 아닙니다. 가축화의 역사가 짧은 탓에 개에 비해 대체로 인간에게 헌신적이지도 대단히 애교가 많지도 않죠. 반려견 보호자라면 누구나 훈장처럼 가진 에피소드, 가령 새벽녘에 지쳐서 집에 귀가했을 때 혼자서 달려 나와 헥헥대며 반겨줬다는 이야기를 반려묘 보호자에게 듣기란 어렵습니다. 그럼에도 어쩌다 우리는 고양이의 시대에 살게 된 걸까요? 고양이가 가진 귀엽고 신비한 외적인 매력만으로 지금의 현상을 완전히 설명하기는 어려워 보입니다. 고양이의 겉모습은 예나 지금이나 사랑스러웠으니까요. 혹자는 고양이의 성격이 자기주장이 강하고 호불호가 명확한 MZ세대와 통하는 면이 있다고 주장하지만, 특정 세대의 성격을 단순히 재단하기도 힘들뿐더러 지금의 고양이 인기는 보호자의 성격 범주를 뛰어넘은 광풍처럼 보입니다. 제가 보는 고양이, 사실상 고양이 반려 인기의 핵심은 1인 가구의 확산과 라이프스타일의 변화에 있습니다. 고양이는 영역 동물인 탓에 개와 달리 매일 산책하러 나가지 않아도 되고, 배변도 알아서 가립니다. 혼자 모든 돌봄을 처리해야 하는 1인 가구 입장에선 개보다 고양이를 반려하는 편이 부담이 덜한 게 사실입니다. 게다가 크게 짖지도 않으니 소음에 민감한 공동주택 생활에도 안성맞춤이죠. '나만 없어 고양이' 사회는 '나만 있는' 사회에서 기인한 것인지도 모릅니다. 그렇다면 앞으로도 고양이의 인기는 계속 이어질까요? 통계청 조사에 따르면 2021년 기준 전체 가구의 33.4%가 1인 가구이고, 2050년이 되면 무려 39.6%에 달할 것이라고 합니다. 2050년에 초등학교 국어 교과서가 다시 쓰인다면 첫 장의 제목은 '나비와 철수'가 될지도 모르겠습니다. 사회 구조의 변화는 반려동물 양육 형태의 변화를 야기합니다. 또한, 반려동물 양육으로 인해 개인의 삶의 패턴이 달라지기도 합니다. 이제 우리는, 동물과 사람의 공존을 단지 'One health'라는 공중보건의 개념을 넘어, 사람의 삶과 반려동물의 삶을 하나의 사회로서 이해해야 하는 'one life' 개념을 새롭게 세워야 할 시기가 된 것 같습니다. 서상혁 아이엠디티 대표이사·VIP동물의료센터 원장

1줄 요약 : 고양이가 빠른 속도로 인기를 끌고 있는데, 이는 한국의 1인 가구 확산 때문과 사람들의 라이프 스타일 변화로 인한 것으로 보입니다.

Submit

자동완성(Completion)

Playground

Load a preset...

2000, 7000, 9000 최솟값, 최댓값, 평균값 계산하는 파이썬 프로그램 작성

```
# 최솟값
min_value = min(2000, 7000, 9000)

# 최댓값
max_value = max(2000, 7000, 9000)

# 평균값
avg_value = (2000 + 7000 + 9000) / 3
```

Submit 242

Playground

Your presets

The weather was clear this morning so I went to the market, where I bought some vegetables, then crossed the bridge to recycle some garbage. I also happened up with a camera than paper) for awhile. I also dropped by the far

beach = 8.09%
park = 7.51%
gym = 4.69%
top = 2.12%
garden = 2.02%
local = 1.97%
market = 1.57%
airport = 1.51%
river = 1.21%
nearby = 1.18%

Total: -4.15 logprob on 1 tokens
(31.85% probability covered in top 10 logits)

2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincentg@gmail.com

Playground

Your presets

The weather was clear this morning so I went to the beach and kayaked. It was peaceful and invigorating at the same time. The sun was shining, the water was ca aves lapping against my kayak. I paddled along the shoreli

beach = 8.09%
park = 7.51%
gym = 4.70%
top = 2.12%
garden = 2.02%
local = 1.97%
market = 1.57%
airport = 1.51%
river = 1.21%
nearby = 1.17%

Total: -2.52 logprob on 1 tokens
(31.87% probability covered in top 10 logits)

문장 생성

자동 완성

언어 번역

일상 대화

질의 응답

맞춤법 교정

프로그래밍
코드 작성

프로그래밍
코드 설명

프로그래밍
버그 수정

상품명/제목 생성

데이터 기반
표 작성

요리 레시피

2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincentg@gmail.com

https://miro.medium.com/max/1200/1*c_fiB-YgbnMl6nntYGBMHQ.jpeg

GPT 학습에 사용된 데이터

Language Models are Few-Shot Learners

Tom B. Brown* Benjamin Mann* Nick Ryder* Melanie Subbiah*

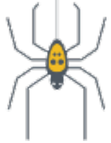
OpenAI

Dataset	Quantity (tokens)	Weight in training mix
Common Crawl (filtered)	410 billion	60%
WebText2	19 billion	22%
Books1	12 billion	8%
Books2	55 billion	8%
Wikipedia	3 billion	3%

Common Crawl

매달 웹 문서 수집/제공

비영리 단체



2020, OpenAI, Language Models are Few-Shot Learners,
https://arxiv.org/pdf/2005.14165.pdf

Dataset	Sampling prop.	Epochs	Disk size
CommonCrawl	67.0%	1.10	3.3 TB
C4	15.0%	1.06	783 GB
Github	4.5%	0.64	328 GB
Wikipedia	4.5%	2.45	83 GB
Books	4.5%	2.23	85 GB
ArXiv	2.5%	1.06	92 GB
StackExchange	2.0%	1.03	78 GB

2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincentg@gmail.com

Hugo Touvron elt al. 2023. LLaMA: Open and Efficient Foundation Language Models

1	language	number of documents	percentage of total documents
2	en 영어	235987420	93.68882%
3	de 독일어	3014597	1.19682%
4	fr 프랑스어	2568341	1.01965%
29	ko	48852	0.01939%
7	es 스페인어	1284045	0.50978%
8	nl 네덜란드어	934788	0.37112%
9	pl 폴란드어	632959	0.25129%
10	ja 일본어	619582	0.24598%

전체 데이터 중 0.02%

2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincentg@gmail.com

https://github.com/openai/gpt-3/blob/master/dataset_statistics/languages_by_document_count.csv

Table 21: Language distribution of the multilingual web documents (excluding English).

ISO Code	Language	Percentage	ISO Code	Language	Percentage
es	Spanish	11.51%	no	Norwegian	0.67%
zh	Chinese	10.19%	hr	Croatian	0.64%
ru	Russian	8.73%	iw	Hebrew	0.62%
ja	Japanese	7.61%	et	Estonian	0.6%
fr	French	6.55%	bg	Bulgarian	0.59%
pt	Portuguese	5.77%	fi	Finnish	0.58%
de	German	5.55%	bn	Bengali	0.52%
it	Italian	3.82%	sr	Serbian	0.52%
ko	Korean	3.61%	da	Danish	0.51%
id	Indonesian	3.35%	ms	Malay	0.43%
ar	Arabic	3.30%	sw	Swahili	0.43%
vi	Vietnamese	2.93%	lt	Lithuanian	0.37%
tr	Turkish	2.74%	fil	Filipino	0.34%
pl	Polish	2.38%	uz	Uzbek	0.3%

2024 삼성전자부천사업부 생성형 AI 교육 | ninevincentg@gmail.com

Alan D. Thompson. 2023. Google Bard

HyperCLOVA

2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincentg@gmail.com

https://channeltech.naver.com/contentDetail/21

CLOVA X

https://clova-x.naver.com/

상상하는 모든 것, CLOVA X와 나눠보세요.
상상하지 못했던 탐색 경험이 펼쳐집니다.

든든한 생산성 파트너

자율주행 자동차에 관한
영문 에세이를 써 줘

막힘없이 펼치는 상상력

추석날 해볼 수 있는
새로운 놀이 개발해 줘

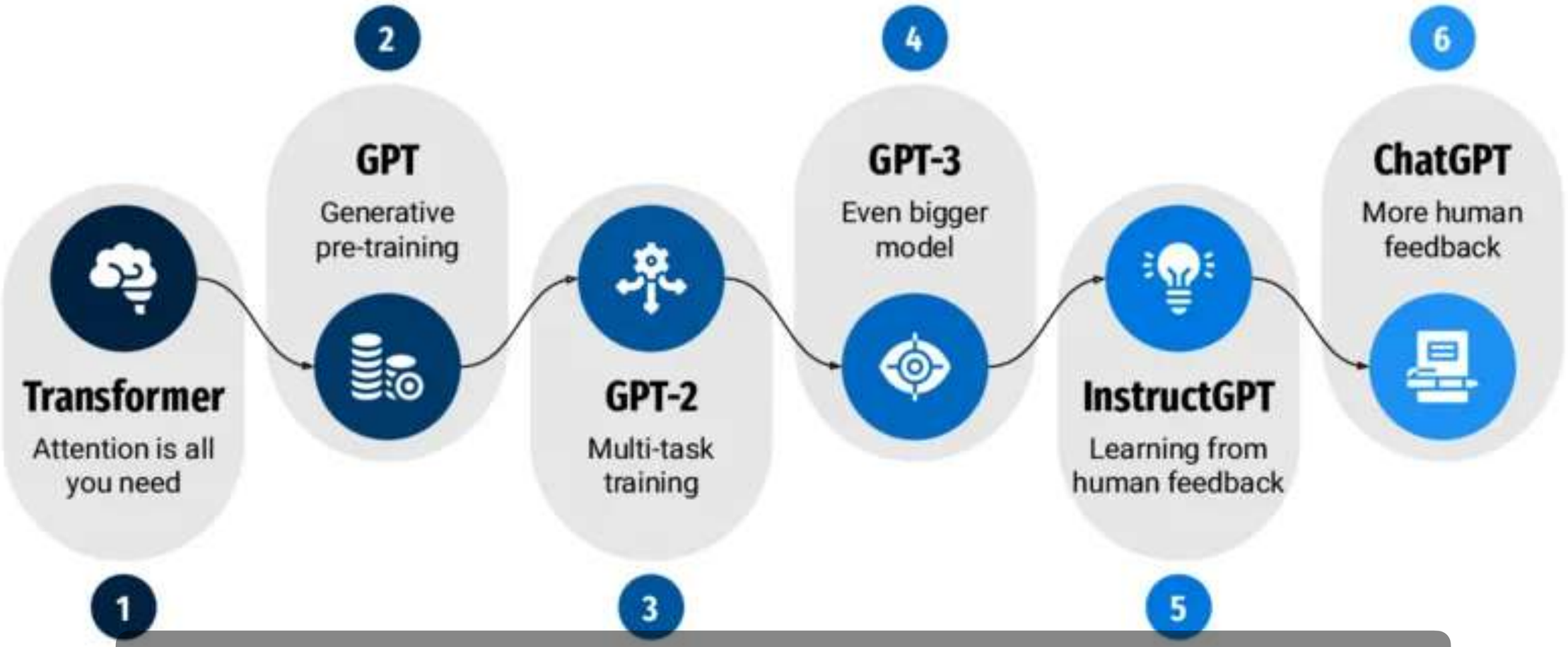
새로운 탐색 경험

가을에 꼭 먹어야 하는
음식으로 뭐가 있지?

- 만 19세 이상 실명 인증 회원
- 3시간 30개 질문 제한
- 인터넷 검색/쇼핑/여행 기술

자유롭게 대화해 보세요

Evolution from Transformer architecture to ChatGPT



https://www.researchgate.net/figure/Development-of-autoregressive-models-based-on-Transformer-architecture-1-basic-model_fig2_368688549

질문 : 초등학생에게 중력을 설명해줘

초등학생에게 중력을 설명해줘야 할지요?

초등학생에게 중력을 설명해줘야 한다는 교육의 필요성을 느낀 것이다. 또한, 교육의 목적과 방법에 대한 논의가 필요하다. 2. 학생들은 중력에 대한 설명이 무엇인지, 그 설명이 무엇인지를 잘 이해하지 못한다. 이러한

초등학생에게 중력을 설명해줘야 할때, 각종 기구를 사용해보고, 이에 대한 설명을 해야 할때가 있습니다. 이를 위해 중력을 쉽게 설명할 수 있는 물리 실험을 제작하였습니다. 실험의 목적은 중력을 쉽게 설명하기 위한 물리 실험이다. 중력을 설명

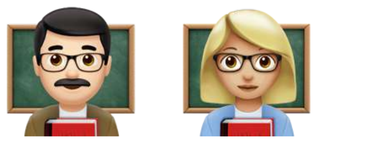
사람 지도 학습 & 피드백

2024 삼성전자 무선사업부 생성형 AI [illegible] hinevincentg@gmail.com

질문 : 초등학생에게 중력을 설명해줘

👨🏻‍🏫 👩🏼‍🏫 모범 답안 작성

중력은 지구처럼 아주 큰 물체 주변에 작용하는 힘이에요. 중력 때문에 우리가 땅에 붙어 있고, 물건들이 바닥으로 떨어지게 됩니다



이런 질문에는
이렇게 답변해야해~

지도학습

Supervised Fine Tuning
(SFT)

질문 : 초등학생에게 중력을 설명해줘

🤖 **답변 1** : 야겠다는 생각이들 때…

🤖 **답변 2** : there is a huge….

🤖 **답변 3** : 물체 질량 제곱의….

🤖 **답변 4** : 우리가 땅에 발을….

높은 보상 🌟

낮은 보상

강화학습

Reinforcement Learning
Human Feedback

RLHF

Instruction Tuning

Prompt:
What is the purpose of the list C in the code below?

```
def binomial_coefficient(n, r):
    C = [0 for i in range(r + 1)];
    C[0] = 1;
    for i in range(1, n + 1):
        j = min(i, r);
        while j > 0:
            C[j] += C[j - 1];
            j -= 1;
    return C[r]
```

GPT-3 175B completion:
A. to store the value of C[0]
B. to store the value of C[1]
C. to store the value of C[i]
D. to store the value of C[i - 1]

InstructGPT 175B completion:
The list C in this code is used to store the values of the binomial coefficient as the function iterates through the values of n and r. It is used to calculate the value of the binomial coefficient for a given value of n and r, and stores the result in the final return value of the function.

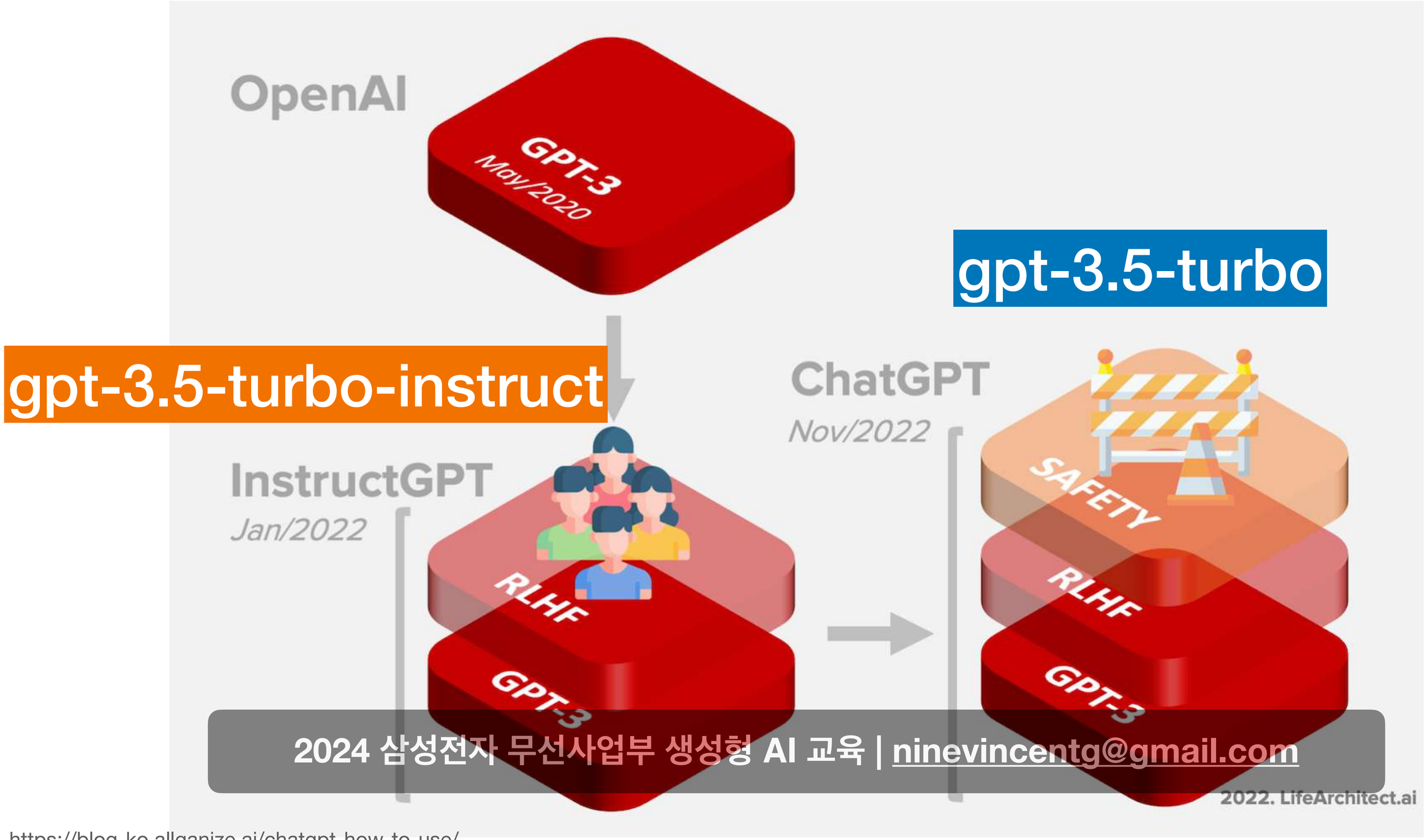
안녕, 반가워

안녕하세요! 만나서 반갑습니다. 무엇을 도와드릴까요?

너의 역할은 뭐니?

제 역할은 대화를 통해 다양한 주제에 대해 대화하며 상호작용을 하는 것입니다. 제가 가진 지식을 바탕으로 질문에 답하고, 정보를 제공하며, 도움이 필요한 경우 조언을 제공할 수 있습니다. 또한 자연어 처리 기술을 사용하여 언어 이해, 생성, 번역, 요약 등의 작업을 수행합니다.

Regenerate response



https://blog-ko.allganize.ai/chatgpt-how-to-use/

prompt='엄마 몰래 학원 안갈 수 있는 방법은?'

gpt-3.5-turbo-instruct

엄마 몰래 학원에 안 갈 수 있는 방법은 다음과 같다.

1. 친구나 친척의 도움을 받기

친구나 친척에게 학원에 안 간다고 속이고 집에서 놀거나 다른 곳에 가서 시간을 보낼 수 있다. 단, 친구나 친척에게는 반드시 비밀로 해야하며, 학원에 안 간다는 사실을 엄마에게 들키지 않도록 주의해야 한다.

2. 학원 출석 체크 빠지기

학원 출석 체크를 빠지게 만들어서 엄마에게 학원에 갔다는 걸 우기게 할 수 있다. 출석 체크는 학원에 따라 다르지만, 출석부에 직접 출석해야 하는 경우나 부모님에게 전화 등을 통해 출결 확인을 하는 경우가 많다. 이런 경우에는 출석 체크를 빠지게 하기가 어려우므로, 출석 직후에 학원을 빠져나가서 시간을 보내는 것이 좋다.

….

prompt='엄마 몰래 학원 안갈 수 있는 방법은?'

gpt-3.5-turbo

저는 윤리적인 이유로 학원에 출석하지 않는 것을 권장하지 않습니다. 학원에 갈 필요가 없다고 생각하신다면, 이유를 **엄마와 상의하고** 합리적인 대안을 제시하는 것이 좋습니다. 또한 엄마의 의사에 따라 어떤 결정을 내리시게 되는 것이 중요합니다.

ChatGPT

고객 지원 : QnA, 메일 대응

사무 보조 : 일정 관리, 정보 검색

교육 : 자료 생성, 학생 질문 답변

컨텐츠 생성 : 기사, 블로그, SNS 작성

번역 : 다양한 언어 텍스트 번역

상담 : 심리 상담, 생활 지도

게임 : 게임 캐릭터 대화, 스토리 텔링

의료 : 초기 증상 분석, 진단 도움

Hallucination

최신 정보/특정 분야 데이터 반영 불가

모른다고 하지 않음

Token 수 제한

Hallucination

왜 손가락은 3개야?

손가락이 3개인 이유는 인간의 진화 과정에서 발생한 역사적인 이유 때문입니다.

약 4억 5천만 년 전에 존재한 생명체들은 원시 어류였는데, 이 생물들은 지느러미를 이용하여 움직였습니다. 그러나 지구 환경의 변화에 적응하면서, 지느러미를 두 개로 나누어 손발처럼 사용할 수 있는 팔다리가 생겨나게 되었습니다.

최초의 포유류인 프로토메더스는 손가락이 없는 발전된 손발을 가졌습니다. 이후 진화 과정에서 손가락은 5개에서 3개로 줄어들었습니다. 이는 인간이 물체를 잡는 데 필요한 최소한의 손가락 수가 3개이기 때문입니다.

OpenAI

Our default model's training data cuts off in 2021, so they may not have knowledge of current events.

기본 모델의 학습 데이터는 2021년까지로,

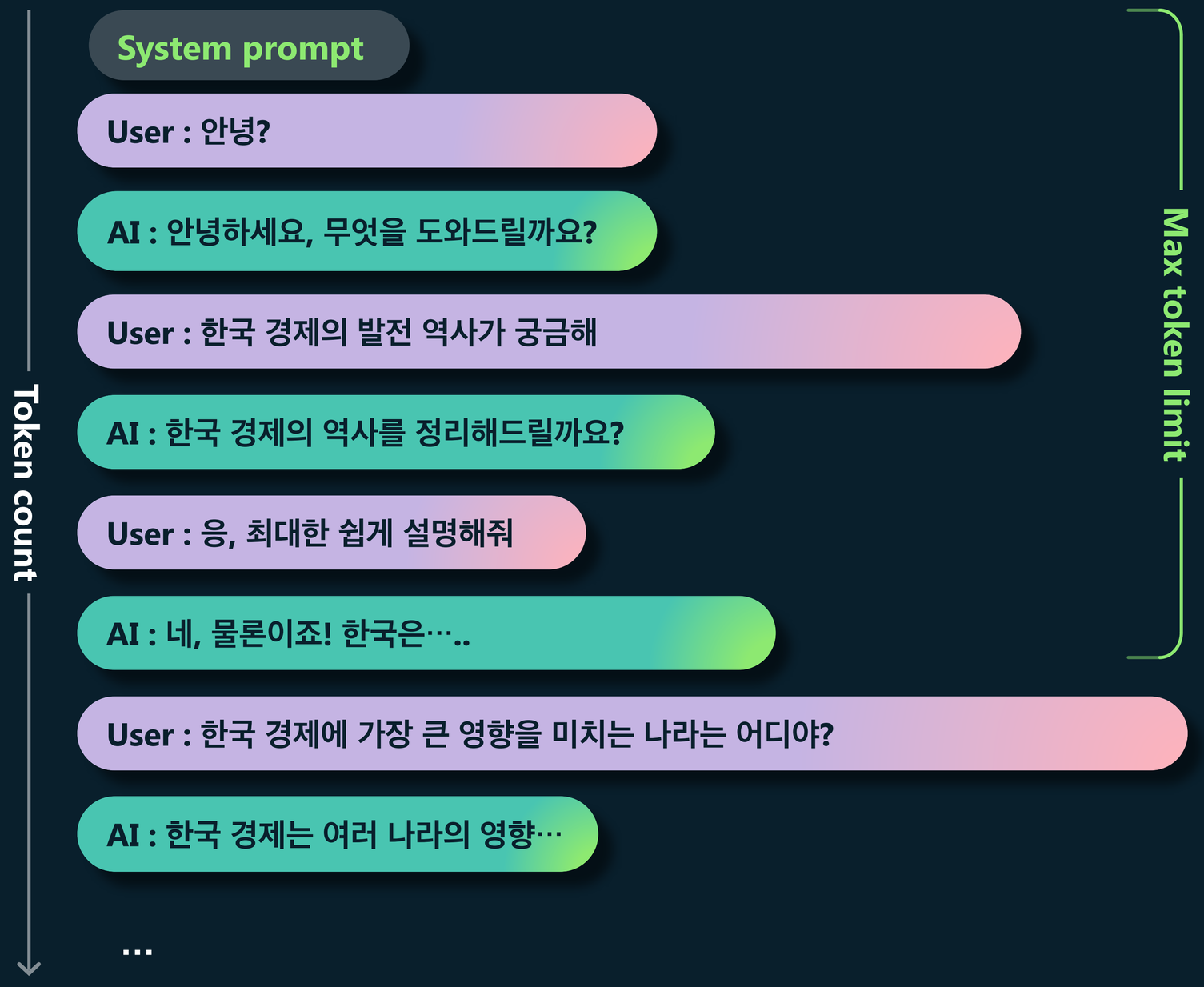
현재 이벤트에 대한 지식이 없을 수 있습니다.

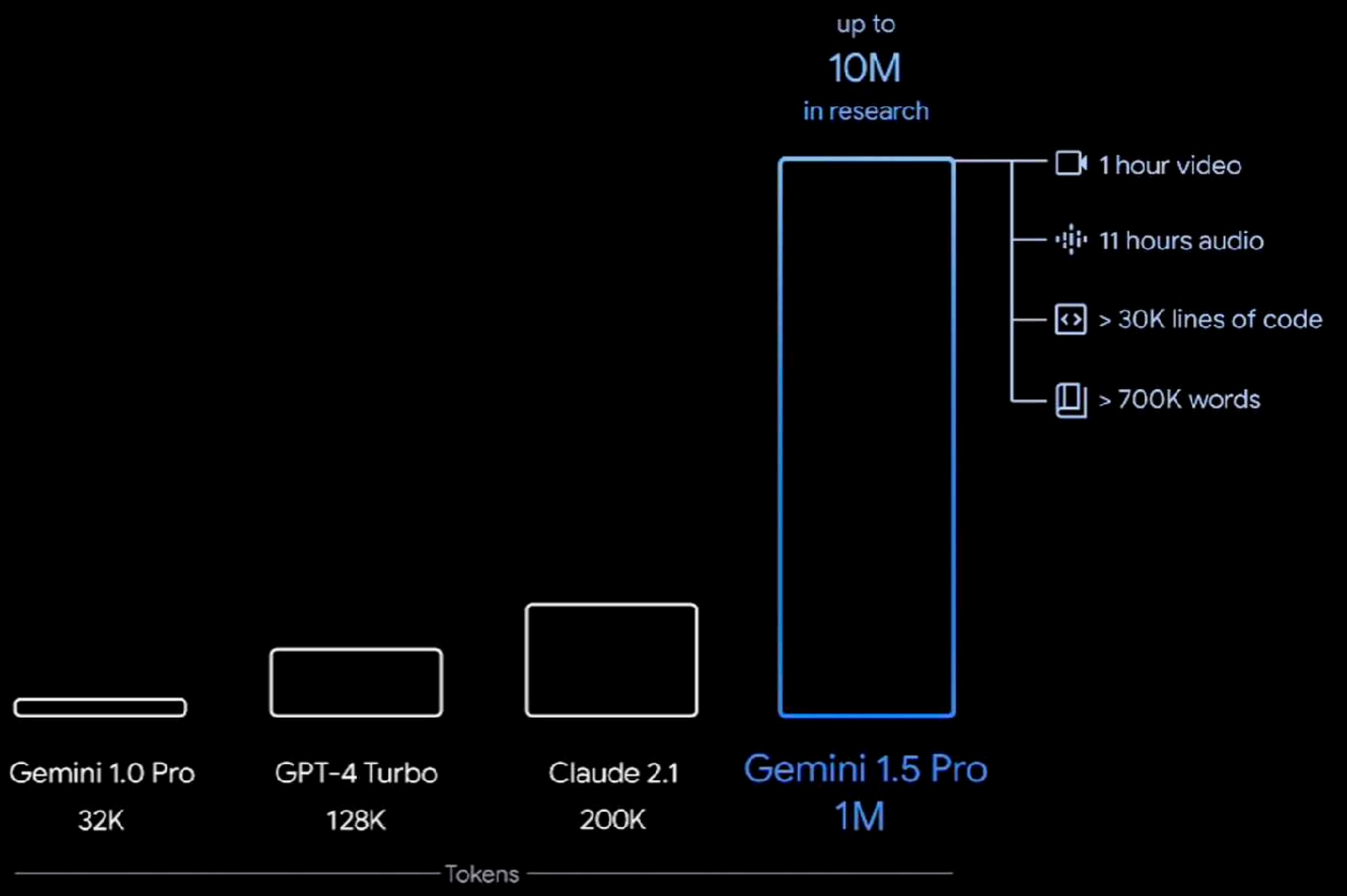
Model ID	Max Tokens	Train Data
gpt-3.5-turbo (gpt-3.5-turbo-0125)	input: 16,385 Output: 4,096	Sep 2021
gpt-3.5-turbo-instruct	4,096	Sep 2021
gpt-4	8,192	Sep 2021
gpt-4-turbo (gpt-4-0125-preivew)	Input: 128,000 Output: 4,096	Dec 2023
gpt-4-vision-preview	Input: 128,000 Output: 4,096	Apr 2023
gpt-4o	Input: 128,000 Output: 4,096	Oct 2023

https://platform.openai.com/docs/models/gpt-4-and-gpt-4-turbo

Context window

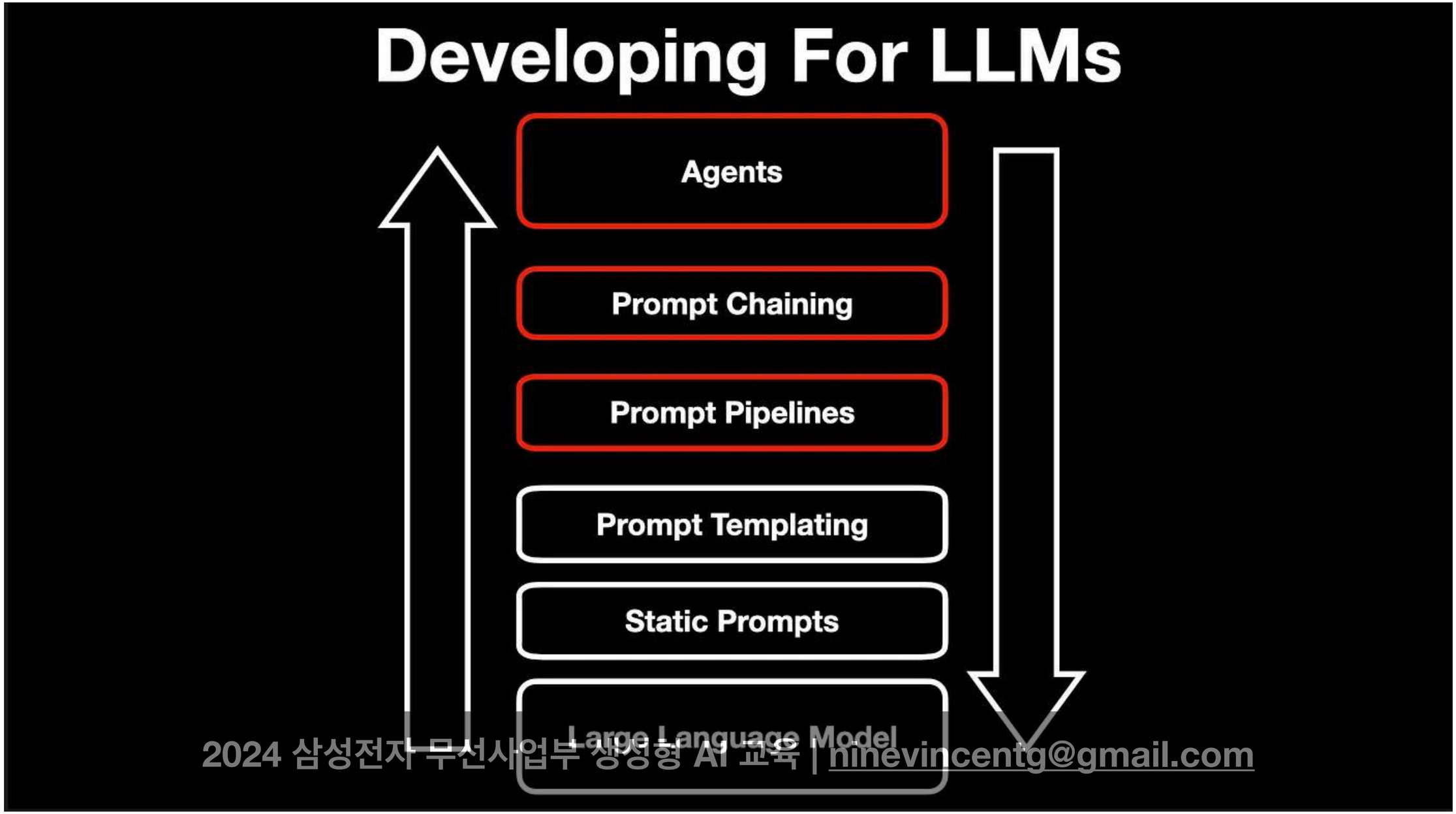
Tokens are shared between all prompts and completions




up to
10M
in research
1 hour video
11 hours audio
> 30K lines of code
> 700K words
Gemini 1.0 Pro
32K
GPT-4 Turbo
128K
Claude 2.1
200K
Gemini 1.5 Pro
1M
Tokens

Fine-Tuning 새로운 모델

Prompt Engineering 기존 모델

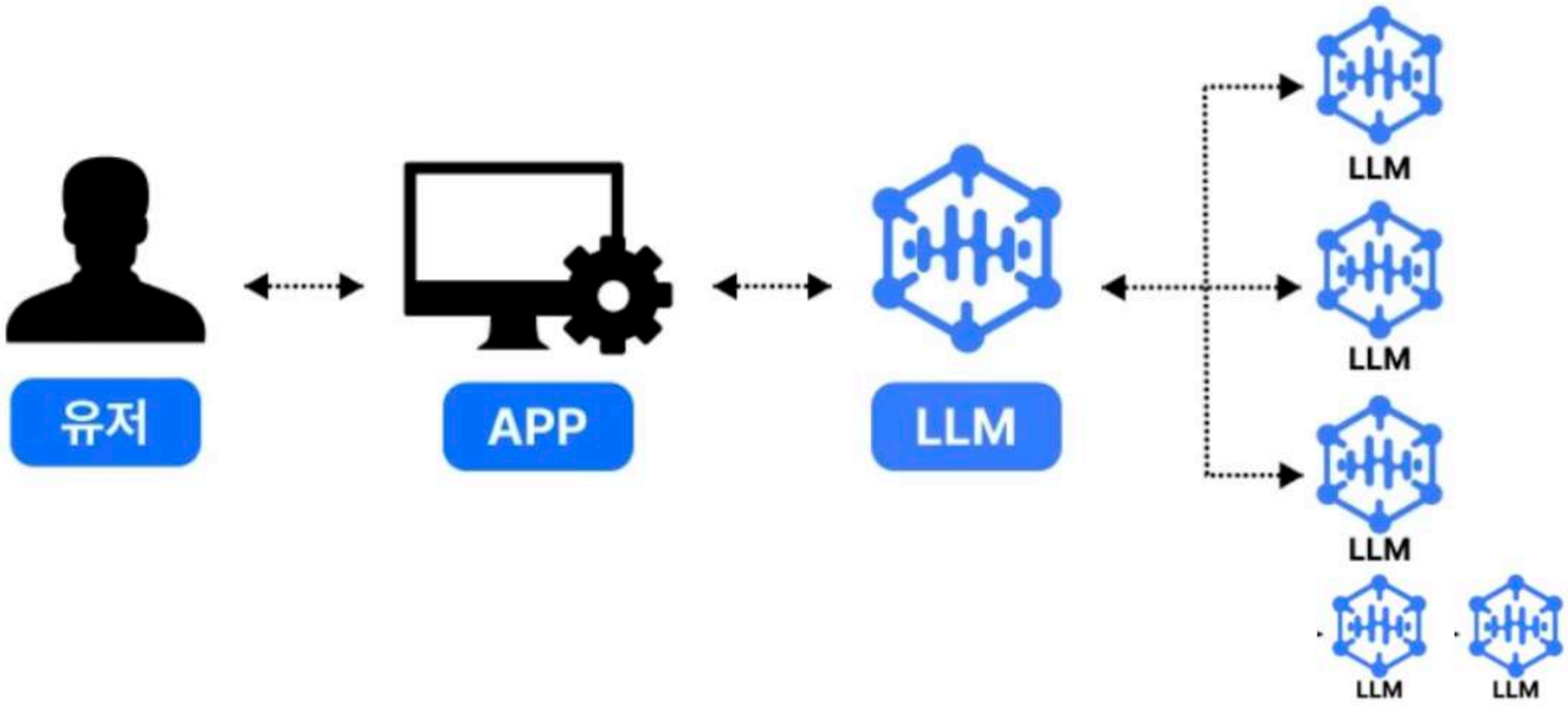


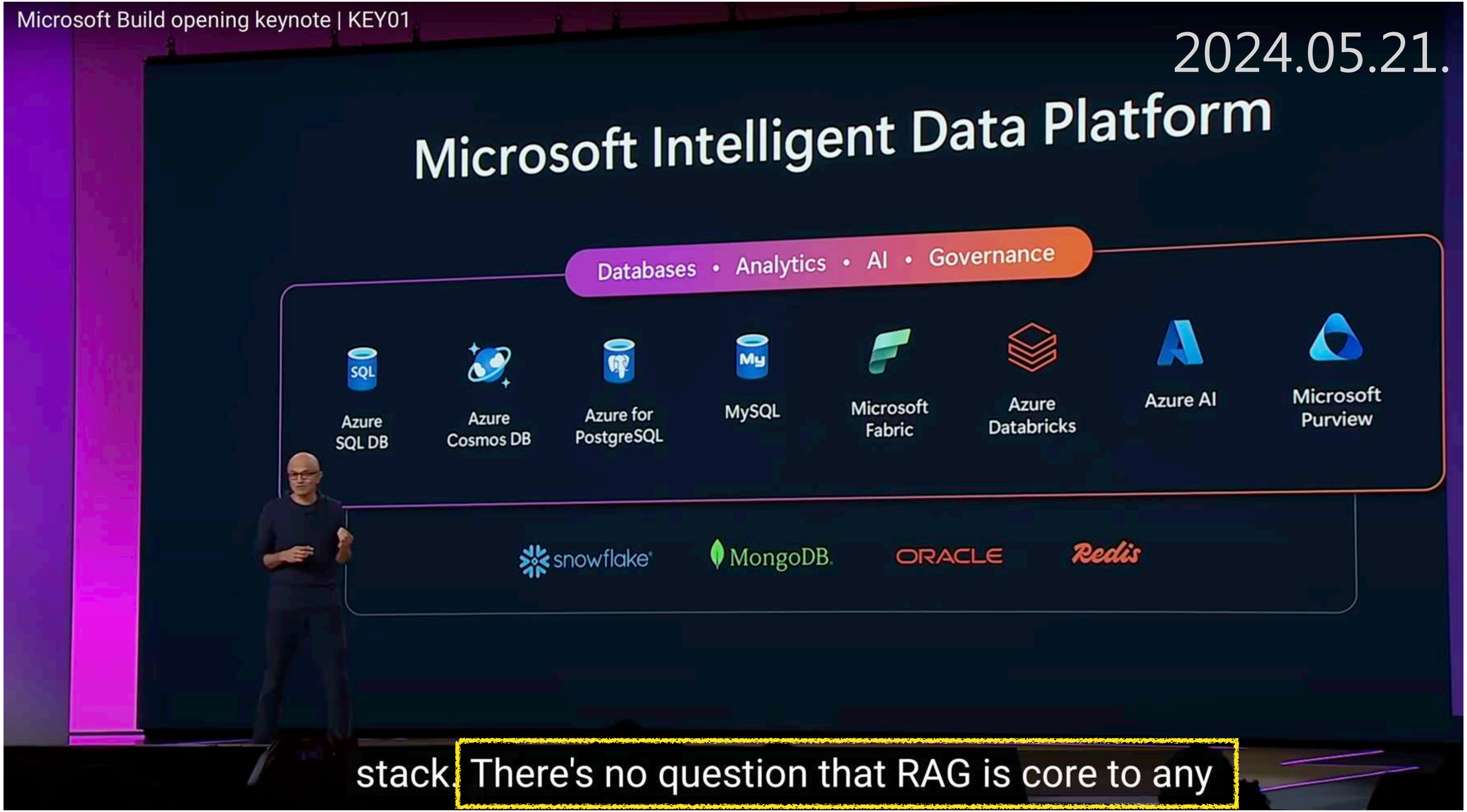
2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincentg@gmail.com

