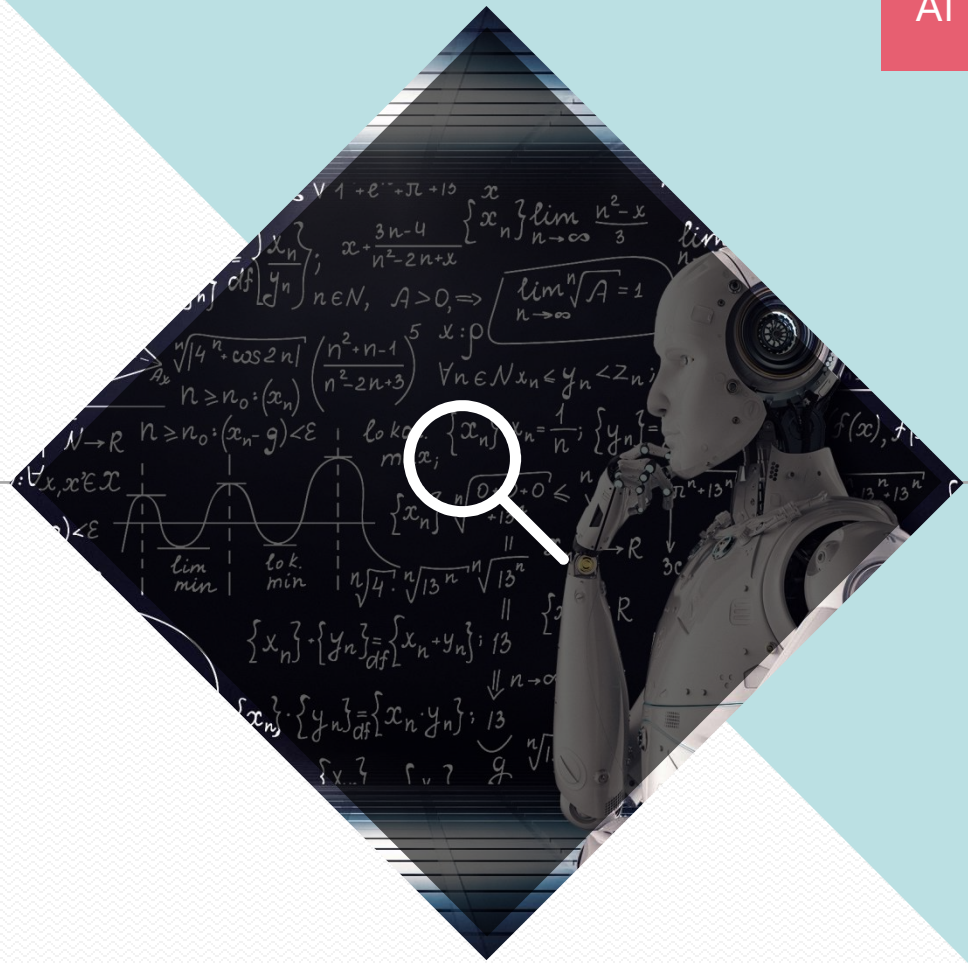
AI TRAINING SERIES

LLM & Edge AI



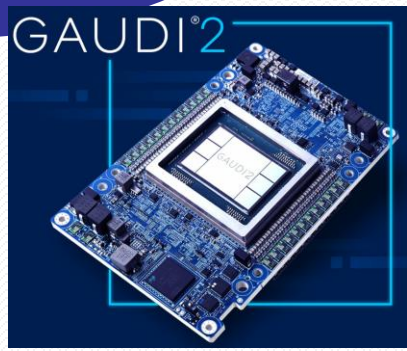
Contents

AI Trends

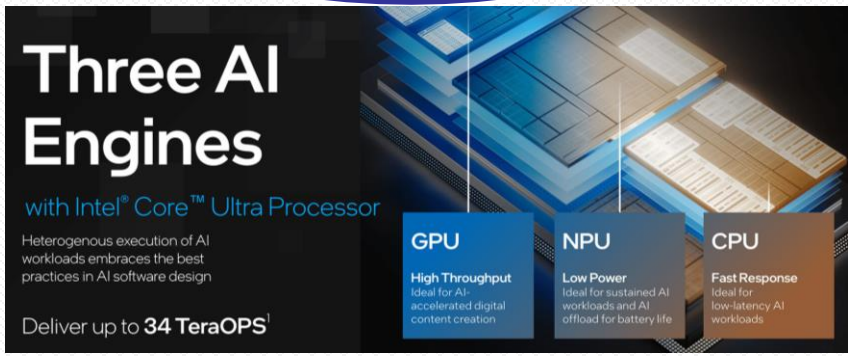
https://www.youtube.com/watch?v=hw-kesm-ktI

