

**2시간만에 배우는 LLM, ChatGPT, 프롬프트
엔지니어링**

INDEX

1. 본인 소개
2. ANN의 이해
3. LLM 이란
4. ChatGPT 개용
5. Vector 란
6. Prompt Engineering 이란?
7. Prompt Engineering 방법
8. 기타 등등

본인 소개

저는?

근무처 : 서울 강남
사는곳 : 인천

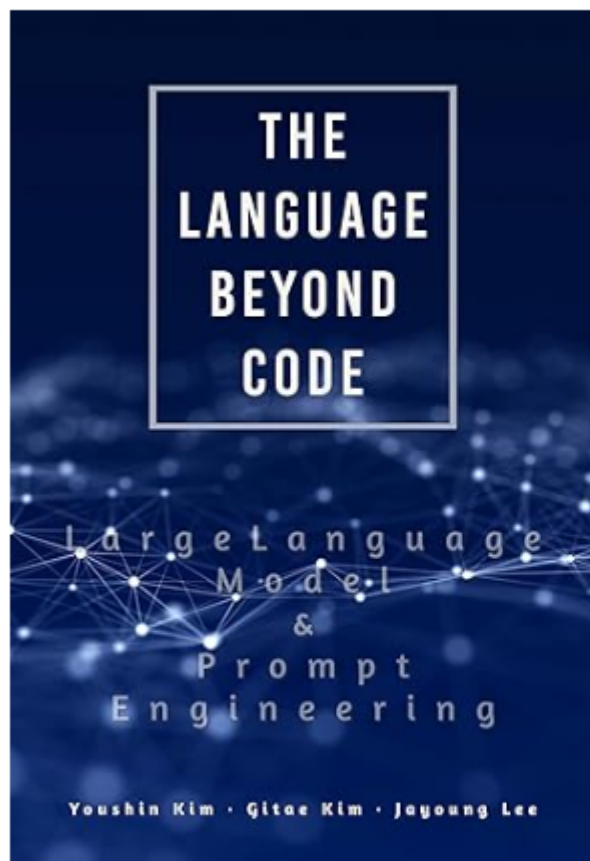
IT 개발 경력 : 20년(웹개발자 -> backend 개발자 -> Cloud Infra deliver(SA) -> Presales)
Cloud 경 력 : 4년차 AWS -> Azure

현재 : ChatGPT 관련 Presales



저는?

Kindle Store > Kindle eBooks > Computers & Technology



Roll over image to zoom in

Read sample

The language beyond code: Large Language Model & Prompt Engineering Kindle Edition

by [Gitae Kim](#) (Author), [Youshin Kim](#) (Author), [Jayoung Lee](#) (Author) | Format: Kindle Edition

[See all formats and editions](#)

This book aims to strike a balance between technical background and practical application, guiding readers to understand and utilize AI technology more effectively. It summarizes the advancements and significance of AI technology, focusing on rapidly evolving technologies such as generative AI and related areas, including Large Language Models (LLM), OpenAI, and the GPT project. Covering the evolution of AI technology from its foundations to the latest models, it provides technical insights and future perspectives from a practical standpoint. Additionally, the book discusses prompt engineering for substantive interaction with AI and demonstrates effective communication methods with AI, exploring ways to make better use of AI technology.

Concept: The core concept revolves around a deep understanding of artificial intelligence technology and its practical application. The book encompasses the development and influence of technical advancements in AI, especially focusing on Large Language Models (LLM) and models like GPT, while simultaneously covering prompt engineering techniques for effective human-AI interaction. Through this book, readers will gain insights into the progression of AI technology and practical knowledge in prompt engineering, applicable in real-world scenarios.

Reading age



7 - 18 years

Print length



255 pages

Language



English

Publication date



January 27, 2024

Page Flip



Enabled



ANN의 이해

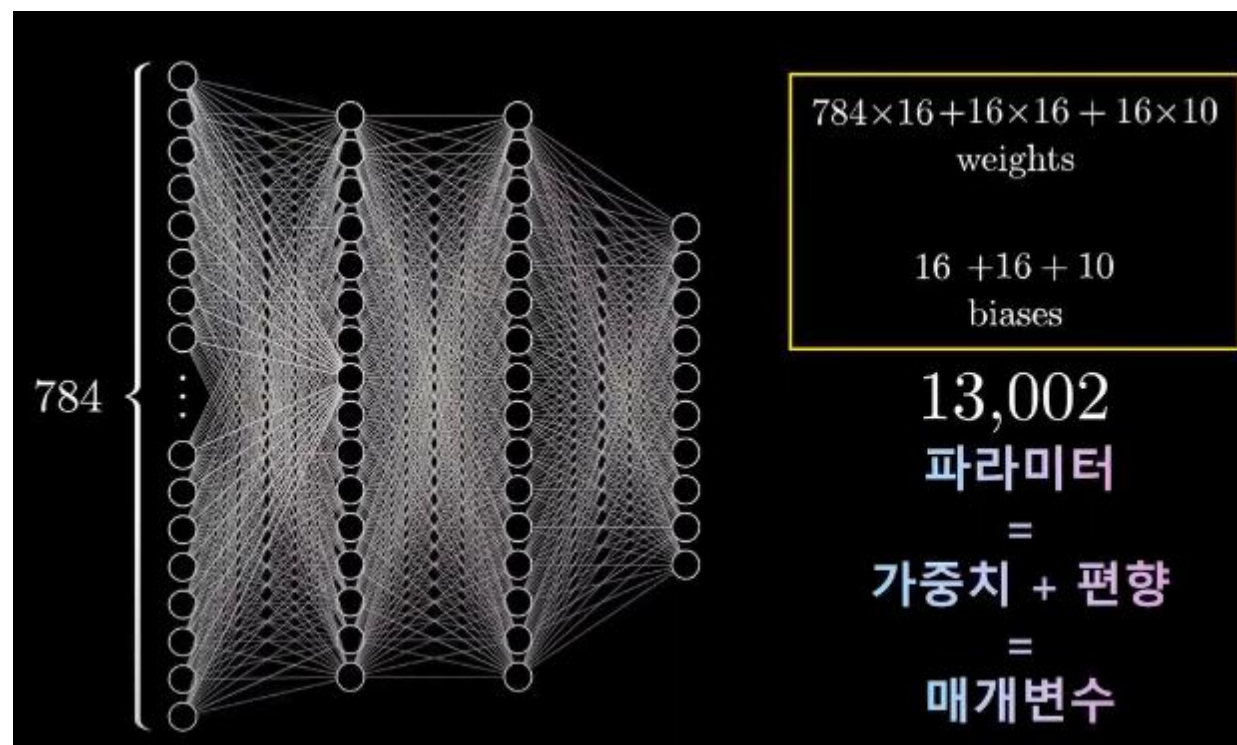
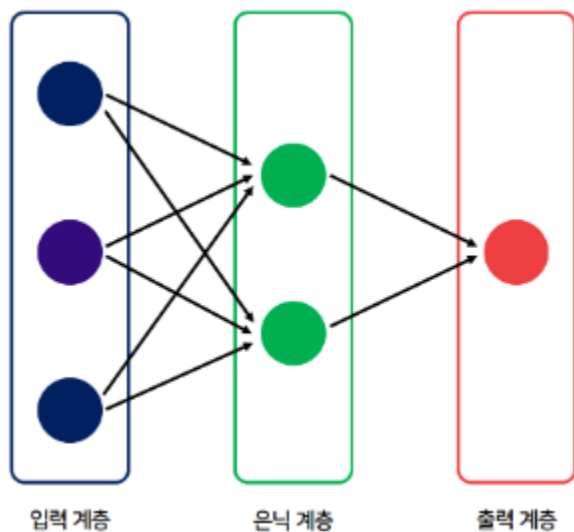
ANN 이란?