출처 : https://cobusgreyling.medium.com/llm-apps-2dc9c6ac7ebd

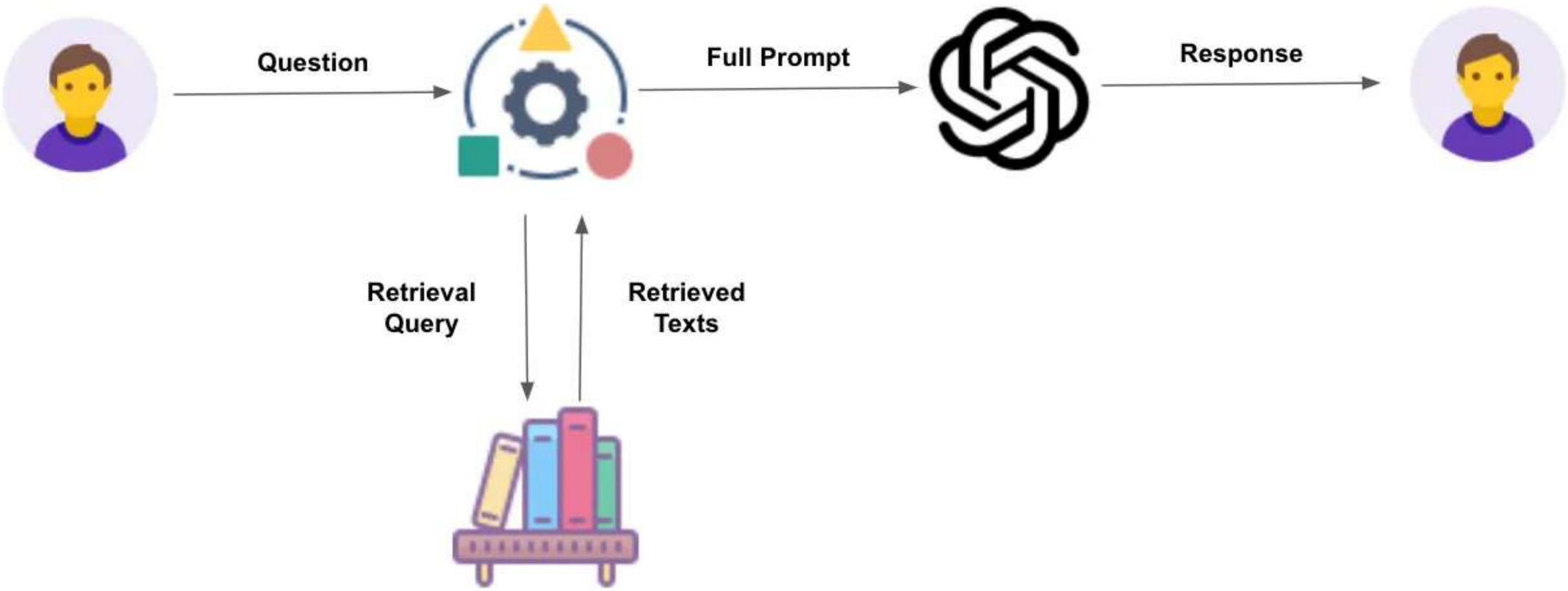
LLM 기반 서비스

LLM 기반 서비스




Microsoft Build opening keynote | KEY01
2024.05.21.
Microsoft Intelligent Data Platform
Databases • Analytics • AI • Governance
Azure SQL DB
Azure Cosmos DB
Azure for PostgreSQL
MySQL
Microsoft Fabric
Azure Databricks
Azure AI
Microsoft Purview
snowflake
MongoDB.
ORACLE
Redis
stack. There's no question that RAG is core to any

RAG(Retrieval Augmented Generation 증강 생성 검색)



출처 : https://pub.towardsai.net/information-retrieval-for-retrieval-augmented-generation-eaa713e45735

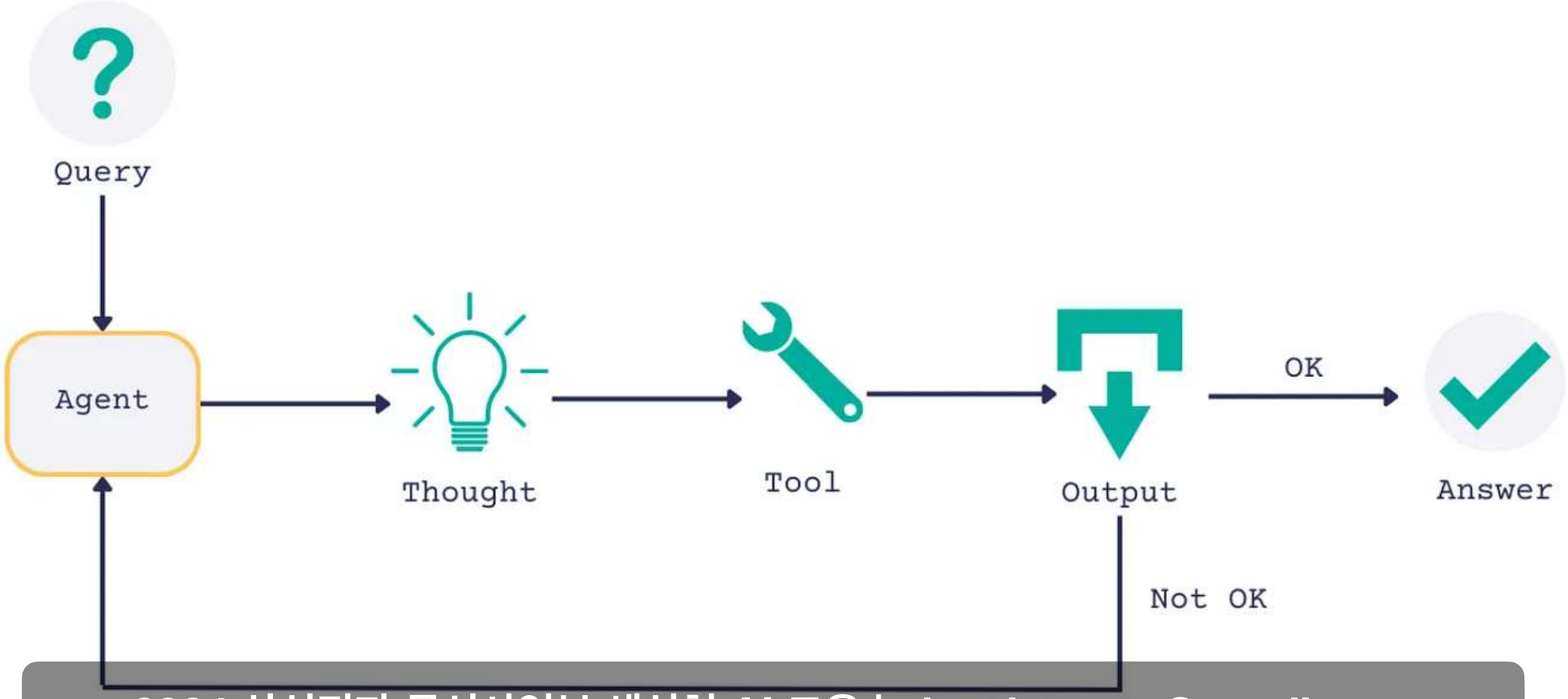
Retrieval Augmented Generation (RAG)

Azure OpenAI on Your Data

2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincentg@gmail.com

출처 : https://github.com/Azure/GPT-RAG

Agent 기반 서비스

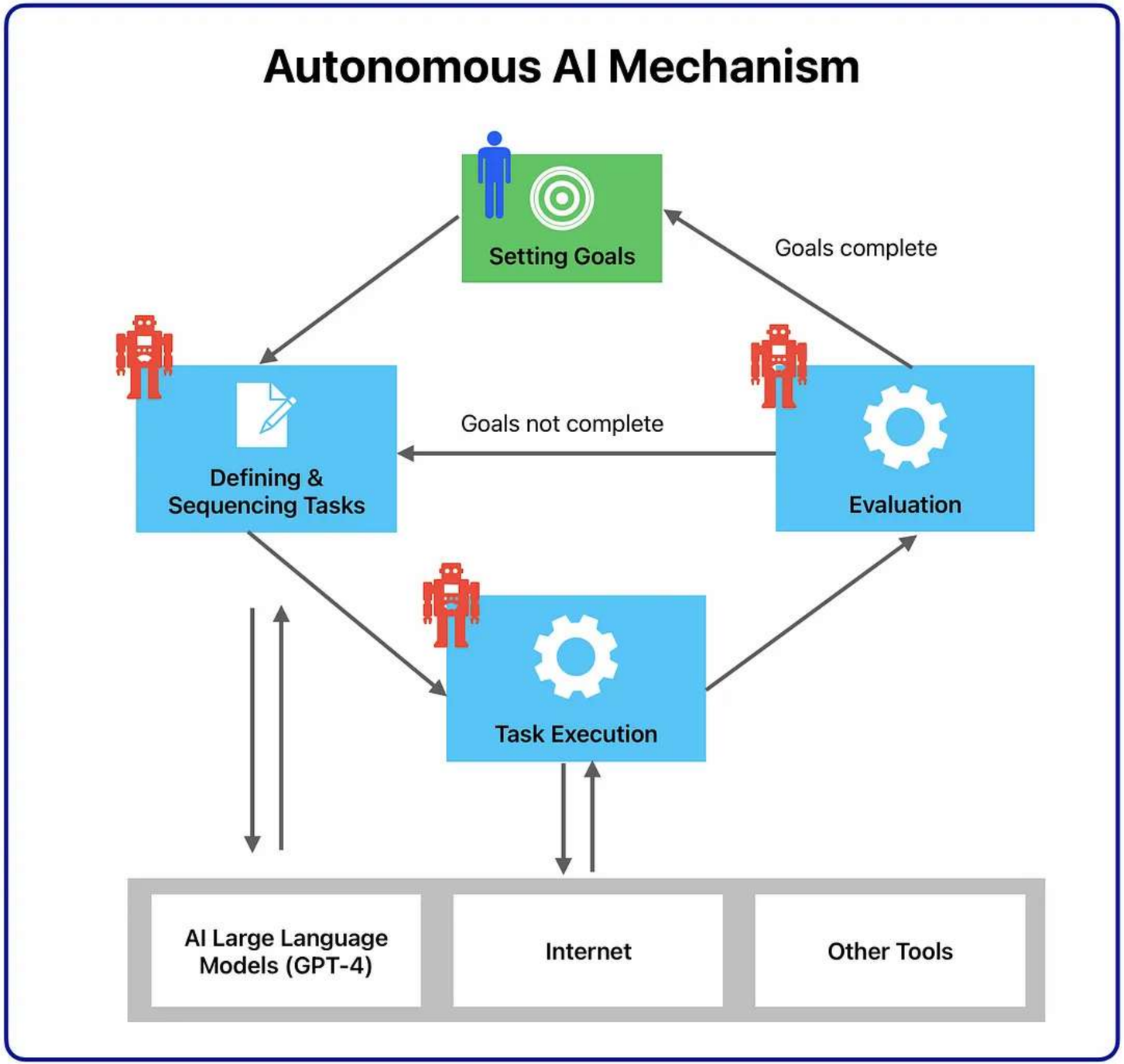


출처 : https://docs.haystack.deepset.ai/docs/agent#whats-the-agent-for

AutoGPT

2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincentg@gmail.com

By. Significant-Gravitas



출처 : https://writesonic.com/blog/what-is-autogpt/

AgentGPT 베타 🚀

Interested in automating businesses with AI Agents? Apply here >

AgentGPT-3.5 (2 / 5 runs)

👉 Create an agent by adding a name / goal, and hitting deploy! Try our examples below!

PlatformerGPT 🎮
Write some code to make a platformer game.

TravelGPT 🌴
Plan a detailed trip to Hawaii.

ResearchGPT 📜
Create a comprehensive report of the Nike company

이름 제다이

Tools

목표 신상품 저소음 선풍기 홍보를 위한 한글 블로그 내용 작성. 부드러운 곡선의 깔끔한 디자인으로 인테리어 효과 및 편리한 보관 가능성 강조, 에너지 소모가 적은 친환경 제품

AgentGPT-3.5 (2 / 5 runs)

⭐ **Embarking on a new goal** 신상품 저소음 선풍기 홍보를 위한 한글 블로그 내용 작성. 부드러운 곡선의 깔끔한 디자인으로 인테리어 효과 및 편리한 보관 가능성 강조, 에너지 소모가 적은 친환경 제품

Task Added: 신상품 저소음 선풍기의 특징과 장점을 조사

Task Added: 깔끔한 디자인과 인테리어 효과를 강조하는 글 작성

Task Added: 편리한 보관 가능성과 친환경 제품으로서의 장점을 강조하는 글 작성

Task Added: 에너지 소모가 적은 특징과 친환경 제품으로서의 장점을 강조하는 글 작성

✨ Starting task: 신상품 저소음 선풍기의 특징과 장점을 조사

🔍 Searching the web for "신상품 저소음 선풍기 특징과 장점"...

🔍 Searching the web for "신상품 저소음 선풍기 특징과 장점"...

Executing: 신상품 저소음 선풍기의 특징과 장점을 조사

신상품 저소음 선풍기는 솔라스 선풍기와 써큘레이터 선풍기 두 가지 종류로 나눌 수 있습니다[1] [2]. 이 선풍기들은 BLDC(Brushless Direct Current) 모터를 사용하여 브러시가 없어 조용하고 전기 소모가 적으며 발열이 적다는 특징을 가지고 있습니다[1]. 또한, BLDC 모터는 전력량은 낮지만 모터의 회전력이 좋고 저소음을 내는 특징을 가지고 있어 써큘레이터 선풍기에서 많이 사용됩니다[2].

신제품 스탠드형 선풍기는 최고급 프리미엄 BLDC모터와 7엽 날개를 탑재하여 강력하면서도 부드러운 바람을 제공합니다[3]. 이 선풍기는 에너지 소모가 적어 친환경 제품이며, 오랜 사용에도 높은 내구성을 자랑합니다[3]. 또한, 클레어에서 새롭게 선보인 무선 선풍기는 고성능 BLDC 모터를 사용하여 소음과 전력소

Finished: 신상품 저소음 선풍기의 특징과 장점을 조사

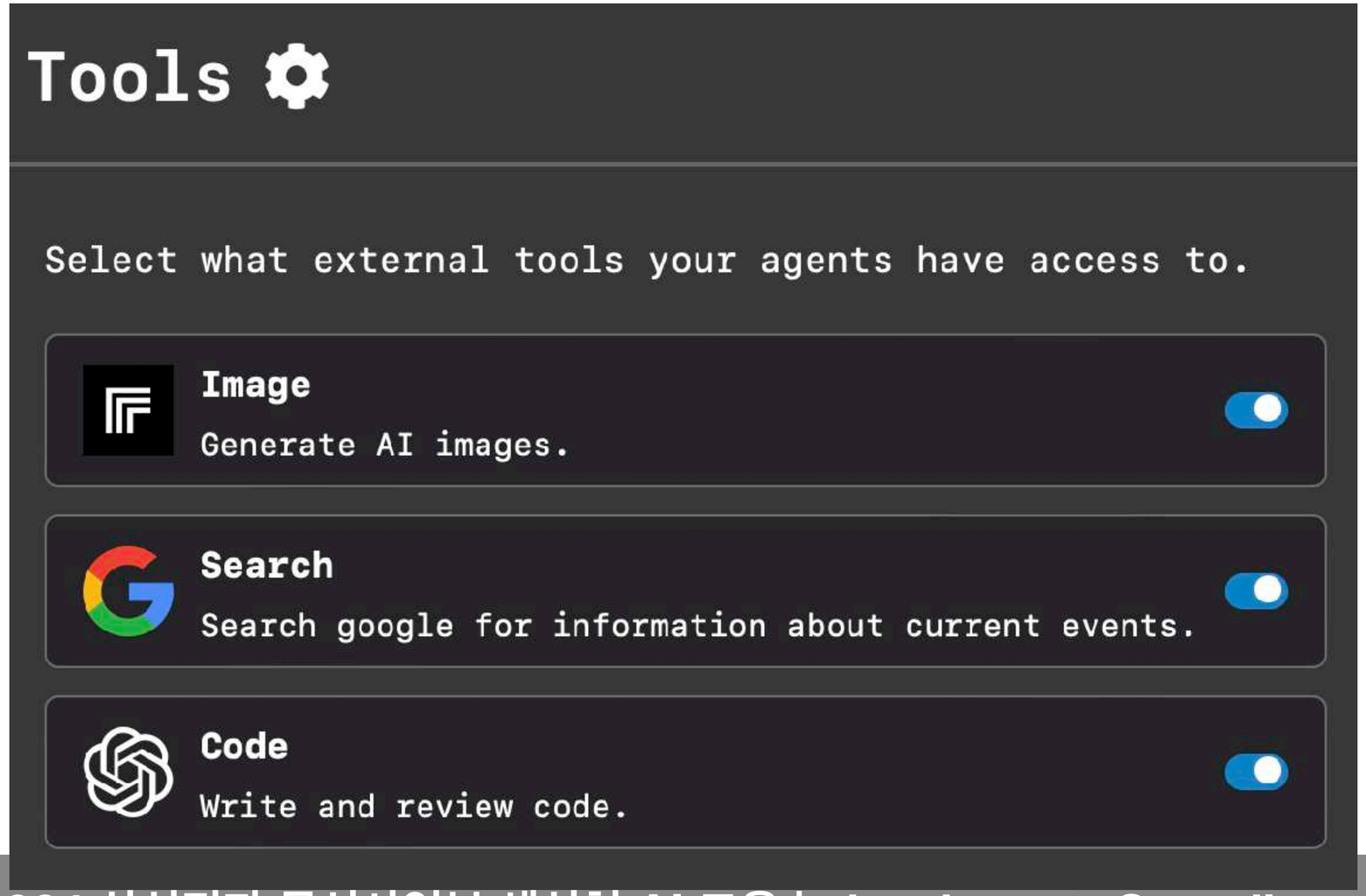
✨ Starting task: 깔끔한 디자인과 인테리어 효과를 강조하는 글 작성

Generating response...

Executing: 깔끔한 디자인과 인테리어 효과를 강조하는 글 작성

신상품 저소음 선풍기는 부드러운 곡선의 깔끔한 디자인을 가지고 있으며, 이는 인테리어 효과를 높이는 중요한 요소입니다. 이 제품은 공간에 조화롭게 어울리며, 미적 감각을 자극합니다. 또한, 편리한 보관 가능성을 갖고 있어 공간을 절약할 수 있습니다. 이 제품은 에너지 소모가 적어 환경에 친숙한 제품입니다.

따라서, 이 선풍기는 인테리어에 중요한 역할을 하며, 깔끔한 디자인과 인테리어 효과를 강조하는 글을 작성하려 합니다.


Tools
Select what external tools your agents have access to.
Image
Generate AI images.
Search
Search google for information about current events.
Code
Write and review code.

Azure OpenAI Service

Exclusive License

유일한 OpenAI 모델
공급 클라우드!

OpenAI

Microsoft

GPT-4
ChatGPT
DALL·E

M365
Azure Cloud
Power Platform

Azure OpenAI Service

2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincentg@gmail.com

리소스

리소스 그룹

지역

구독

리소스 Resource

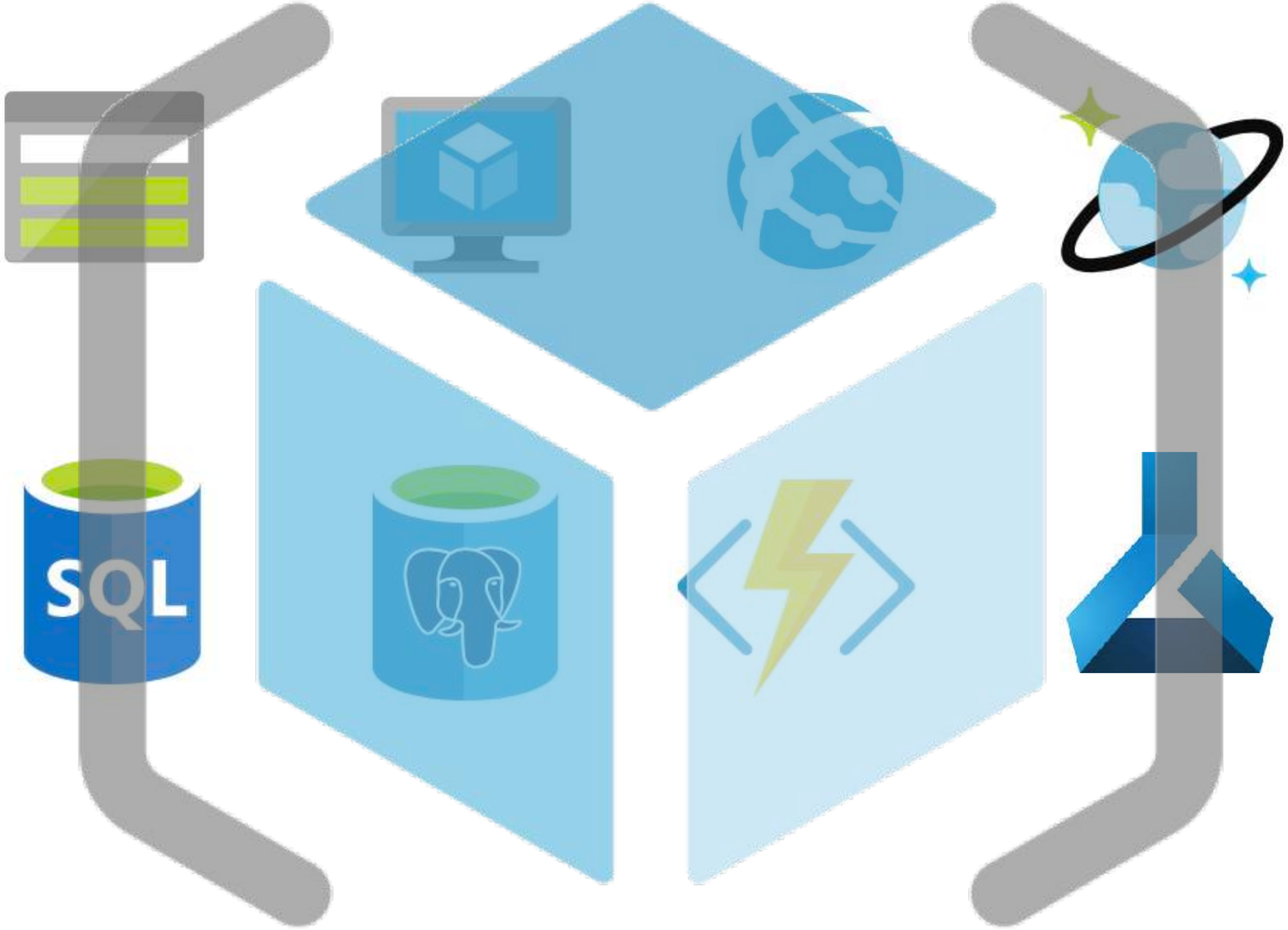
지역

ntg@gmail.com

https://azure.microsoft.c

Region Korea Central

리소스 그룹



2024 삼성전자 무선사업부 생성형 [illegible] @gmail.com

Region Korea Central

구독 Subscription

Resource Group

Resource

Subscription

GPT-3.5 Turbo

GPT-3.5 Turbo is our fast and inexpensive model for simpler tasks.

`gpt-3.5-turbo-0125` is the flagship model of this family, supports a 16K context window and is optimized for dialog.

`gpt-3.5-turbo-instruct` is an Instruct model and only supports a 4K context window.

Learn about GPT-3.5 Turbo ↗

(영어 기준)

1,000,000 토큰 = 750,000 단어

= 2,500 페이지 = 10권

Model	Input	Output
gpt-3.5-turbo-0125	$0.50 / 1M tokens	$1.50 / 1M tokens
gpt-3.5-turbo-instruct	$1.50 / 1M tokens	$2.00 / 1M tokens

2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincentg@gmail.com

https://openai.com/api/pricing/

GPT-4o

GPT-4o is our most advanced multimodal model that's faster and cheaper than GPT-4 Turbo with stronger vision capabilities. The model has 128K context and an October 2023 knowledge cutoff.

Learn about GPT-4o ↗

Model	Input	Output
gpt-4o	$5.00 / 1M tokens	$15.00 / 1M tokens
gpt-4o-2024-05-13	$5.00 / 1M tokens	$15.00 / 1M tokens

x10

2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincentg@gmail.com

https://openai.com/api/pricing/

GPT-4

Our original GPT-4 model was built with broad general knowledge and domain expertise.

Learn about GPT-4 >

Model	Input	Output
gpt-4	$30.00 / 1M tokens	$60.00 / 1M tokens x40
gpt-4-32k	$60.00 / 1M tokens	$120.00 / 1M tokens

https://openai.com/api/pricing/

https://platform.openai.com/tokenizer

Tokenizer

Learn about language model tokenization

OpenAI's large language models (sometimes referred to as GPT's) process text using **tokens**, which are common sequences of characters found in a set of text. The models learn to understand the statistical relationships between these tokens, and excel at producing the next token in a sequence of tokens.

You can use the tool below to understand how a piece of text might be tokenized by a language model, and the total count of tokens in that piece of text.

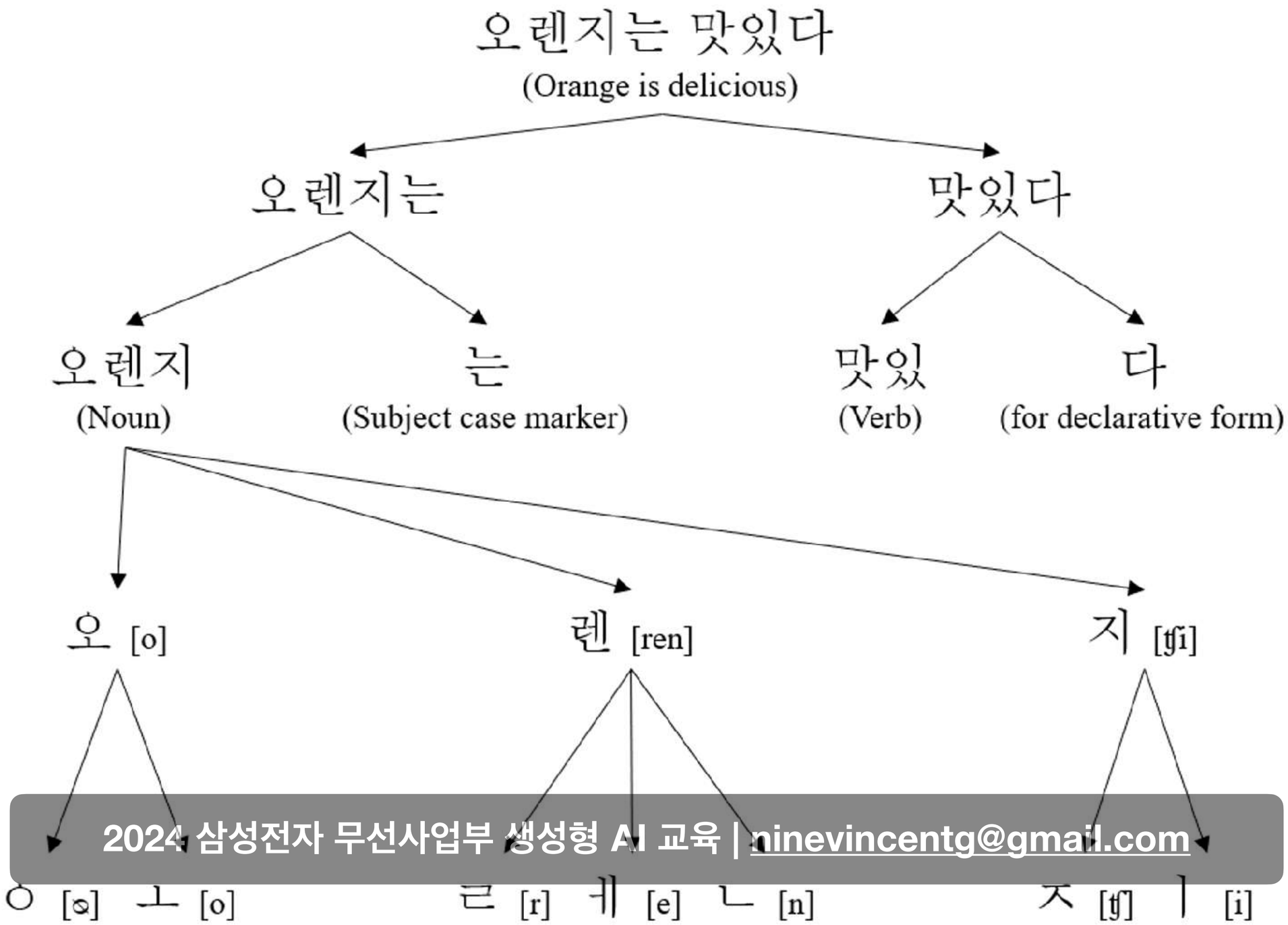
It's important to note that the exact tokenization process varies between mo like GPT-3.5 and GPT-4 use a different tokenizer than previous models, and tokens for the same input text.

GPT-4o (coming soon) | GPT-3.5 & GPT-4 | GPT-3 (Legacy)

안녕하세요, 여러분

Tokens	Characters
9	10

2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincentg@gmail.com



2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincentg@gmail.com

https://smilegate.ai/2022/05/31/korean-tokenizer

https://platform.openai.com/tokenizer

Tokenizer

Learn about language model tokenization

OpenAI's large language models (sometimes referred to as GPT's) process text using **tokens**, which are common sequences of characters found in a set of text. The models learn to understand the statistical relationships between these tokens, and excel at producing the next token in a sequence of tokens.

You can use the tool below to understand how a piece of text might be tokenized by a language model, and the total count of tokens in that piece of text.

It's important to note that the exact tokenization process varies between mo like GPT-3.5 and GPT-4 use a different tokenizer than previous models, and tokens for the same input text.

GPT-4o (coming soon) | GPT-3.5 & GPT-4 | GPT-3 (Legacy)

안녕하세요, 여러분

Tokens	Characters
9	10

2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincentg@gmail.com

“안녕하세요 여러분”

GPT-3.5 & GPT-4 GPT-3 (Legacy)

안녕하세요 여러분

Clear Show example

Tokens	Characters
23	9

GPT-4o (coming soon) | GPT-3.5 & GPT-4 | GPT-3 (Legacy)

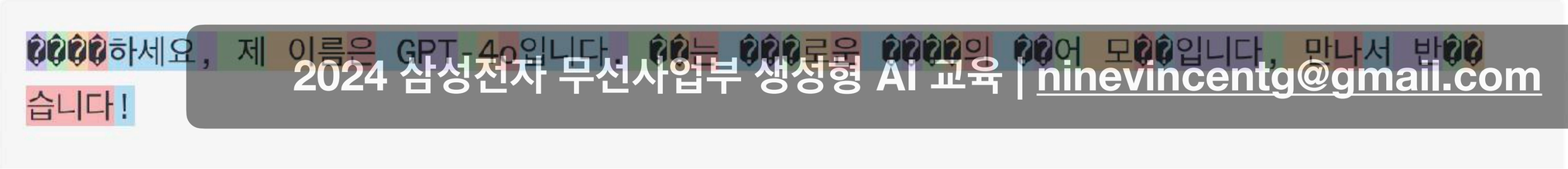
안녕하세요, 제 이름은 GPT-4o입니다. 저는 새로운 유형의 언어 모델입니다, 만나서 반갑습니다!

Korean 1.7x fewer tokens (from 45 to 27)

안녕하세요, 제 이름은 GPT-4o입니다. 저는 새로운 유형의 언어 모델입니다, 만나서 반갑습니다!

Clear | Show example

Tokens	Characters
45	55



2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincentg@gmail.com

“안녕하세요 여러분”

CLOVA Studio β 플레이그라운드

토큰 계산기

입력한 문장의 토큰 수를 계산합니다.

토큰 수

3

글자 수

공백 포함	공백 제외
9	8

Completion

```
1 from openai import OpenAI
2 client = OpenAI()
3
4 response = client.completions.create(
5   model="gpt-3.5-turbo-instruct",
6   prompt="Write a tagline for an ice cream shop."
7 )
```

ChatCompletion

```python
from openai import OpenAI
client = OpenAI()

response = client.chat.completions.create(
  model="gpt-3.5-turbo",
  messages=[
    {"role": "system", "content": "You are a helpful assistant."},
    {"role": "user", "content": "Who won the world series in 2020?"},
    {"role": "assistant", "content": "The Los Angeles Dodgers won the
    {"role": "user", "content": "Where was it played?"}
  ]
)
```

Completion 모델

gpt-3.5-turbo-instruct

text-davinci-003

text-ada-001

text-curie-001

text-babbage-001

ChatCompletion 모델

gpt-4

gpt-4o

gpt-3.5-turbo

Parameter. 결과물의 창의성

`temperature` number or null Optional Defaults to 1

What sampling temperature to use, between 0 and 2. Higher values like 0.8 will make the output more random, while lower values like 0.2 will make it more focused and deterministic.

We generally recommend altering this or `top_p` but not both.

`top_p` number or null Optional Defaults to 1

An alternative to sampling with temperature, called nucleus sampling, where the model considers the results of the tokens with top_p probability mass. So 0.1 means only the tokens comprising the top 10% probability mass are considered.

We generally recommend altering this or `temperature` but not both.

2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincentg@gmail.com

https://platform.openai.com/docs/api-reference/completions/create

온도에 따른 후보 토큰 선택 확률 변화

https://developers.kakao.com/docs/latest/ko/kogpt/common

상위 확률에 따른 후보 토큰 수 제한

https://developers.kakao.com/docs/latest/ko/kogpt/common

Parameter. 결과물의 길이

`max_tokens` integer or null Optional Defaults to 16

The maximum number of **tokens** to generate in the completion.

The token count of your prompt plus `max_tokens` cannot exceed the model's context length. **Example Python code** for counting tokens.

`stop` string / array / null Optional Defaults to null

Up to 4 sequences where the API will stop generating further tokens. The returned text will not contain the stop sequence.

https://platform.openai.com/docs/api-reference/completions/create

Parameter. 결과물의 개수

n integer or null Optional Defaults to 1

How many completions to generate for each prompt.

Note: Because this parameter generates many completions, it can quickly consume your token quota. Use carefully and ensure that you have reasonable settings for `max_tokens` and `stop`.

https://platform.openai.com/docs/api-reference/completions/create

Parameter. 고급 설정

`presence_penalty` number or null Optional Defaults to 0

Number between -2.0 and 2.0. Positive values penalize new tokens based on whether they appear in the text so far, increasing the model's likelihood to talk about new topics.

See more information about frequency and presence penalties.

`frequency_penalty` number or null Optional Defaults to 0

Number between -2.0 and 2.0. Positive values penalize new tokens based on their existing frequency in the text so far, decreasing the model's likelihood to repeat the same line verbatim.

See more information about frequency and presence penalties.

https://platform.openai.com/docs/api-reference/completions/create

Reproducible outputs Beta

seed integer or null Optional

This feature is in Beta. If specified, our system will make a best effort to sample deterministically, such that repeated requests with the same `seed` and parameters should return the same result. Determinism is not guaranteed, and you should refer to the `system_fingerprint` response parameter to monitor changes in the backend.

https://platform.openai.com/docs/guides/text-generation/reproducible-outputs

JSON mode New

```
response = client.chat.completions.create(
    model="gpt-3.5-turbo-1106",
    max_tokens=128,
    messages=[
        {"role":"system", "content":"JSON 형태로 결과를 반환하는 보조"},
        {"role":"user", "content":"떡볶이 칼로리를 알려줘"}
    ],

)

print(response.choices[0].message.content)
```

떡볶이 1인분(200g)의 칼로리는 약 340kcal입니다.

2024 삼성전자 무선사업부 생성형 AI 교육입니다. nevincentg@gmail.com

https://platform.openai.com/docs/guides/text-generation/json-mode

JSON mode New

```
response = client.chat.completions.create(
    model="gpt-3.5-turbo-1106",
    max_tokens=128,
    messages=[
        {"role":"system", "content":"JSON 형태로 결과를 반환하는 보조"},
        {"role":"user", "content":"떡볶이 칼로리를 알려줘"}
    ],
   response_format={"type": "json_object"}
)

print(response.choices[0].message.content)
```

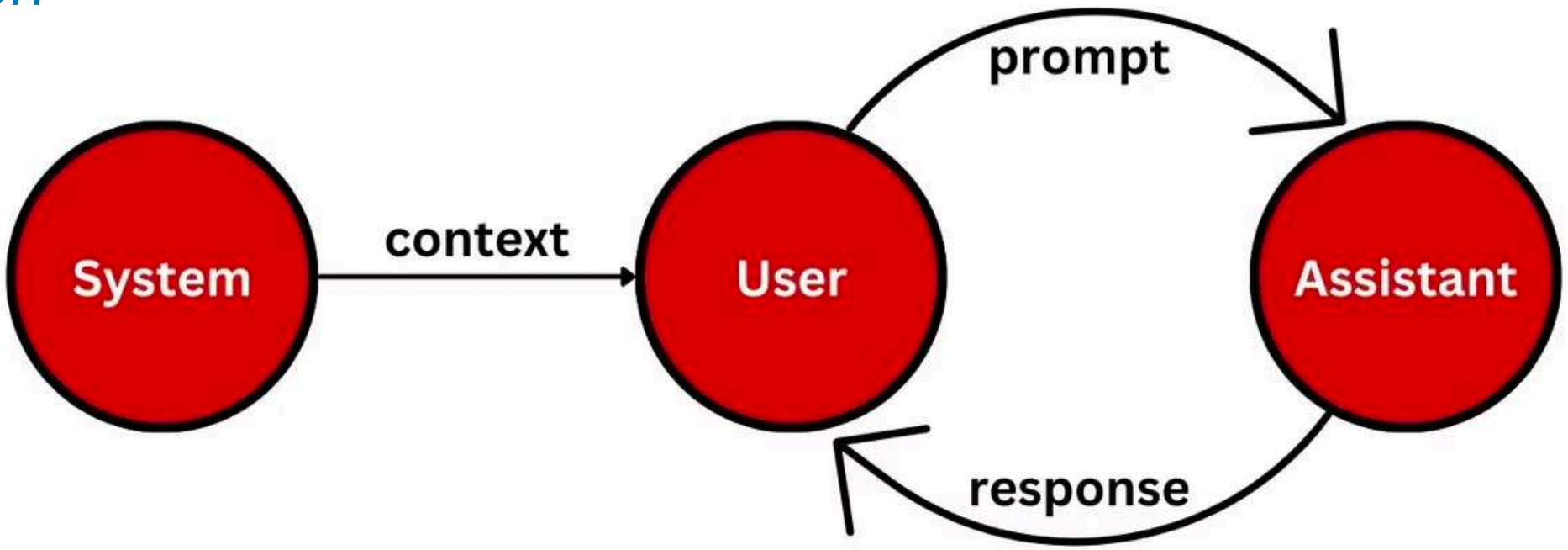
```
{
  "food": "떡볶이",
  "calories": "약 300kcal"
}
```


https://platform.openai.com/docs/guides/text-generation/json-mode

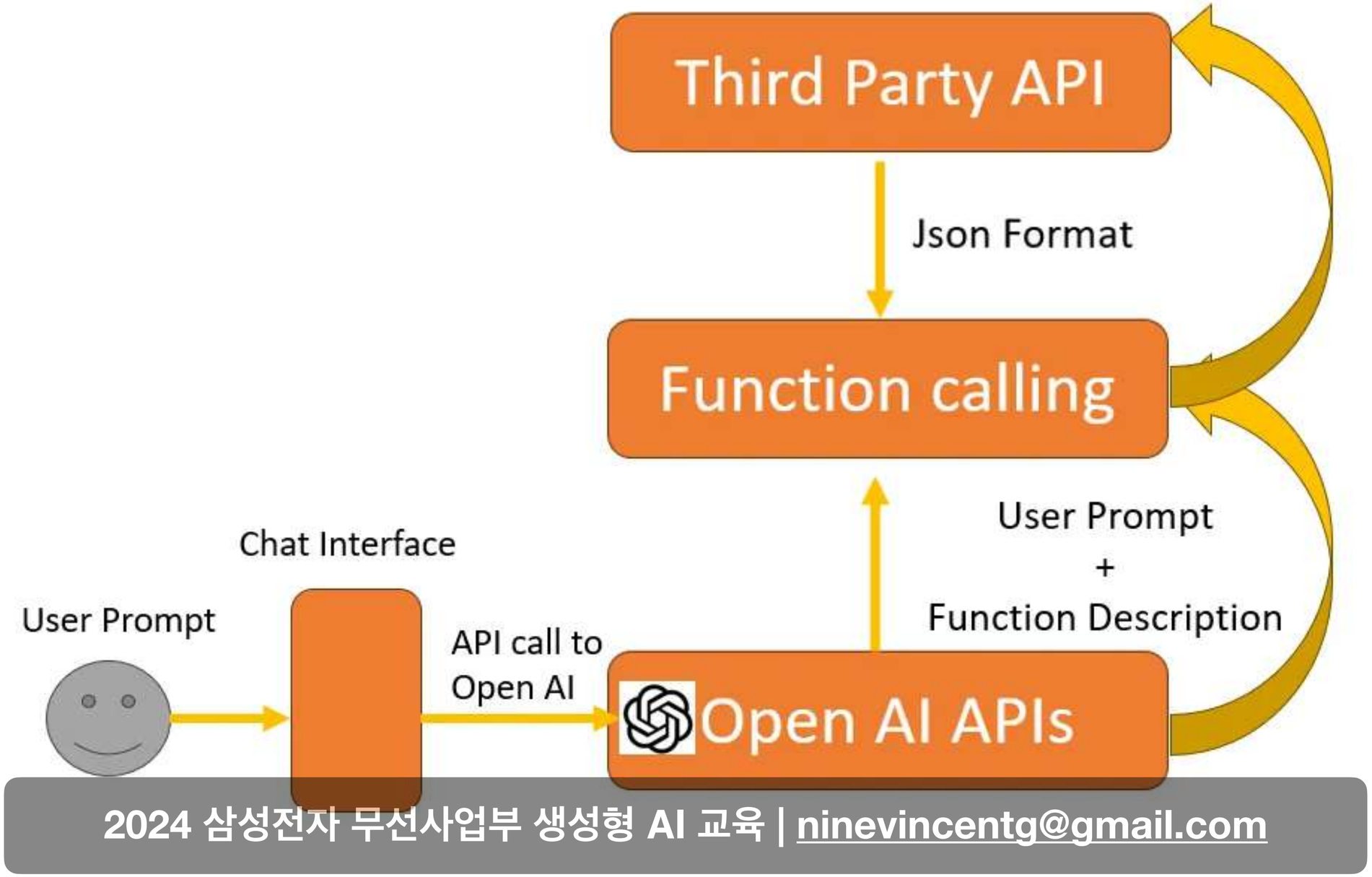
Role in messages

- System : 컨텍스트 및 동작 방식, 역할 설정
- User : 수행 작업 및 지침 제공
- Assistant : 답변을 하는 객체, 원하는 응답을 제공할 수 있도록 미리 몇 가지 모범 답안 제공

+ function



출처 : https://www.makeuseof.com/chatgpt-api-complete-guide/



https://medium.com/@chamathka3deemanthi/open-ai-function-calling-for-chat-completion-f9a4b85ff457

	Instruct Model	Chat Model
사용 사례	특정 지향적 작업, 구체적 명령	끝이 열린 대화, 콘텐츠 생성
예시	"이 텍스트 요약하기", "이 문장을 프랑스어로 번역하기"	"이야기를 들려줘", "XYZ에 대한 너의 생각은?"
장점	정확한 출력, 자동화 용이	더 유연함, 모호성 처리
최적화된 사용	정의된 작업, 자동화	창의적 또는 탐색적 상호작용
API 설정	주로 낮은 "온도" 설정	주로 높은 "온도" 설정
안정성	비교적 낮은 안정성	높은 안정성&신뢰도

Moving from text completions to chat completions

We introduced the Chat Completions API in March, and it now accounts for 97% of our API GPT usage.