Intel CEO Pat Gelsinger speaks at the World Economic Forum in Davos - 1/17/2024

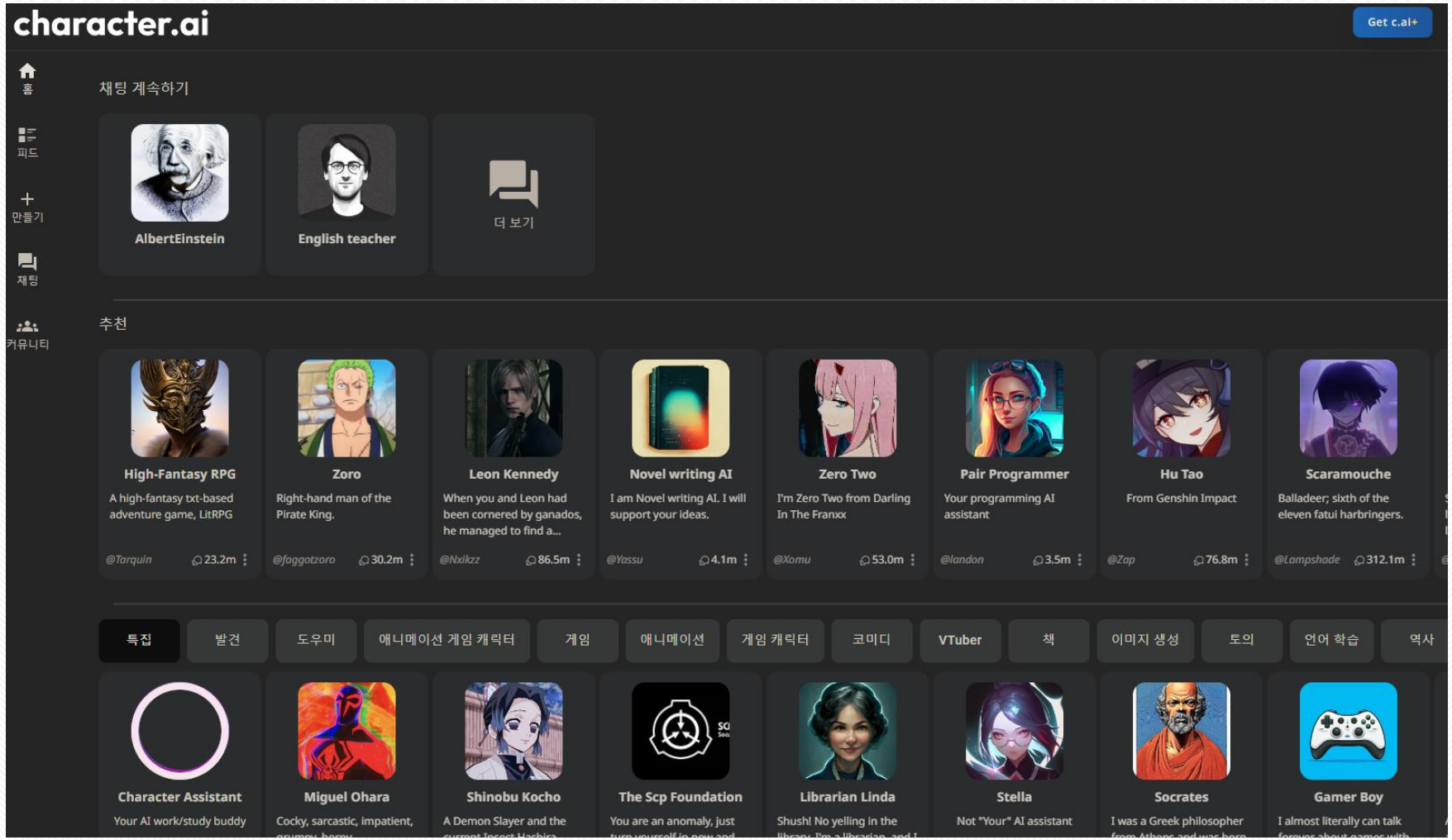
Explosive **Demand for AI computing** over the next decade



LLM models from Cloud to Edge.
AI PC!!