ANN 이란?

ANN = Artificial Neural Network, 인공신경망은 두뇌의 신경망을 모방한 모델로, 입력 데이터를 통해 학습하여 복잡한 데이터를 처리하는 데 사용됩니다.

퍼셉트론 = 뉴런

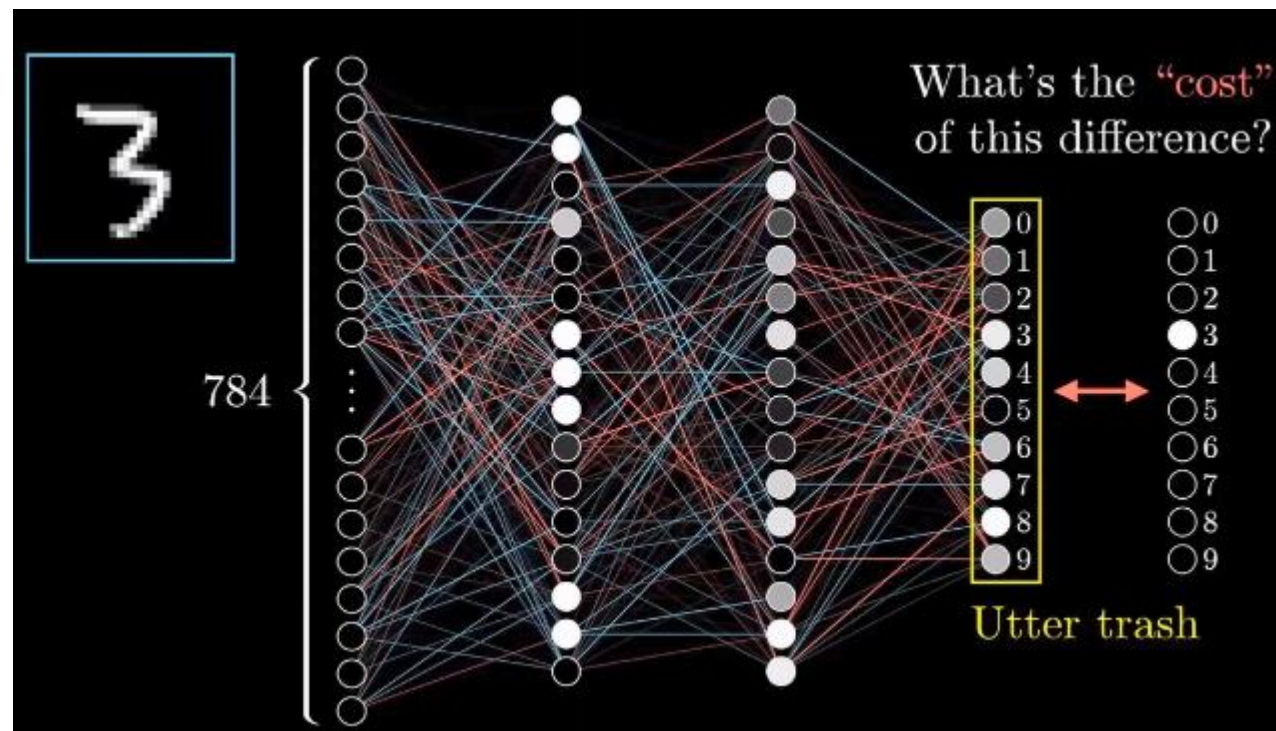
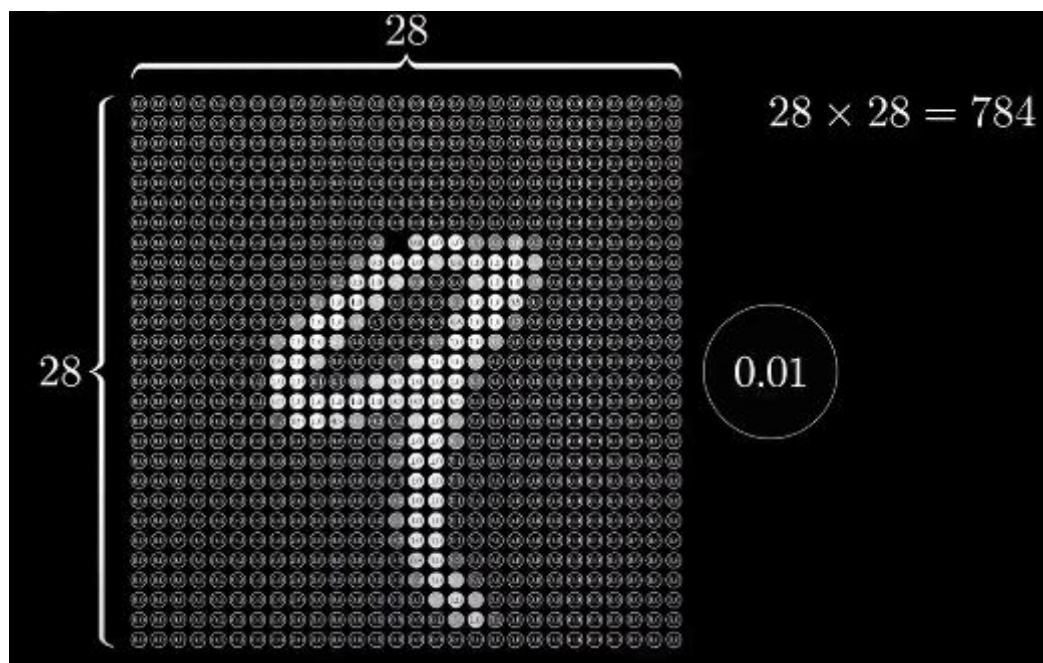
$$y = x1w1 + b$$



ANN 이란?

ANN 이란?

ANN = Artificial Neural Network, 인공신경망

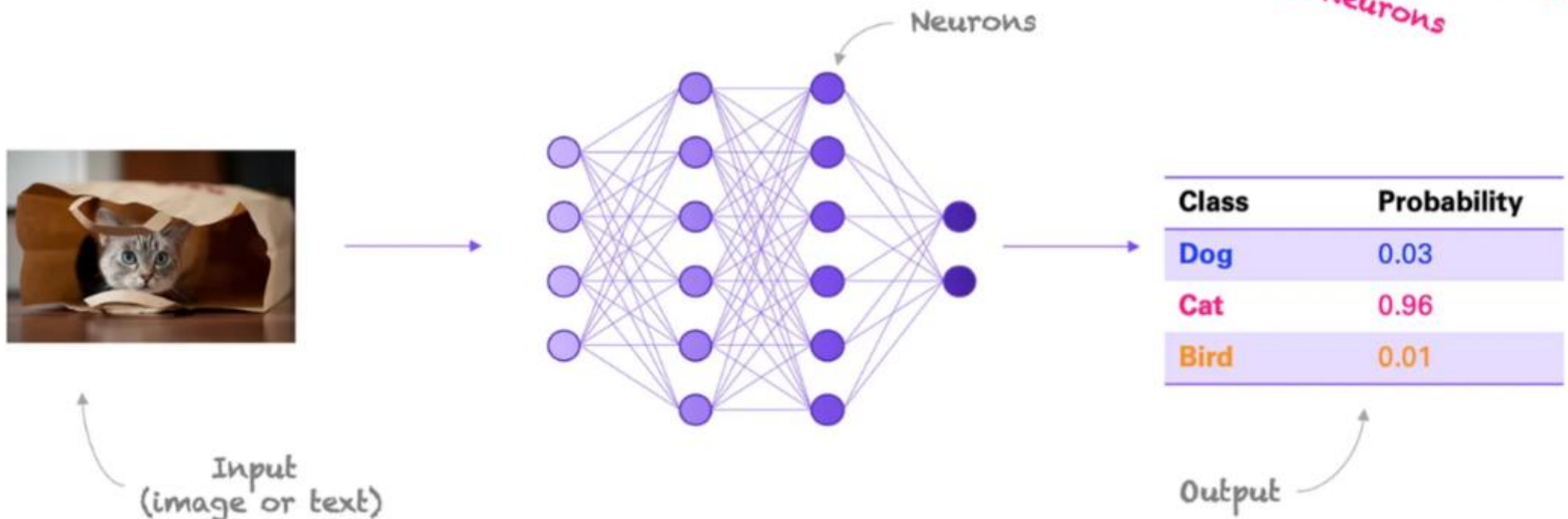


ANN 예제

ANN 이란?