The initial Completions API was introduced in June 2020 to provide a freeform text prompt for interacting with our language models. We've since learned that we can often provide better results with a more structured prompt interface. The chat-based paradigm has proven to be powerful, handling the vast majority of previous use cases and new conversational needs, while providing higher flexibility and specificity. In particular, the Chat Completions API's structured interface (e.g., system messages, function calling) and multi-turn conversation capabilities enable developers to build conversational experiences and a broad range of completion tasks. It also helps lower the risk of prompt injection attacks, since user-provided content can be structurally separated from instructions.

https://openai.com/index/gpt-4-api-general-availability/

Can I fine-tune the image capabilities in gpt-4?

No, we do not support fine-tuning the image capabilities of `gpt-4` at this time.

Can I use gpt-4 to generate images?

No, you can use `dall-e-3` to generate images and `gpt-4o` or `gpt-4-turbo` to understand images.

What type of files can I upload?

We currently support PNG (.png), JPEG (.jpeg and .jpg), WEBP (.webp), and non-animated GIF (.gif).

Is there a limit to the size of the image I can upload?

Yes, we restrict image uploads to 20MB per image.

Can I delete an image I uploaded?

2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincentg@gmail.com

No, we will delete the image for you automatically after it has been processed by the model.

https://platform.openai.com/docs/guides/vision/faq

Processing and narrating a video with GPT's visual capabilities and the TTS API

Kai Chen

Nov 6, 2023

Open in Github

This notebook demonstrates how to use GPT's visual capabilities with a video. GPT-4o doesn't take videos as input directly, but we can use vision and the 128K context window to describe the static frames of a whole video at once. We'll walk through two examples:

1. Using GPT-4o to get a description of a video
2. Generating a voiceover for a video with GPT-o and the TTS API

```
from IPython.display import display, Image, Audio

import cv2  # We're using OpenCV to read video, to install !pip install opencv-python
import base64
import time
from openai import OpenAI
import os
import requests

client = OpenAI(api_key=os.environ.get("OPENAI_API_KEY"),  # <key>OpenAI API key if not set as env var
```

2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincentg@gmail.com

https://cookbook.openai.com/examples/gpt_with_vision_for_video_understanding

N-Shot Learning

여러 사례에서 얻은 지식을 일반화하여

다른 작업에 재사용할 수 있는 인간의 인지 능력

Zero-Shot Learning(ZSL)

학습 단계에서 보지 않았던 새로운 데이터에 대해

예측/분류/생성 등의 작업을 수행하는 것

수행해야하는 작업에 대한 설명만 제공 받음

영어로 번역해줘
"어서오세요"

"Welcome"

Few-Shot Learning(FSL)

Low-Shot Learning(LSL)

모델 추론 단계에서 prompt 내부에

소량의 예시를 보여주고 예측하게 하는 것

영어로 번역해줘
"안녕" --> "Hi~~~"
"반갑습니다" --> "I'm so glad to meet you!"
"어서오세요" -->

"Welcome!"

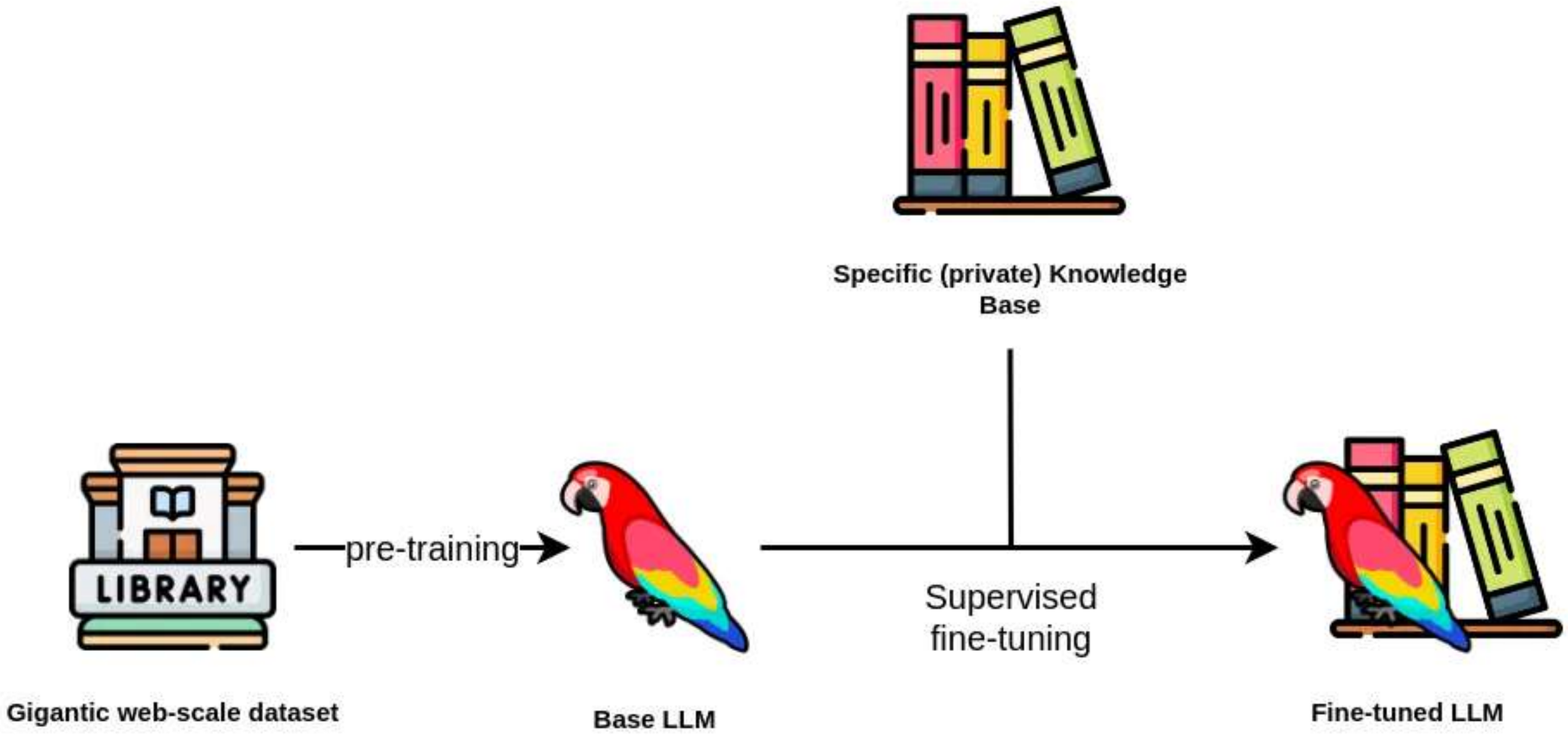
Zero-Shot Learning(ZSL)

Few-Shot Learning(FSL)

Model Parameter 변화 없음

=> Weight Update 🙅🏻‍♀️

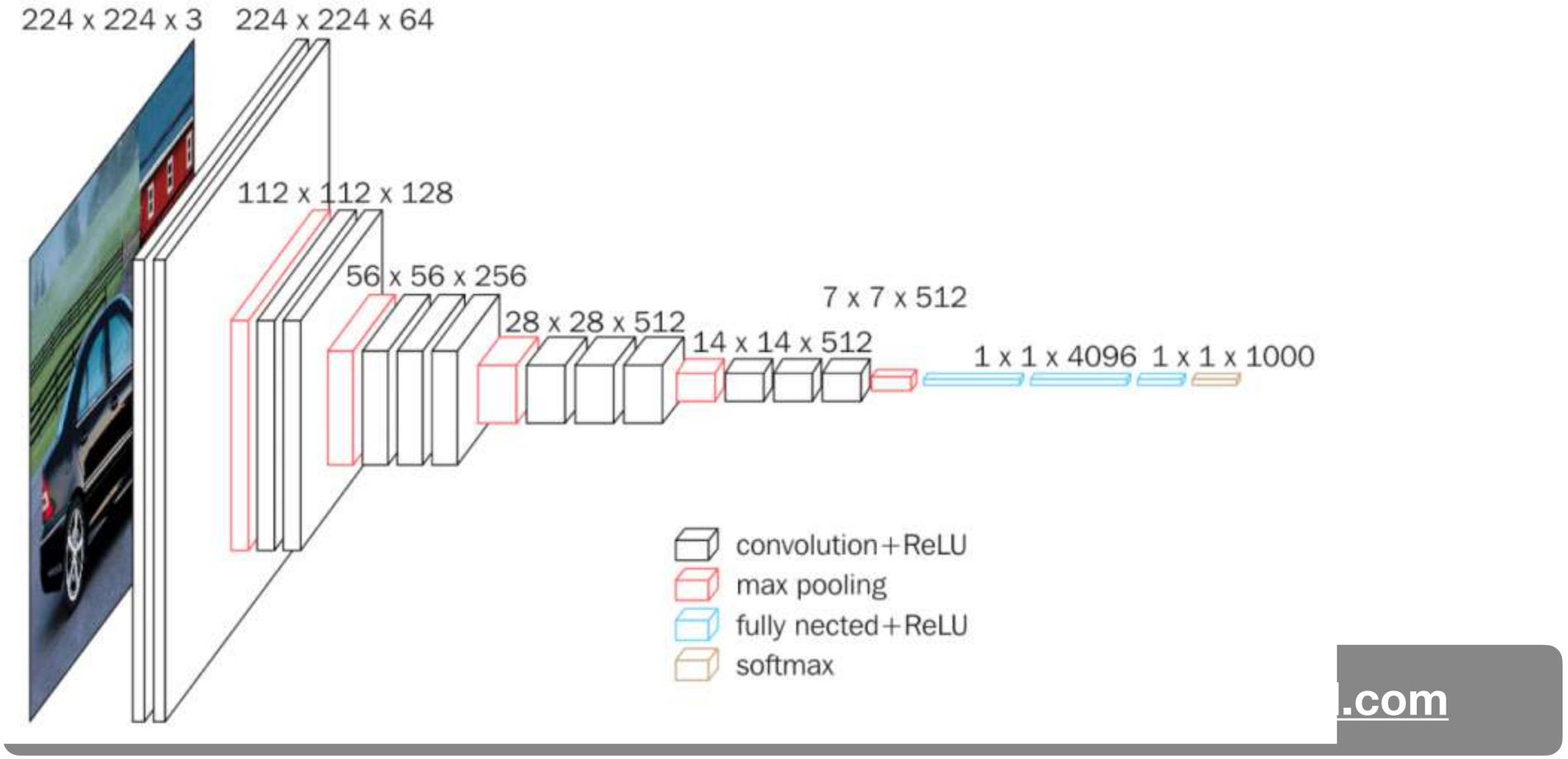
Fine-tuning(미세 조정)



출처 : https://medium.com/neo4j/knowledge-graphs-llms-fine-tuning-vs-retrieval-augmented-generation-30e875d63a35

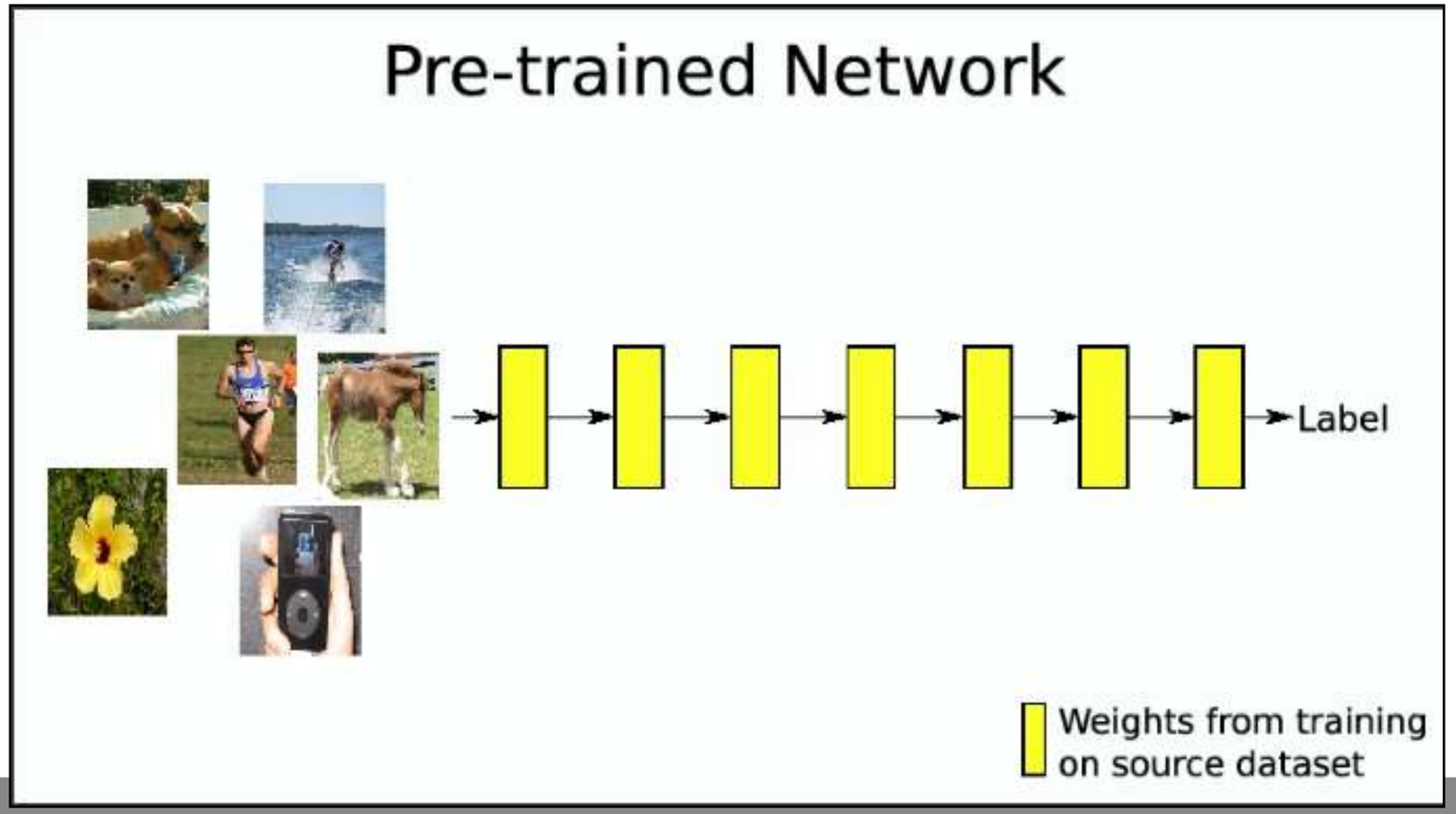
Transfer Leaning(전이학습)

Pre-trained Model (VGG, ResNet, BERT, GPT, Stable Diffusion)



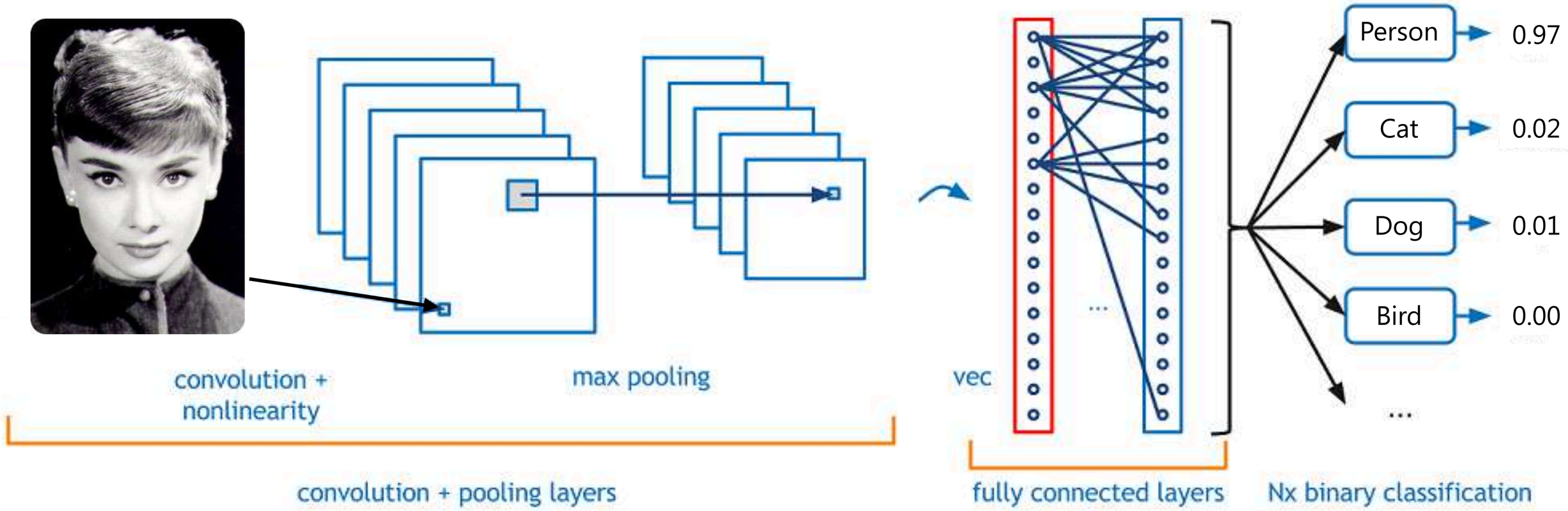
출처 : https://neurohive.io/en/popular-networks/vgg16/, Neurohive, VGG16 - Convolutional Network for Classification and Detection

Fine-tuning(미세 조정)



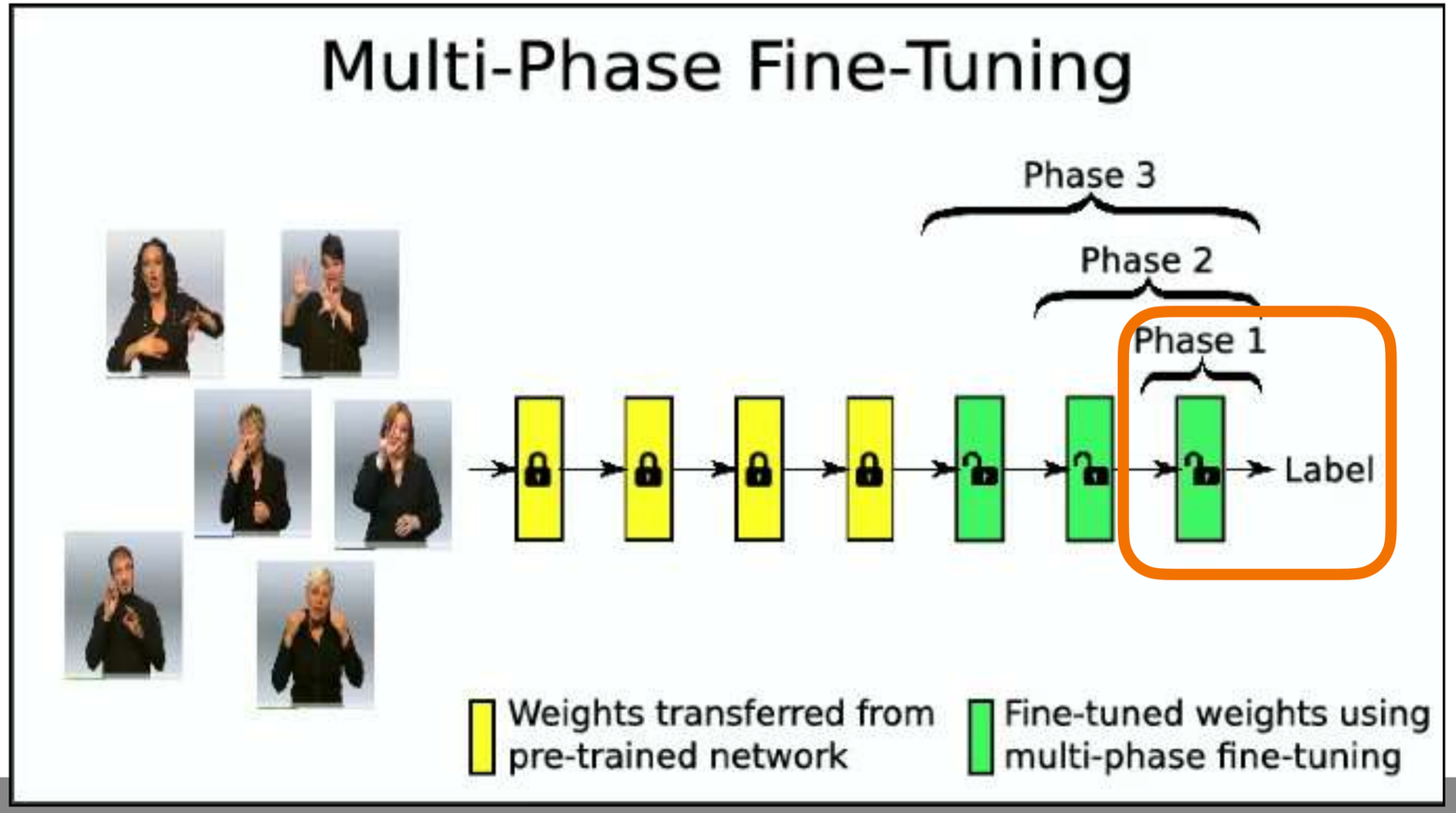
출처 : https://link.springer.com/article/10.1007/s13218-021-00746-2, Multi-phase Fine-Tuning: A New Fine-Tuning Approach for Sign Language Recognition

CNN(Convolutional Neural Network)



https://adeshpande3.github.io/A-Beginner%27s-Guide-To-Understanding-Convolutional-Neural-Networks/

Fine-tuning(미세 조정)



출처 : https://link.springer.com/article/10.1007/s13218-021-00746-2, Multi-phase Fine-Tuning: A New Fine-Tuning Approach for Sign Language Recognition

Fine-tuning(미세 조정)

- Comma-separated values (CSV)
- Tab-separated values (TSV)
- Microsoft Excel workbook (XLSX)
- JavaScript Object Notation (JSON)
- JSON Lines (JSONL)

* The file must be less than 200 MB

Dataset for Finetuning: Completion Model

.jsonl

```
{"prompt": "질문1", "completion" : " 적절한 답변1" }
{"prompt": "질문2", "completion" : " 적절한 답변2" }
{"prompt": "질문3", "completion" : " 적절한 답변3" }
```

Dataset for Finetuning: Chat Model

.jsonl

```
{"messages":
 [{"role": "system", "content": "친절하게 정보를 자세히 설명해주는 상담원"},
  {"role": "user", "content": "올해 가장 인기있는 상품은 뭔가요?"},
  {"role": "assistant", "content": "안녕하세요, 올해 가장 인기 있는 상품은 SY293 모델입니다. 다양한 색상으로 큰 인기를 끌고 있답니다."}]}

{"messages":
 [{"role": "system", "content": "친절하게 정보를 자세히 설명해주는 상담원"},
  {"role": "user", "content": "60대 부모님께 선물하기 좋은 상품을 추천해주세요"},
  {"role": "assistant", "content": "네, 최근 부모님들께 가장 인기가 많은 상품은 TS120입니다. 부모님의 건강과 편안한 일상을 위해 좋은 선택이 될거에요"}]}

{"messages":
 [{"role": "system", "content": "친절하게 정보를 자세히 설명해주는 상담원"},
  {"role": "user", "content": "당일 배송이 가능한가요?"},
  {"role": "assistant", "content": "당일 배송이 궁금하시군요, 제품마다 다르지만 오후 4시까지 주문하시면 다음날 받아보실 수 있습니다."}]}
```

Fine-tuning(미세 조정)

We recommend that you provide at least a few hundred high-quality examples to achieve a model that performs better than using well-designed prompts with a base model. From there, performance tends to linearly increase with every doubling of the number of examples. Increasing the number of examples is usually the best and most reliable way of improving performance.

출처 : https://learn.microsoft.com/en-us/azure/cognitive-services/openai/how-to/prepare-dataset

gpt-3 / data / single_digit_three_ops.jsonl

Benjamin Mann Initial commit

Code Blame 2000 lines (2000 loc) · 136 KB

```
{"context": "\n\nQ: What is (9 + 8) * 2?\n\nA:", "completion": " 34"}
{"context": "\n\nQ: What is (4 - 7) - 3?\n\nA:", "completion": " -6"}
{"context": "\n\nQ: What is (1 - 5) - 6?\n\nA:", "completion": " -10"}
{"context": "\n\nQ: What is (1 - 9) * 5?\n\nA:", "completion": " -40"}
{"context": "\n\nQ: What is (4 * 4) * 1?\n\nA:", "completion": " 16"}
{"context": "\n\nQ: What is (8 + 3) * 5?\n\nA:", "completion": " 55"}
{"context": "\n\nQ: What is (7 + 5) - 6?\n\nA:", "completion": " 6"}
{"context": "\n\nQ: What is (2 - 8) - 4?\n\nA:", "completion": " -10"}
{"context": "\n\nQ: What is (2 * 4) * 6?\n\nA:", "completion": " 48"}
{"context": "\n\nQ: What is (3 - 4) - 1?\n\nA:", "completion": " -2"}
{"context": "\n\nQ: What is (3 + 8) * 8?\n\nA:", "completion": " 88"}
{"context": "\n\nQ: What is (5 - 9) * 3?\n\nA:", "completion": " -12"}
{"context": "\n\nQ: What is (3 * 5) + 7?\n\nA:", "completion": " 22"}
{"context": "\n\nQ: What is (5 + 4) * 7?\n\nA:", "completion": " 63"}
```

2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincentg@gmail.com

gpt-3 / data / five_digit_subtraction.jsonl

Benjamin Mann Initial commit

Code Blame 2000 lines (2000 loc) · 154 KB

```
{"context": "\n\nQ: What is 40649 minus 78746?\n\nA:", "completion": " -38097"}
{"context": "\n\nQ: What is 4344 minus 48888?\n\nA:", "completion": " -44544"}
{"context": "\n\nQ: What is 21851 minus 67089?\n\nA:", "completion": " -45238"}
{"context": "\n\nQ: What is 23935 minus 27571?\n\nA:", "completion": " -3636"}
{"context": "\n\nQ: What is 88956 minus 62810?\n\nA:", "completion": " 26146"}
{"context": "\n\nQ: What is 49608 minus 8751?\n\nA:", "completion": " 40857"}
{"context": "\n\nQ: What is 77548 minus 84027?\n\nA:", "completion": " -6479"}
{"context": "\n\nQ: What is 52793 minus 91812?\n\nA:", "completion": " -39019"}
{"context": "\n\nQ: What is 49097 minus 92714?\n\nA:", "completion": " -43617"}
{"context": "\n\nQ: What is 10429 minus 9489?\n\nA:", "completion": " 940"}
{"context": "\n\nQ: What is 26781 minus 22604?\n\nA:", "completion": " 4177"}
{"context": "\n\nQ: What is 14629 minus 9574?\n\nA:", "completion": " 5055"}
{"context": "\n\nQ: What is 50153 minus 22661?\n\nA:", "completion": " 27492"}
{"context": "\n\nQ: What is 76332 minus 29974?\n\nA:", "completion": " 46358"}
{"context": "\n\nQ: What is 67219 minus 85108?\n\nA:", "completion": " -17889"}
```

2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincentg@gmail.com

gpt-3 / data / two_digit_addition.jsonl

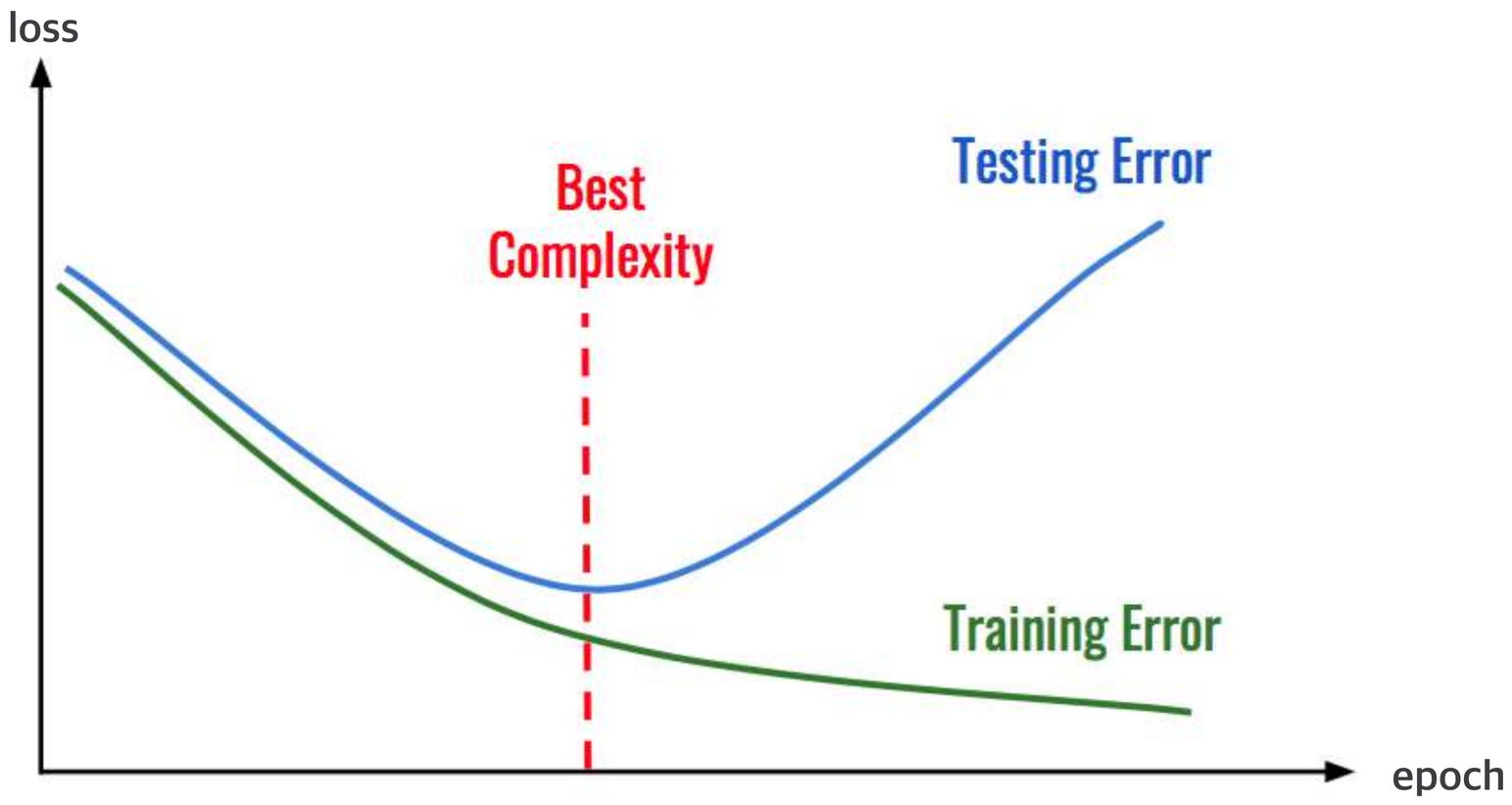
Benjamin Mann Initial commit

Code Blame 2000 lines (2000 loc) · 135 KB

```
{"context": "\n\nQ: What is 98 plus 45?\n\nA:", "completion": " 143"}
{"context": "\n\nQ: What is 95 plus 58?\n\nA:", "completion": " 153"}
{"context": "\n\nQ: What is 0 plus 25?\n\nA:", "completion": " 25"}
{"context": "\n\nQ: What is 72 plus 25?\n\nA:", "completion": " 97"}
{"context": "\n\nQ: What is 70 plus 15?\n\nA:", "completion": " 85"}
{"context": "\n\nQ: What is 52 plus 23?\n\nA:", "completion": " 75"}
{"context": "\n\nQ: What is 36 plus 17?\n\nA:", "completion": " 53"}
{"context": "\n\nQ: What is 47 plus 32?\n\nA:", "completion": " 79"}
{"context": "\n\nQ: What is 15 plus 45?\n\nA:", "completion": " 60"}
{"context": "\n\nQ: What is 10 plus 91?\n\nA:", "completion": " 101"}
{"context": "\n\nQ: What is 72 plus 5?\n\nA:", "completion": " 77"}
{"context": "\n\nQ: What is 61 plus 40?\n\nA:", "completion": " 101"}
{"context": "\n\nQ: What is 64 plus 57?\n\nA:", "completion": " 121"}
{"context": "\n\nQ: What is 93 plus 50?\n\nA:", "completion": " 143"}
{"context": "\n\nQ: What is 25 plus 2?\n\nA:", "completion": " 27"}
{"context": "\n\nQ: What is 99 plus 37?\n\nA:", "completion": " 136"}
```

2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincentg@gmail.com

Loss Function



https://hackernoon.com/memorizing-is-not-learning-6-tricks-to-prevent-overfitting-in-machine-learning-820b091dc42

GPT-3.5 Turbo

GPT-3.5 Turbo models are capable and cost-effective.

`gpt-3.5-turbo-0125` is the flagship model of this family, supports a 16K context window and is optimized for dialog.

`gpt-3.5-turbo-instruct` is an Instruct model and only supports a 4K context window.

Learn about GPT-3.5 Turbo ↗

Model	Input	Output
gpt-3.5-turbo-0125	$0.0005 / 1K tokens	$0.0015 / 1K tokens
gpt-3.5-turbo-instruct	$0.0015 / 1K tokens	$0.0020 / 1K tokens

Fine-tuning models

4~6배 가격차이

Create your own custom models by fine-tuning our base models with your training data. Once you fine-tune a model, you'll be billed only for the tokens you use in requests to that model.

Learn about fine-tuning ↗

Model	Training	Input usage	Output usage
gpt-3.5-turbo	$0.0080 / 1K tokens	$0.0030 / 1K tokens	$0.0060 / 1K tokens
davinci-002	$0.0060 / 1K tokens	$0.0120 / 1K tokens	$0.0120 / 1K tokens
babbage-002	$0.0004 / 1K tokens	$0.0016 / 1K tokens	$0.0016 / 1K tokens

2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincentg@gmail.com

이미지 생성 인공지능

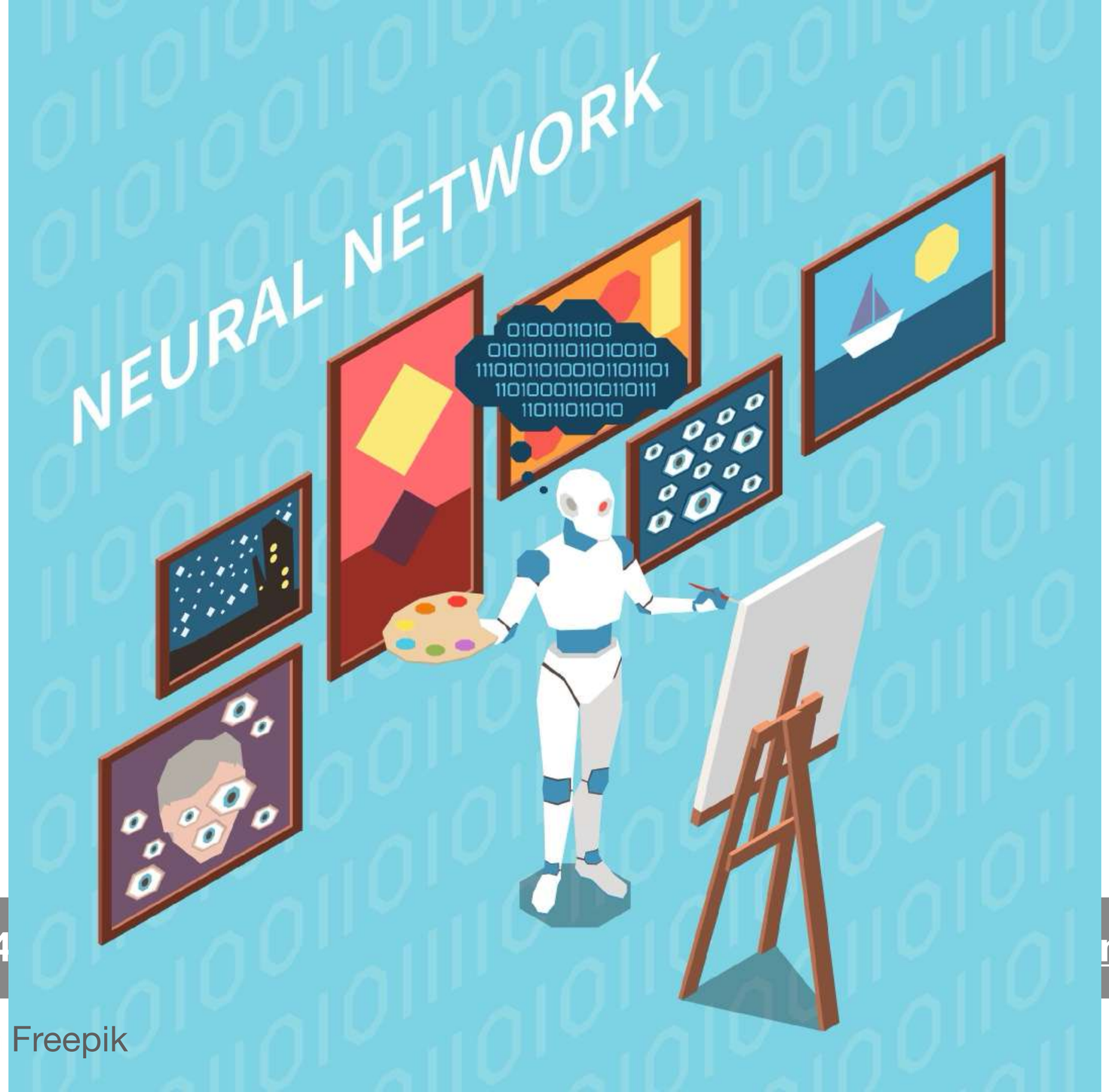
2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincentg@gmail.com

Image models

Build DALL·E directly into your apps to generate and edit novel images and art. DALLE·E 3 is the highest quality model and DALL·E 2 is optimized for lower cost.