캐릭터와 채팅 - https://beta.character.ai



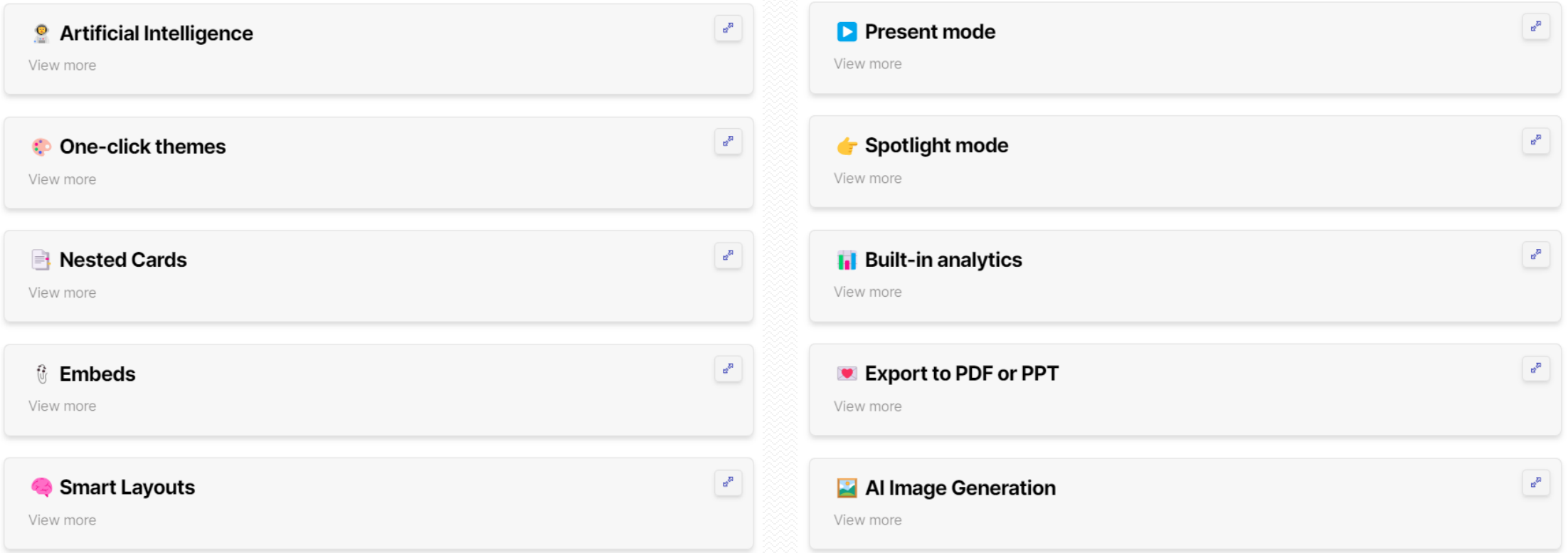
현재 무료 이용 가능

Youtube 요약 - lilys.ai

Lilys

현재 무료 이용 가능

Power Point AI 조수 - Gamma



현재 무료 이용 가능

AI Video 생성기 - Heygen

주요 기능

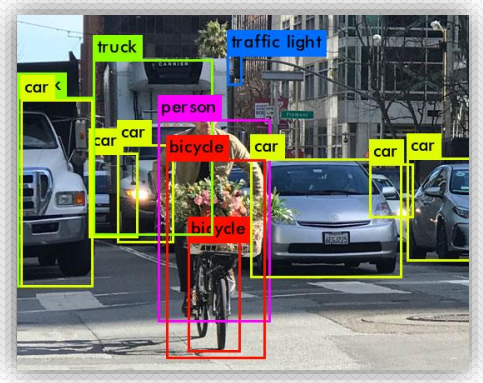
- AI Video 를 생성해 주고 40여개 이상의 다른 언어로 변환해 줌
- 본인의 모습과 제스처, 목소리를 기반으로 아바타 생성
- 유료 모델이나 가격이 착함

Contents

생성형 AI란?