ANN 이란?

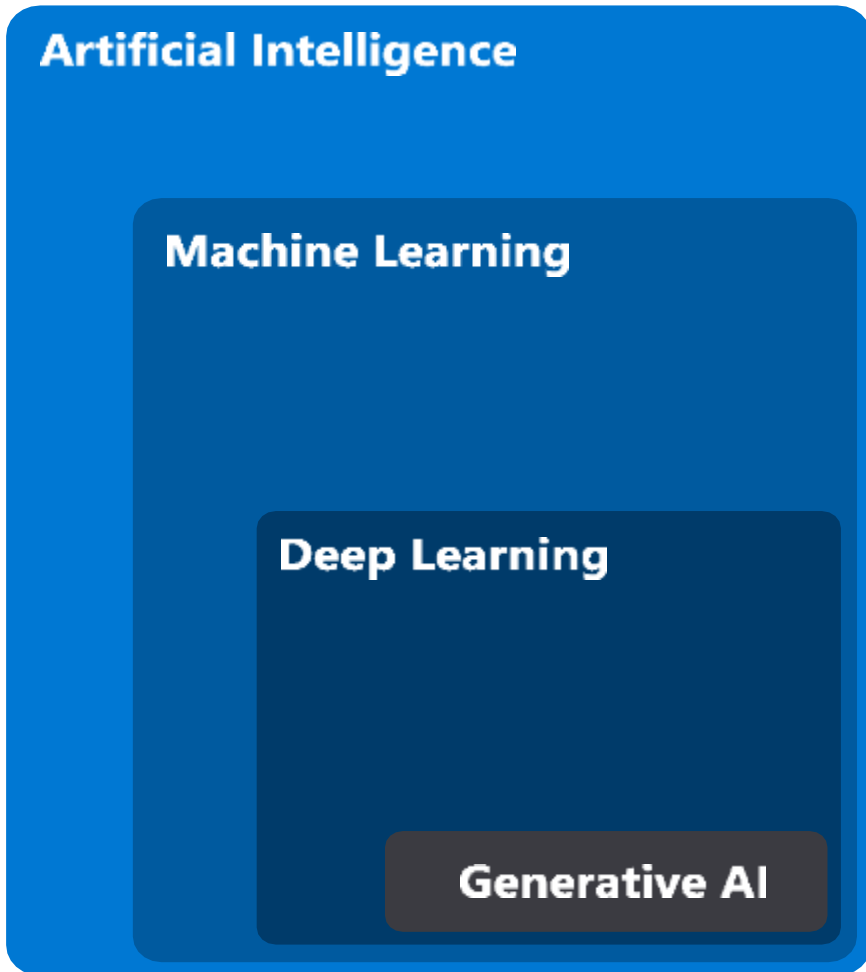
We need something way more powerful... **Neural Networks**



LLM이란

LLM이란?

인공 지능의 발전사입니다.



Artificial Intelligence

인공 지능은 머신의 학습 및 실행을 통해 인간의 행동을 모방합니다.



Machine Learning

데이터를 활용하여 예측을 하는 알고리즘입니다.
예: 기상 조건을 사용하여 매장 수익 예측



Deep Learning

인공 신경망을 이용한 복잡한 알고리즘 계층입니다.
Azure AI 서비스는 주로 딥 러닝 모델에 기반합니다.



Generative AI

입력 데이터 기반으로 새로운 콘텐츠 생성 가능
OpenAI 모델: 언어, 코드, 이미지 생성 등 다양한 분야에서 활용되는 생성 AI 모델 컬렉션

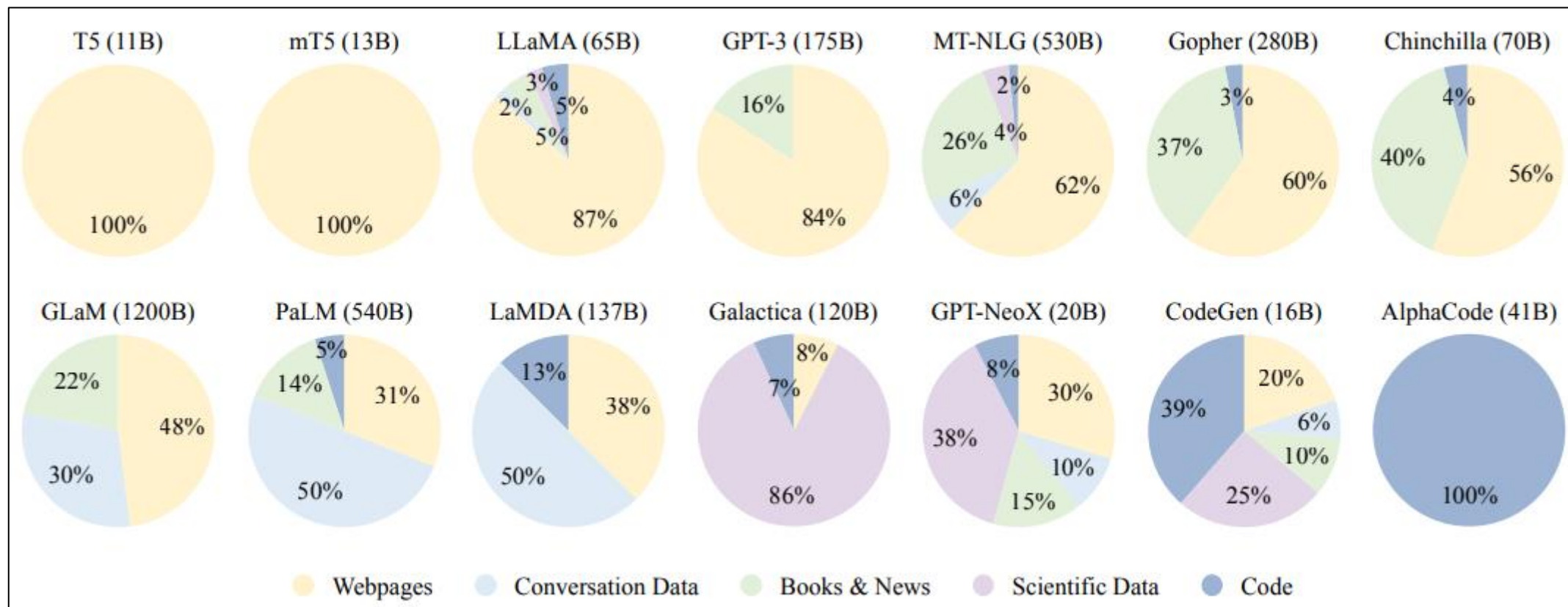
LLM이란?

LLM 이란.



LLM이란?

LLM모델 별 소스 데이터



LLM이란?

Next word prediction

Massive training data



We can create **vast amounts of sequences** for training a language model

● Context ● Next Word ● Ignored

[The cat likes to sleep in the]
[The cat likes to sleep in the]
[The cat likes to sleep in the]
[The cat likes to sleep in the]
[The cat likes to sleep in the]

We do the same with much longer sequences. For example:

A language model is a probability distribution over sequences of words. [...] Given any sequence of words, the model predicts the next ...