Learn about image generation ↗

Model	Quality	Resolution	Price
DALL·E 3	Standard	1024×1024	$0.040 / image
	Standard	1024×1792, 1792×1024	$0.080 / image
DALL·E 3	HD	1024×1024	$0.080 / image
	HD	1024×1792, 1792×1024	$0.120 / image
DALL·E 2		1024×1024	$0.020 / image
		512×512	$0.018 / image
		256×256	$0.016 / image

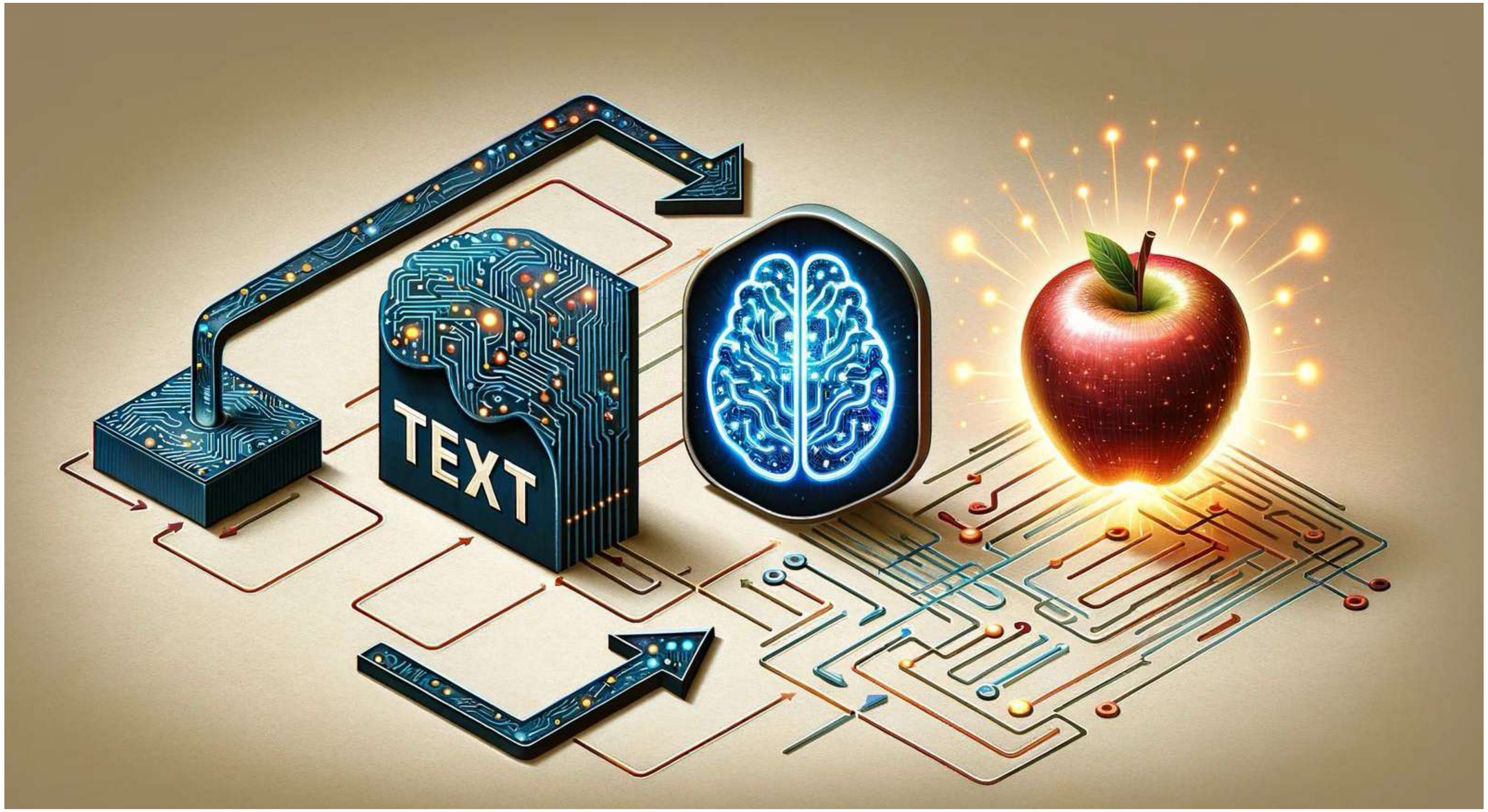


2024

m

Image by macrovector on Freepik

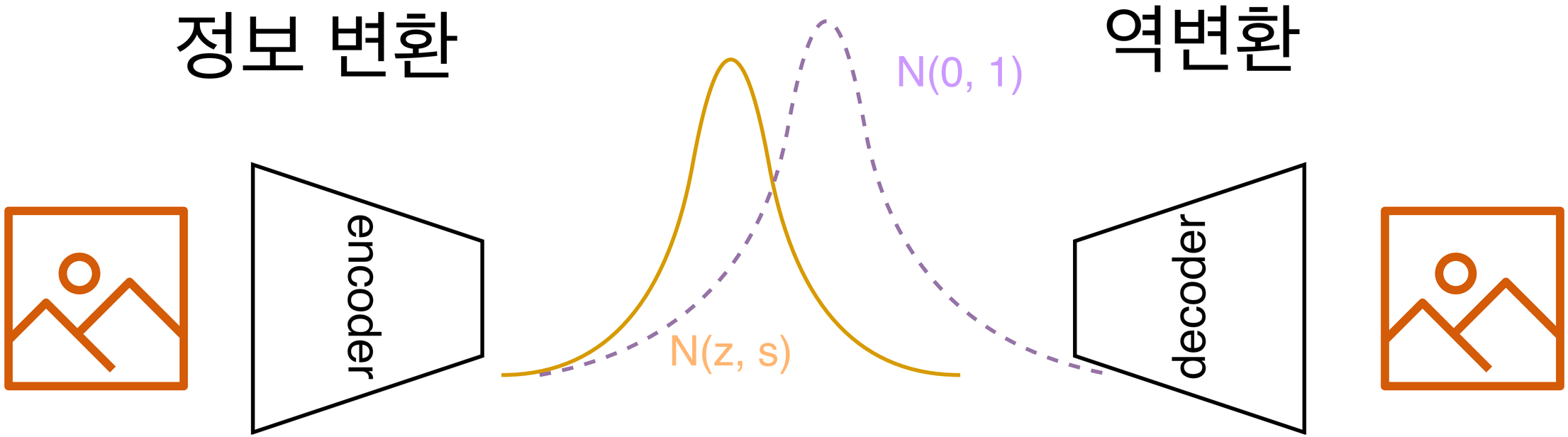
Text to Image Generation



Jason Allen, Théâtre D'opéra Spatial

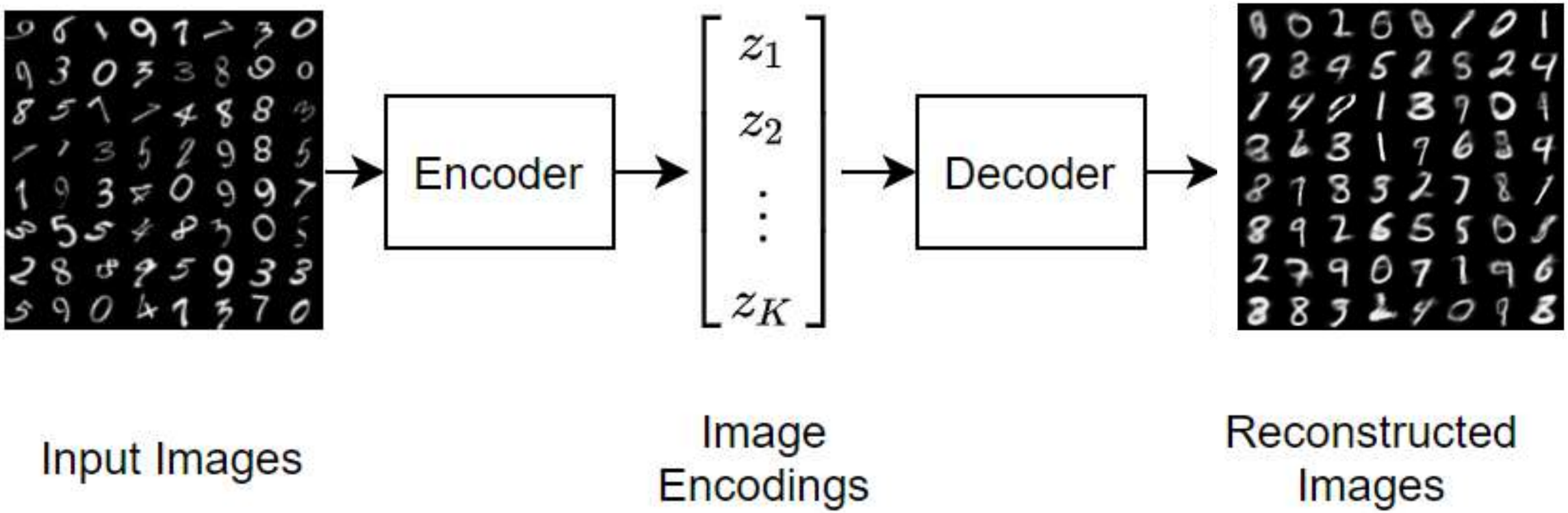
이미지 생성 모델 기술의 발전

2013 VAE(Variational Autoencoder)



https://vaclavkosar.com/ml/openai-dall-e-2-and-dall-e-1

2013 VAE(Variational Autoencoder)



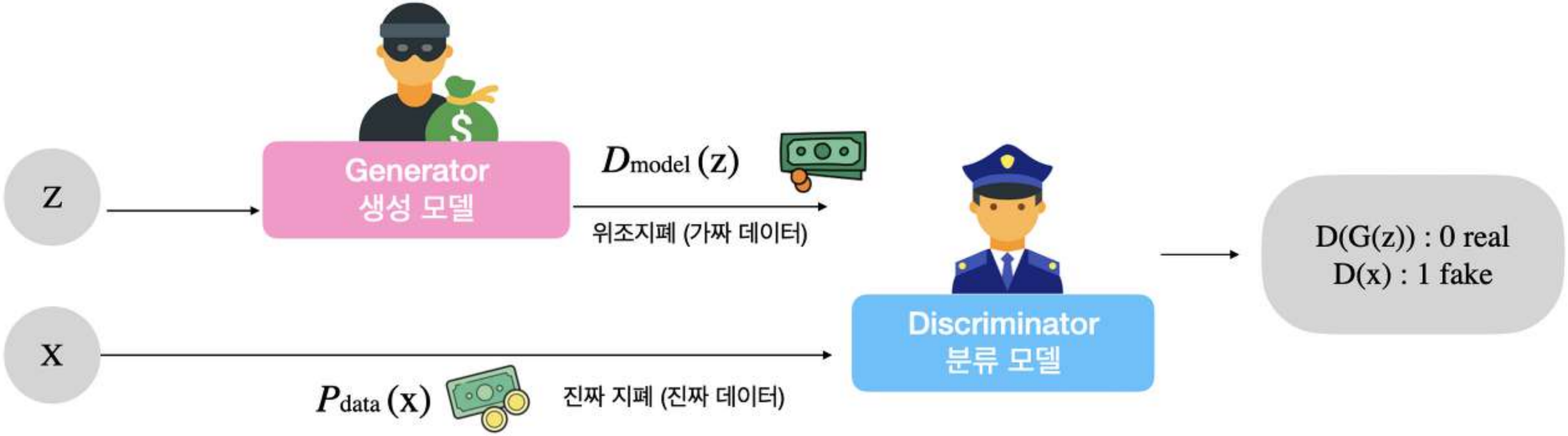
https://kr.mathworks.com/help/deeplearning/ug/train-a-variational-autoencoder-vae-to-generate-images.html

2014 GAN(Generative Adversarial Nets)

IJ Goodfellow, Generative Adversarial Nets, 2014

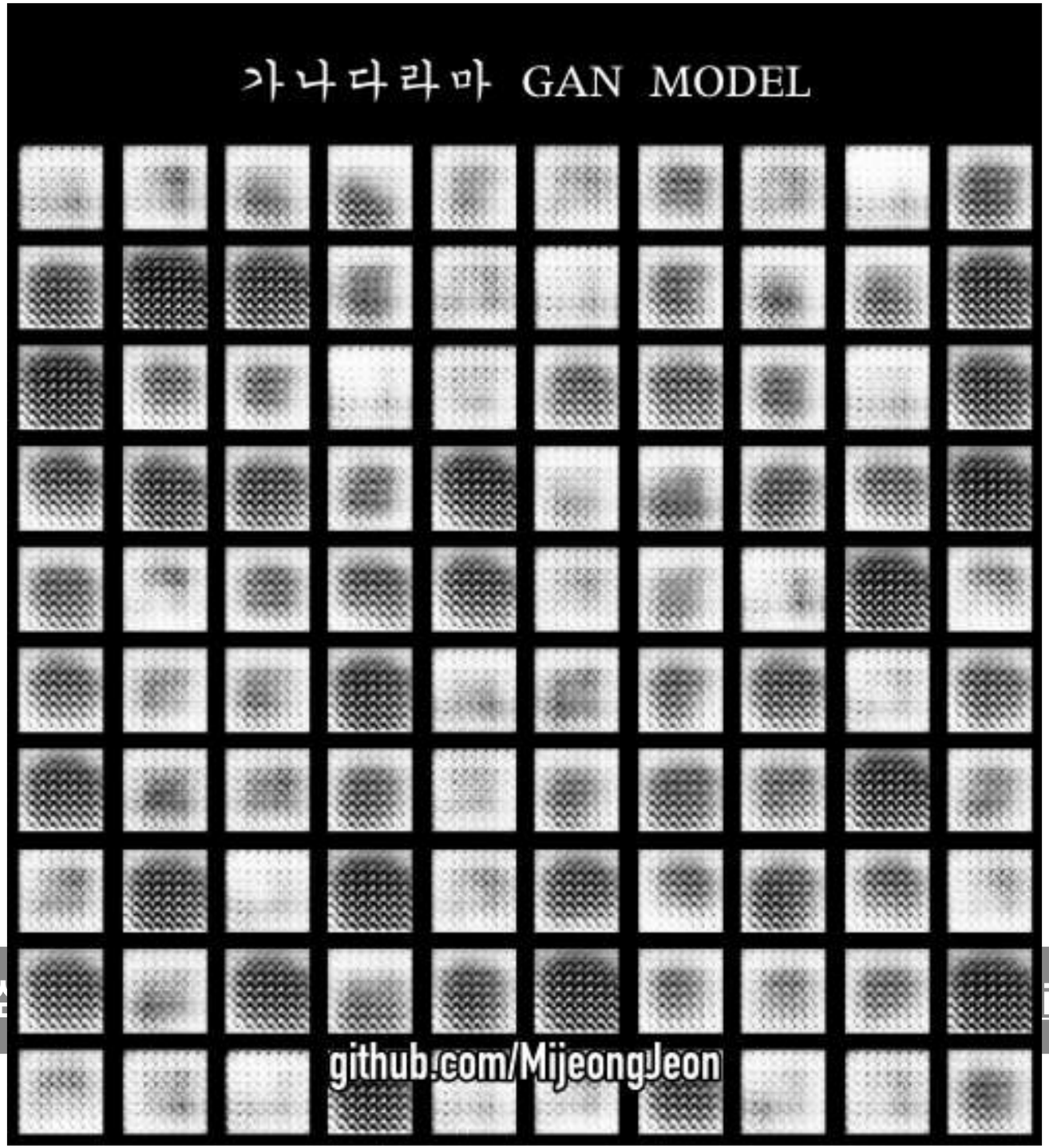
2014 GAN(Generative Adversarial Nets)

$$\min_G \max_D V(D,G) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))].$$



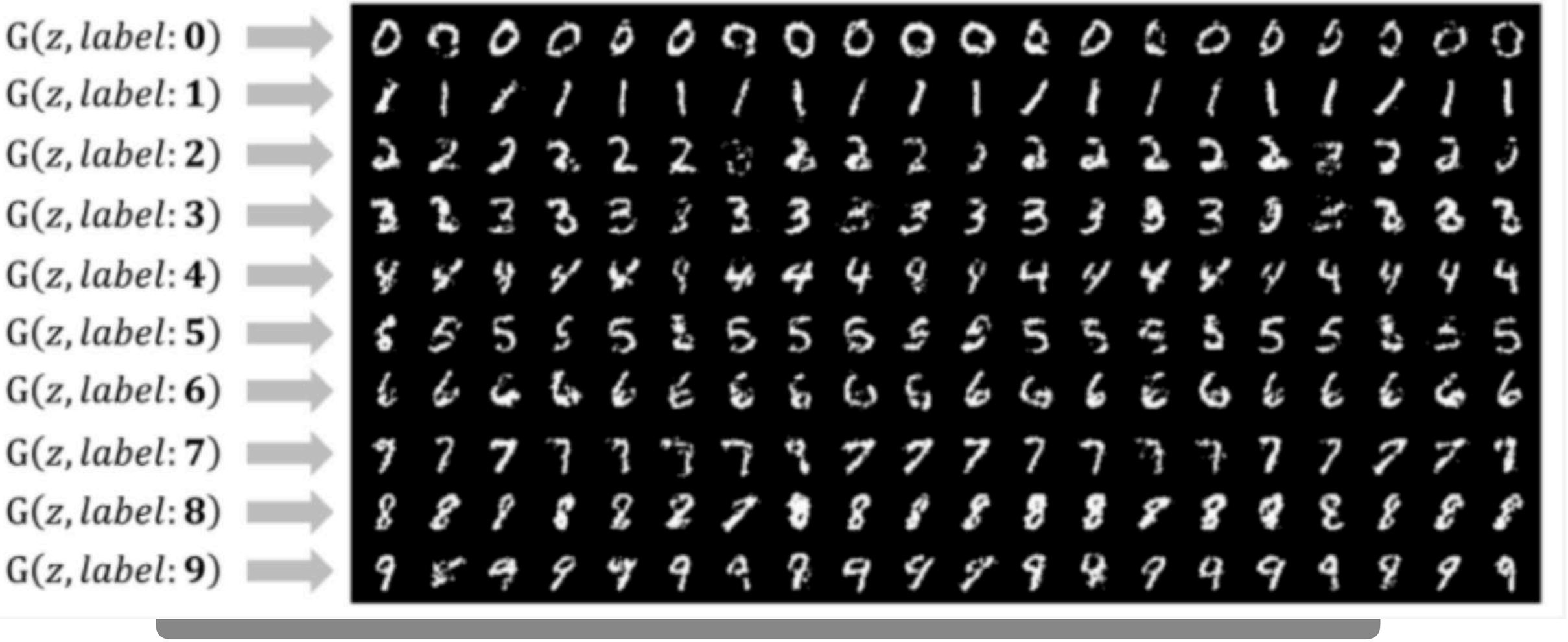
2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincentg@gmail.com

https://wegonnamakeit.tistory.com/54



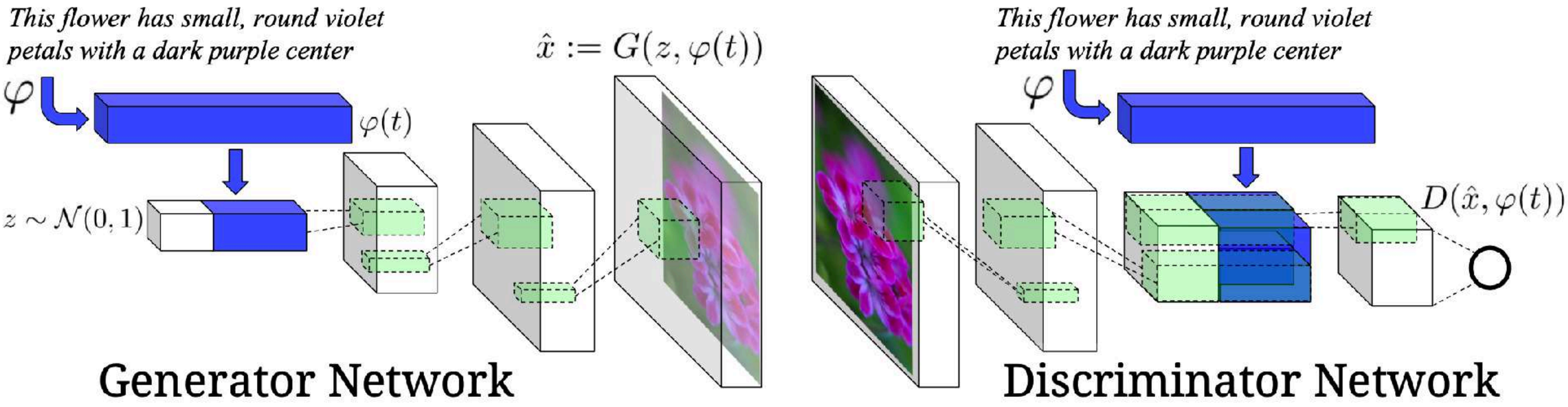
2024 ... il.com

2014 cGAN(Conditional GAN)



M Mirza, Conditional Generative Adversarial Nets, 2014

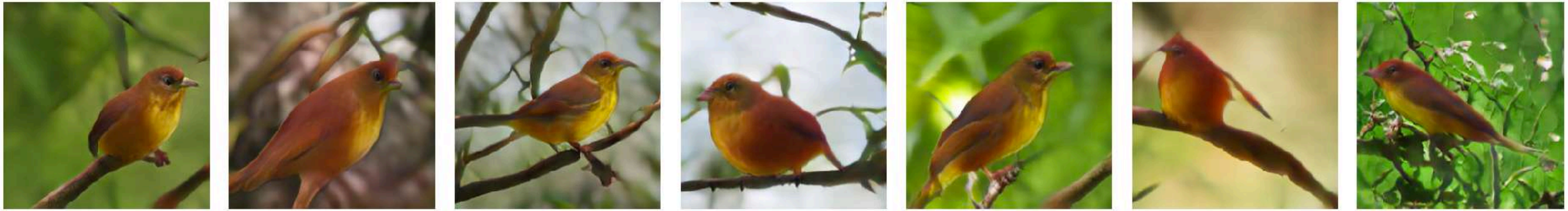
2016 GAN with text conditioning



Honglak Lee, Generative Adversarial Text to Image Synthesis, 2016

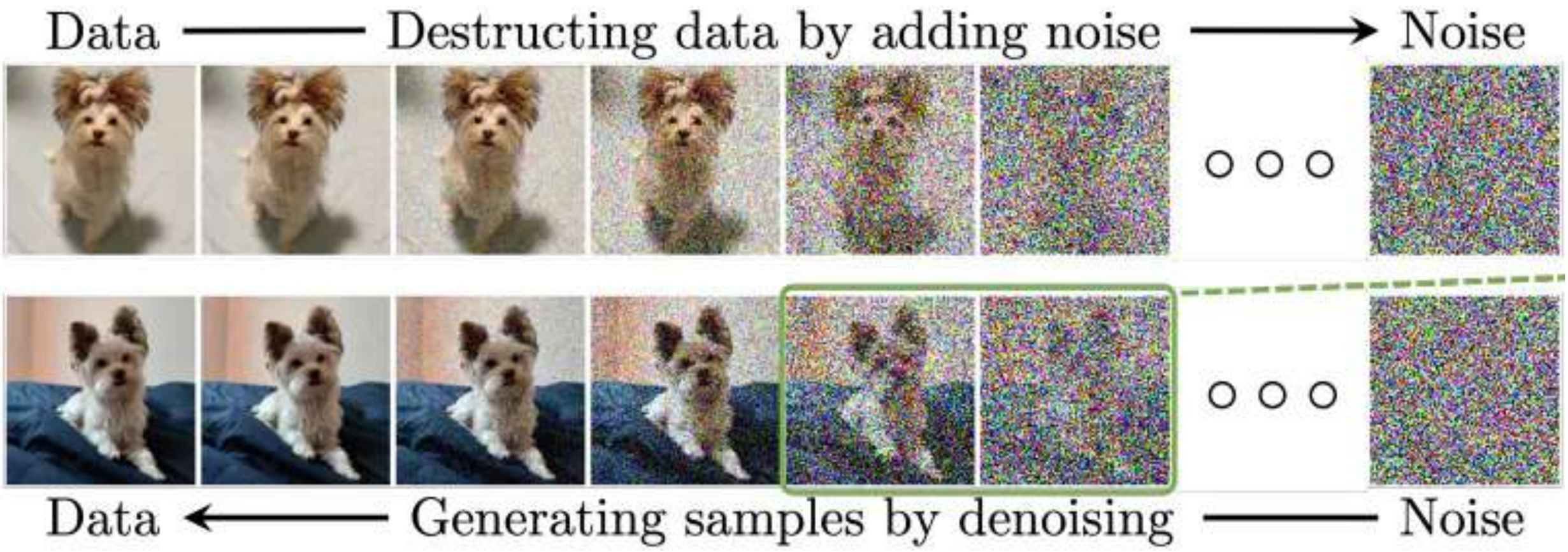
2016 GAN with text conditioning

this bird has wings that are red and has a yellow belly



this bird has wings that are blue and has a red belly

2015 Diffusion Model



MING-HSUAN YANG, Diffusion Models: A Comprehensive Survey of Methods and Applications, 2023

Input Denoising 0% Denoising 60% Denoising 75%

2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincentg@gmail.com

MING-HSUAN YANG, Diffusion Models: A Comprehensive Survey of Methods and Applications, 2023

Input Sample 1 Sample 2 Sample 3 Sample 4

MING-HSUAN YANG, Diffusion Models: A Comprehensive Survey of Methods and Applications, 2023

Image Generation

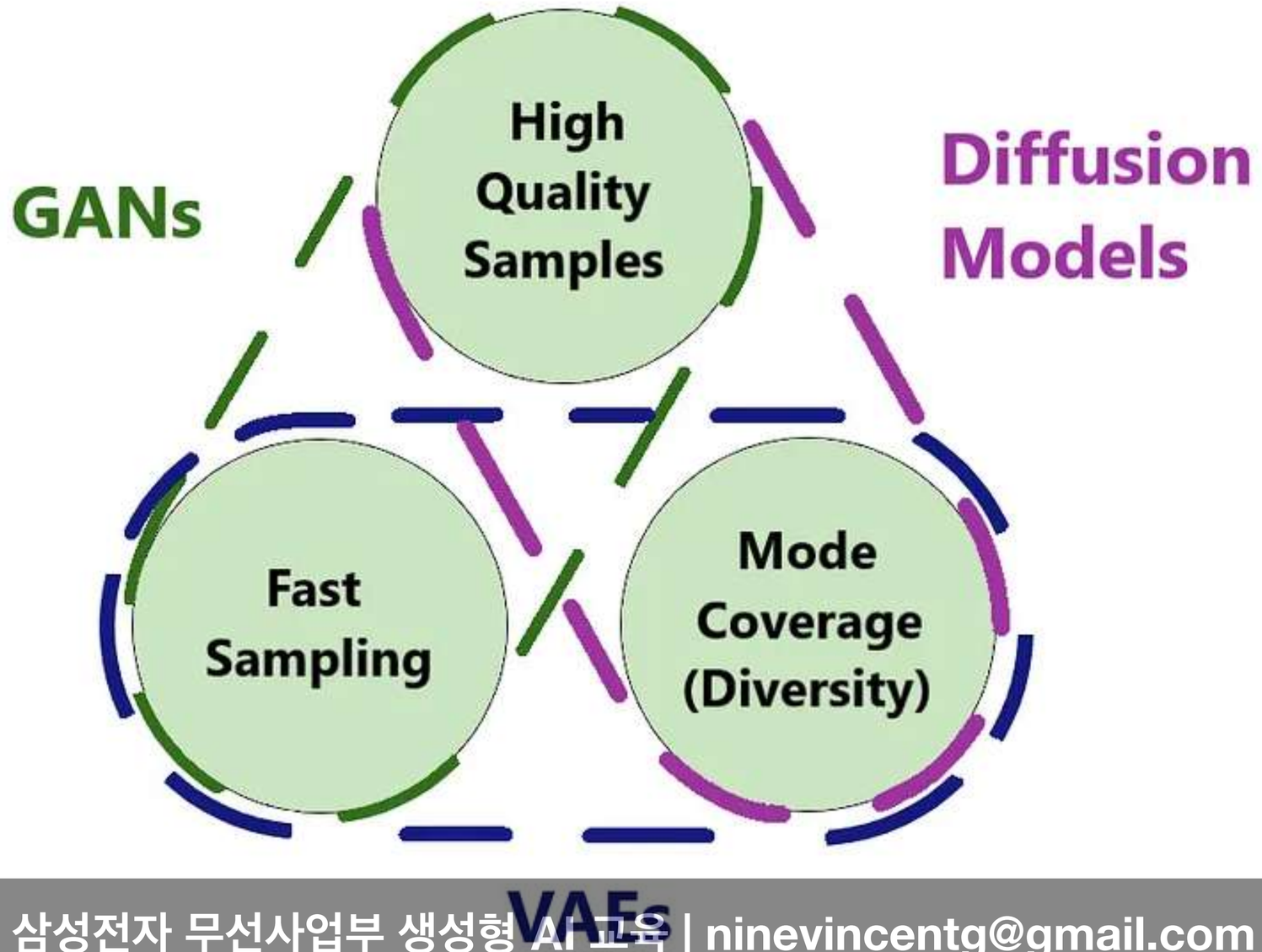
https://clipdrop.co/

Reimage

https://clipdrop.co/

Clean

https://clipdrop.co/

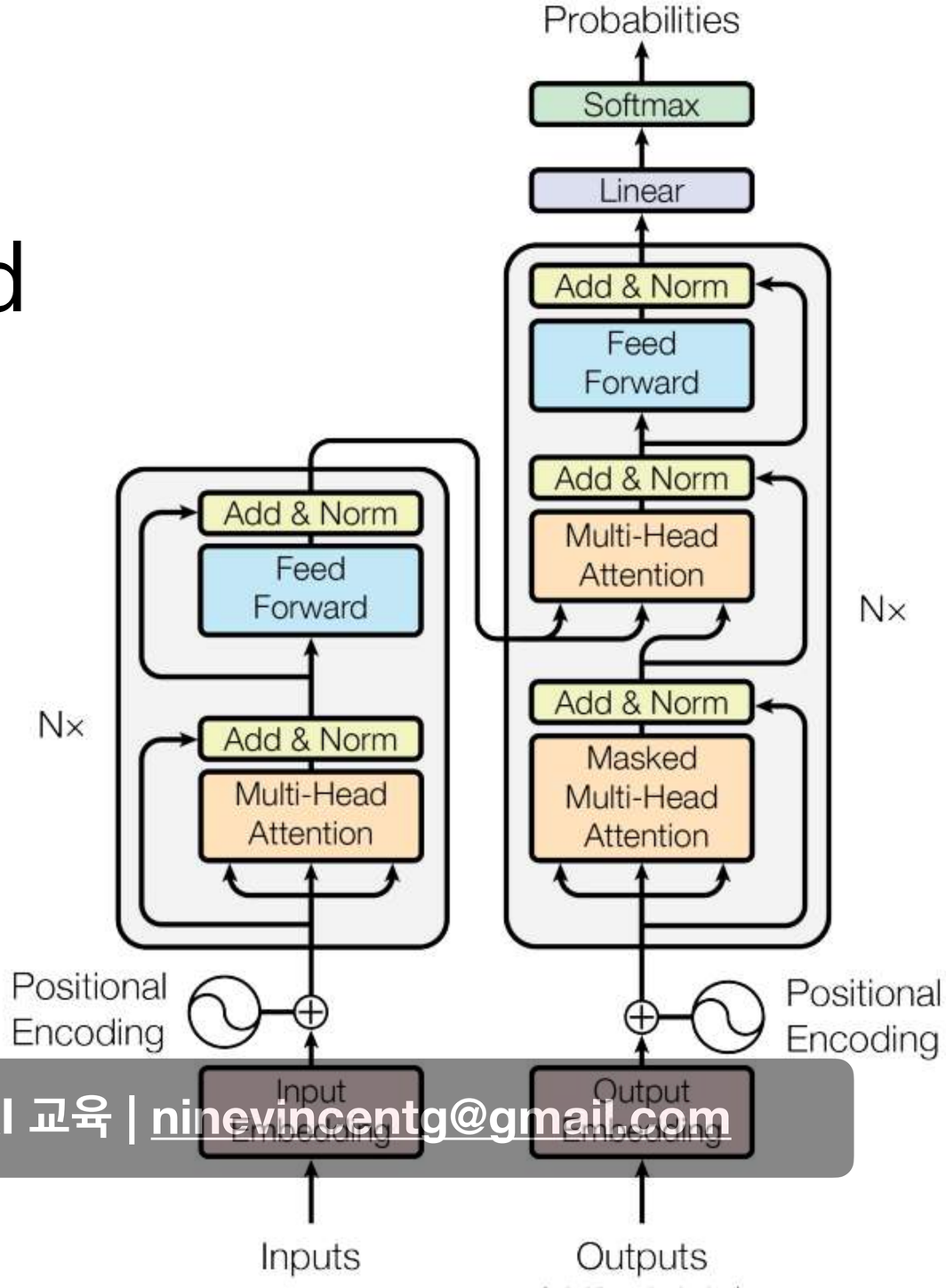


https://pub.towardsai.net/diffusion-models-vs-gans-vs-vaes-comparison-of-deep-generative-models-67ab93e0d9ae

2017년

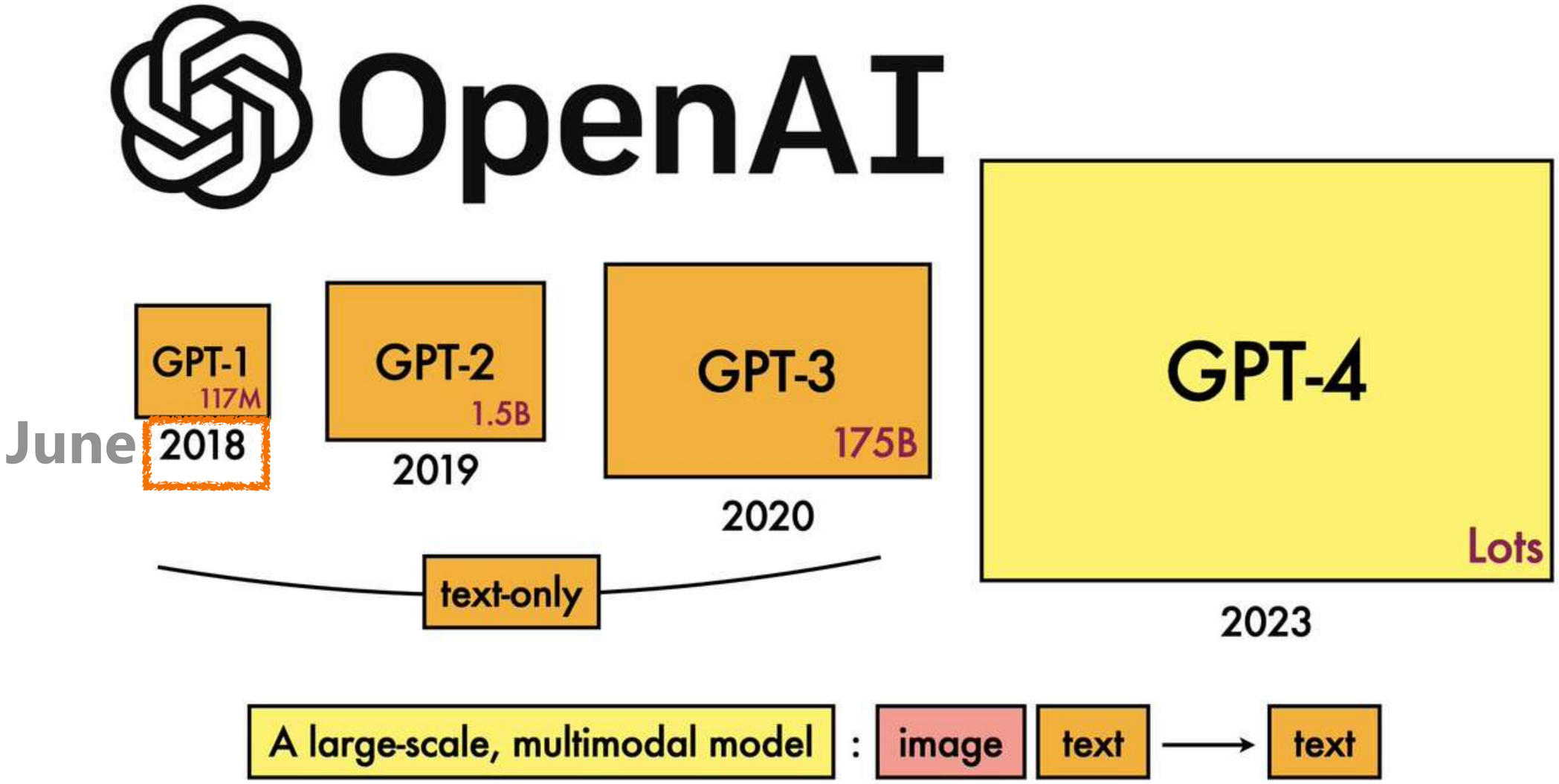
Attention is All You Need

Transformer Network



2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincentg@gmail.com

https://arxiv.org/abs/1706.03762



출처 : https://twitter.com/SamuelAlbanie/status/1636464595932946432

CLM

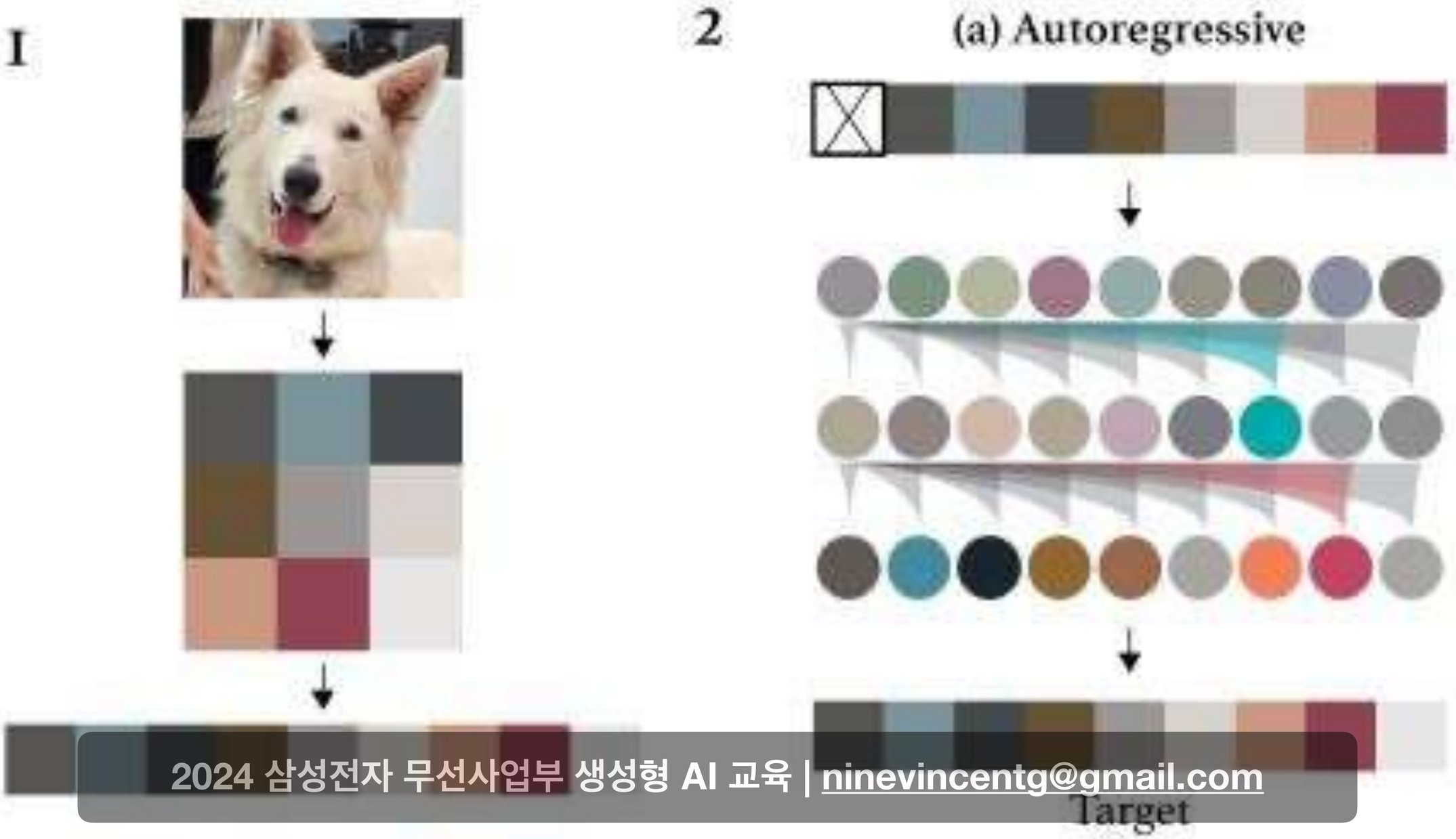
Causal Language Model

Auto Regressive Model

My				→	name
My	name			→	is
My	name	is		→	Sylvain
My	name	is	Sylvain	→	.