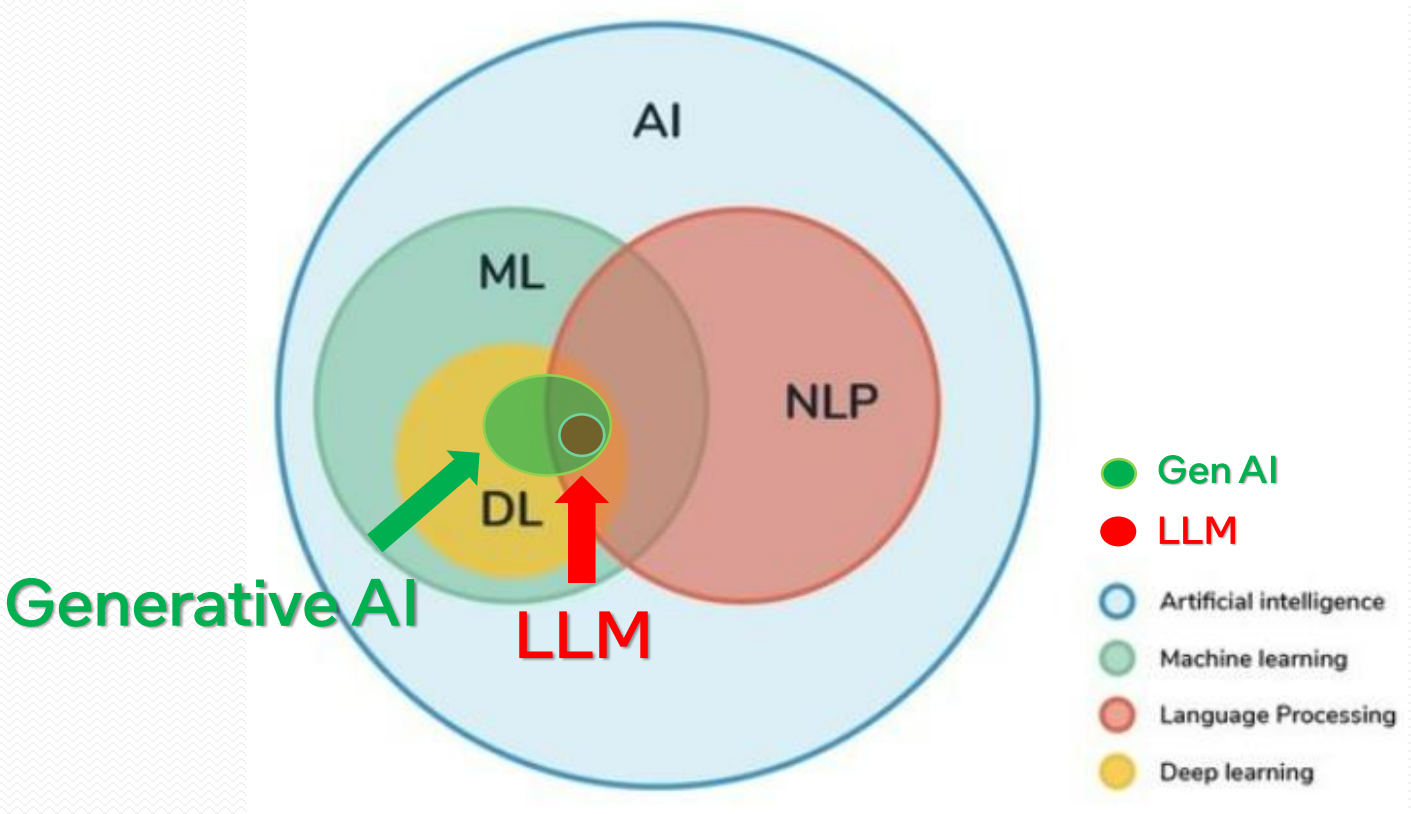
생성형 AI는 데이터를 기반으로 새로운 콘텐츠를 자동 생성하는 인공지능 시스템

Discriminative	Generative
매우 구체적인 규칙을 따라 특정 과업만을 수행	사전에 학습한 데이터를 바탕으로 새로운 것을 창조
레이블링 된 데이터셋을 기반으로 학습	다양한 컨텐츠가 지닌 대량 데이터를 학습
규칙을 배우고, 문제를 해결하는 것에 집중	기존과는 다른 예상 밖의 창의적 행동으로 다양화



LLM 은 무엇이고 NLP 와 생성형 AI 와의 관계

LLM (Large Language Model) 은 엄청나게 많은 데이터를 기반으로 훈련되어 마치 인간처럼 텍스트를 이해하고 생성하는 데 매우 능숙한 생성형 AI 모델



LLM 의 주요 제약사항

거짓 정보 (Hallucination)

LLM은 때때로 부정확하거나 오래된 정보를 제공할 수 있으며, 사실 확인이 없는 경우 오류를 전파할 수 있으며 이를 마치 진짜 정보처럼 전달함.

데이터 편향과 공정성 문제

LLM은 훈련 데이터에 내재된 편향을 학습할 수 있으며, 이는 결과물에도 영향을 줌. 이는 훈련 데이터가 다양성이 부족하거나 특정 인구 집단의 관점을 과대 대표하는 경우 발생하며 현재 LLM 모델 훈련 데이터는 English 가 다수.