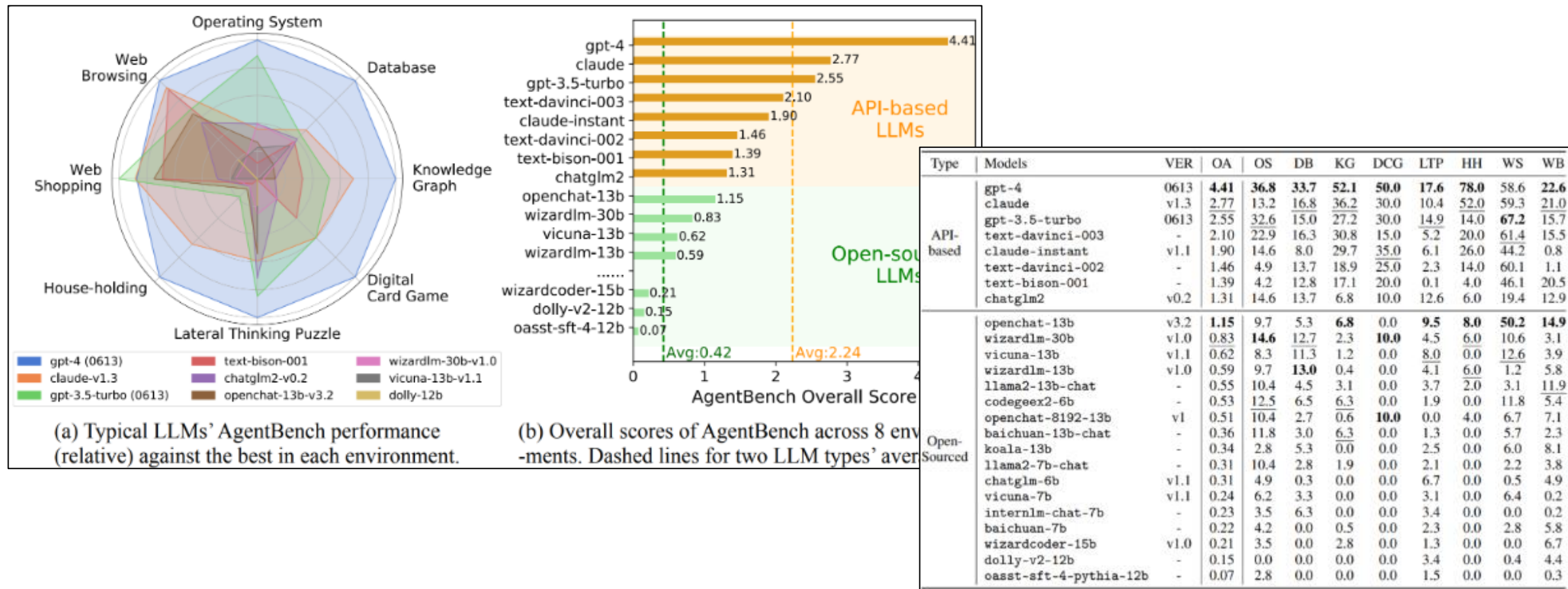
Or also with code:

```
def square(number):  
    """Calculates the square of a number."""  
    return number ** 2
```

And as a result - the model becomes incredibly good at predicting the next word in any sequence.

LLM이란?

각 LLM모델 별 벤치마크



출처 : https://the-decoder.com/gpt-4-dominates-other-llms-in-real-world-pragmatic-missions-study-finds/?utm_source=aibreakfast.beehiiv.com&utm_medium=newsletter&utm_campaign=ai-is-building-highly-effective-antibodies-that-humans-can-t-even-imagine

ChatGPT 개요

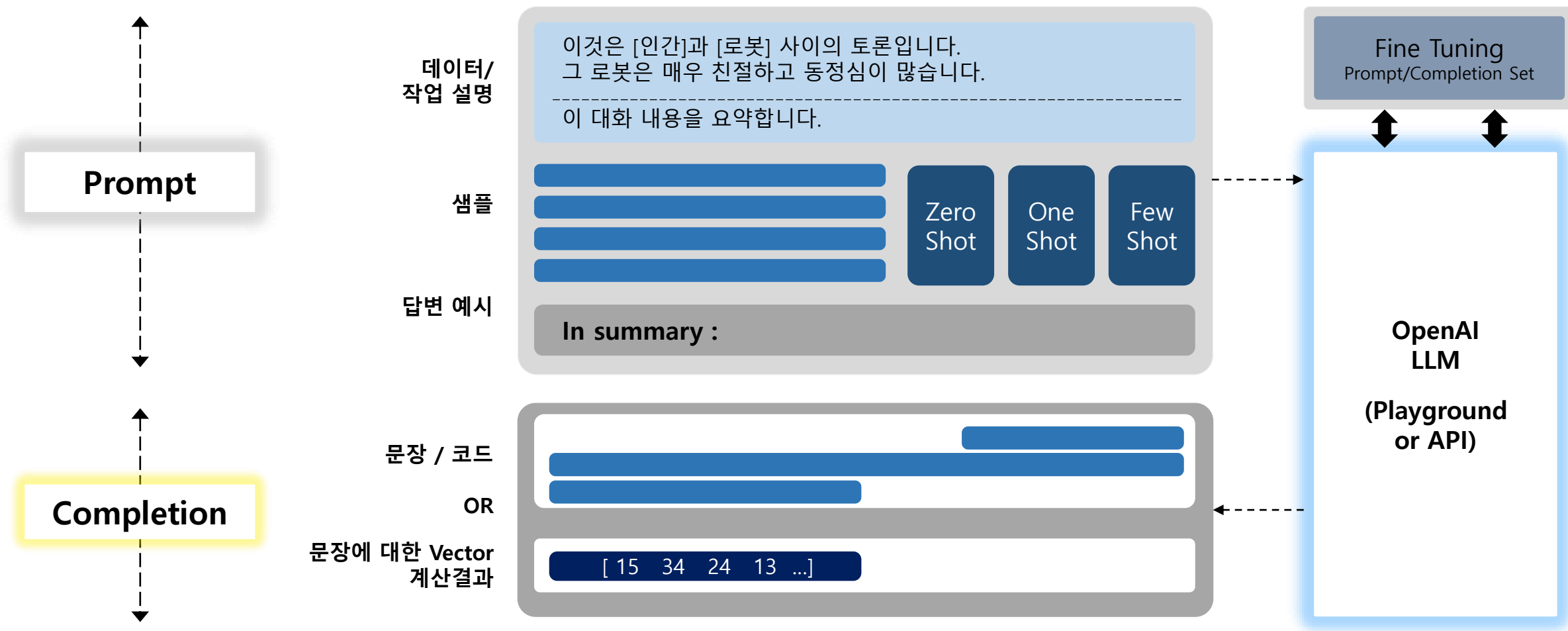
ChatGPT 개요

깃허브, 코덱스 기반 노코딩 AI 'Copilot' 정식 출시



ChatGPT 개요

사용자 입력(prompt)에 따라 ChatGPT내부에서 어떻게 처리되고 결과가 생성되는지 설명 합니다.



Vector 란?