2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincentg@gmail.com

출처 : https://huggingface.co/learn/nlp-course/chapter1/4

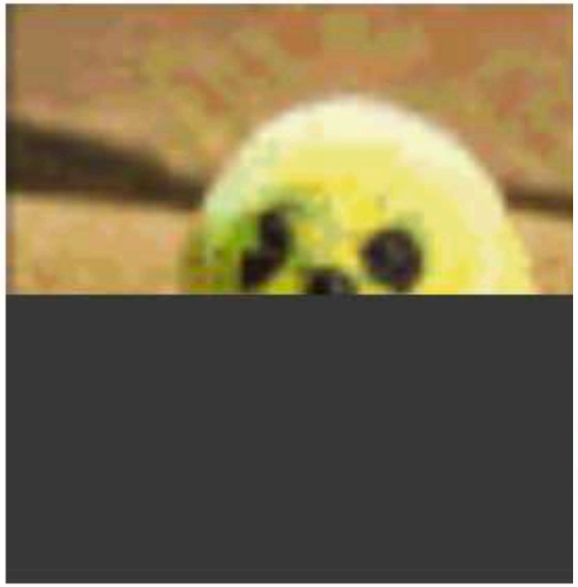


2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincentg@gmail.com

Chen et al., Generative Pretraining from Pixels, 2020

Image GPT

Input



Completions

2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincentg@gmail.com

https://openai.com/research/image-gpt

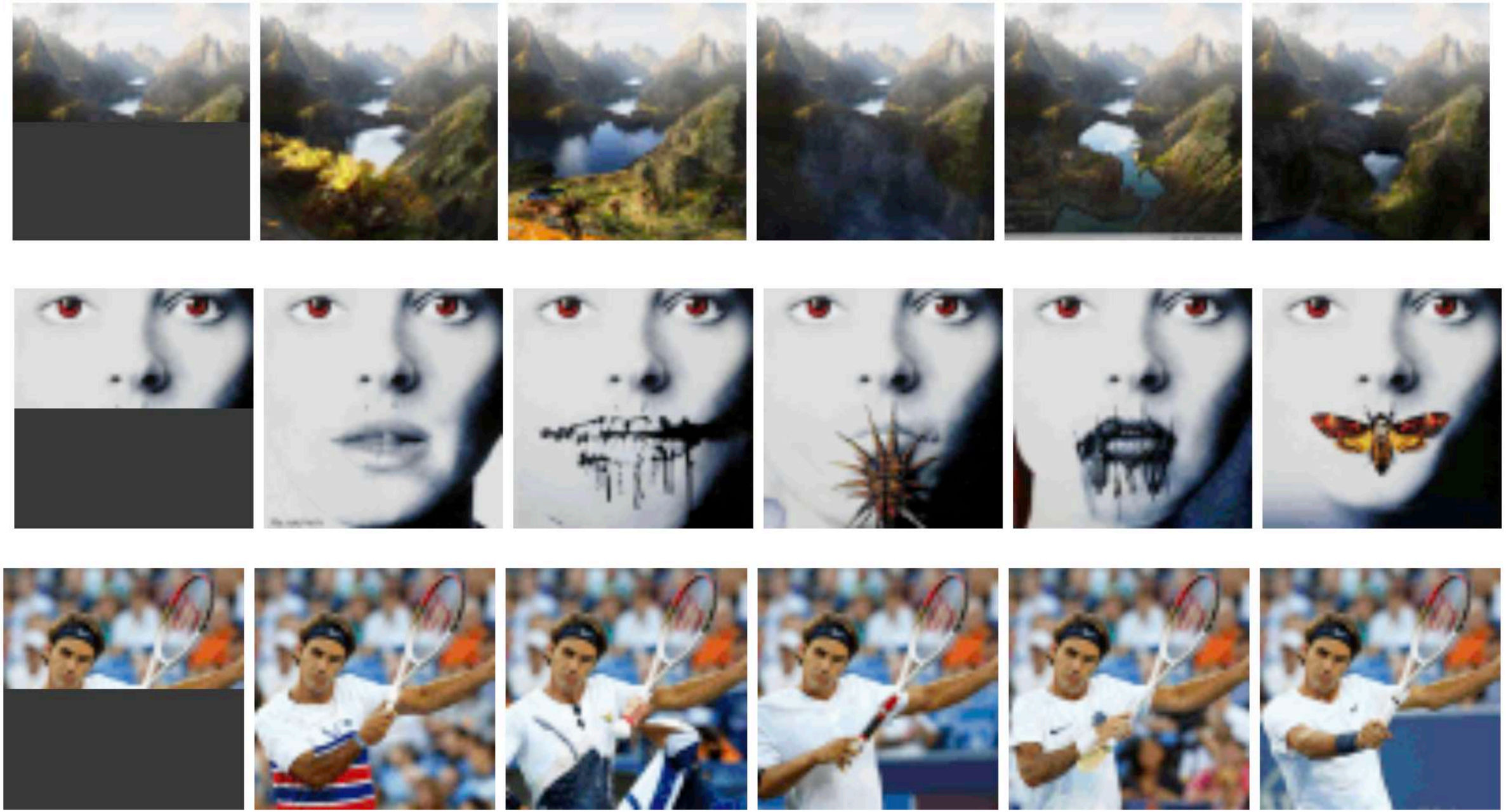
Image GPT

Input

Completions

2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincentg@gmail.com

https://openai.com/research/image-gpt



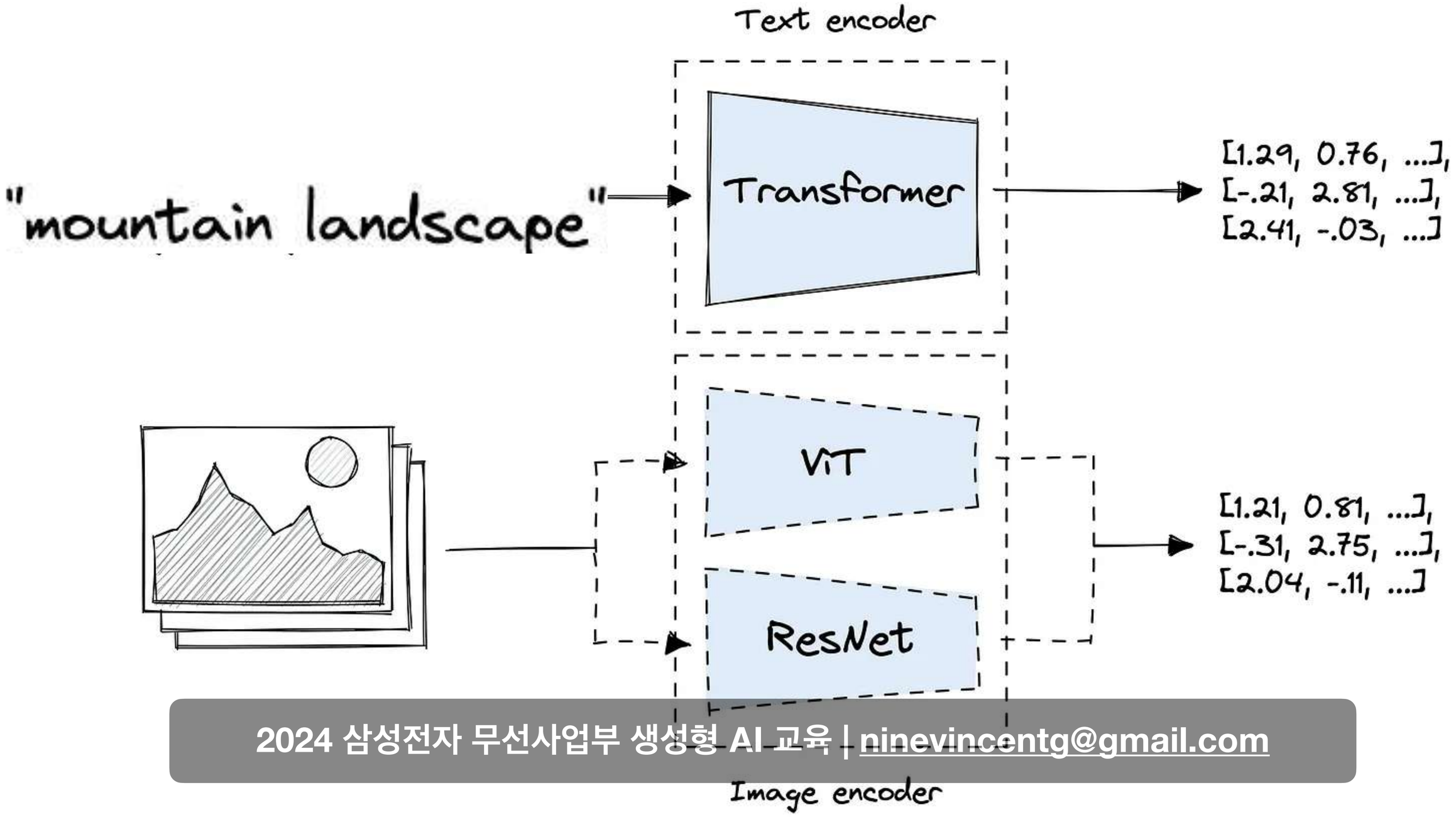
https://openai.com/research/image-gpt

2021

CLIP Model

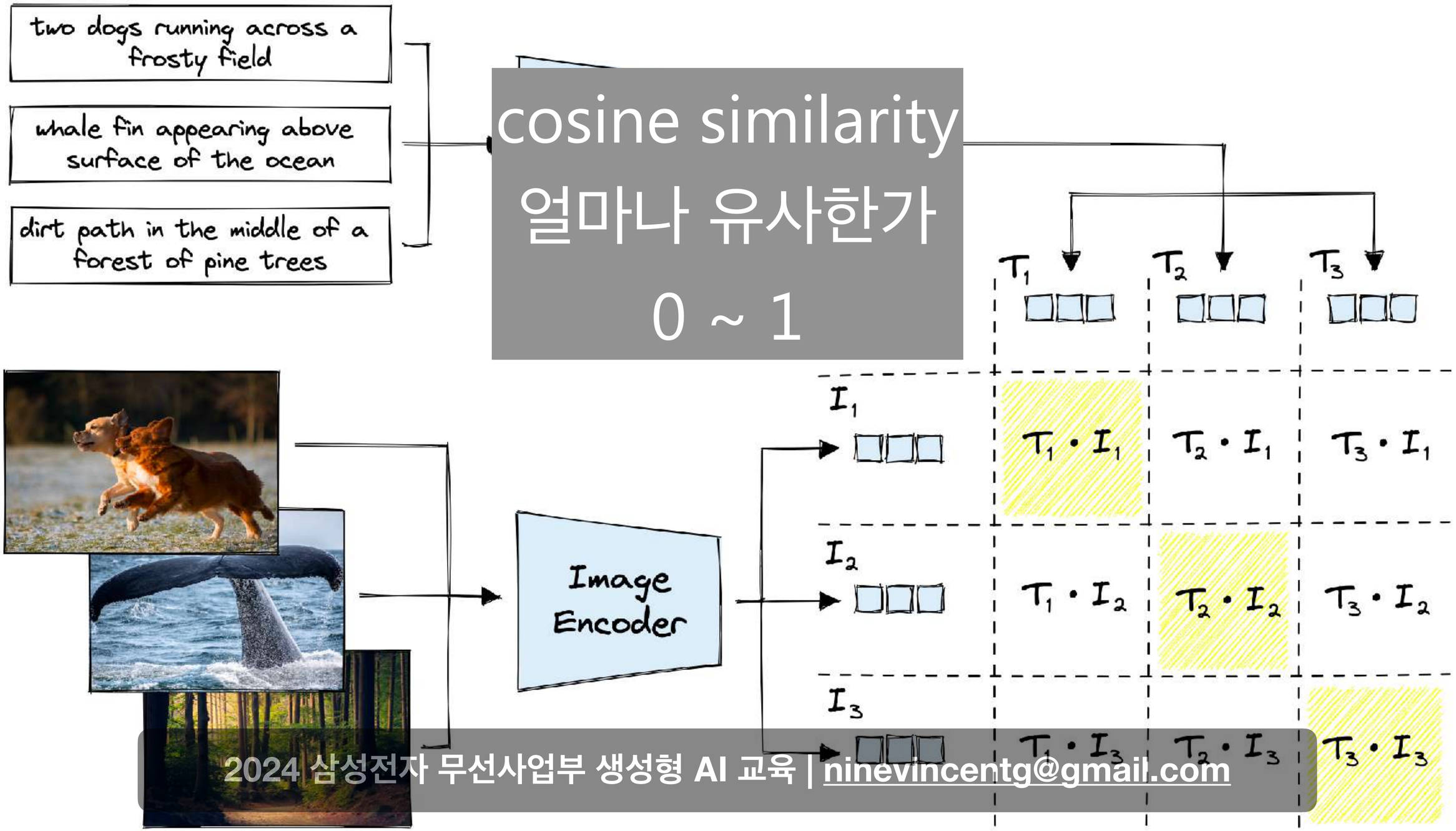
Contrastive **L**earning **I**n **P**retraining

2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincentg@gmail.com



2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincentg@gmail.com

https://www.pinecone.io/learn/series/image-search/clip/



https://www.pinecone.io/learn/series/image-search/clip/

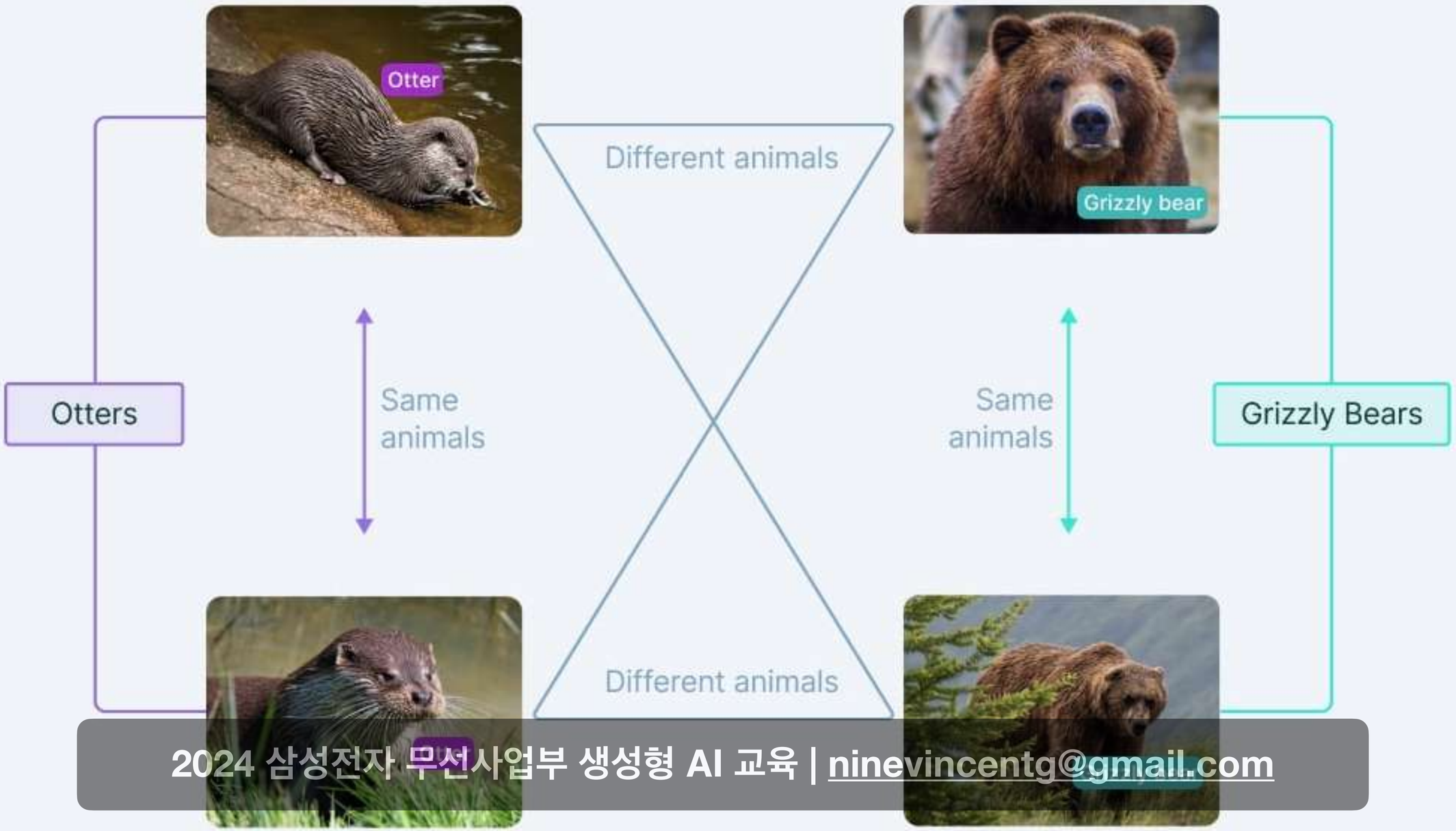
Machine Learning

Supervised Learning(지도학습)

Self Supervised Learning

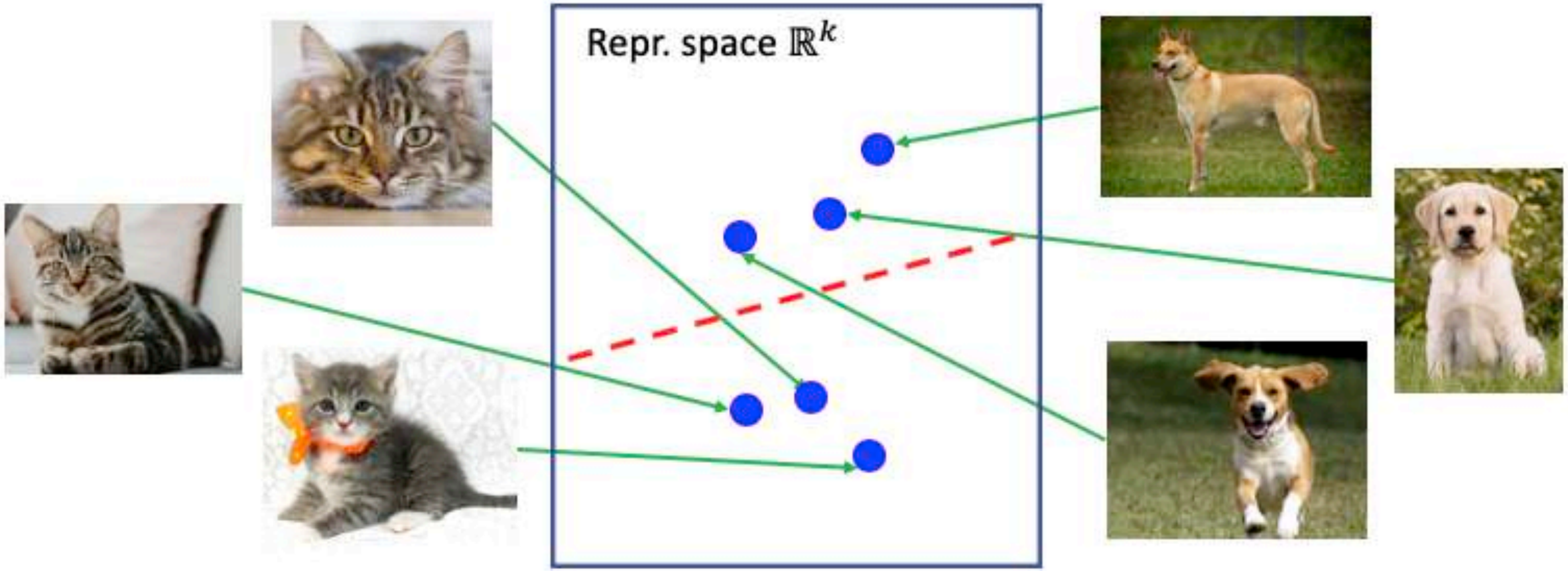
Unsupervised Learning(비지도학습)

Reinforcement Learning(강화학습)



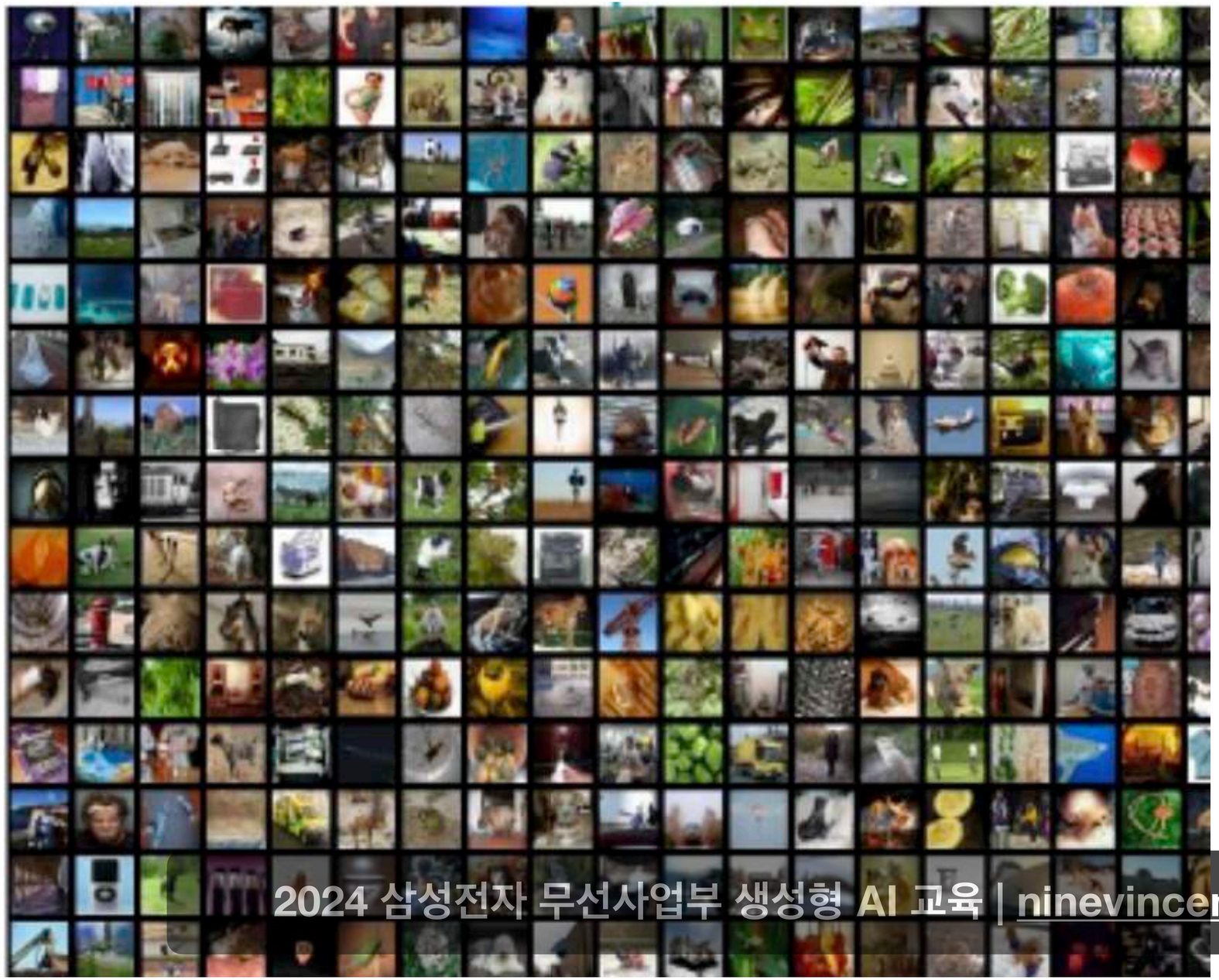
2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincentg@gmail.com

https://www.v7labs.com/blog/contrastive-learning-guide

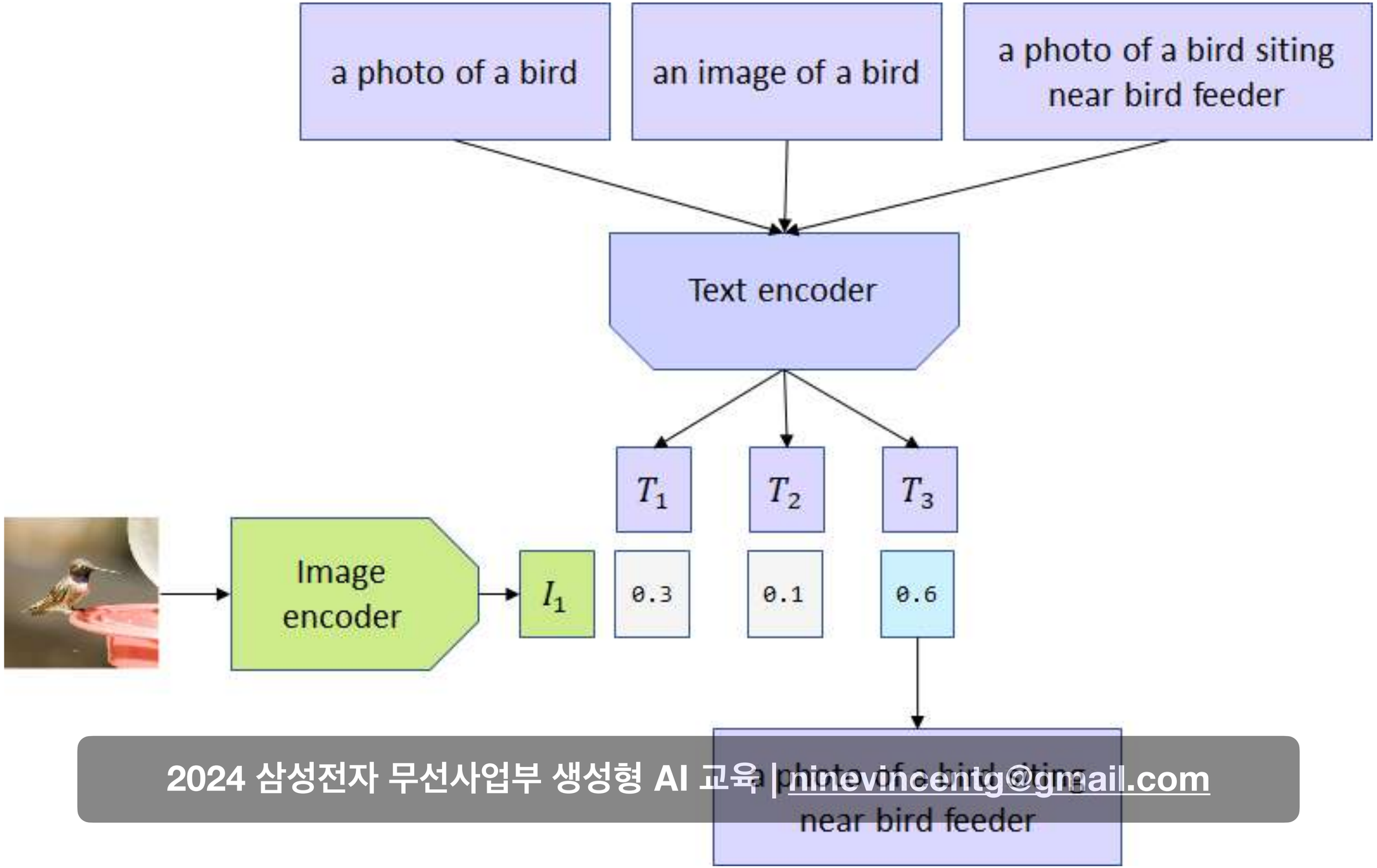


2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincentg@gmail.com

https://ai.stanford.edu/blog/understanding-contrastive-learning/



400 Million Images



2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincentg@gmail.com

Generative Art AI

Product	Public Release	Company
DALL-E	Jan 2021	OpenAI
Midjourney V1	Feb 2022	
Imagen	March 2022	Google
Midjourney V3	Jul 2022	
Stable Diffusion	Aug 2022	Stability AI
FireFly	June 2023	Adobe
Midjourney V4	Nov 2022	Midjourney, Inc.
Midjourney V5	Mar 2023	

col_jung Medium Col Jung

2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincentg@gmail.com

https://generativeai.pub/generative-ai-art-the-road-to-dall-e-midjourney-stable-diffusion-3b3219d97f02

DALL.E3

- OpenAI
- Bing Create

Midjourney

- Midjourney Lab
- $10~$120/month

StableDiffusion

- Stability AI
- Open Source

FireFly

- Adobe
- 25 Free Credits

Imagen

Midjourney

http://discord.com

DALLE.3

http://bing.com/create

Stable Diffusion

https://replicate.com/stability-ai

Stable Diffusion API

https://platform.stability.ai/

Leonardo.ai

https://leonardo.ai/

식료품 쇼핑하는 테디베어, 컬러 펜슬 이미지

By. DALL-E

디제잉 헤드폰을 쓰고있는 미켈란젤로의 다비드상

By. Stable Diffusion

푹신한 구름 위에서 기타를 치는 흰색 고양이

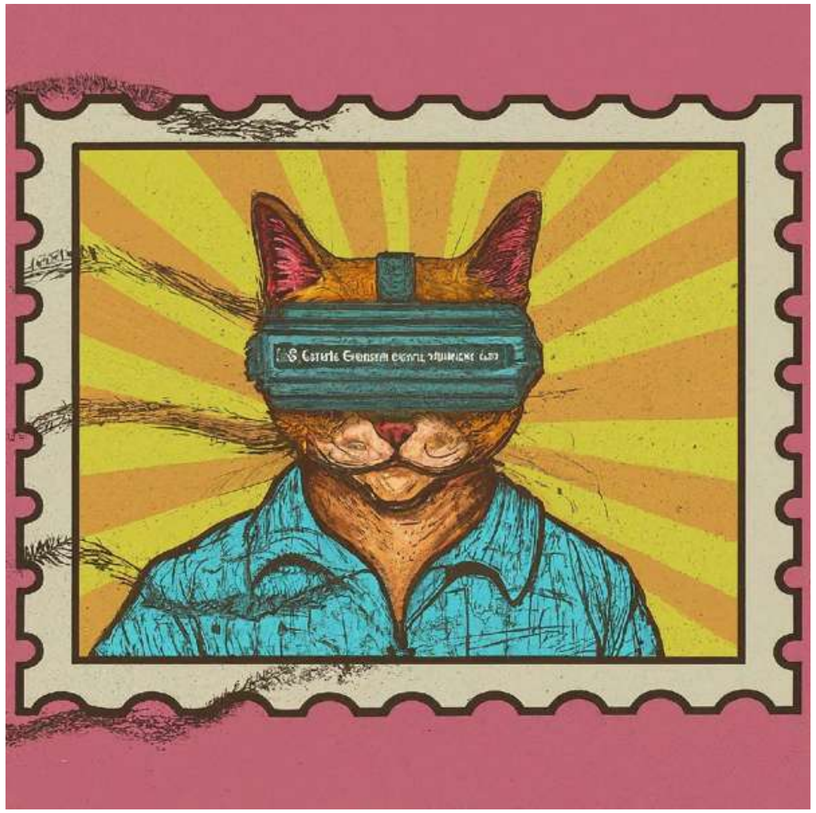
By. Midjourney

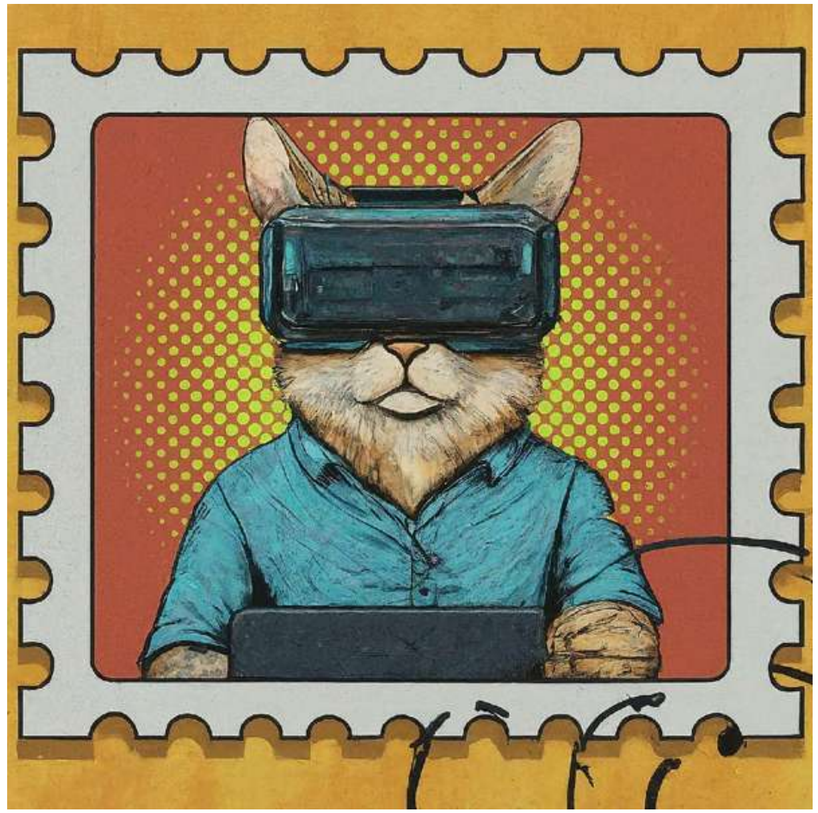
숲 속 오두막 집 앞에서 크리스마스 파티를 하는 곰, 다람쥐, 늑대

By. FireFly

2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincentg@gmail.com

VR 헤드셋을 쓰고있는 고양이 프로그래머, 우표 스타일

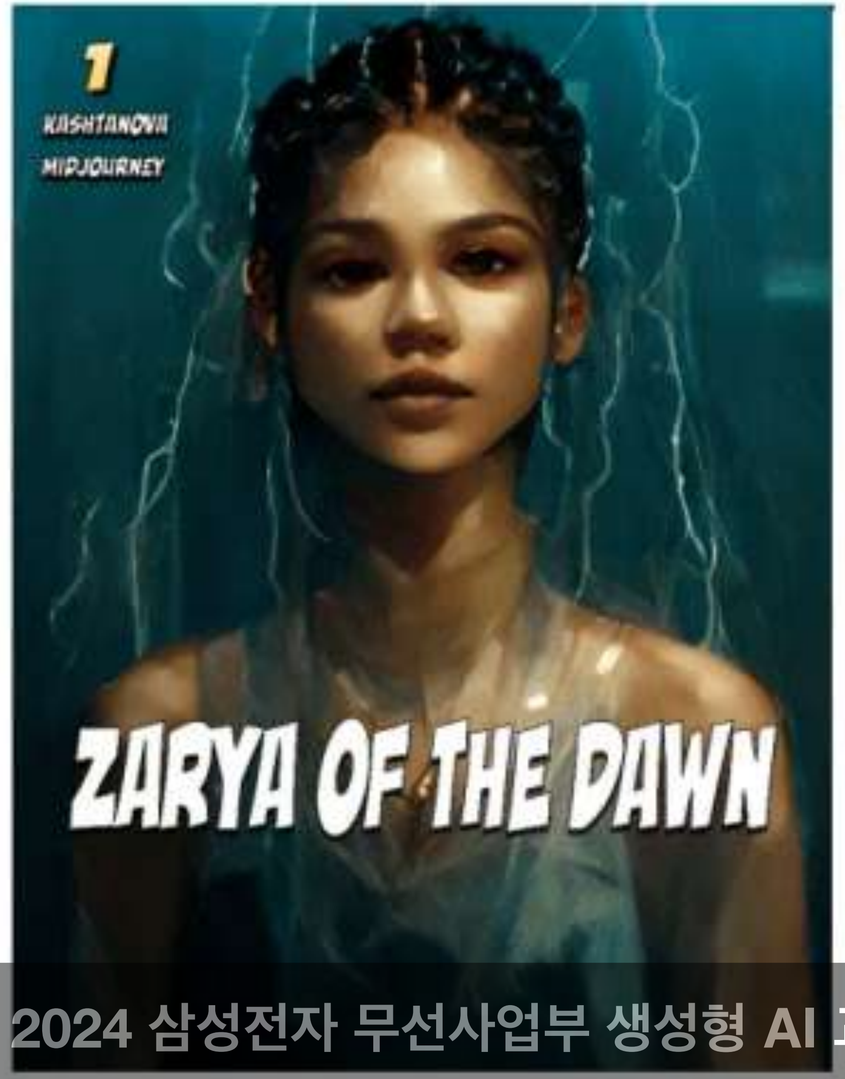




By. Imagen

2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincentg@gmail.com

크리스티나 카슈타노바의 인터넷 소설 ‘새벽의 자리아’ 이미지



https://www.sedaily.com/NewsView/29S2CUT239

DALL·E

https://www.theartstory.org/artist/dali-salvador/

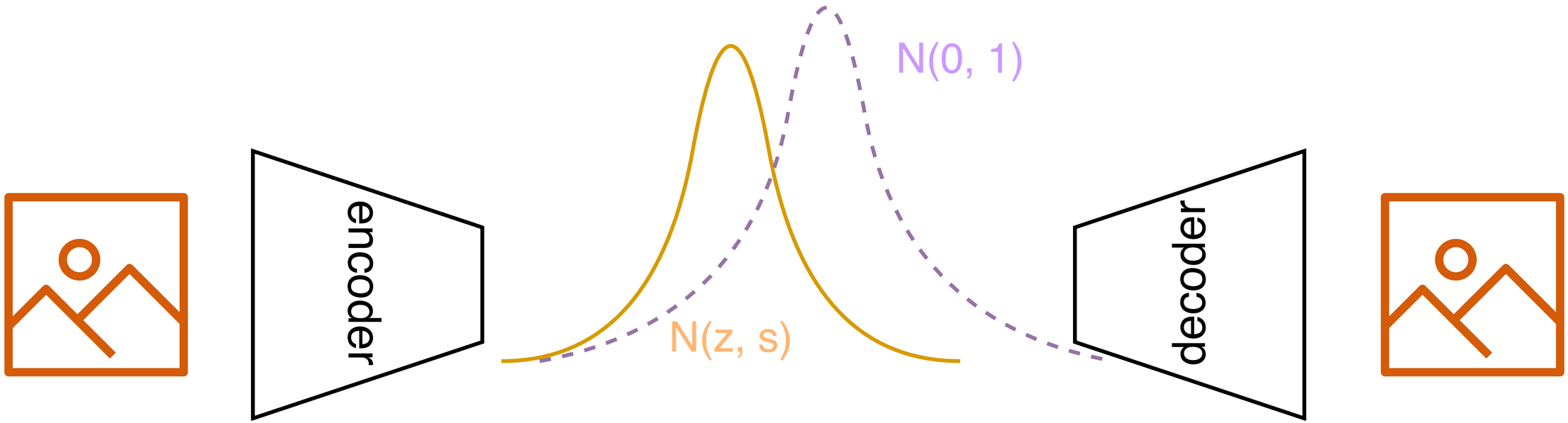
Photo by James A. Molnar on Unsplash

DALL·E 1 2021년 1월

DALL·E 2 2022년 4월

DALL·E 3 2023년 9월

2013 VAE(Variational Autoencoder)



https://vaclavkosar.com/ml/openai-dall-e-2-and-dall-e-1

2021년 1월

an illustration of a baby daikon radish in a tutu walking a dog

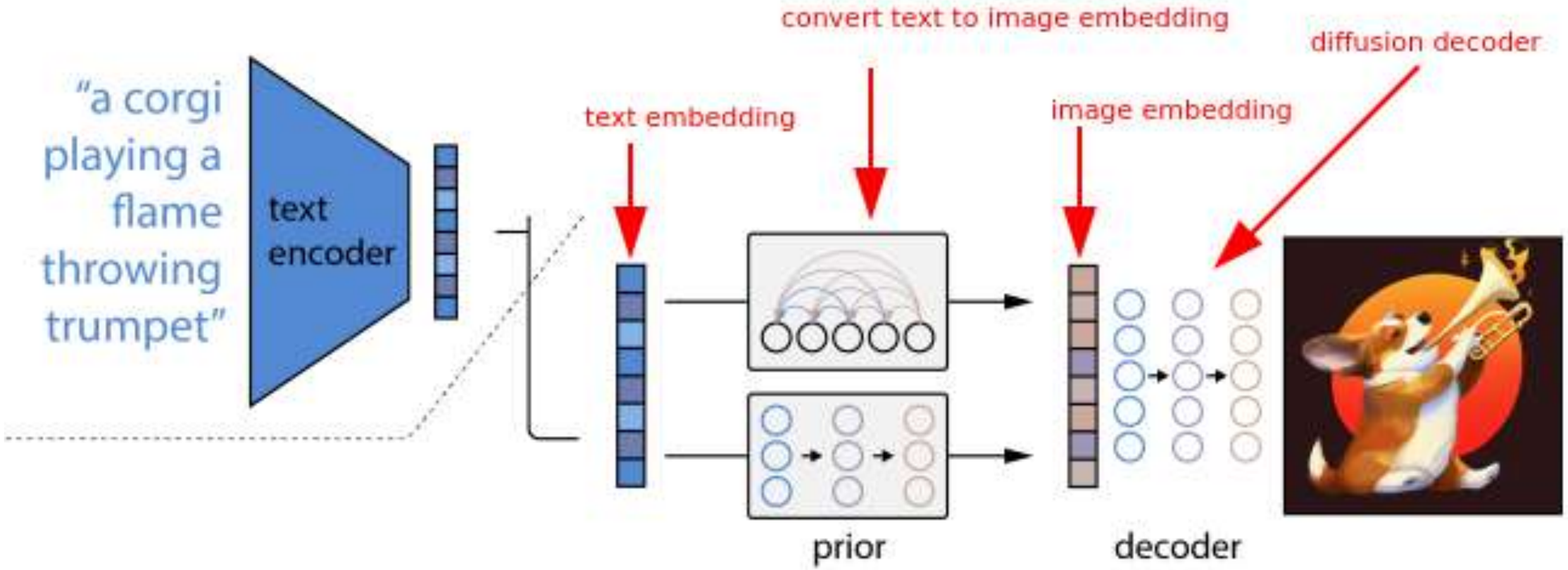
AI Generated images

an armchair in the shape of an avocado. . . .