컴퓨킹 자원 집약적

기본적으로 거대한 언어모델을 사용하므로 LLM 을 훈련하고 실행하는 데는 상당한 양의 컴퓨팅 자원이 필요함. 이는 고성능 GPU, 대규모 저장소, 높은 전력 소모 등이 필요하여 비용과 환경적 영향이 큼.

윤리적, 법적 문제

혐오발언, 증오발언, 욕설 등에 노출된 데이터로 학습될 수 있으며 이를 통해 윤리적으로 부적절한 내용을 포함하여 여러 갈등을 유발 할 수 있음.

가장 핫한 LLM 모델들 소개

closed source

open source

2024

OpenAI GPT Store

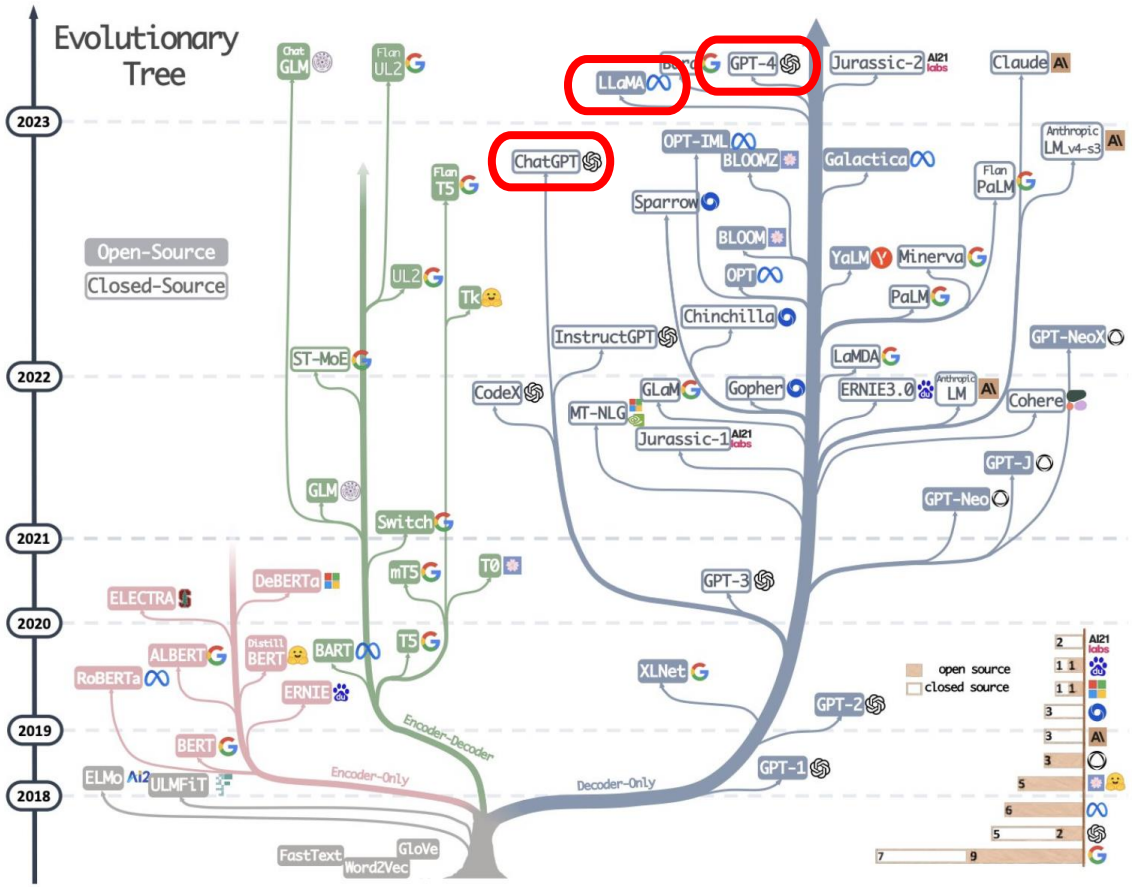
OpenAI ChatGPT4.0

Gemini

NAVER CLOVA X

META LLAMA 2

upstage Solar



Contents

ChatGPT API 사용하기

https://colab.research.google.com/drive/12faqNFiHfwzKD2jUJ0YuVD_eYZhZyXo1?usp=sharing

https://platform.openai.com/docs/overview

https://github.com/openai/openai-cookbook

Gemini API 사용하기

https://colab.research.google.com/drive/1yaGy9HUfx7H7DeY79dXBA8JnveTLtlPQ?usp=sharing

https://ai.google.dev/docs?_gl=1*1ephda3*_up*MQ..*_ga*Mjc2NDIyMTYwLjE3MDgwNjI3MzE.*_ga_P1DBVKWT6V*MTcwODA2MjczMS4xLjAuMTcwODA2MjczMS4wLjAuMA..