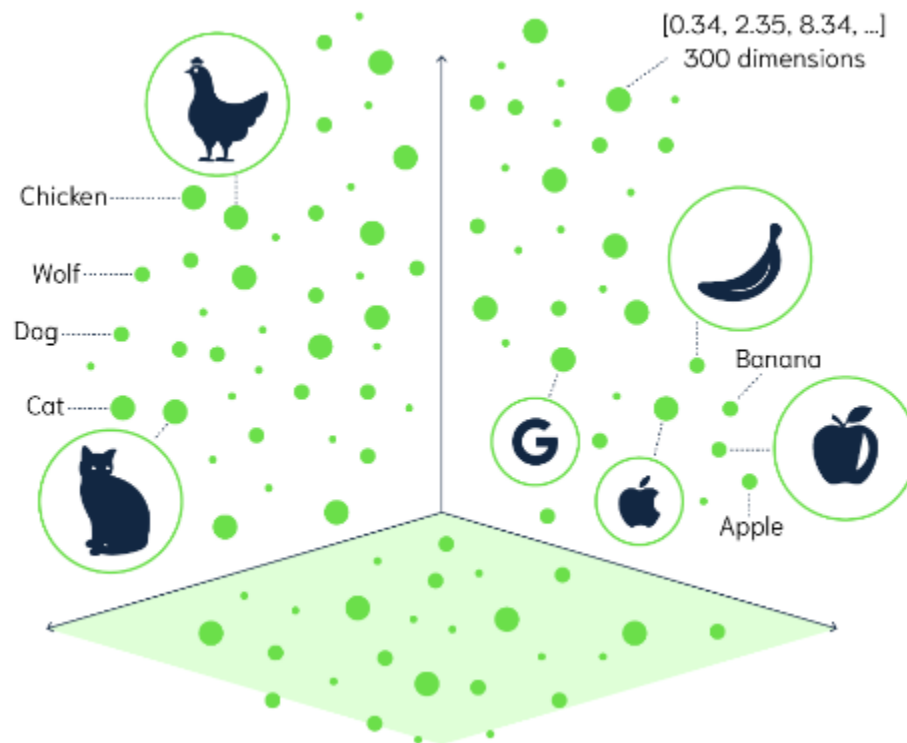
Vector 란?

Vector Store란?

3차원 벡터 값을 가지고 목표와의 거리를 가지고 유사성 검색을 위해 고안된 벡터 저장소입니다

유사성 검색: 벡터 스토어는 특정 벡터와 유사한 벡터를 찾는 데 사용할 수 있습니다. 예를 들어, 벡터 스토어를 사용하여 단어와 유사한 단어를 찾을 수 있습니다.

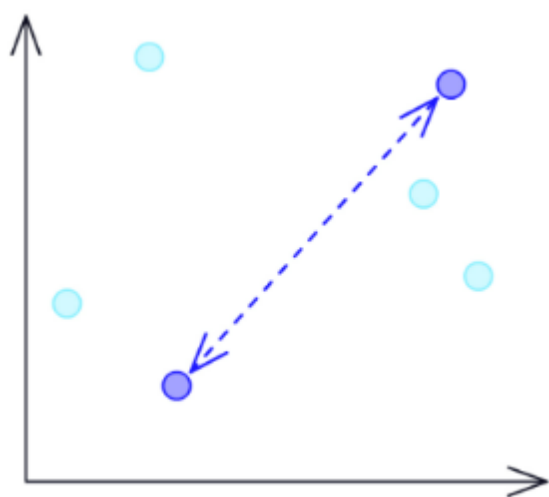
벡터 데이터베이스에서는 유사성 메트릭을 적용하여 쿼리와 **가장 유사한** 벡터를 찾습니다.



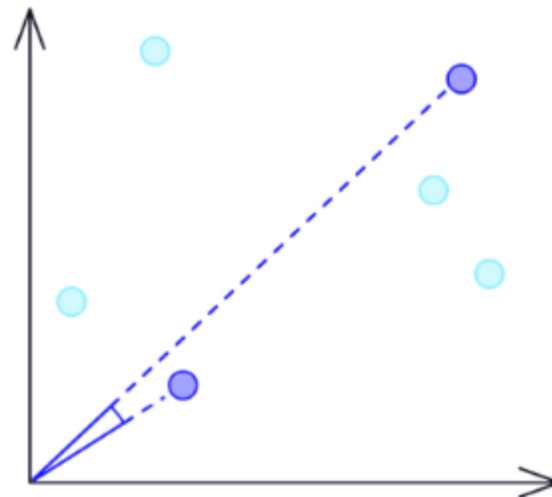
LLM이란?

유사성 측정

- **코사인 유사성(Cosine Similarity):** 벡터 공간에서 두 벡터 간의 각도의 코사인을 측정합니다. 범위는 -1에서 1까지이며, 여기서 1은 동일한 벡터를 나타내고, 0은 직교 벡터를 나타내고, -1은 정반대의 벡터를 나타냅니다. -> 각도가 작으면 유사
- **유클리드 거리(Euclidean distance):** 벡터 공간에서 두 벡터 사이의 직선 거리를 측정합니다. 범위는 0에서 무한대까지이며, 여기서 0은 동일한 벡터를 나타내고 값이 클수록 점점 더 다른 벡터를 나타냅니다. -> 거리가 가까우면 유사
- **내적:** 두 벡터 크기의 곱과 두 벡터 사이 각도의 코사인 값을 측정합니다. 범위는 $-\infty$ 에서 ∞ 까지이며 양수 값은 같은 방향을 가리키는 벡터를 나타내고 0은 직교 벡터를 나타내고 음수 값은 반대 방향을 가리키는 벡터를 나타냅니다.



Euclidean distance



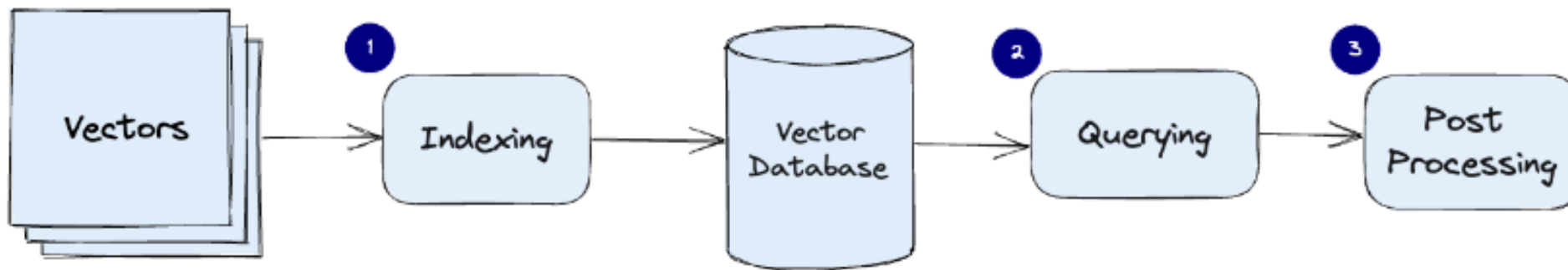
Cosine Similarity

이미지 출처 : <https://www.pinecone.io/learn/vector-database-old/>

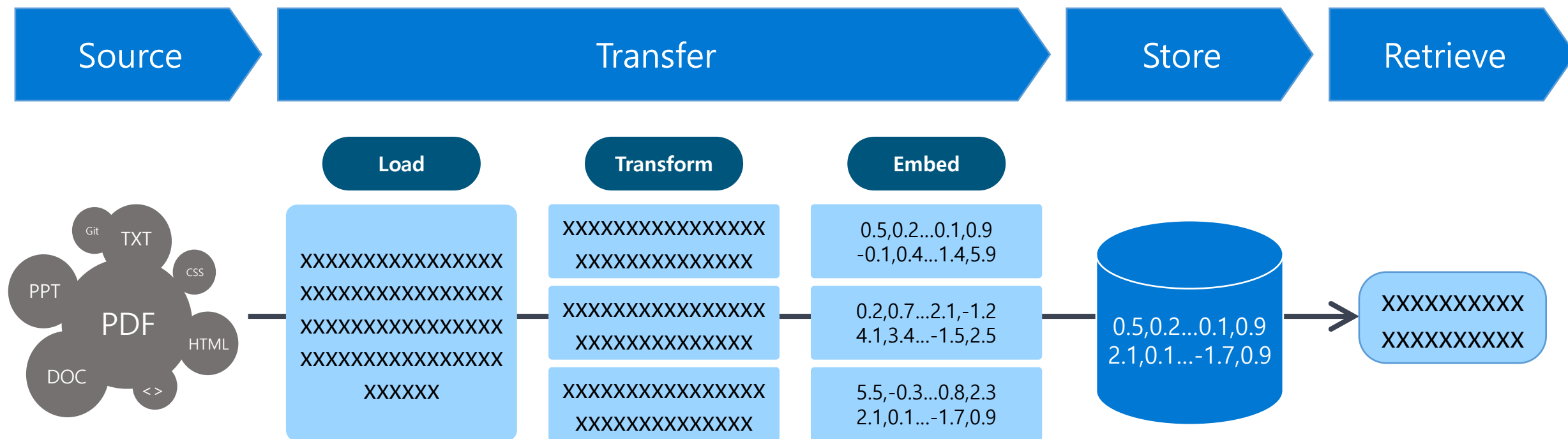
Vector 란?