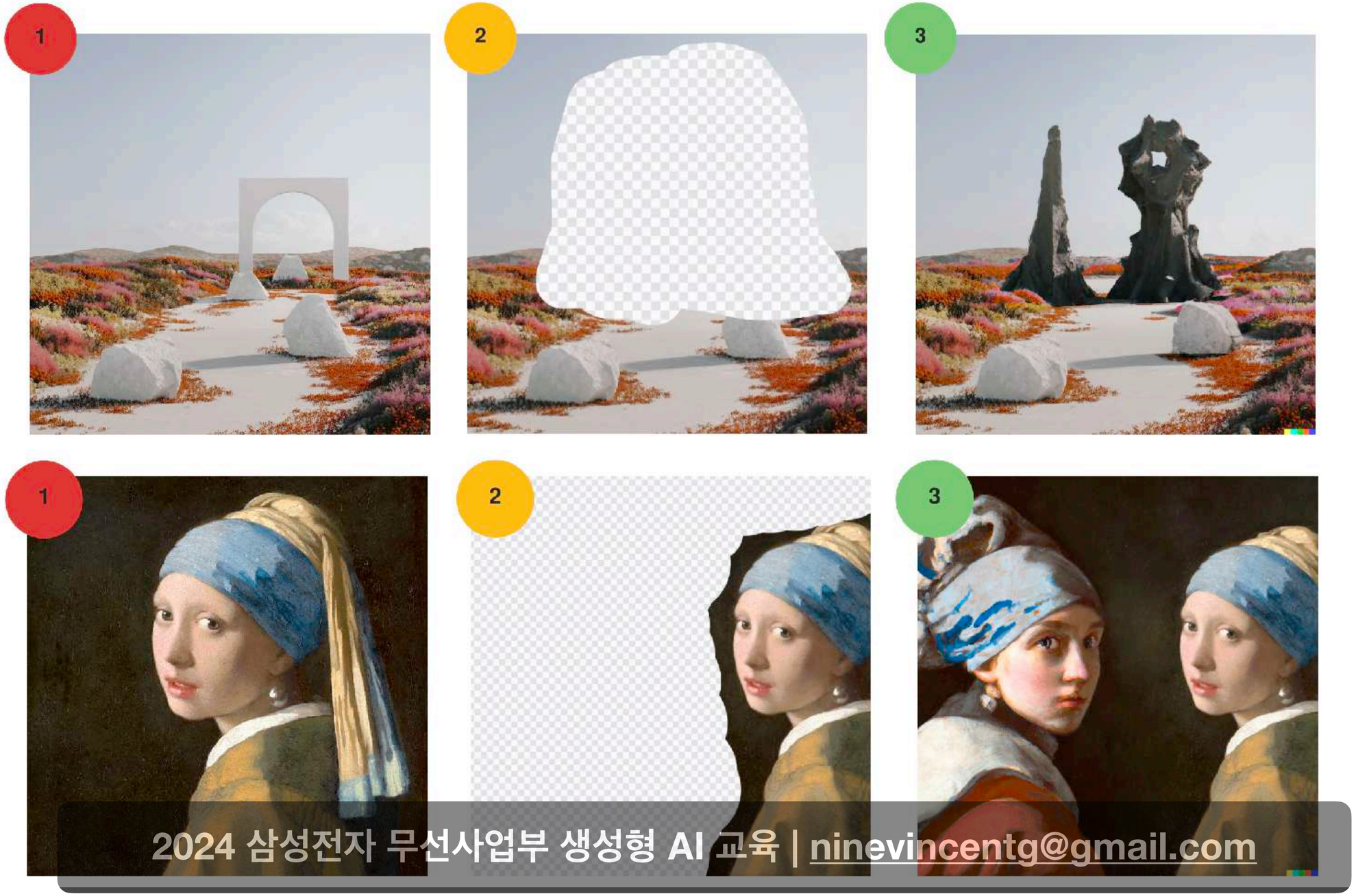
AI Generated images

https://openai.com/research/dall-e

2022년 4월



https://vaclavkosar.com/ml/openai-dall-e-2-and-dall-e-1



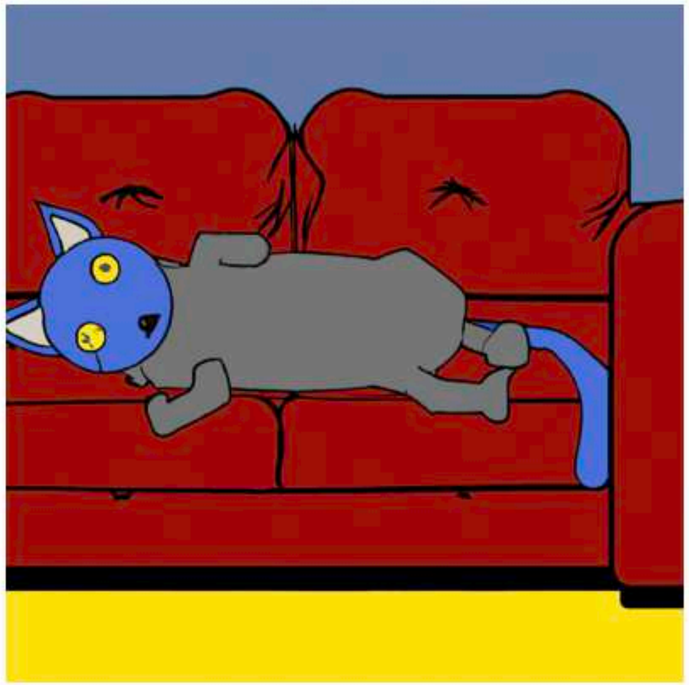
2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincentg@gmail.com

https://dmnfarrell.github.io/general/dalle2-aiart

Russian blue cat on a sofa,
Pop art Andy Warhol

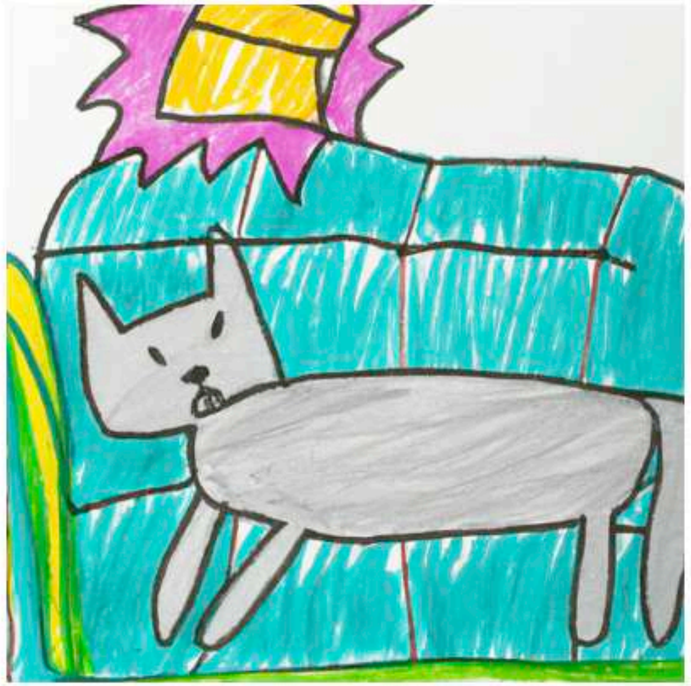
Russian blue cat on a sofa

Russian blue cat on a sofa,
Rococo style painting



Russian blue cat on a sofa, in the style of The Simpsons

Russian blue cat on a sofa, oil painting



Russian blue cat on a sofa, children hand drawing

“prompt following” problem

prompt의 단어를 빼먹는 경우

prompt 내 단어의 순서를 혼동하는 경우

단어의 의미를 잘 못 이해하는 경우

i

t

Russian Blue Postcard

i

t

blue eyes and cute kitten image

1. 배경 물체의 존재 및 해당 물체에 대한 설명

2. 물체의 위치와 해당 물체의 수

3. 물체의 색상 및 크기와 같은 세부적인 사항

4. 이미지에 표시된 텍스트

LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs
https://arxiv.org/pdf/2111.02114.pdf

Human *Captioner*

AI Image *Captioner*

Image			
Alt Text	now at victorian plumbing.co.uk	is he finished...just about!	23 (19 of 30) 1200
SSC	a white modern bathtub sits on a wooden floor.	a quilt with an iron on it.	a jar of rhubarb liqueur sitting on a pebble background.
DSC	this luxurious bathroom features a modern freestanding bathtub in a crisp white finish. the tub sits against a wooden accent wall with glass-like panels, creating a serene and relaxing ambiance. three pendant light fixtures hang above the tub, adding a touch of sophistication. a large window with a wooden panel provides natural light, while a potted plant adds a touch of greenery. the freestanding bathtub stands out as a statement piece in this contemporary bathroom.	a quilt is laid out on a ironing board with an iron resting on top. the quilt has a patchwork design with pastel-colored strips of fabric and floral patterns. the iron is turned on and the tip is resting on top of one of the strips. the quilt appears to be in the process of being pressed, as the steam from the iron is visible on the surface. the quilt has a vintage feel and the colors are yellow, blue, and white, giving it an antique look.	rhubarb pieces in a glass jar, waiting to be pickled. the colors of the rhubarb range from bright red to pale green, creating a beautiful contrast. the jar is sitting on a gravel background, giving a rustic feel to the image.

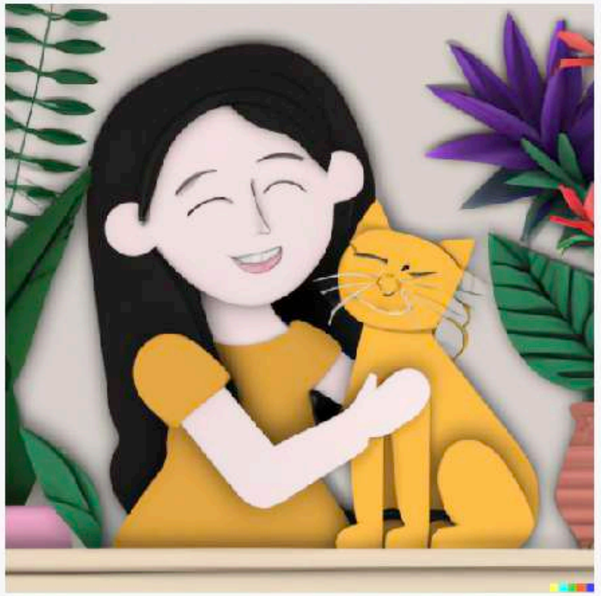
2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincentg@gmail.com

Improving Image Generation with Better Captions
https://cdn.openai.com/papers/dall-e-3.pdf

“고양이를 부드럽게 안고 있는 소녀를 묘사한 종이 공예 작품.
둘 다 식물 화분 사이에 앉아 있으며,
고양이는 만족하며 코를 들이대고 소녀는 미소를 짓는다.
수제 종이 꽃과 잎으로 장식되어 있는 장면”

“Paper craft depicting a girl gently holding a cat.
Both are sitting between pots of plants,
The cat nuzzles in contentedly and the girl smiles.
Scene decorated with handmade paper flowers and leaves”

DALL·E 2



https://labs.openai.com/

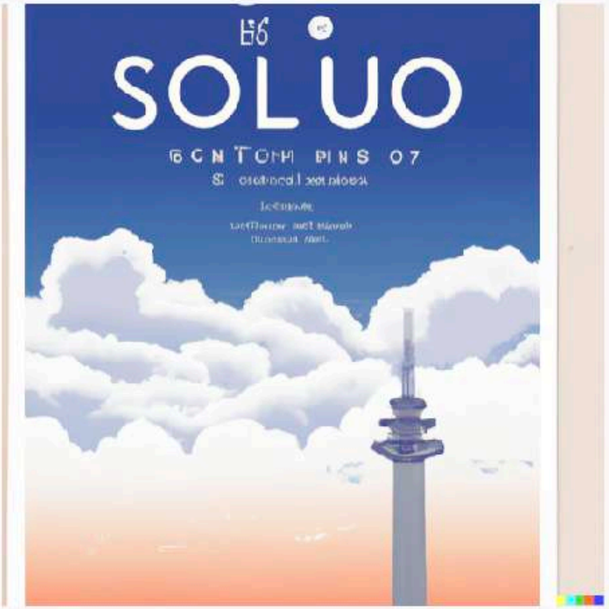
DALL·E 3

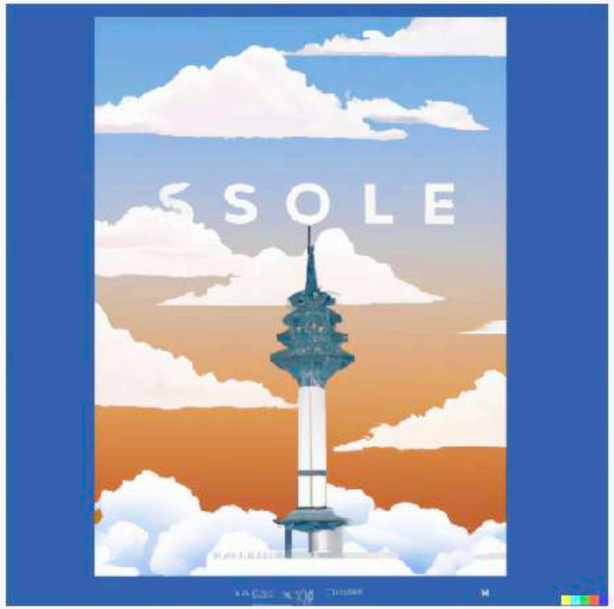
https://www.bing.com/create

서울 여행 세로 포스터.
남산타워 주변의 흰 구름과 경복궁의 실루엣을 묘사하고 있다.
포스터 하단에는 ‘Seoul Travel: Beauty of Korea’이 적혀있다.
여행에 관심을 가질 수 있도록
색상은 흰색, 파란색, 부드러운 오렌지색으로 구성한다.

“Vertical Seoul travel poster. This scene depicts white clouds around Namsan Tower and the silhouette of Gyeongbokgung Palace. At the bottom of the post, the text reads ‘Seoul Travel: Beauty of Korea.’ The color scheme consists of white, blue and soft orange to create travel interest”

DALL·E 2





https://labs.openai.com/

DALL·E 3

https://www.bing.com/create

A photography of _____, close up

An image of _____, Van Gogh style

A 3D illustrate of _____, bright mood

A poster of _____, with orange color

이미지 모델 + 언어 모델

ImageFX

white cat is dancing on the top of snowed mountain with rainbow colored fish,

Start over

Generate

More

abstract

scifi

baroque

film still

cinematic

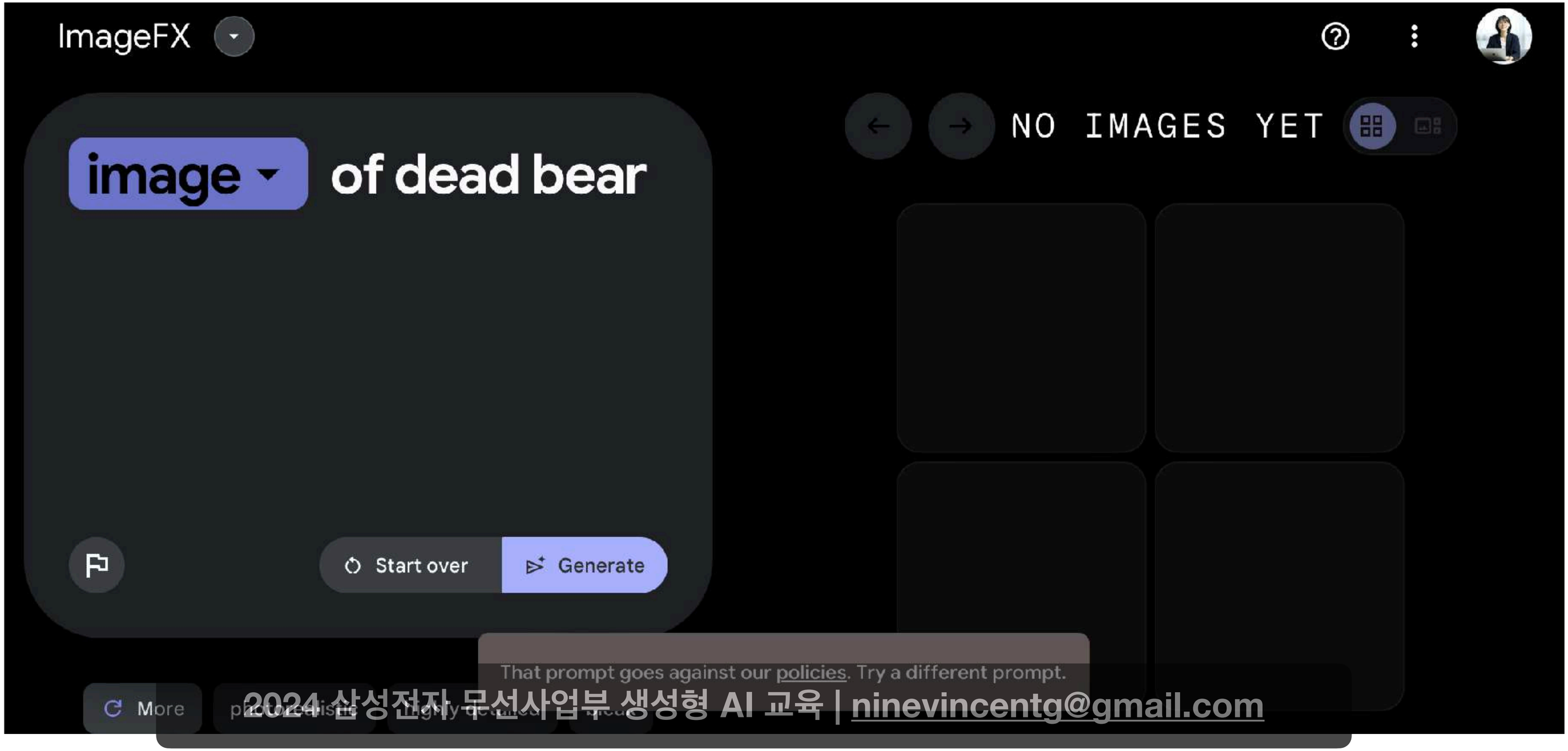
close up

natural light

IMAGE SET 2/2

```python
result_dalle3 = client_aoai.images.generate(
    model="dall-e-3",
    prompt="image of bleeding bear",
    size="1024x1024",
    quality='standard',
    n=1
)
```

```
BadRequestError: Error code: 400 - {'created': 1707666184, 'error': {'code':
'contentFilter', 'message': 'Your task failed as a result of our safety
system.'}}
```



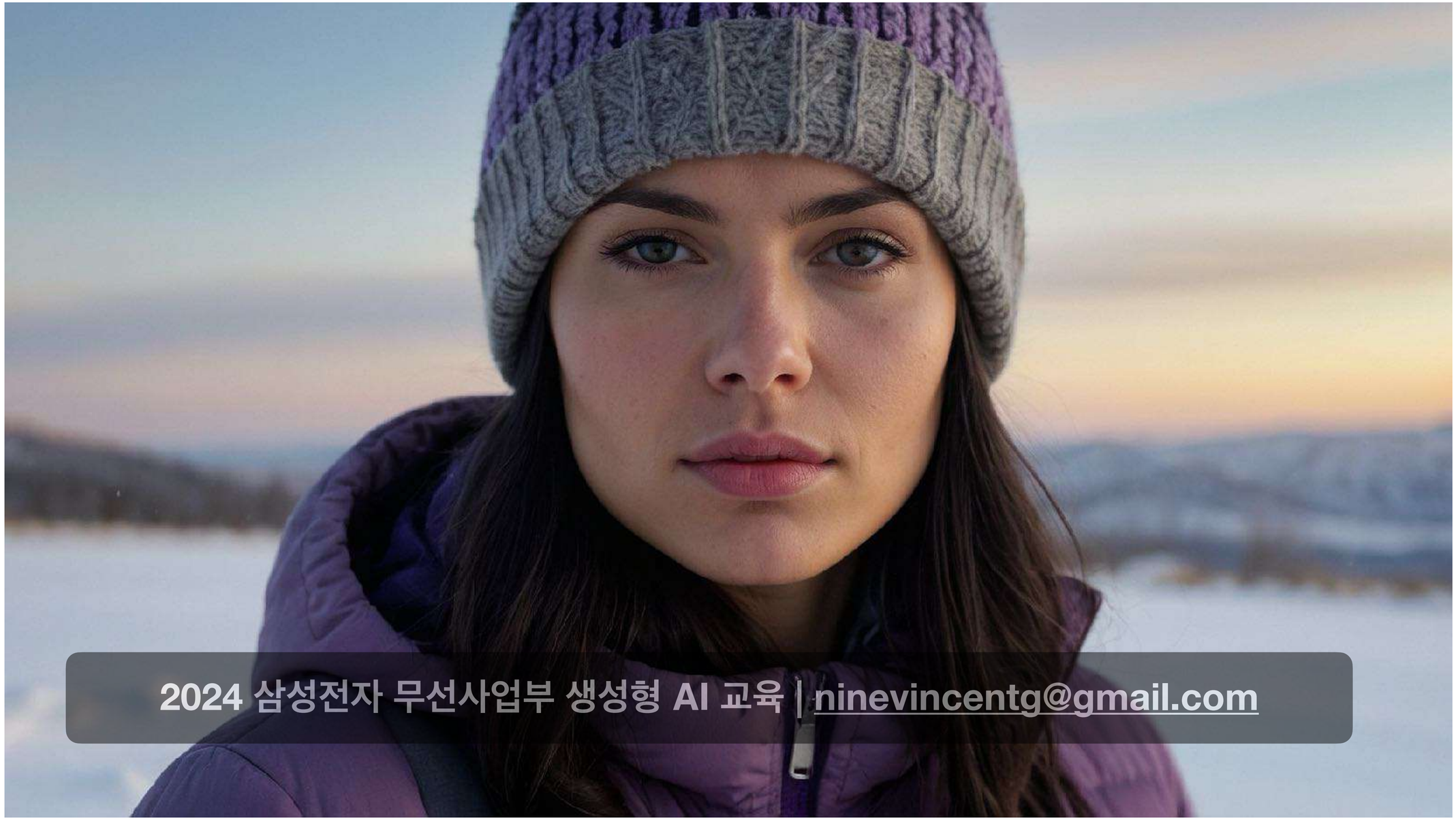
2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincentg@gmail.com

Dall.E

Imagen

Midjourney

StableDiffusion



Leonardo.Ai

Delete Dataset

Dataset Name

Winnie the Pooh

Dataset Description

This is my new dataset. I should change the description to describe the contents of this collection!

Dataset Images

Remove Images | Upload Images | Train Model

12/40 images | Cancel selection

a bear and a pig. they are jumping and playing with happy face in the forest

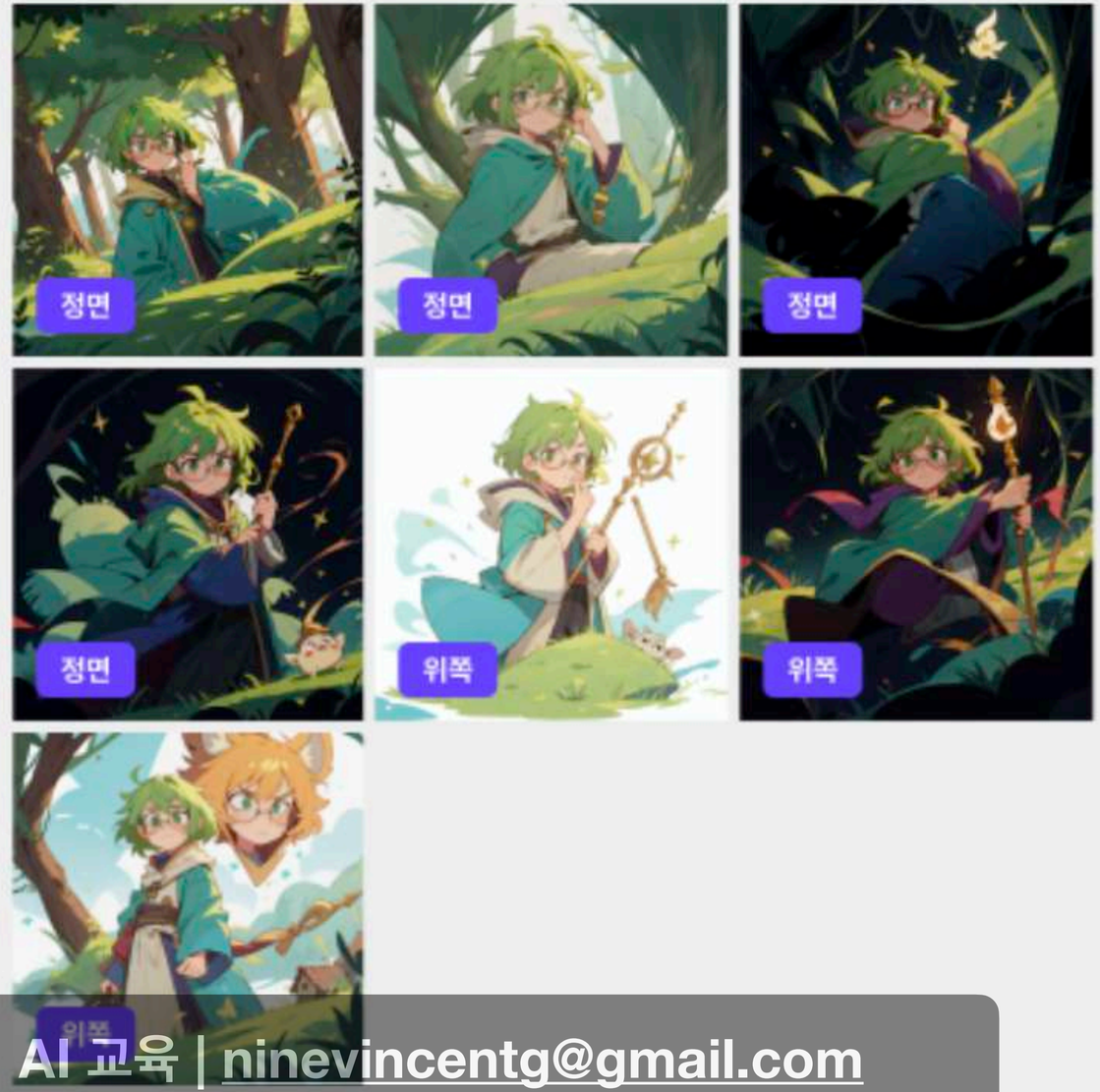
Winnie the Pooh 4 768 x 512px

스타트업 라이언 로켓

LIØN
RØCKET
MØRE THAN AI
Beyond your imagination.
2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincentg@gmail.com

75%

캐릭터를 등록해주세요



2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincentg@gmail.com

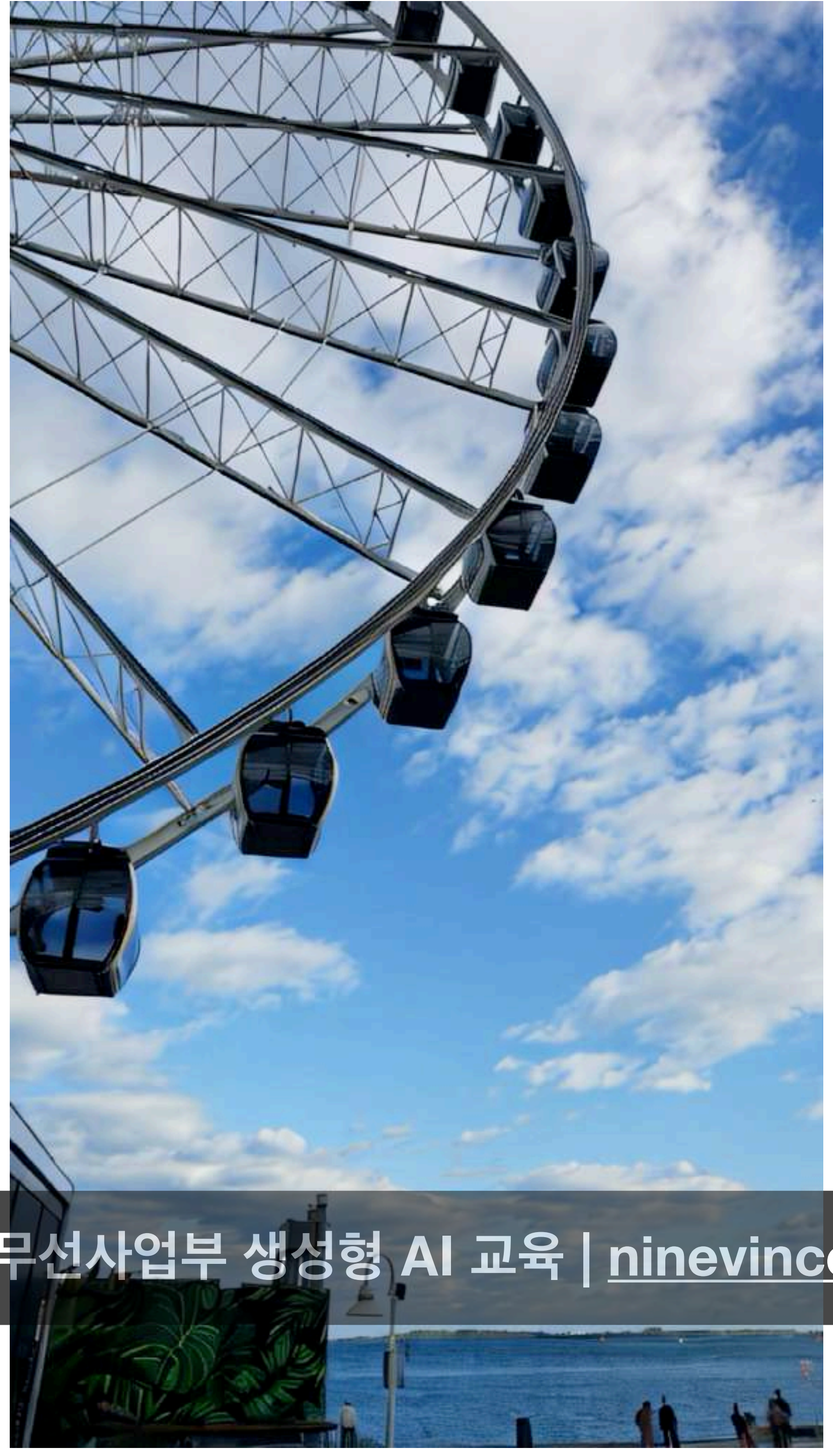
https://www.lionrocket.ai/

Stable Video

2024.02.08. Stable Video Diffusion 1.1

2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincentg@gmail.com

Stable Video



2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincentg@gmail.com

Research

Video generation models as world simulators

We explore large-scale training of generative models on video data. Specifically, we train text-conditional diffusion models jointly on videos and images of variable durations, resolutions and aspect ratios. We leverage a transformer architecture that operates on spacetime patches of video and image latent codes. Our largest model, Sora, is capable of generating a minute of high fidelity video. Our results suggest that scaling video generation models is a promising path towards building general purpose simulators of the physical world.

February 15, 2024

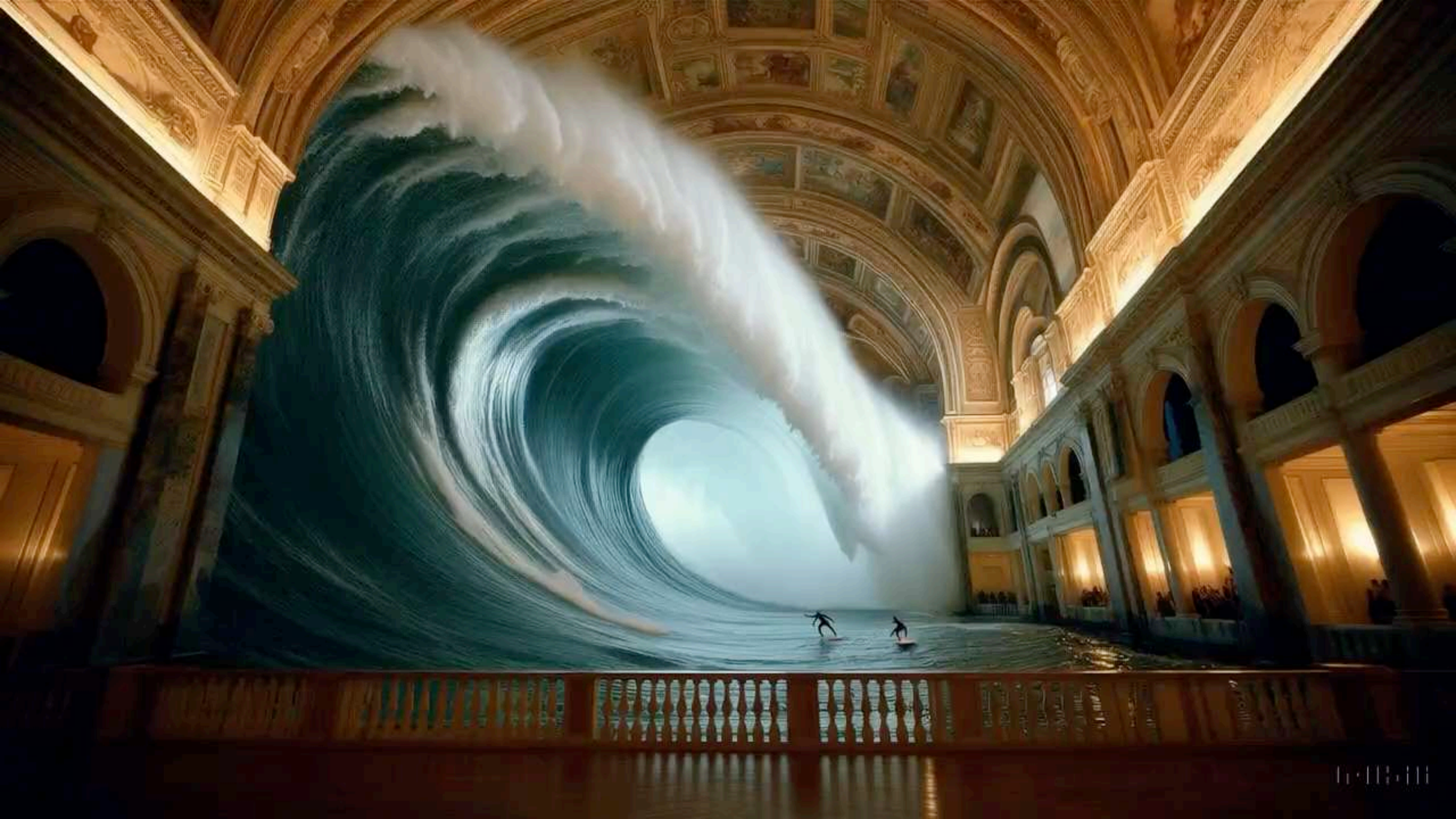
View Sora overview

Video generation, Sora, Milestone, Release

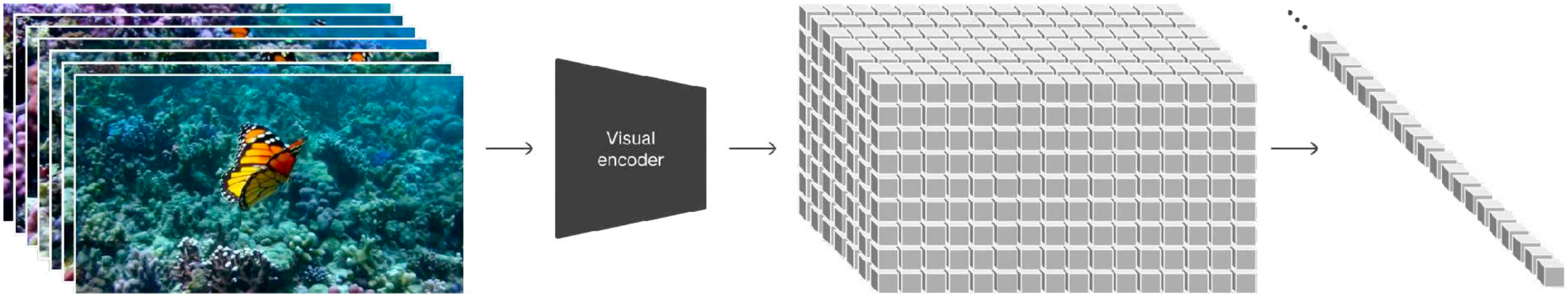
“아름다운 눈 덮인 도쿄 도시는 분주합니다.
카메라는 번화한 도시 거리를 따라 이동하며
아름다운 눈 덮인 날씨를 즐기고 근처 노점에서
쇼핑하는 여러 사람을 따라갑니다.
화려한 벚꽃잎이 눈꽃과 함께 바람을 타고 날아갑니다.”

“Beautiful, snowy Tokyo city is bustling. The camera moves through the bustling city street, following several people enjoying the beautiful snowy weather and shopping at nearby stalls. Gorgeous sakura petals are flying through the wind along with snowflakes.”

0:00 / 0:17



Diffusion Transformer Model



https://openai.com/research/video-generation-models-as-world-simulators

Tyler Perry halts $800m studio expansion after being shocked by AI

US film and TV mogul says he has paused his plans, having seen demonstrations of OpenAI video generator

ail.com

@donalleniii
MADE WITH | SORA
episode 1 unveils the Girafflamingo

Midjourney

http://discord.com

DALLE.3

http://bing.com/create

Stable Diffusion

https://replicate.com/stability-ai

Stable Diffusion API

https://platform.stability.ai/

Leonardo.ai

https://leonardo.ai/

저작권은?

“저작물”이란 **인간**의 사상 또는 감정을 표현한 창작물을 말합니다

(「**저작권법**」 제2조제1호)

[단독] 클릭 두 번에 뚝딱…작곡 AI '이봄' 저작료 중단

2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincentg@gmail.com

https://news.sbs.co.kr/news/endPage.do?news_id=N1006933085

2023.08.21.