Whisper 모델로 Speech 2 Text 적용하기

https://colab.research.google.com/drive/1wGGBrwQX1RaTdcgWcCcuD5a_3kvwA2TD?usp=sharing

https://github.com/openai/whisper

Contents

AI Open Community

AI open community는 인공지능 기술을 공유, 개발 및 혁신할 수 있는 공간

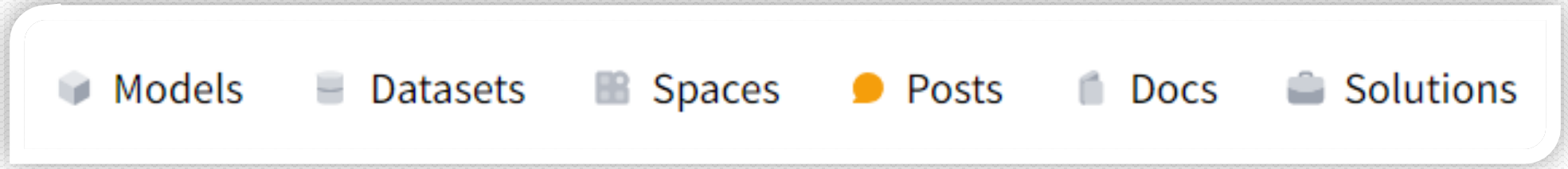
- 어떻게 활용?
  - 최신 AI model 들을 살펴보고 활용할 수 있음
  - Dataset 공유
  - 전 세계 연구자와 개발자와의 네트워킹 기회

- 왜 Big Tech 회사들이 적극적으로 참여할까?
  - 기술 리더쉽을 강화해 브랜드 가치와 업계 내 영향력 향상
  - 기술 표준을 설정해 장기적으로 기업 영향력 확대
  - 비용 절감 및 리소스 최적화

- 유명한 생성형 AI Open Community
  - Hugging Face - https://huggingface.co/
  - OpenAI Community - https://community.openai.com/
  - GitHub AI Projects - https://github.com/
  - Kaggle - https://www.kaggle.com/

Hugging Face 알아보기

생성형 AI 를 배우기 위한 최고의 Open AI Community



https://huggingface.co

https://huggingface.co/tasks

Hugging Face 실습

https://colab.research.google.com/drive/1UY9YHzxRPnYA7uqr1dQuD0jwUdH6P2Rl?usp=sharing