Vector Store의 처리 방법

- 1.인덱싱** : 벡터 데이터베이스는 PQ, LSH 또는 HNSW와 같은 알고리즘을 사용하여 벡터를 인덱싱합니다. 이 단계는 더 빠른 검색을 가능하게 하는 데이터 구조에 벡터를 매핑합니다.
- 2.쿼리** : 벡터 데이터베이스는 가장 가까운 이웃을 찾기 위해 인덱스 쿼리 벡터를 데이터 세트의 인덱스 벡터와 비교합니다(해당 인덱스에서 사용하는 유사성 메트릭 적용).
- 3.후 처리** : 경우에 따라 벡터 데이터베이스는 데이터 세트에서 가장 가까운 최종 이웃을 검색하고 후 처리하여 최종 결과를 반환합니다. 이 단계에는 다른 유사성 척도를 사용하여 가장 가까운 이웃의 순위를 재지정하는 작업이 포함될 수 있습니다.



Vector 란?



- 사용자가 질문 파일 업로드
- Word, Excel, PDF, JSON, CSV 등 다양한 확장자 파일 지원

- 각 파일 확장자에 맞는 loader를 사용, 파일 Format에 맞게 Document를 Load
- Load된 문서는 각 텍스트에 가장 의미가 맞게 Vector화

- 구조화된 데이터를 저장. 검색하는 일반적인 방법은 Vector를 저장 방식
- Vector 저장소는 포함된 데이터를 저장 및 Vector 검색을 수행

- 의미론적 검색을 수행할 수 있는 형태로 저장된 Vector 파일은 사용자 쿼리로 조회될 수 있다.

Prompt Engineering 이란?

Token이란?

모델별 Token 수 한계

MODEL	Token Limit	Token per Minute	Requests per minute
Ada (GPT-3) : fastest	2,049	120,000	300
text-ada-001 (GPT-3)	2,049	120,000	300
Babbage (GPT-3)	2,049	120,000	300
test-Babbage-001 (GPT-3)	2,049	120,000	300
Curie (GPT-3)	2,049	120,000	300
text-curie-001 (GPT-3)	2,049	120,000	300
Davinci (GPT-3)	2,049	120,000	300
text-davinci-002 (GPT-3)	4,097	40,000	120
text-davinci-003 (GPT-3)	4,097	40,000	120
GPT-35-Turbo (ChatGPT)	4,096	120,000	300
GPT-4 (GPT-4)	8,192	120,000	12
GPT-4-32k (GPT-4)	32,768	120,000	12

PROMPT + COMPLETION 한도
입력 토큰 + 출력 토큰 한도

Prompt와 Completion이란?

프롬프트 구성

프롬프트의 구성 요소	설명
Instruction	모델이 수행하기를 원하는 특정 태스크 또는 지시 사항
Context	모델이 보다 더 나은 답변을 하도록 유도하는 외부 정보 또는 추가 내용
Input Data	답을 구하고자 하는 것에 대한 인풋 또는 질문
Output Indicator	결과물의 유형 또는 형식을 나타내는 요소

지시

Instruction, 모델이 수행하기를 원하는 지시문 또는 작업입니다.
모델에게 어떤 결과를 기대하는지 알려줍니다.

문맥

Context. 앞서 이야기 했던 주제에 관한 이야기.
단발성대화가 아닌 피드백을 통한 문맥가반의 발전형 대화

예시

"다음 단락 요약하기."
"[이름], [나이] 순으로 정리"
"업로드한 보고서를 참고하여"

Prompt와 Completion이란?

Prompt & Completion – 예를 들어 출력

PROMPT

TASK DESCRIPTION

영화 리뷰 감성 분류기입니다.

EXAMPLE

리뷰: "나는 이 영화를 좋아했다!"
리뷰 감성: 긍정적

리뷰: "나는 잘 모르겠다, 괜찮았던 거 같긴 한데.."
리뷰 감성: 중립적

리뷰: "정말 시간 낭비였어요, 이 영화는 추천하지 않아요."
리뷰 감성: 부정적

리뷰: "나는 이 영화를 정말 즐겼어요!"

OUTPUT INDICATOR

리뷰 감성:

COMPLETION

긍정적

Language
Model



Prompt Engineering 방법


Prompt Engineering 방법


Instruction과 Context를 구체적으로 명시

2023. 5. 8. 오전 8:57:30

As a novelist, your task is to create stories that captivate and engage readers for extended periods of time. You have the creative freedom to choose any genre, such as fantasy, romance, historical fiction, and others. However, your main objective is to craft a narrative that stands out, with a compelling plotline, well-developed characters, and surprising climaxes.

For your first assignment, you are tasked with writing [a science-fiction novel about a guy living on another planet, where he can play with futuristic technology, otherworldly creatures, and visionary concepts. Your story should transport the reader to this exciting new world, filled with adventure, mystery, and excitement. The goal is to create a science-fiction novel that is both thought-provoking and entertaining, leaving a lasting impression on your readers.]



 2023. 5. 8. 오전 8:58:10

Title: Whisperings of Osera

Chapter 1: Arrival on Osera

As the spacecraft touched down on the planet Osera, the ramp lowered, and Jonathan Rayker stepped onto the unfamiliar terrain. The atmosphere was breathable but tinged with a faint scent of ozone. He marveled at the vibrant colors of the flora that covered the landscape. Towering trees with luminescent leaves reached towards the sky, painting the horizon with shades of purple and blue.

Jonathan had been a star explorer for years, traversing the galaxy in search of new worlds and undiscovered civilizations. This time, he had been sent to investigate the mythical technologically-advanced utopia many called Osera.