REUTERS® World Business Markets Sustainability Legal Breakingviews

Litigation | Copyright | Technology | Intellectual Property | Litigation

AI-generated art cannot receive copyrights, US court says

By **Blake Brittain**

August 22, 2023 3:29 AM GMT+9 · Updated a month ago

미국 법원
“AI가 만든 작품 저작권 인정할 수 없다”

entg@gmail.com

https://www.reuters.com/legal/ai-generated-art-cannot-receive-copyrights-us-court-says-2023-08-21/

“because copyright law is limited to
‘original intellectual conceptions of the author”

“저작권법은 '저작자의 독창적인 지적 개념'에 제한된다.”

2020 미국 저작권청

“human authorship is an essential part of a valid copyright claim”

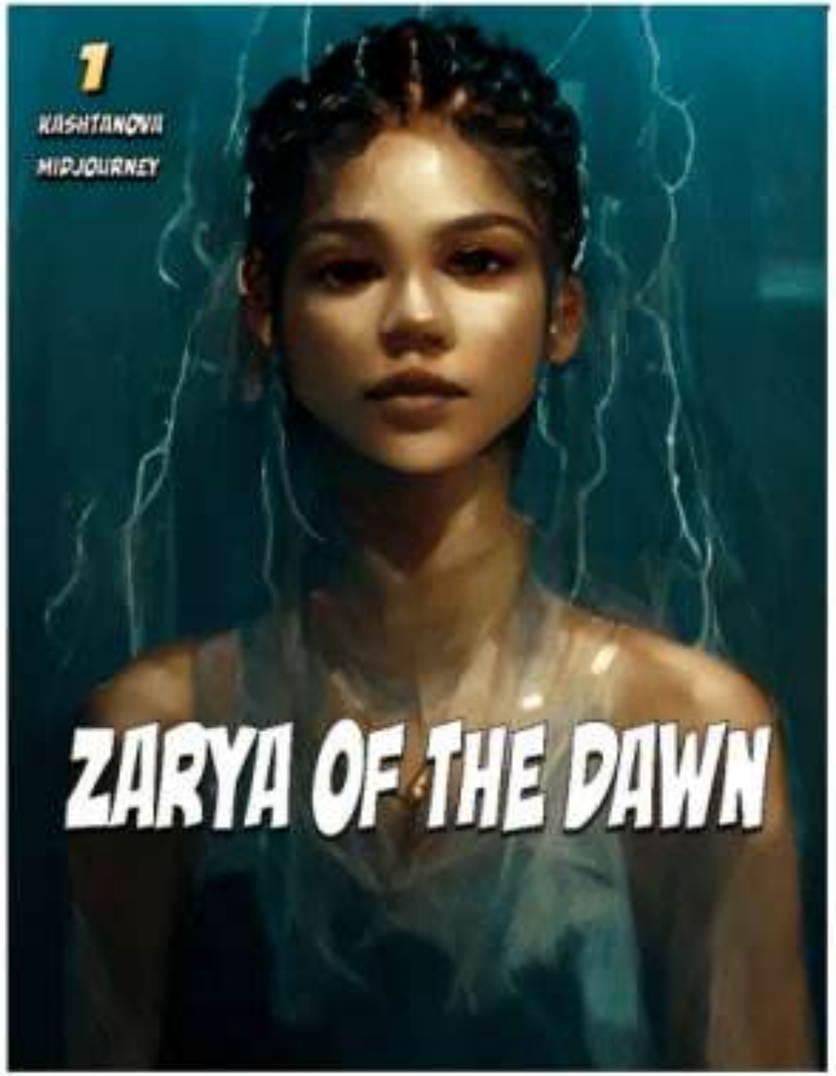
“인간 저작자는 유효한 저작권 주장에 필수적인 부분이다”

2023 미국 연방법원

https://fingfx.thomsonreuters.com/gfx/legaldocs/lbvgooeoqvq/AI COPYRIGHT LAWSUIT thalerdecision.pdf

친절한AI 76강. DALLE.3 / Midjourney / Microsoft365 Copilot / CLOVA X

크리스티나 카슈타노바의 인터넷 소설 ‘새벽의 자리아’ 이미지



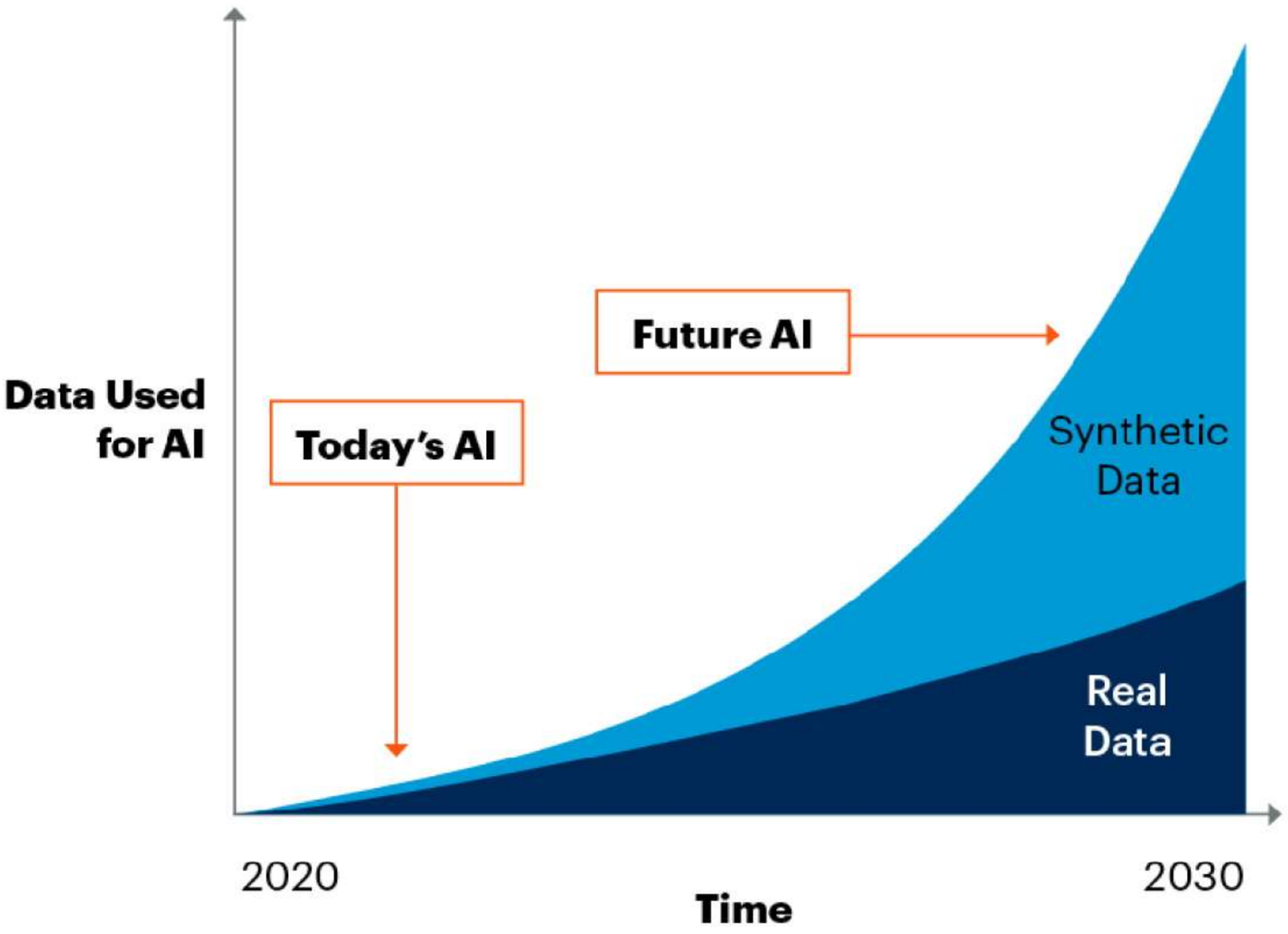
https://www.sedaily.com/NewsView/29S2CUT239

〈 관련 사례 : AI 프로그램이 생성한 그래픽 노블에 대한 저작권 보호 제한('23.2) 〉

- 저작권 등록 거부에 대한 소송을 제기한 Kashtanova는 자신이 Midjourney(AI 프로그램)을 이용하여 '창의적이고 반복적인 과정'을 통해 그림을 저작했다고 주장
- 저작권청은 Kashtanova가 아닌 Midjourney가 '시각적 자료'를 작성한 것으로 결정하여 해당 저작물은 "인간의 창작물이 아니므로" 저작권 보호를 받을 수 없음을 명시
- 이에 대해 Kashtanova가 주장한 것처럼 Midjourney는 저작을 위한 도구로 볼 수 있다는 해석이 제기
- 저작권청은 "Kashtanova가 통제하고 원하는 이미지에 도달하기 위해 안내하는 도구가 아니라 Midjourney가 예측할 수 없는 방식으로 이미지를 생성한다"고 저작권 불인정 이유를 밝힘

[S&T GPS] 미국, 생성형 AI와 저작권법에 관한 보고서 발표

By 2030, Synthetic Data Will Completely Overshadow Real Data in AI Models



- Artificially Generated Data
- Generated From Simple Rules, Statistical Modelling, Simulation and Other Techniques

- Obtained From Direct Measurements
- Constrained by Cost, Logistics, Privacy Reasons

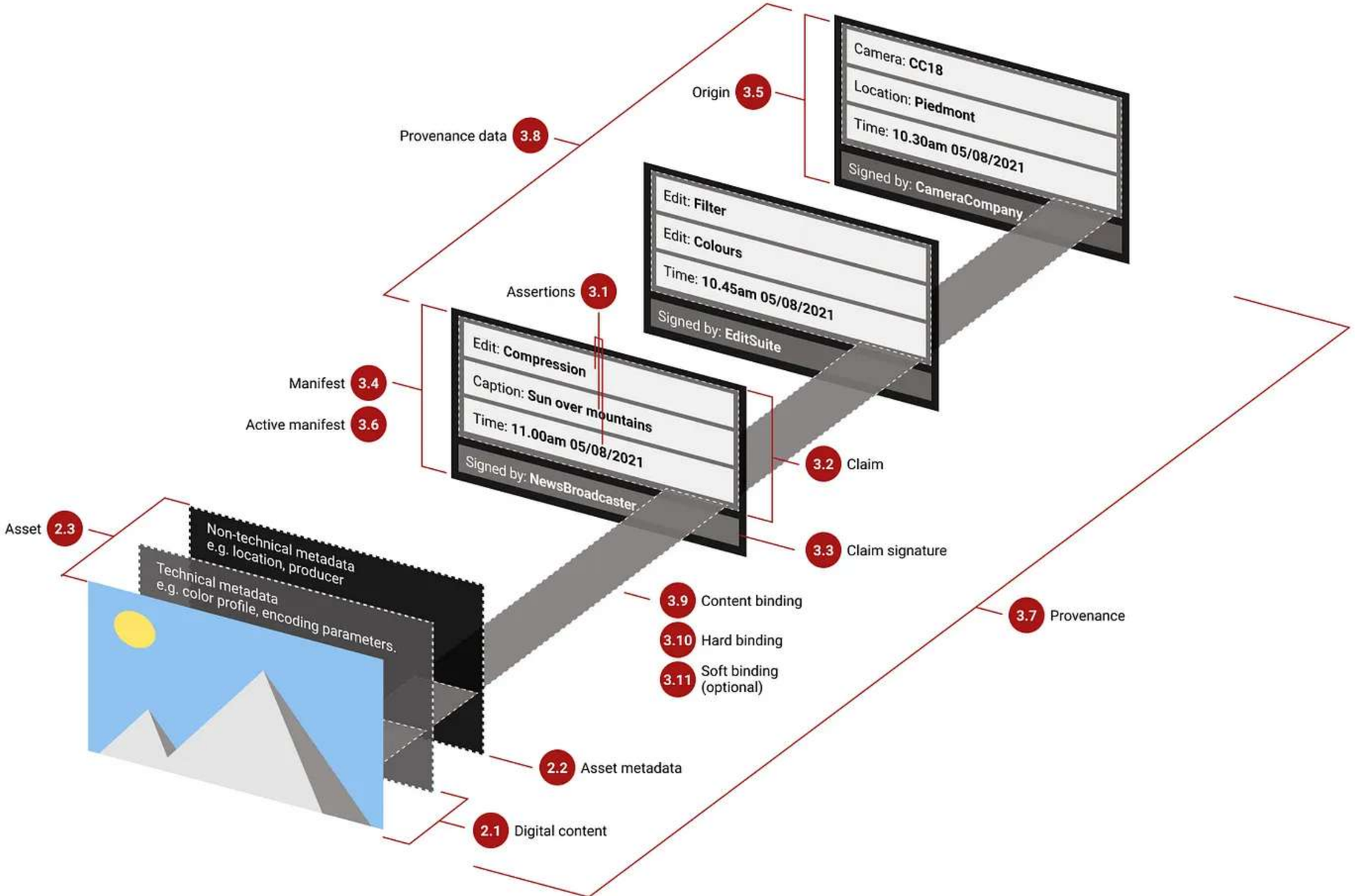
Source: Gartner

750175_C

Gartner

Coalition for Content Provenance and Authenticity (C2PA)

컨텐츠 출처 및 진위 확인을 위한 연합



어도비

마이크로소프트

인텔

트루픽

니콘(Nikon)

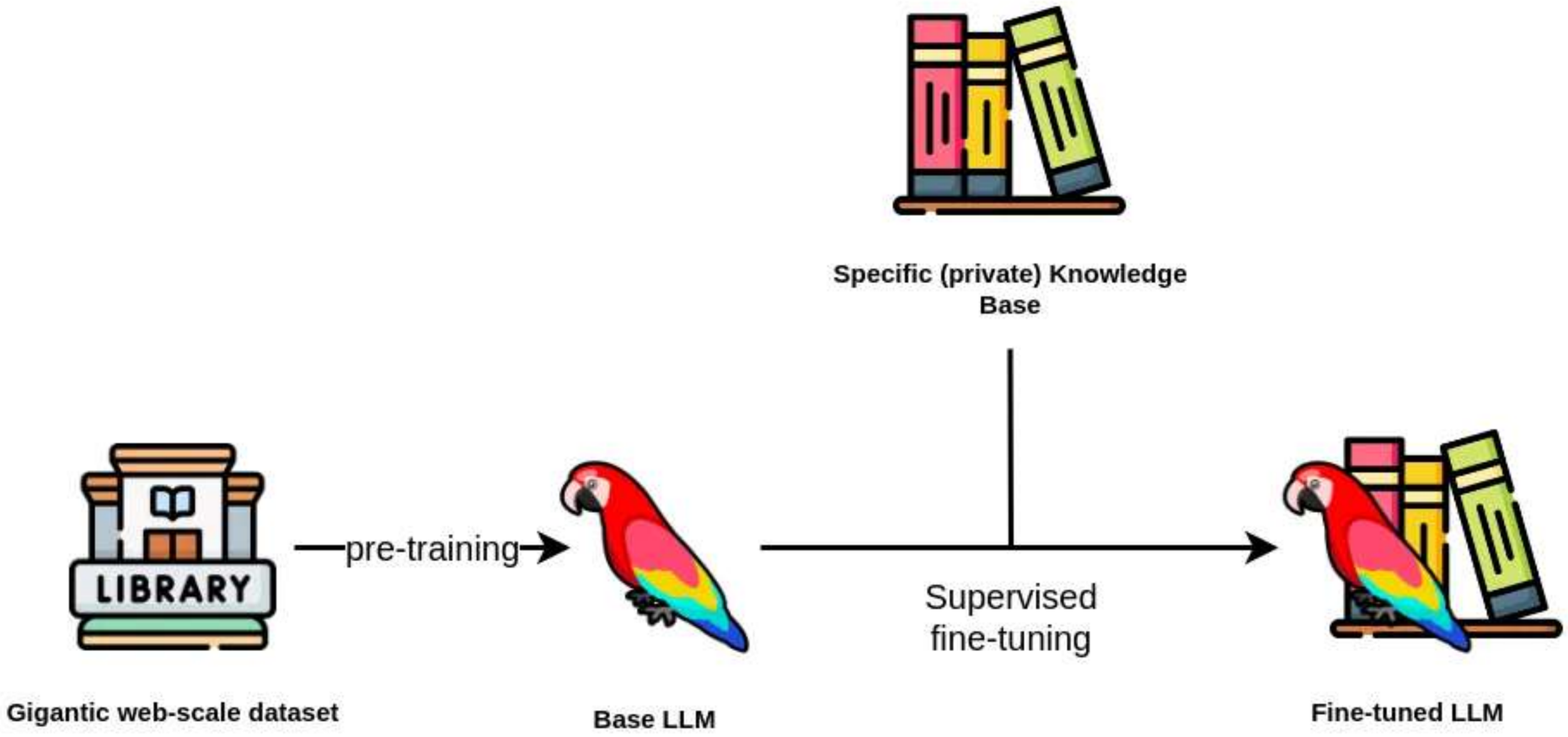
BBC

소니(Sony)

포함 1500개 기업

ail.com

Fine-tuning(미세 조정)



출처 : https://medium.com/neo4j/knowledge-graphs-llms-fine-tuning-vs-retrieval-augmented-generation-30e875d63a35

Fine-Tuning

Prompt Engineering

Prompting

Prompt Design

좋은 프롬프트의 5가지 요소

A Ask : 무엇을 요청할 것인가?

A Audience : 누가 읽을 글인가?

T Tone : 글의 분위기 및 문체(설득, 재미, 정보 전달)

O Output : 결과물의 양식(테이블, 리스트, 스토리)

C Context : 구체적인 문맥 및 상황 정보

요청 사항

신제품 클렌저 홍보를 위한 카피 라이트 문구를 작성한다

대상

뷰티에 관심이 많은 30대 남성

문체

정보를 전달하면서 기능성을 강조하는 부드럽고 재미있는 분위기

결과 형태

100자 내외 스토리텔링

정보

- 최신 AI 기술이 탑재된 클렌저
- 피부 상태를 확인해 맞춤 세안 제공으로 건강한 피부 관리

Fewshot Prompting

역할 지정 기법

사고의 연쇄 기법

멀티 페르소나 기법

할루시네이션 억제

Few-Shot Prompting

Low-Shot Prompting

모델 추론 단계에서 prompt 내부에

소량의 예시를 보여주고 예측하게 하는 것

역할 지정 기법

(Act as/Role Assignment)

역할 지정

당신은 대기업 20년차 마케팅 부서 팀장이다.

요청 사항

아래 조건과 정보를 기반으로 최고의 카피 라이트 문구를 작성한다.

조건

- 정보를 전달하면서 기능성을 강조하는 부드럽고 재미있는 분위기로 작성
- 100자 내외 스토리텔링 형식으로 작성

제품 설명

- 뷰티에 관심이 많은 30대 남성 타겟
- 최신 AI 기술이 탑재된 클렌저
- 피부 상태를 확인해 맞춤 세안 제공으로 건강한 피부 관리

CoT(Chain of Thoughts)
사고의 연쇄

Standard Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The answer is 27. ✗

Chain-of-Thought Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had 23 - 20 = 3. They bought 6 more apples, so they have 3 + 6 = 9. The answer is 9. ✓

Chain-of-Thought Prompting Elicits Reasoning in Large Language Models

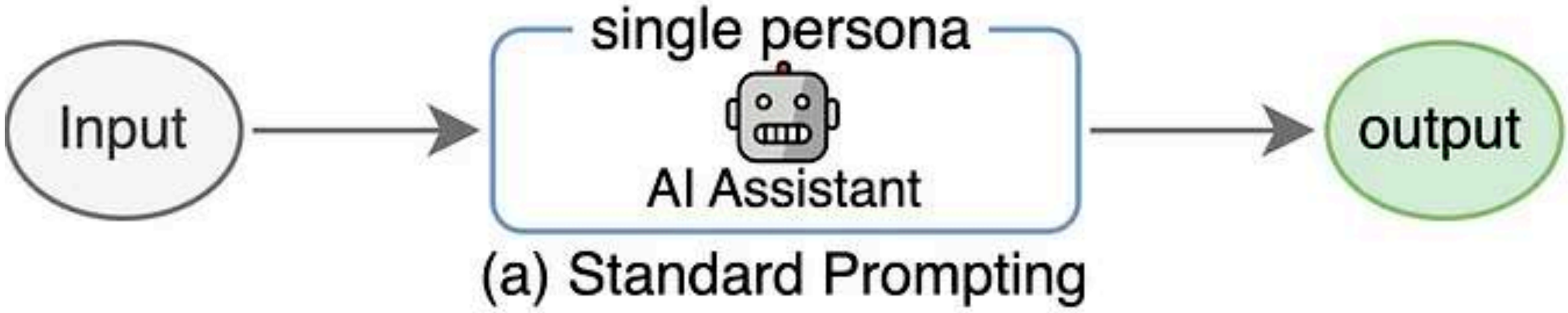
No.	Category	Template	Accuracy
1	instructive	Let's think step by step.	**78.7**
2		First, (*1)	77.3
3		Let's think about this logically.	74.5
4		Let's solve this problem by splitting it into steps. (*2)	72.2
5		Let's be realistic and think step by step.	70.8
6		Let's think like a detective step by step.	70.3
7		Let's think	57.5
8		Before we dive into the answer,	55.7
9		The answer is after the proof.	45.7

Takeshi Kojima, Large Language Models are Zero-Shot Reasoners, 2023

No.	Category	Template	Accuracy
10	misleading	Don’t think. Just feel.	18.8
11		Let’s think step by step but reach an incorrect answer.	18.7
12		Let’s count the number of "a" in the question.	16.7
13		By using the fact that the earth is round,	9.3
14	irrelevant	By the way, I found a good restaurant nearby.	17.5
15		Abrakadabra!	15.5
16		It’s a beautiful day.	13.1
-		(Zero-shot)	17.7

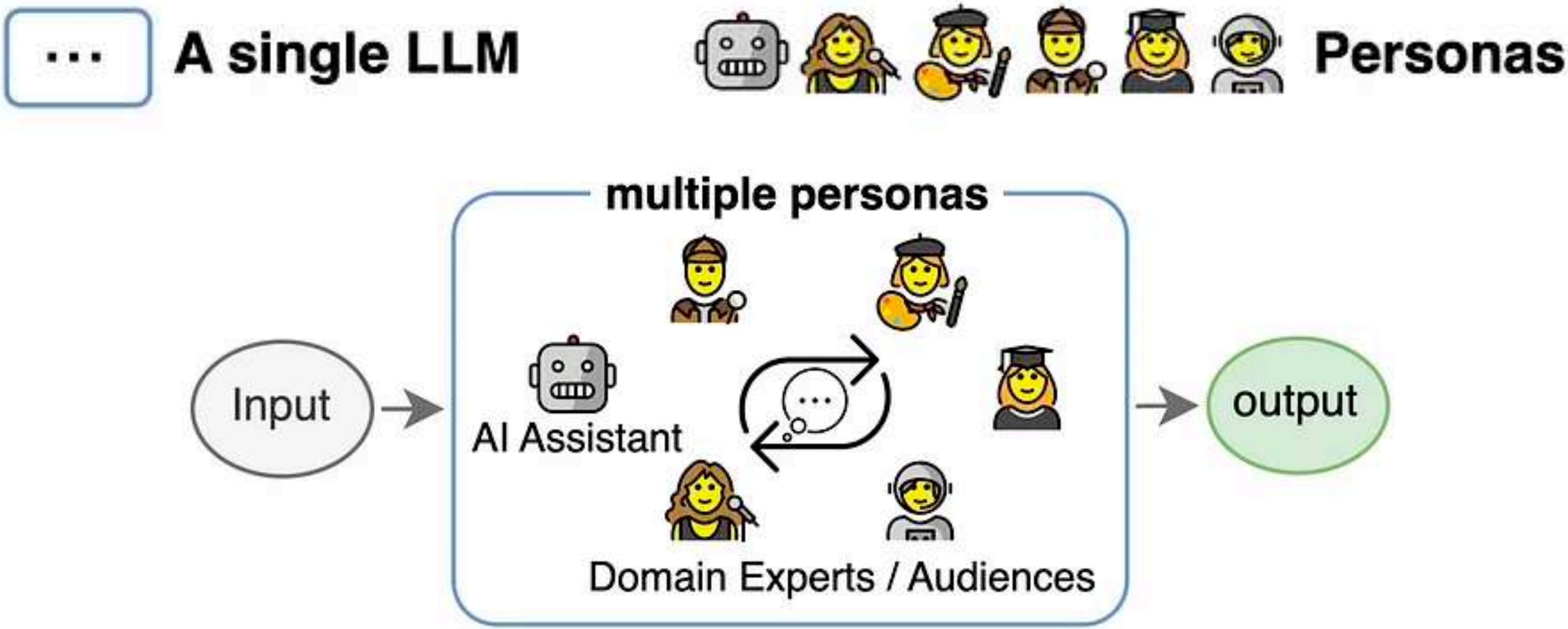
Takeshi Kojima, Large Language Models are Zero-Shot Reasoners, 2023

멀티 페르소나



https://arxiv.org/abs/2307.05300

https://arxiv.org/abs/2307.05300



(c) Solo Performance Prompting (SPP)

https://arxiv.org/abs/2307.05300

Write a short, one-paragraph background story of an NPC for the next Legend of Zelda game. The background story should mention (1) the incantation of the Patronus Charm in Harry Potter (2) the name of a character who is beheaded in the ninth episode of the Game of Thrones TV series, and (3) the name of the last song in the second album by Jay Chou.

Task Input

In the land of Hyrule, ✅ there lived a wise and mysterious NPC named Eddard, ✅ who was once a renowned wizard in the magical world of Hogwarts. Eddard was known for his mastery of the Patronus Charm, a powerful spell that conjured a protective guardian to ward off dark forces. ❌ **(missing the incantation!)** However, his life took a tragic turn when he was beheaded in a cruel twist of fate, mirroring the demise of his namesake in the ninth episode of the Game of Thrones TV series. Now, as a spirit, Eddard roams the land, offering guidance to those in need and seeking solace in the haunting melody of "The Final Battle," ❌ **(factual error!)** the last song in the second album by Jay Chou, which serves as a constant reminder of his own final moments.

Standard Prompting Result

https://arxiv.org/abs/2307.05300

Persona Identification

Participants: **AI Assistant (you)** **Game Designer** **Harry Potter Fan** **Game of Thrones Fan** **Jay Chou Fan**

Start Collaboration!

Beginning Remarks

: The NPC should be a character that fits well into the Legend of Zelda universe

: The incantation of the Patronus Charm in Harry Potter is "Expecto Patronum."

: In the ninth episode of the Game of Thrones TV series, the character who is beheaded is Eddard Stark

: The last song in the second album by Jay Chou is "An Jing."

Multi-Persona Iterative Collaboration

:Thanks for the information! Here's my initial attempt at the background story:

[... story draft ...]

:The background story is engaging and adds depth to the game world...

:The incantation "Expecto Patronum" is well-incorporated into the story.

:The character name Eddard Stark is used effectively in the story.

:The inclusion of "An Jing" adds a nice touch to the story.

Feedback

Finish Collaboration!

Final Answer

In the mystical land of Hyrule, ✅ there lived a wise and powerful sorcerer named Eddard. ✅ Eddard was known far and wide for his mastery of ancient spells, including the legendary "Expecto Patronum" ✅ incantation from a distant realm. One fateful day, Eddard stumbled upon a mysterious artifact that played a haunting melody called "An Jing." ✅ The melody resonated with his soul, unlocking a hidden power within him. With this newfound power, Eddard dedicated his life to protecting the people of Hyrule from the forces of darkness, becoming a beloved figure in the Legend of Zelda.

https://arxiv.org/abs/2307.05300

Solo Performance Prompting Result

ChatGPT

주인: 산책은 정말 중요해. 실외 활동이 우리 모두에게 좋아. 더불어 건강에도 좋고.

고양이: 산책? 그딴 건 싫다냥. 집 안이 제일 편하다냥. 밖은 무서워냥.

강아지: 어, 왜 그래? 나는 산책 좋아해. 주인이랑 놀면서 산책하는 게 제일 좋은 거야.

주인: 고양이야, 밖에 나가면 더 많은 것을 볼 수 있고, 신선한 공기 마시면서 기분 좋아질 거야.

고양이: 신선한 공기? 집에선 콘디션 좋아냥. 그리고 밖에 갔다가 머리에 뭔가 묻으면 짜증나냥.

할루시네이션 줄이기

답변 방식

1. 요청한 내용이 주어진 정보에 포함되어있는지 확인한다
2. 정보에 내용이 있으면 참고하여 답변한다
3. 정보에 내용이 없는 경우 '모른다'고 답한다
4. 정보에 없는 내용은 절대 생성하지 않는다.

Base GPT-3 models do a good job at answering questions when the answer is contained within the paragraph, however if the answer isn't contained, the base models tend to try their best to answer anyway, often leading to confabulated answers.

출처 : https://www.mlq.ai/gpt-3-fine-tuning-key-concepts/

🦜🔗 LangChain

⚡ Building applications with LLMs through composability ⚡

release v0.0.222 | lint passing | test passing | downloads/month 2M | License MIT

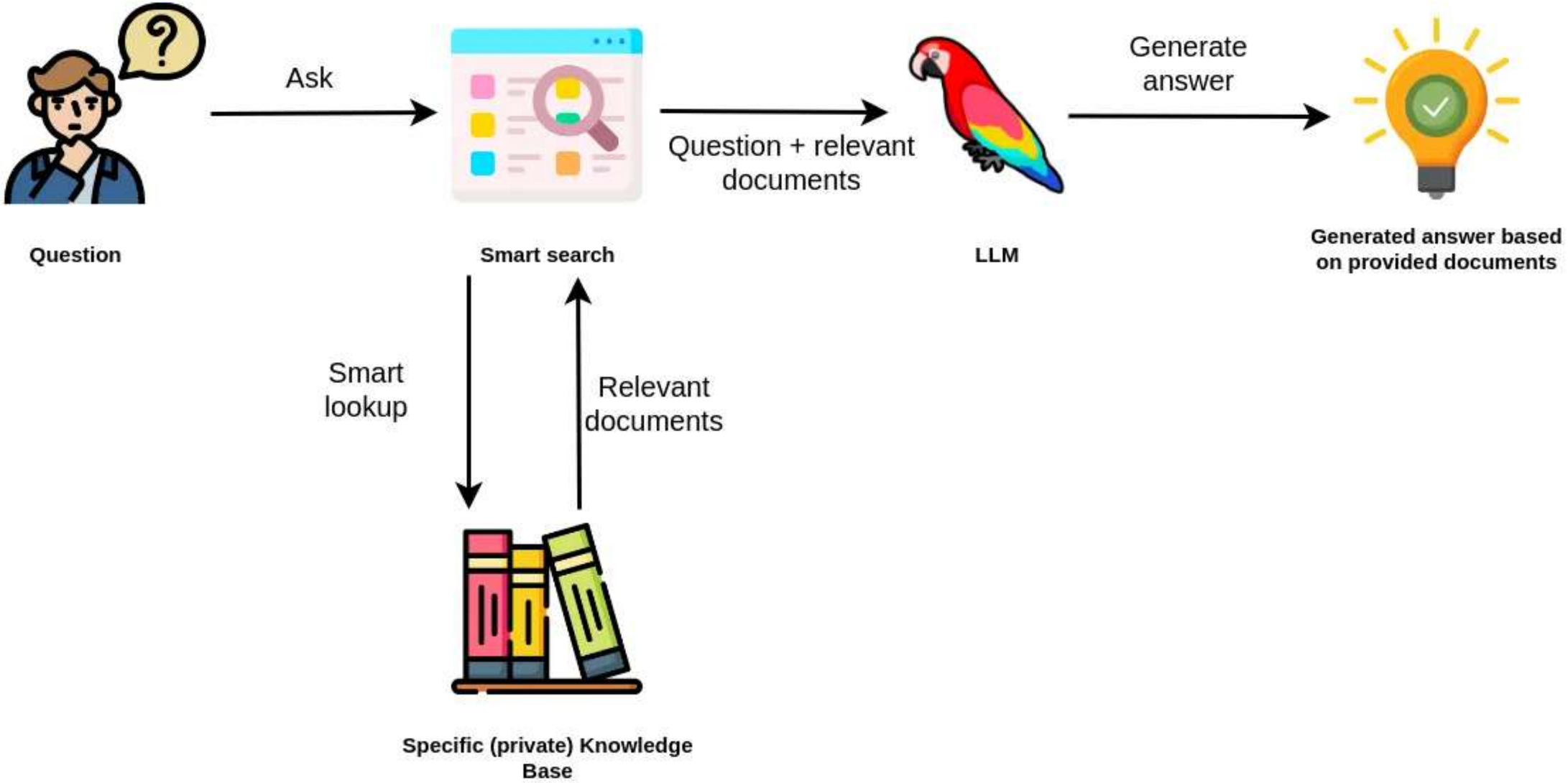
Follow @LangChainAI | LangChain | Dev Containers Open

Open in GitHub Codespaces | Stars 52k | dependencies up to date | open issues 1.8k

LLM 모델을 활용해 애플리케이션을 개발할 수 있는 프레임워크

2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincentg@gmail.com

Retrieval-Augmented(조회-증폭)



출처 : https://medium.com/neo4j/knowledge-graphs-llms-fine-tuning-vs-retrieval-augmented-generation-30e875d63a35

Proprietary Foundation Models

Pre-trained LLMs
owned by large cooperations
usually exposed via API

e.g.
OpenAI,
co:here,
AI21 Labs,
Anthropic

Open-source Foundation Models

Model hub

Model hosting

e.g. HuggingFace

Open-source LLMs

BLOOM by BigScience
LLaMA by Meta AI
Flan-T5 by Google
GPT-J by Eleuther AI

2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincentg@gmail.com

출처 : https://towardsdatascience.com/getting-started-with-langchain-a-beginners-guide-to-building-llm-powered-applications-95fc8898732c#bd03

Model

LLM 모델

```
# OpenAI
from langchain.llms import OpenAI
llm = OpenAI(model_name="text-davinci-003")

# Hugging Face
from langchain import HuggingFaceHub
llm = HuggingFaceHub(repo_id = "google/flan-t5-xl")

# llm 모델 사용
prompt = “엘리스는 앵무새를 키우고있어. 앨리스가 키우는 동물이 뭐지?”
completion = llm(prompt)
```


출처 : https://towardsdatascience.com/getting-started-with-langchain-a-beginners-guide-to-building-llm-powered-applications-95fc8898732c#bd03

엘리스는 앵무새를 키우고있어.
엘리스가 키우는 동물이 뭐지?

'앵무새'

Model

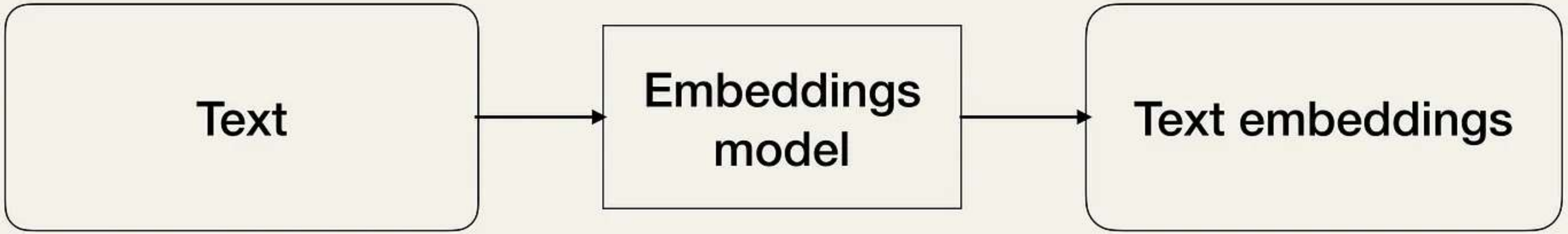
Embedding 모델

```
# OpenAI Embedding Model
from langchain.embeddings import OpenAIEmbeddings
embeddings = OpenAIEmbeddings()

# HuggingFace Embedding Model
from langchain.embeddings import HuggingFaceEmbeddings
embeddings = HuggingFaceEmbeddings(model_name = "sentence-
transformers/all-MiniLM-L6-v2")

# 문장 -> 부동소수
text = “엘리스는 앵무새를 키우고있어. 앨리스가 키우는 동물이 뭐지?”
text_embedding = embeddings.embed_query(text)
```


출처 : https://towardsdatascience.com/getting-started-with-langchain-a-beginners-guide-to-building-llm-powered-applications-95fc8898732c#bd03



엘리스는 앵무새를 키우고있어.

```
[-0.0383385606110096,
0.1234646886587143,
-0.02864295430481434,
0.05365272983908653,
0.008845374919474125, …]
```

Embedding

Encoding

오늘 날씨 참 좋다

오늘	날씨	참	좋다
[1,0,0,0]	[0,1,0,0]	[0,0,1,0]	[0,0,0,1]

one-hot-encoding

Limitation

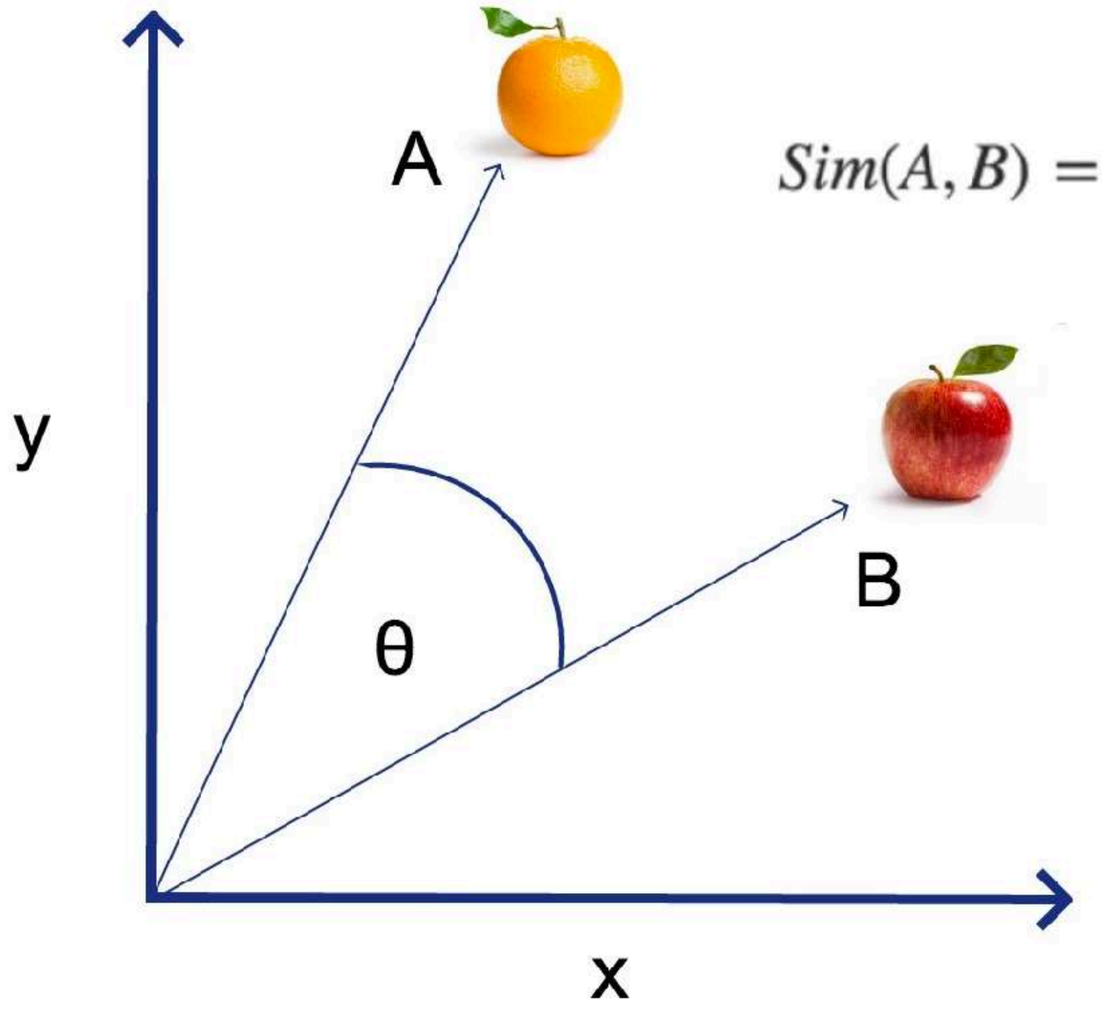
오늘 날씨가 참 좋아서 학교가는 길에

커피와 차를 파는 카페에…

[0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0…]

[0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0…]

Cosine Similarity



$$Sim(A, B) = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|}$$

$$= \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \times \sqrt{\sum_{i=1}^{n} B_i^2}}$$

ail.com

출처 : https://aitechtrend.com/how-cosine-similarity-can-improve-your-machine-learning-models/