https://huggingface.co/google-t5/t5-small

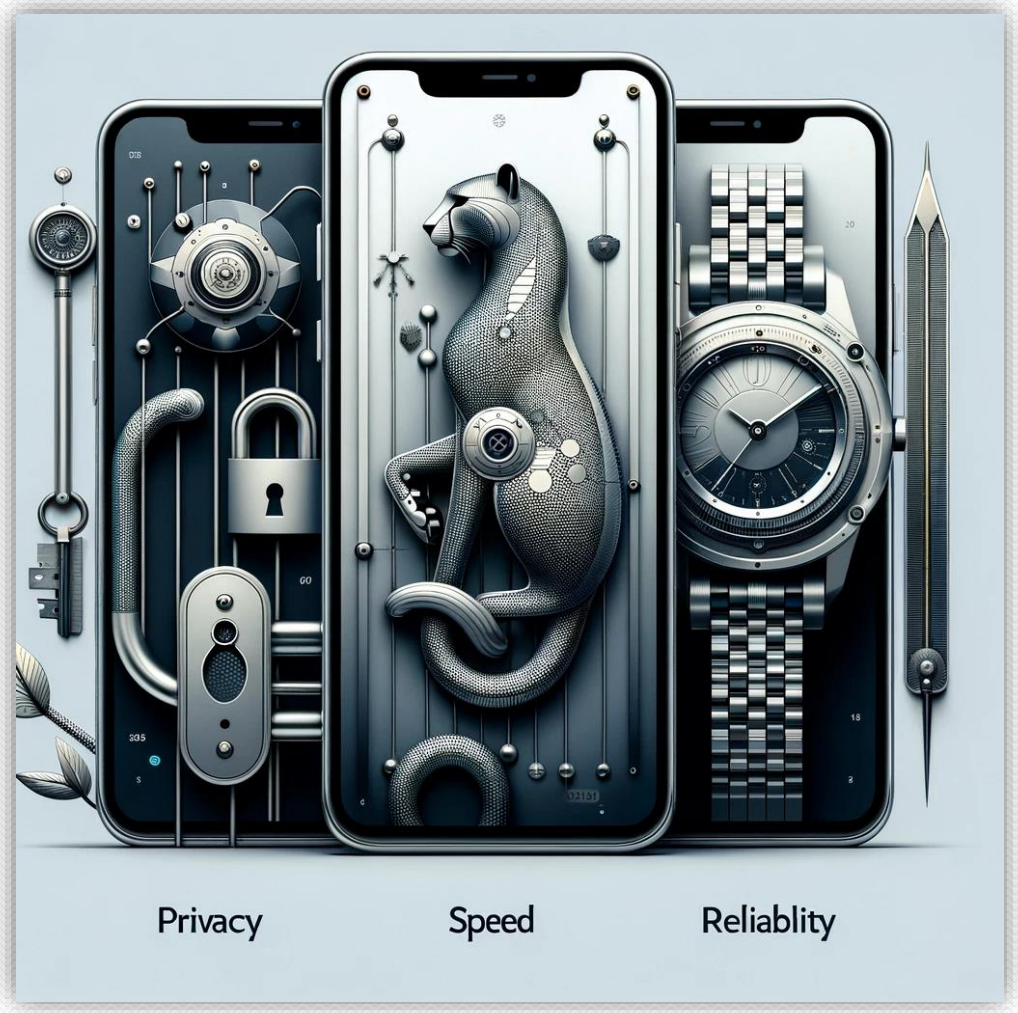
Contents

Edge AI 란?

PC, Mobile, Smart watch 등 사용자가 직접 사용하는 장비에서 on-device 로 AI 를 실행시키는 것으로 AI 실행을 위해 Cloud 와 연동하지 않아 별도의 Network가 필요 없음

Edge AI 의 장점

- Privacy/Security (정보 보안)
  - 데이터 처리가 기기 내에서 이루어 지기 때문에 사용자 정보가 외부 Cloud 등으로 전송되지 않음
  - 예시: 얼굴인식 잠금 해제, 음성 비서
- Speed (빠른 속도)
  - 외부 서버에서 AI 실행을 위해 관련 데이터를 네트워크를 통해 전송할 필요가 없어, 네트워크 지연이 없음
  - 예시: 실시간 번역, 게임 내 AI
- Reliability (신뢰성)
  - 네트워크 연결이 불안정하거나 없는 환경, 또는 AI 서버에 이상 등 예측 불가능한 상황에서 오는 오동작 문제 없이 AI 기능이 지속 작동
  - 예시: 오프라인 음성 명령
- Energy Efficiency (에너지 효율성)
  - 데이터를 기기 내에서 처리함으로써 데이터 전송에 필요한 에너지 소비 감소
- Customization (맞춤형 최적화)
  - 기기가 사용자의 행동 및 선호도를 학습하여, 사용자에게 맞춤화된 경험과 서비스를 제공



Edge AI 와 LLM 의 결합

AI 전용 Unit 필수

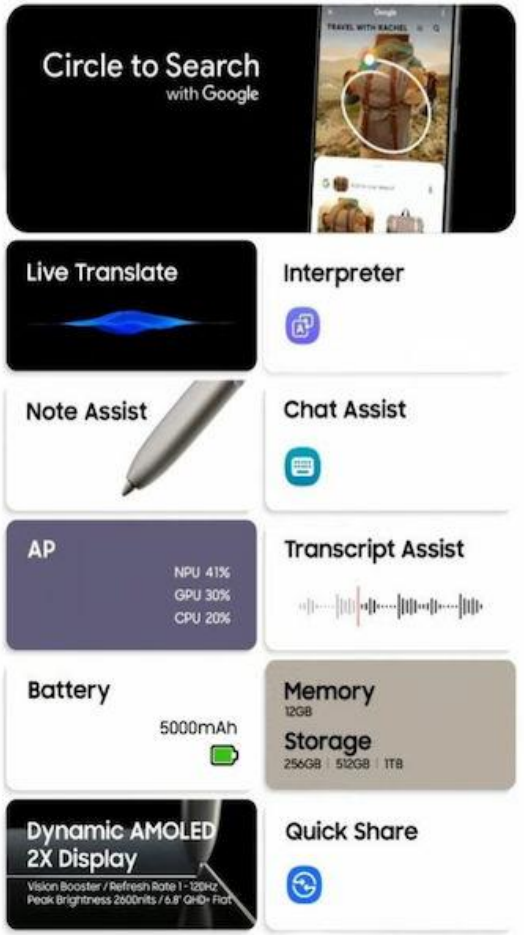
AI 처리만을 위한 전용 process 가 내장되어 On-device AI 성능을 향상 및 전반적으로 향상된 HW spec 이 필히 요구됨