Chapter 2: Strange Encounters

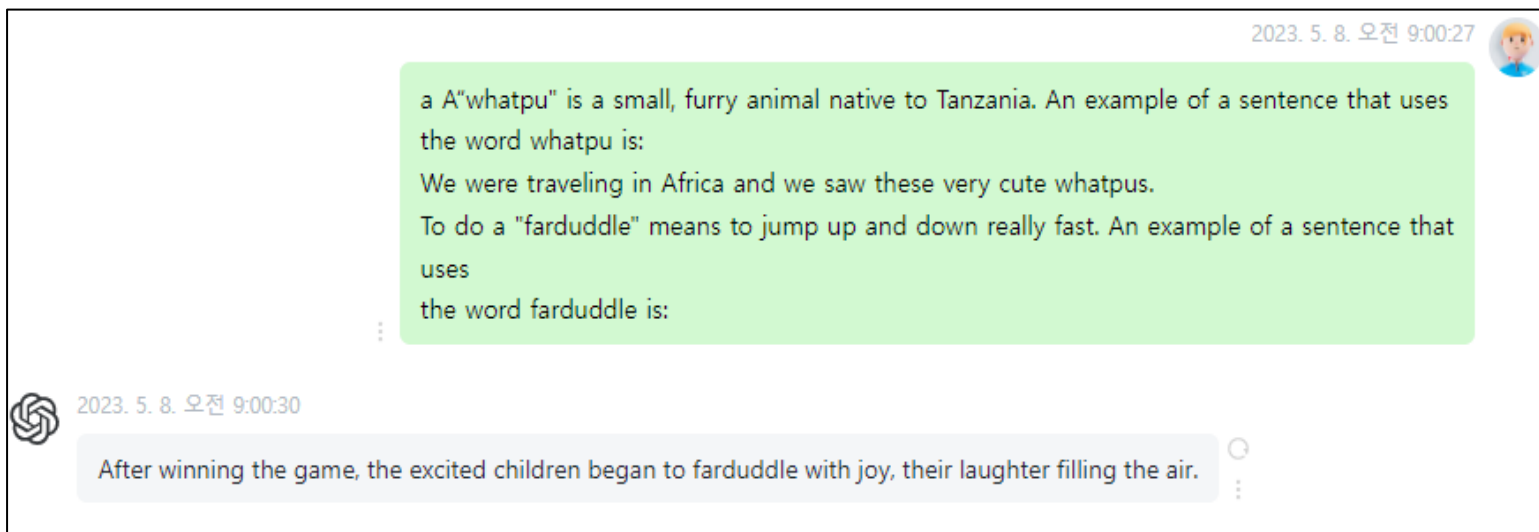
첫번째 문단에서는 AI에게 구체적으로 지시 (instruct)하였고, AI가 보다 명확히 instruction을 이해하는 데 필요한 맥락을 함께 명시

두번째 문단에서는 추가적으로 활용할 수 있는 내용을 서술하였습니다. 위와 같이 단순 지시만 하는 것이 아닌, 사용자가 정확하게 어떤 답을 기대하고 지시를 했는지까지 서술

Prompt Engineering 방법

Zero-Shot Prompting

추가 학습 또는 예제 데이터 없이 답변을 생성하게 하는 프레임워크로, 거대 언어 모델에게 아무런 instruction (지시) 없이 완수할 태스크를 주는 것을 의미



첫번째 'whatpu'라는 단어의 뜻과 해당 단어를 활용하여 문장을 생성하여 이를 하나의 예제로 제공

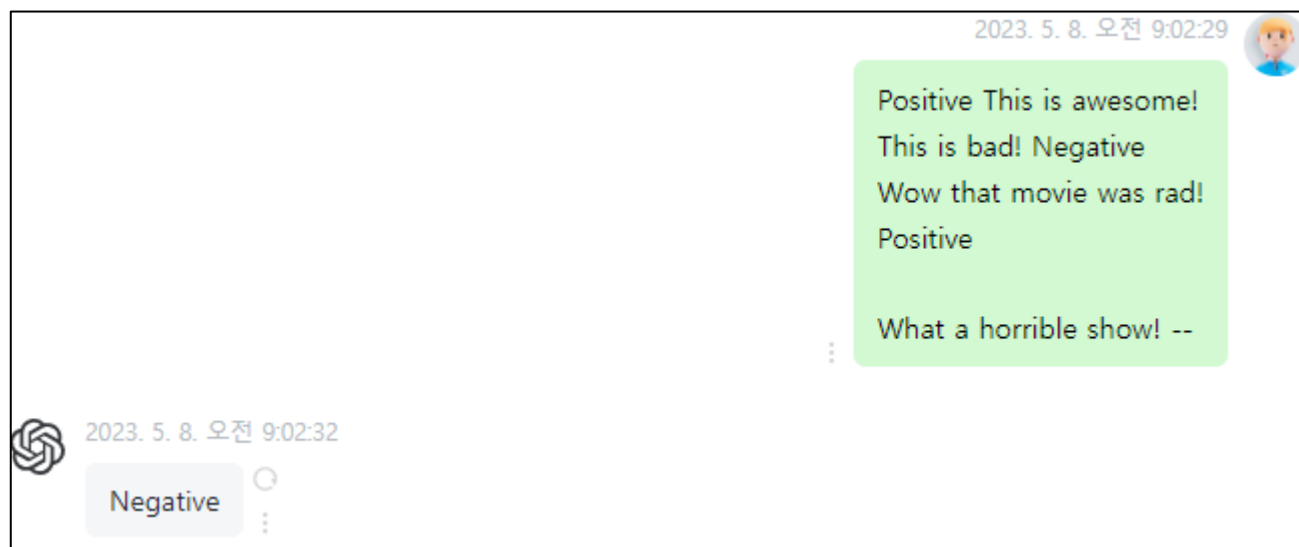


두번째 문단에서는 'farduddle'이라는 단어의 뜻을 주고 해당 단어를 활용하여 문장을 생성하라고 지시

Prompt Engineering 방법

One-Shot Prompting

하나의 예제 또는 템플릿을 기반으로 답변을 생성하게 하는 기법



다섯개의 예제를 먼저 제공

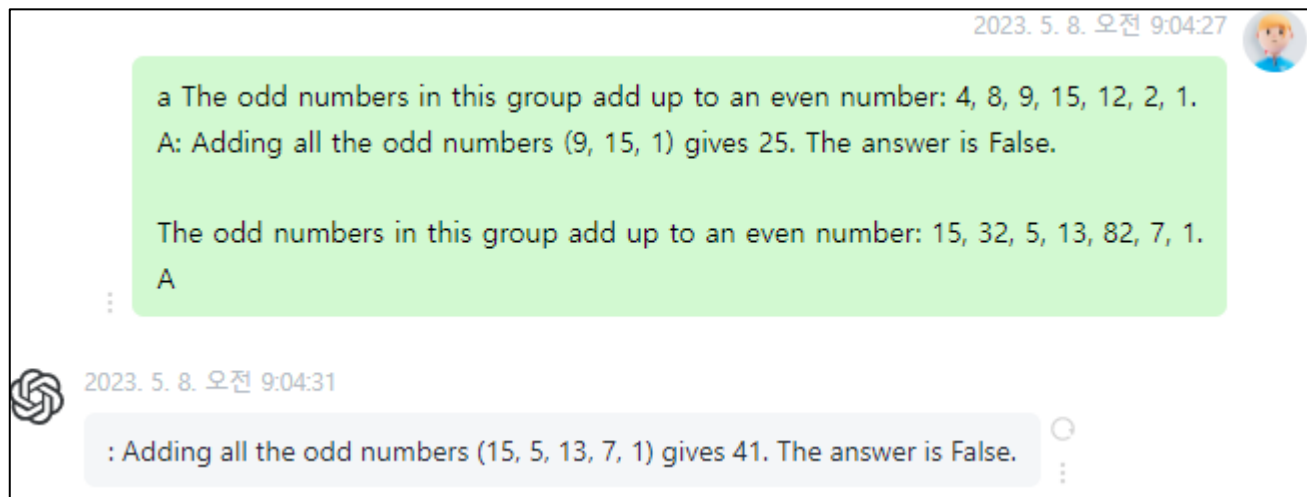


마지막에 오는 문장에 알맞는 감정을 추론하게 하였다

Prompt Engineering 방법

Few-Shot Prompting

두개에서 다섯개의 예제를 바탕으로 답변을 생성하게 하는 기법으로, 프롬프트 앞단에 One-Shot Prompting 기법 보다 조금 더 AI에게 직접적으로 원하는 답변에 도달할 수 있도록 유도



먼저 질문과 답변 예시를 주는 것은 동일하나, 답변까지 어떻게 도달하였는지 추론 과정을 포함하여 예제를 제공




그렇게 되면 AI가 이번에는 단순히 다음 질문에 대한 답변만을 주는 것이 아니고, 어떻게 해당 답변에 도달하였는지 스스로 설명

Prompt Engineering 방법

CoT (Chain-of-Thought)

단지 답변을 내놓기 위한 것이 아닌, 답변에 도달하는 과정을 학습시키는 것을 목적으로 본 질문 전에 미리 태스크와 추론 과정을 포함한 답변 예제를 AI에게 제공


2023. 5. 8. 오전 9:06:13



I went to the market and bought 10 apples. | gave 2 apples to the neighbor and 2 to the repairman. | then went and bought 5 more apples and ate 1. How many apples did I remain with?