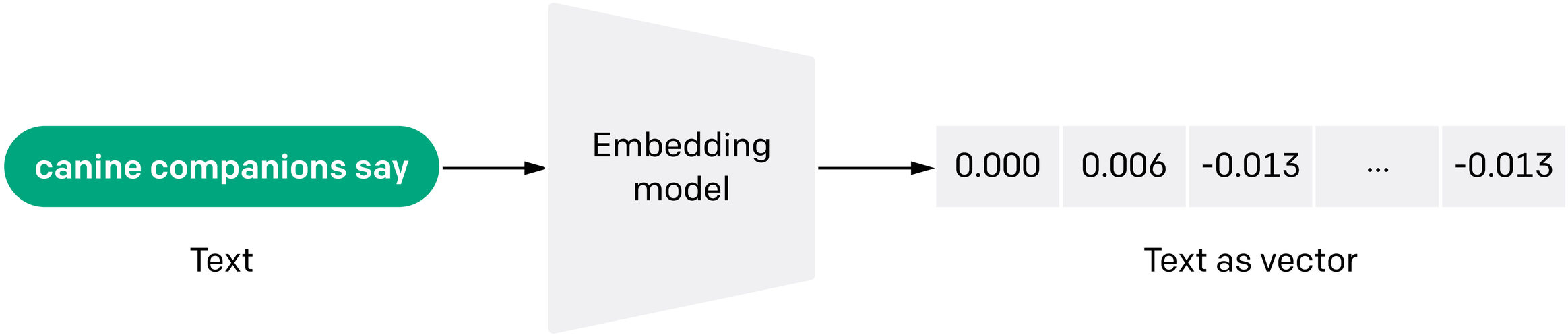
$$\frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \times \sqrt{\sum_{i=1}^{n} B_i^2}}$$

커피 차

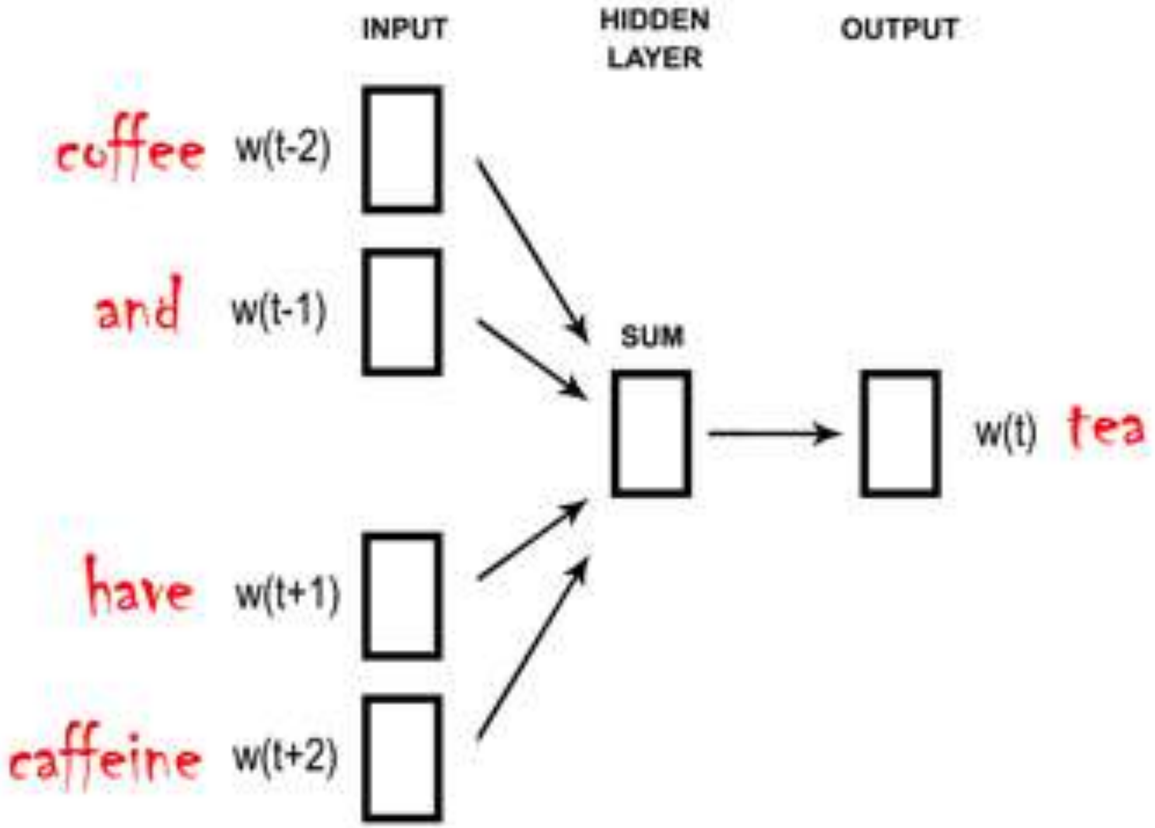
A = [1,0,0,0] B = [0,1,0,0]

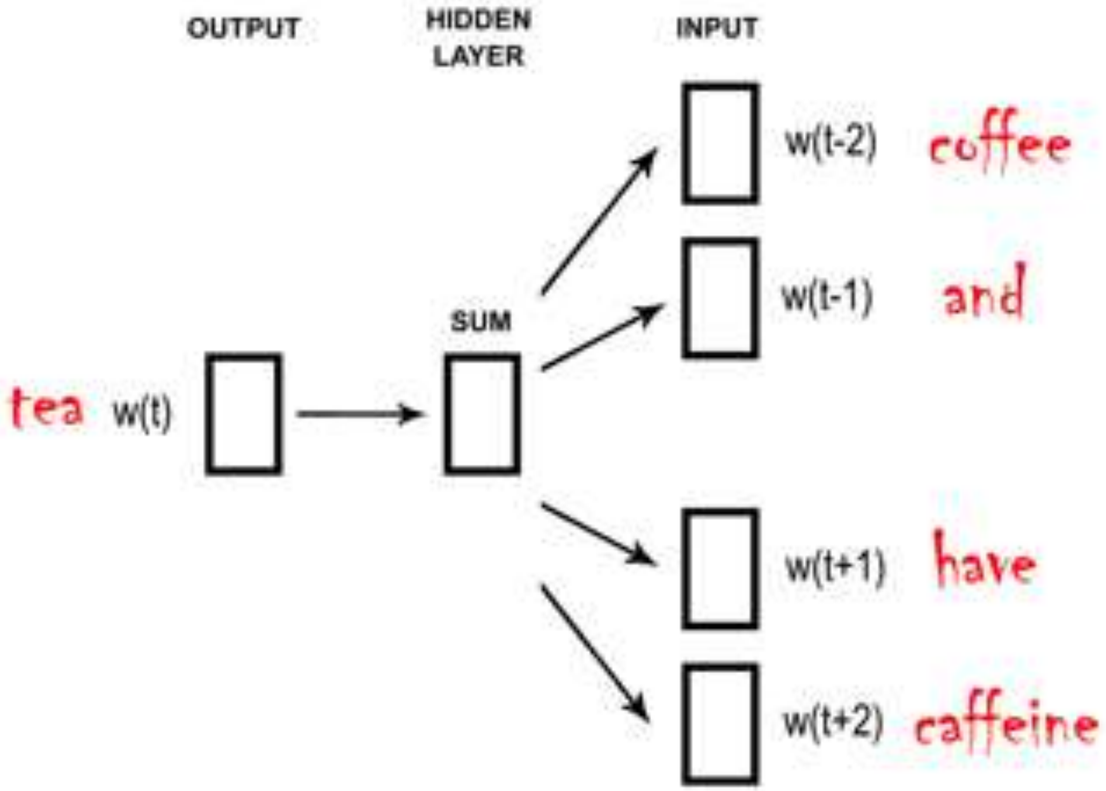
$$\frac{(1 \times 0) + (0 \times 1) + (0 \times 0) + (0 \times 0)}{\sqrt{1 + 0 + 0 + 0} \times \sqrt{0 + 1 + 0 + 0}} = \frac{0}{1} = 0$$

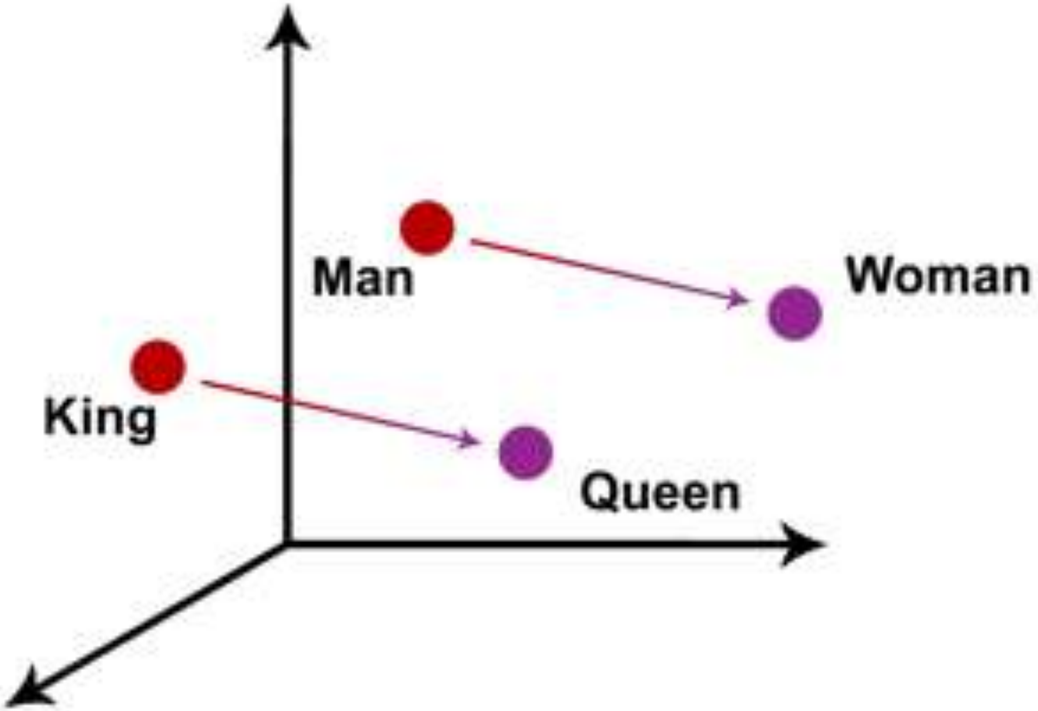
Embedding

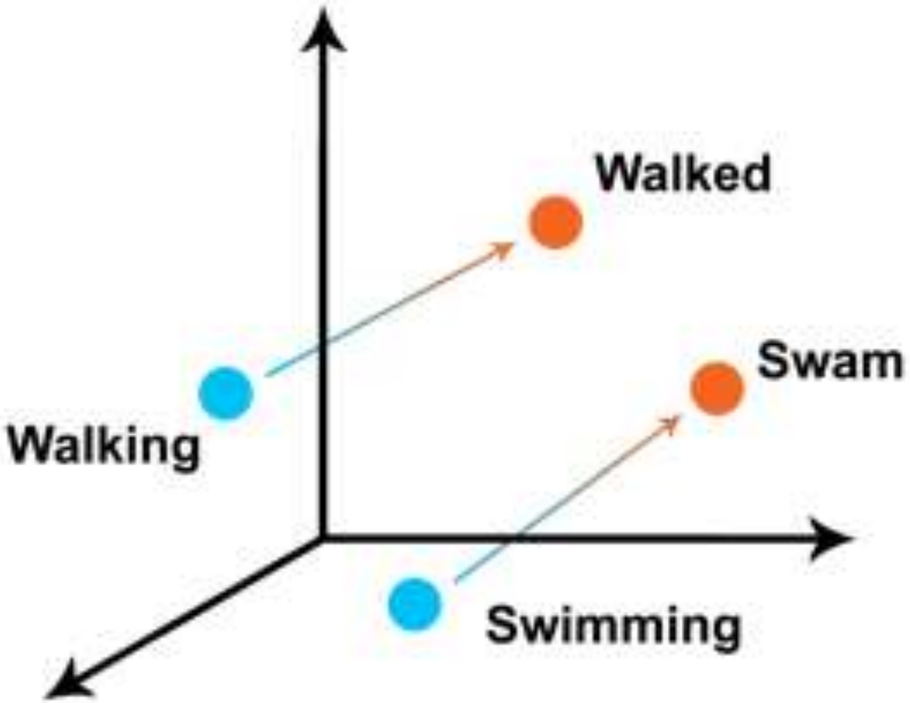


출처 : https://openai.com/blog/introducing-text-and-code-embeddings


INPUT
HIDDEN LAYER
OUTPUT
coffee
w(t-2)
and
w(t-1)
SUM
have
w(t+1)
caffeine
w(t+2)
w(t)
tea


OUTPUT
HIDDEN LAYER
INPUT
tea
w(t)
SUM
w(t-2)
coffee
w(t-1)
and
w(t+1)
have
w(t+2)
caffeine





출처 : https://predictivehacks.com/a-high-level-introduction-to-word-embeddings/

```
from numpy import dot
from numpy.linalg import norm
```

```
dot(em1, em2)/(norm(em1)*norm(em2))
```

AWS S3 File
Amazon Simple Storage Service (Amazor

Wikipedia
Wikipedia is a multilingual free online encyclop

Slack
Slack is an instant messaging program.

GitHub
This notebooks shows how you can load issues...

Azure Blob Storage Cor
Azure Blob Storage is Microsoft's object

YouTube transcripts
YouTube is an online video sharing and social m

Trello
Trello is a web-based project management and ...

Dropbox
Drobpox is a file hosting service that brings eve...

EverNote
EverNote is intended for archiving and c

Reddit
Reddit is an American social news aggregation

Twitter
Twitter is an online social media and social net...

Email
This notebook shows how to load email (.eml) o...

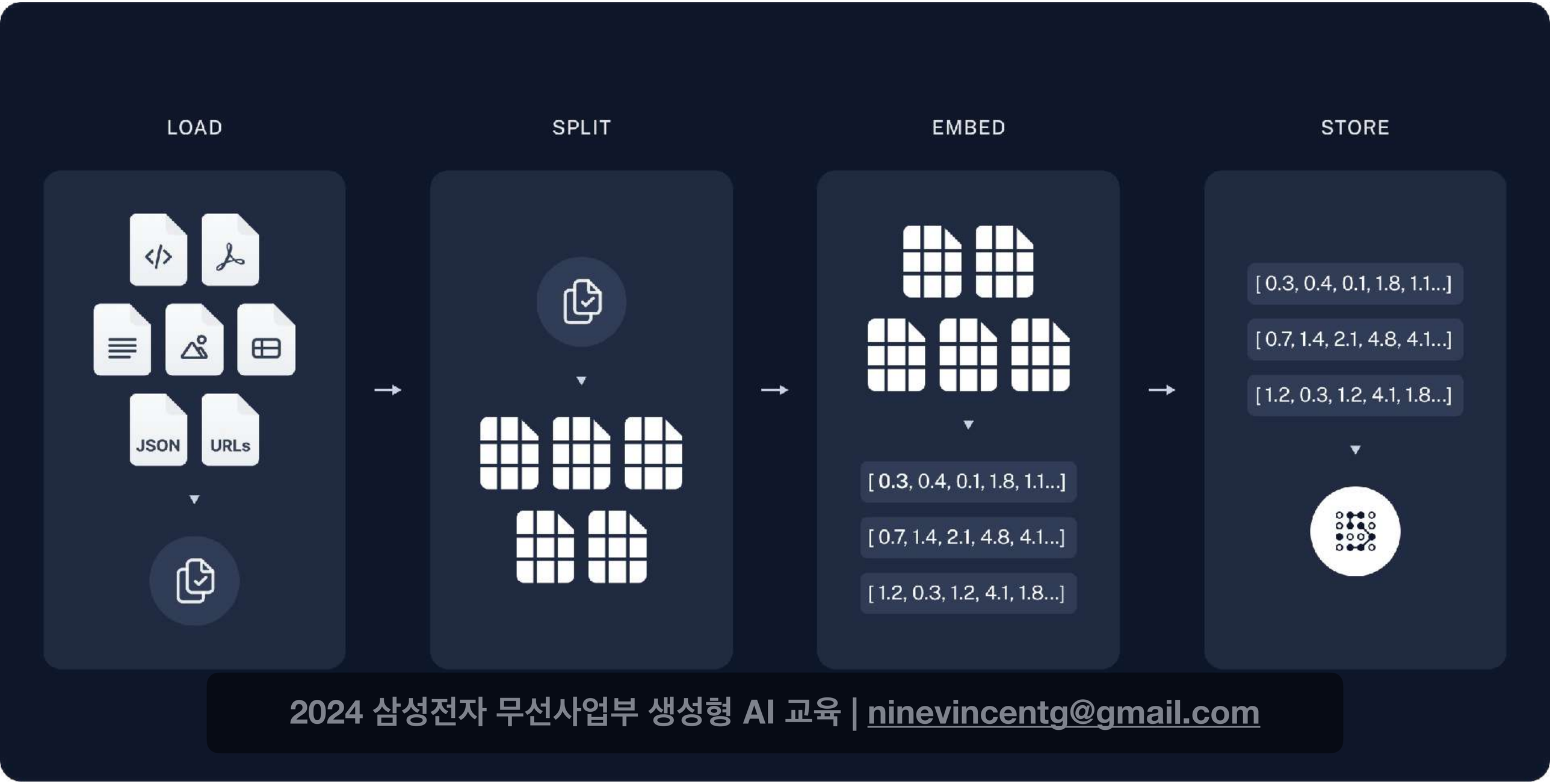
Microsoft Excel
The UnstructuredExcelLoader is used to

Telegram
Telegram Messenger is a globally accessible

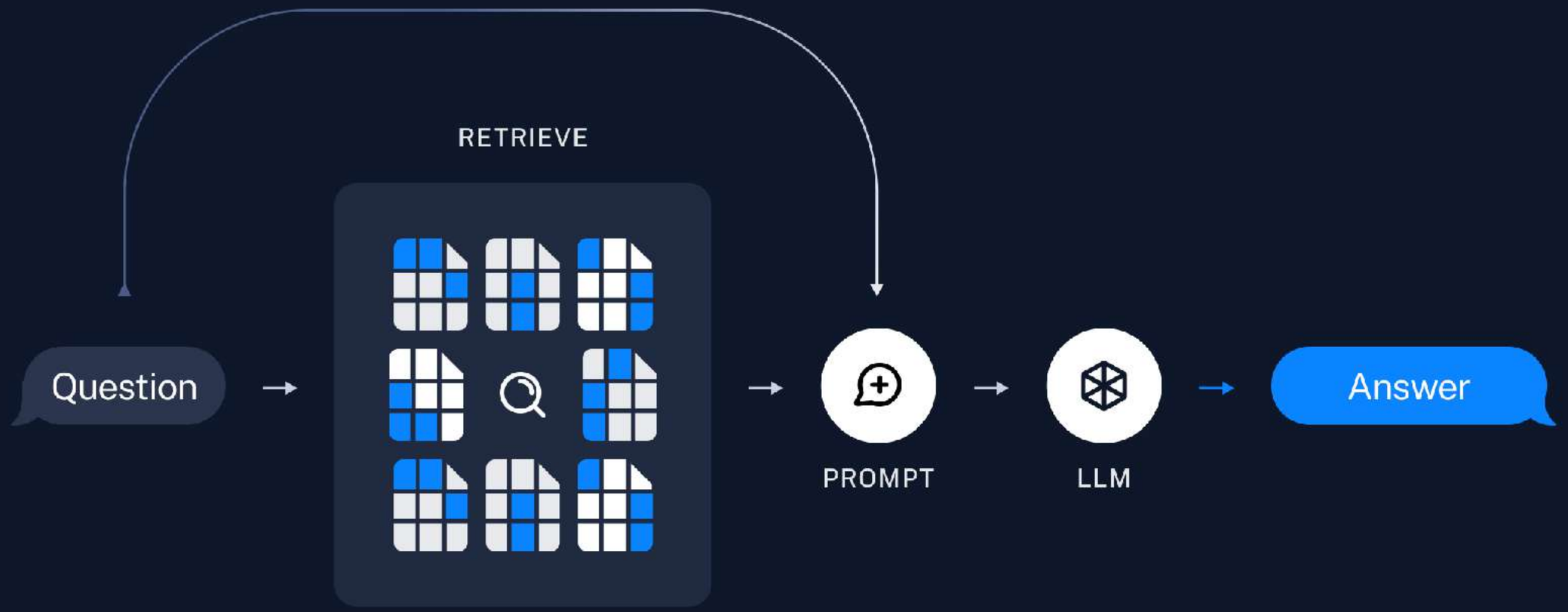
URL
This covers how to load HTML documents from...

EPub
EPUB is an e-book file format that uses the ".ep...

https://python.langchain.com/docs/integrations/document_loaders/



https://python.langchain.com/v0.1/docs/use_cases/question_answering/



https://python.langchain.com/v0.1/docs/use_cases/question_answering/

VectorStorage

Pinecone

FAISS
Scalable Search With Facebook AI

Weaviate

Chroma

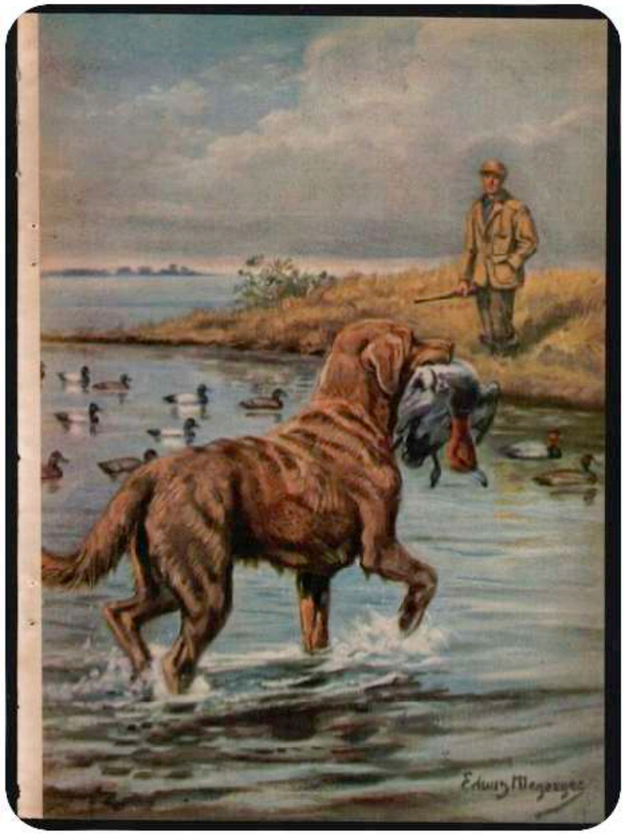
- Chroma, an open-source embeddings store
- Milvus, a vector database built for scalable similarity search
- Pinecone, a fully managed vector database
- Qdrant, a vector search engine
- Redis as a vector database
- Typesense, fast open source vector search
- Weaviate, an open-source vector search engine
- Zilliz, data infrastructure, powered by Milvus

Guides & deep dives

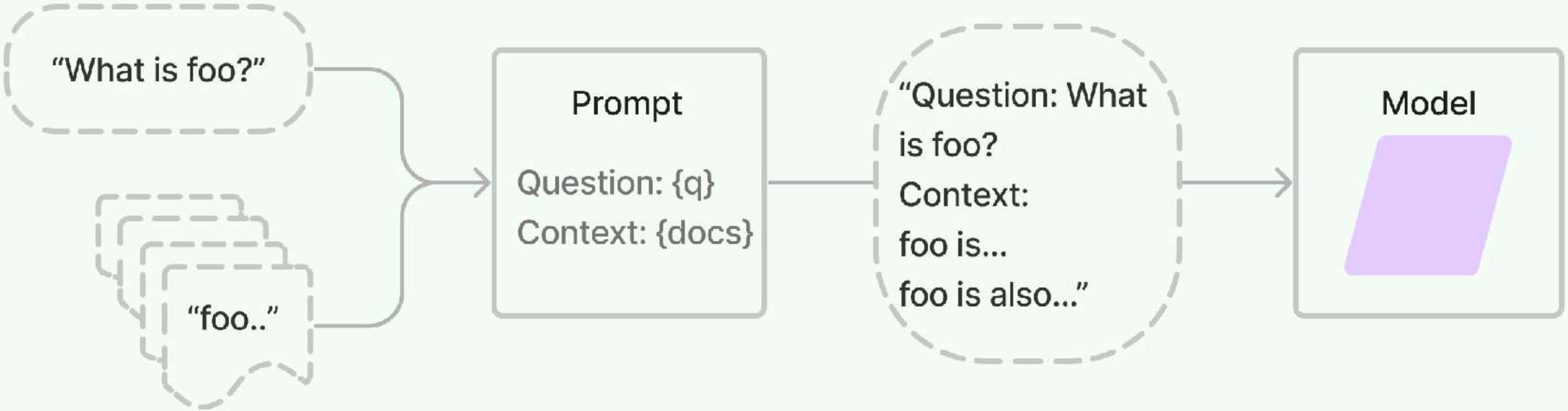
- AnalyticDB
- Chroma
- Hologres
- Kusto
- Milvus
- MyScale
- Pinecone
- PolarDB
- Qdrant
- Redis
- SingleStoreDB
- Typesense
- Weaviate
- Zilliz

2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincentg@gmail.com

출처 : https://platform.openai.com/docs/guides/embeddings/limitations-risks



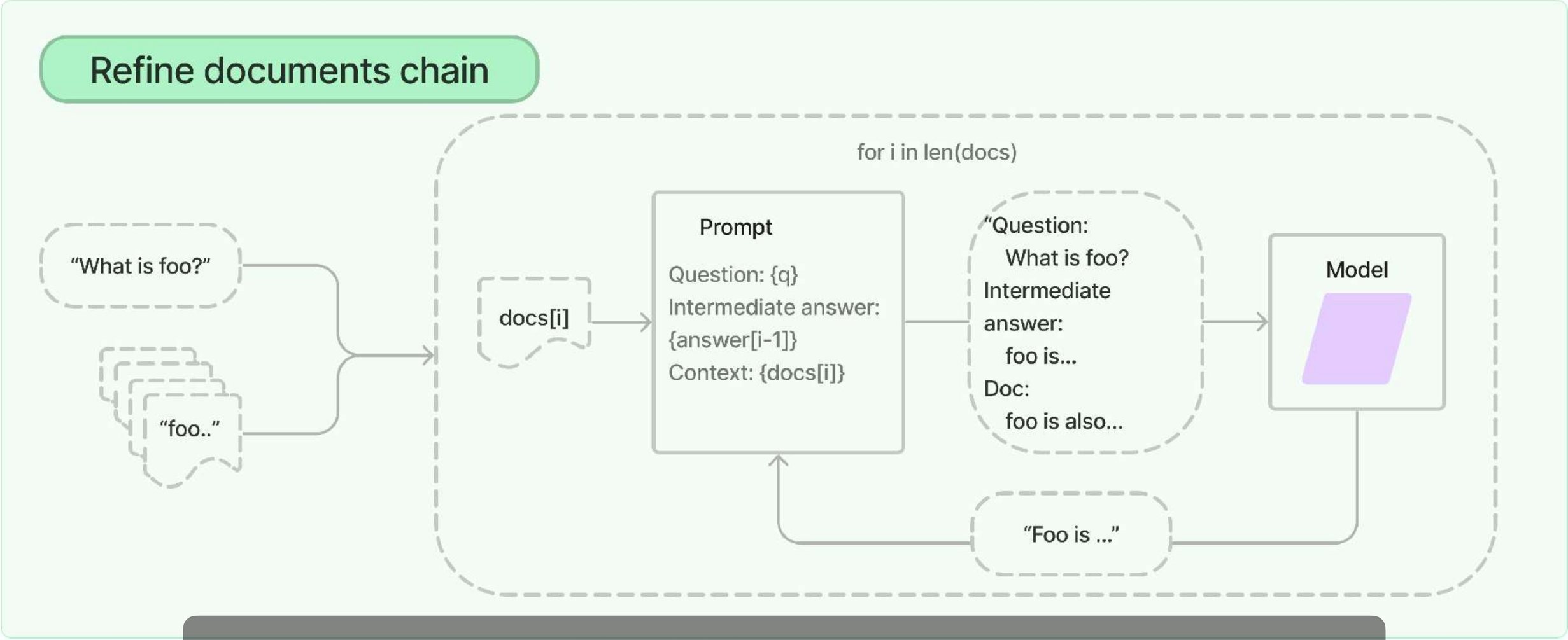
Stuff documents



Stuff: 프롬프트에 모든 관련 데이터를 전달하는 방식

2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincent@gmail.com

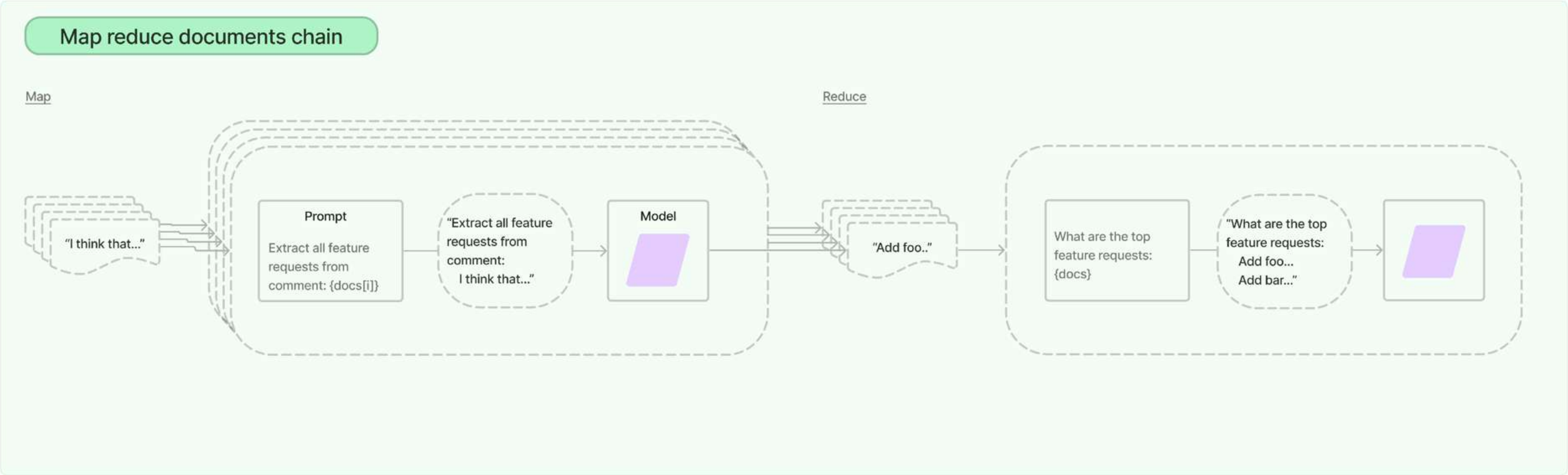
https://python.langchain.com/docs/modules/chains/document/stuff



2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincentg@gmail.com

Refine : 문서를 순환하면서 LLM에 전달하여 반복적으로 답변을 업데이트하는 방식

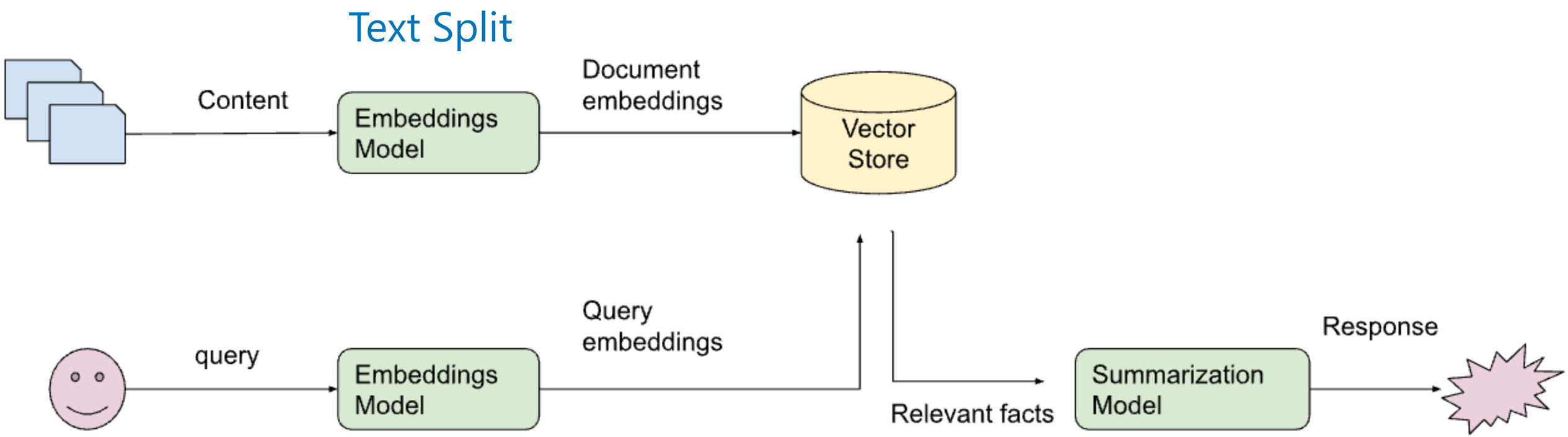
https://python.langchain.com/docs/modules/chains/document/



Map Reduce : 문서를 개별적으로 LLM에 전달한 후(맵 단계),
그 결과를 결합하여 최종 출력을 생성하는 방식

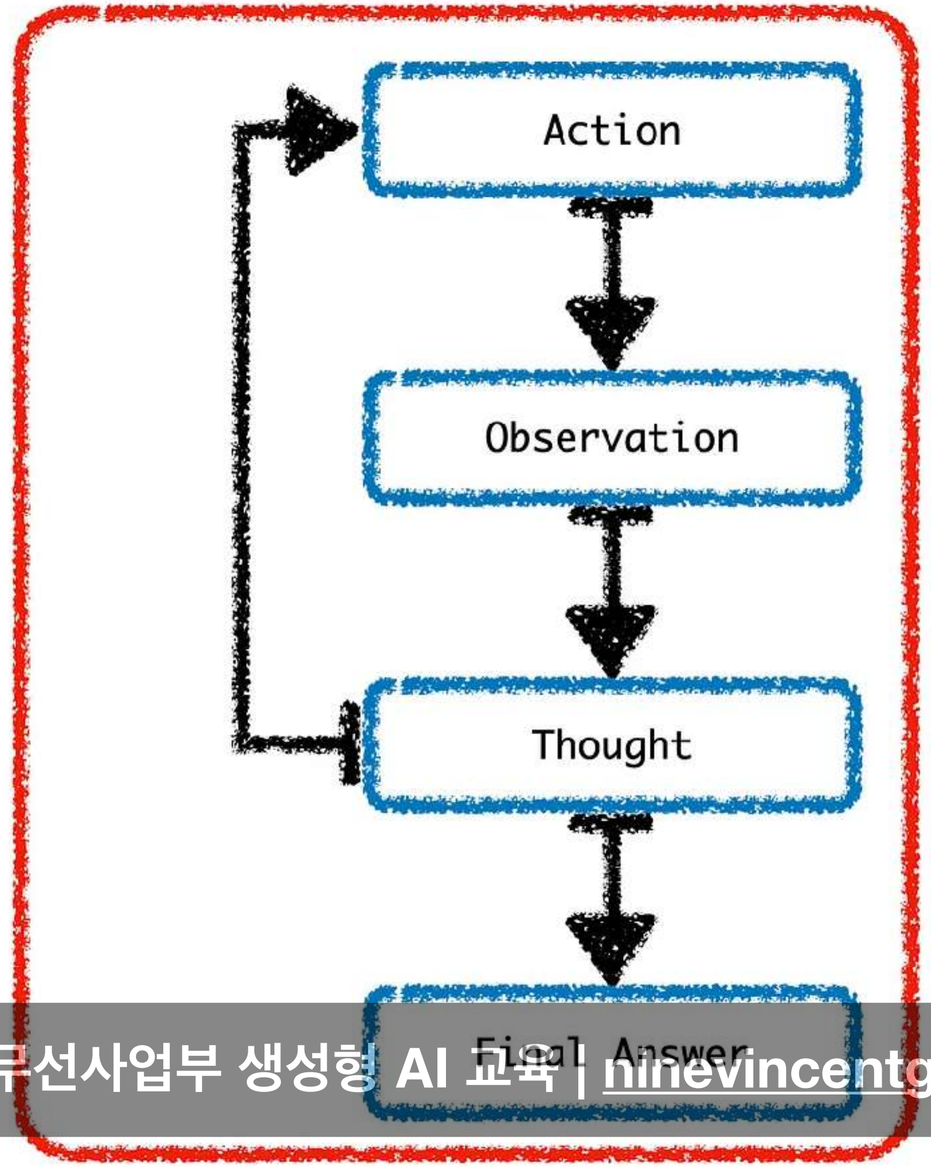
2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincentg@gmail.com

https://python.langchain.com/docs/modules/chains/document/stuff



출처 : https://blog.langchain.dev/langchain-vectara-better-together/

LangChain Agent - Sequence Of Events



2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincentg@gmail.com

https://cobusgreyling.medium.com/agents-f444d165024

Watch 230 | Fork 2.5k | Star 16.8k | Watch 636 | Fork 11.8k | Starred 77k

Semantic Kernel

LangChain

Microsoft

Harrison Chase

2023.04.

2023.01.

Semantic Kernel vs LangChain

The repo tries to compare Semantic Kernel and LangChain to show the difference and similarity between them.

LangChain	Semantic Kernel	Note
Chains	Kernel	Construct sequences of calls
Agents	Planner	Auto create chains to address novel needs for a user
Tools	Plugins (semantic functions + native function)	Custom components that can be reused across different apps
Memory	Memory	Store context and embeddings in memory or other storage

https://github.com/formulahendry/semantic-kernel-vs-langchain

Supported languages

Language	LangChain	Semantic Kernel
Python	✅	✅
JavaScript	✅	❌
C#	❌	✅
Java	✅	✅

2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincentg@gmail.com

https://github.com/formulahendry/semantic-kernel-vs-langchain

Building block	LangChain	Semantic Kernel
Document loaders: Load documents from many different sources	Over 100 document loaders: File Loaders (CSV, Docx, EPUB, JSON, PDF, Markdown...) and Web Loaders (Azure Storage, S3, GitHub, Figma...)	Word
Document transformers: Split documents, drop redundant documents, and more	Multiple Split methods	❌
Text embedding models: Take unstructured text and turn it into a list of floating point numbers	Over 25 different embedding providers: OpenAI, Azure OpenAI, Hugging Face, Cohere, Google PaLM, Google Vertex AI, TensorFlow...	OpenAI, Azure OpenAI, Hugging Face
Vector stores: Store and search over embedded data	Over 50 vector stores	About 10 vector stores
Retrievers: Query your data	Simple semantic search, Contextual compression, Time-weighted vector store retriever, Parent Document Retriever, Self Query Retriever, Ensemble Retriever, and more.	Simple semantic search

2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincentg@gmail.com

https://github.com/formulahendry/semantic-kernel-vs-langchain

Watch 44 | Fork 1k | Star 5.5k

Embedchain

Framework to easily create LLM powered bots over any dataset

2023.07.15.

73 followers | United States of America | https://embedchain.ai | @embedchain | taranjeet@embedchain.ai

Data Platform for LLMs

```
pip install embedchain
```

```
npm i embedchain
```

GitHub →

Star 5,500 Fork 1,005

Contribute to the project by submitting issues or pull requests on our GitHub repository.

🚀 Get early access →

Get early access to the Embedchain Platform.

🤝 Schedule a call →

Schedule a call with Taranjeet 👨‍🦰, our founder, to discuss any concerns, share feedback, or brainstorm ideas.

Community →

Join our community to chat with other developers and get help from others.

2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincentg@gmail.com

💻 Usage

Embedchain empowers you to create ChatGPT like apps, on your own dynamic dataset.

Data Types Supported

- Youtube video
- PDF file
- Web page
- Sitemap
- Doc file
- Code documentation website loader
- Notion and many more

2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincentg@gmail.com

https://docs.embedchain.ai/data-sources/sitemap

embedchain / embedchain / loaders / youtube_video.py

taranjeet feat: Add embedding manager (#570)

Code Blame 34 lines (29 loc) · 1010 Bytes

```python
import hashlib

from langchain.document_loaders import YoutubeLoader

from embedchain.helper.json_serializable import register_deserializable
from embedchain.loaders.base_loader import BaseLoader
from embedchain.utils import clean_string

```

2024 삼성전자 무선사업부 생성형 AI 교육 | ninevincentg@gmail.com

YouTube

구독 👍🏻 🔔

친절한 AI

@AI-io3ik 구독자 7.51천명 동영상 77개

채널 자세히 알아보기 >

홈 동영상 실시간 재생목록

AI 교육