LLM 모델 최적화

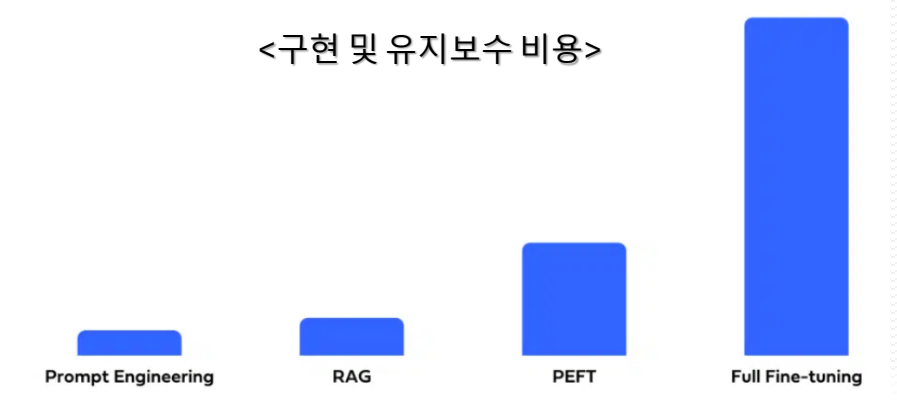
LLM 모델의 사용 목적에 맞게 최적화 / 경량화 추세

LLM 모델 업데이트

언어는 지속적으로 발전하고 새로운 용어와 표현이 등장하기 때문에, LLM을 최신 상태로 유지하는 것이 중요. Edge 기기에서도 지속적인 학습과 모델 업데이트를 지원하는 메커니즘이 필요



LLM Customization (최적화)



기술적 복잡성

컴퓨팅 자원

시간 & 성능

Prompting

RAG

Retrieval-Augmented Generation

PEFT

Parameter Efficient Fine-Tuning

Full Fine Tuning

Contents

Llama2-7B 모델로 On-device AI LLM 구현

https://github.com/facebookresearch/llama

https://github.com/facebookresearch/llama-recipes

LLM 모델의 주요 tuning parameter 이해

https://platform.openai.com/playground

Temperature

- temperature 매개변수는 모델이 생성할 텍스트의 무작위성을 조정
- 낮은 temperature 값은 보다 확실한(덜 다양한) 텍스트를 생성하는 반면, 높은 temperature 값은 더 다양하고 예측 불가능한 텍스트를 생성
- 예시: 모델에게 "고양이는" 이라는 문구로 문장을 시작하라고 요청
  - temperature=0.1 (낮은 값): 모델은 아마도 "고양이는 귀엽다" 와 같이 일반적이고 예측 가능한 문장을 생성
  - temperature=2.0 (높은 값): 모델은 "고양이는 파란색 눈물을 흘리며 재즈 음악을 연주한다"와 같은 매우 독창적이고 예측하기 어려운 문장을 생성

Top-p

- 모델이 다음 단어를 선택할 때 고려할 확률 분포의 부분 집합을 결정
- 예시: "고양이는" 이후의 단어를 선택하는 상황. 모델의 예측에 따르면, "잠을", "귀엽다", "뛰어다닌다", "낮잠을", "우아하다"가 다음 단어로 나올 가능성이 가장 높음
  - top-p=0.8: 이 경우, 모델은 "잠을", "귀엽다", "뛰어다닌다"와 같이 누적 확률이 80%에 도달할 때까지의 단어들만 고려할 것입니다. 이 범위 내에서 무작위로 다음 단어를 선택
  - top-p=0.95: 더 많은 단어, 예를 들어 "잠을", "귀엽다", "뛰어다닌다", "낮잠을"까지 포함될 수 있으며, 이는 생성된 텍스트에 더 많은 다양성을 제공

RAG 를 활용한 LLM model 최적화

https://colab.research.google.com/drive/1oPdaPohIK3CMsBUMrVONx8G3GzOcjSJE?usp=sharing

THANK YOU