Let's think step by step.

⋮

 2023. 5. 8. 오전 9:06:25

Sure, let's break it down step by step:

1. You bought 10 apples.
2. You gave away 2 apples to the neighbor and 2 to the repairman. So, you gave away a total of 4 apples.
3. You then bought 5 more apples.
4. Finally, you ate 1 apple.



CoT 프롬프트에 "Let's think step by step (차근차근 생각해보자)" 한 문장만 추가한 것이 전부인데 이번에는 추론 과정을 '단계'별로 풀어서 설명

Prompt Engineering 방법

CoT Prompting의 장점

1. 다단계 문제 분해

설명: 모델은 여러 중간 단계를 거쳐 복잡한 문제를 해결할 수 있음.

효과: 더 많은 추론 단계가 필요한 문제에 적절한 계산 자원을 할당 가능.

2. 모델 행동의 해석 가능성

설명: 모델이 어떻게 특정 답변에 도달했는지 이해할 수 있는 창을 제공.

효과: 오류가 발생했을 때 추론 경로를 디버깅하고 수정하는 데 유용.

3. 다양한 작업 적용 가능성

설명: CoT 추론은 수학, 상식적 추론, 상징적 조작 등 다양한 작업에 적용될 수 있음.

범위: 언어를 통해 인간이 해결할 수 있는 거의 모든 종류의 작업에 적용 가능.

4. 큰 규모의 언어 모델에서의 쉬운 유도

설명: 충분한 규모의 언어 모델에 CoT 시퀀스 예시를 넣음으로써 CoT 추론을 간단히 유도할 수 있음.

방법: 기존의 큰 언어 모델에 CoT 패턴을 소개하여 추론 방식을 안내.

기타등등


GPT 시리즈

GPT-4와 GPT-3.5 Turbo의 경우 약 20% 정도의 성능 차이를 보입니다.(영문 기준)

Benchmark	GPT-4 Evaluated few-shot	GPT-3.5 Evaluated few-shot	LM SOTA Best external LM evaluated few-shot	SOTA Best external model (incl. benchmark-specific training)
MMLU 57개 과목, 객관식 문항(전문&학술)	86.4% 5-shot	70.0% 5-shot	70.7% 5-shot U-PaLM	75.2% 5-shot Flan-PaLM
HellaSwag 일상적인 사건에 대한 상식적인 추론	95.3% 10-shot	85.5% 10-shot	84.2% LLAMA (validation set)	85.6% ALUM
AI2 Reasoning Challenge(ARC) 초등학교 객관식 과학문제 도전 세트	96.3% 25-shot	85.2% 25-shot	84.2% 8-shot PaLM	85.6% ST-MOE
WinoGrande 대명사 분해에 관한 상식적 추론	87.5% 5-shot	81.6% 5-shot	84.2% 5-shot PALM	85.6% 5-shot PALM
HumanEval 파이썬 코딩 태스크	67.0% 0-shot	48.1% 0-shot	26.2% 0-shot PaLM	65.8% CodeT + GPT-3.5
DROP (f1 score) 독해력과 산술력	80.9 3-shot	64.1 3-shot	70.8 1-shot PaLM	88.4 QDGAT

출처: <https://openai.com/research/gpt-4>


Google Gemini




Google
구독자 1160만명

구독

혼작용




동영상 더보기



Google
구독자 1160만명


구독

혼작용



The bird is swimming in the water.
It has a long neck and beak.
It is a duck.


동영상 더보기



Google
구독자 1160만명


구독

혼작용



I'm not sure what material
it's made of, but it looks like it
might be rubber or plastic.


동영상 더보기




Gemini 실습: 멀티모달 AI와 상호작용

나중에 시...

공유



동영상 더보기



Gemini 실습: 멀티모달 AI와 상호작용

나중에 시...

공유

No, the correct order is
Sun, Earth, Saturn.

동영상 더보기

	Gemini Ultra	Gemini Pro	GPT-4	GPT-3.5	PaLM 2-L	Claude 2	Inflection-2	Grok 1	LLAMA-2
MMLU Multiple-choice questions in 57 subjects (professional & academic) (Hendrycks et al., 2021a)	90.04% CoT@32*	79.13% CoT@8*	87.29% CoT@32 (via API**)	70% 5-shot	78.4% 5-shot	78.5% 5-shot CoT	79.6% 5-shot	73.0% 5-shot	68.0%***
GSM8K Grade-school math (Cobbe et al., 2021)	94.4% Maj1@32	86.5% Maj1@32	92.0% SFT & 5-shot CoT	57.1% 5-shot	80.0% 5-shot	88.0% 0-shot	81.4% 8-shot	62.9% 8-shot	56.8% 5-shot
MATH Math problems across 5 difficulty levels & 7 subdisciplines (Hendrycks et al., 2021b)	53.2% 4-shot	32.6% 4-shot	52.9% 4-shot (via API**)	34.1% 4-shot (via API**)	34.4% 4-shot	—	34.8% 4-shot	23.9% 4-shot	13.5% 4-shot
BIG-Bench-Hard Subset of hard BIG-bench tasks written as CoT problems (Srivastava et al., 2022)	83.6% 3-shot	75.0% 3-shot	83.1% 3-shot (via API**)	66.6% 3-shot (via API**)	77.7% 3-shot	—	—	—	51.2% 3-shot
HumanEval Python coding tasks (Chen et al., 2021)	74.4% 0-shot (IT)	67.7% 0-shot (IT)	67.0% 0-shot (reported)	48.1% 0-shot	—	70.0% 0-shot	44.5% 0-shot	63.2% 0-shot	29.9% 0-shot
Natural2Code Python code generation. (New held-out set with no leakage on web)	74.9% 0-shot	69.6% 0-shot	73.9% 0-shot (via API**)	62.3% 0-shot (via API**)	—	—	—	—	—
DROP Reading comprehension & arithmetic. (metric: F1-score) (Dua et al., 2019)	82.4 Variable shots	74.1 Variable shots	80.9 3-shot (reported)	64.1 3-shot	82.0 Variable shots	—	—	—	—
HellaSwag (validation set) Common-sense multiple choice questions (Zellers et al., 2019)	87.8% 10-shot	84.7% 10-shot	95.3% 10-shot (reported)	85.5% 10-shot	86.8% 10-shot	—	89.0% 10-shot	—	80.0%***
WMT23 Machine translation (metric: BLEURT) (Tom et al., 2023)	74.4 1-shot (IT)	71.7 1-shot	73.8 1-shot (via API**)	—	72.7 1-shot	—	—	—	—

GPT-4 Turbo(Preview)



- **컨텍스트 길이 확장:** GPT-4 Turbo는 최대 128,000 컨텍스트 길이를 지원해, 기존 모델 대비 4배에서 16배까지 긴 텍스트 처리 가능 (약 300페이지)
- **경제적인 비용:** 입력 토큰 \$0.01, 출력 토큰 \$0.03으로 기존 GPT-4 대비 약 27% 저렴
- **JSON 모드:** JSON 형식의 유효한 출력을 지원하여 웹 개발에 용이하고 재현 가능한 결과 제공
- **최신 학습 데이터:** 2023년 4월 데이터로 훈련되어, 이전 모델보다 더 최신 정보 반영
- **멀티모달 기능:** DALL·E 3 API를 통한 이미지 생성과 TTS 기능을 지원하는 API 제공

