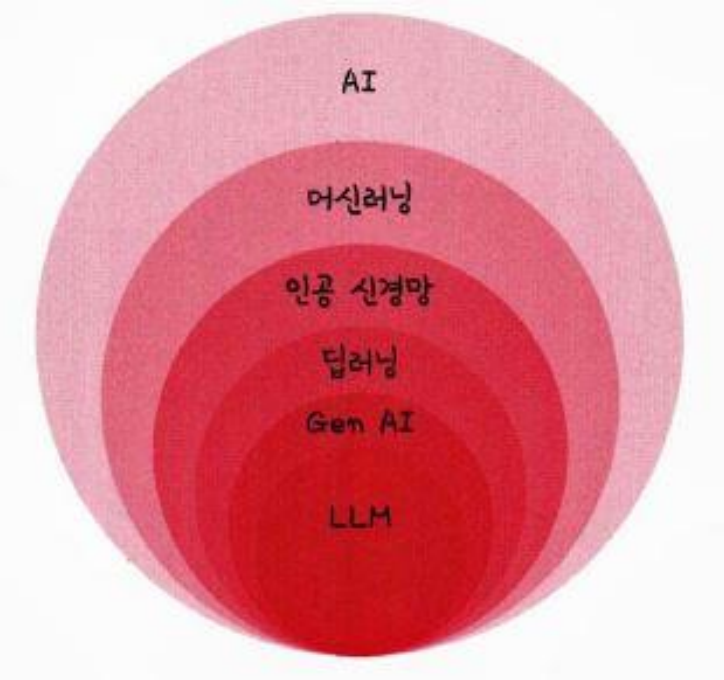
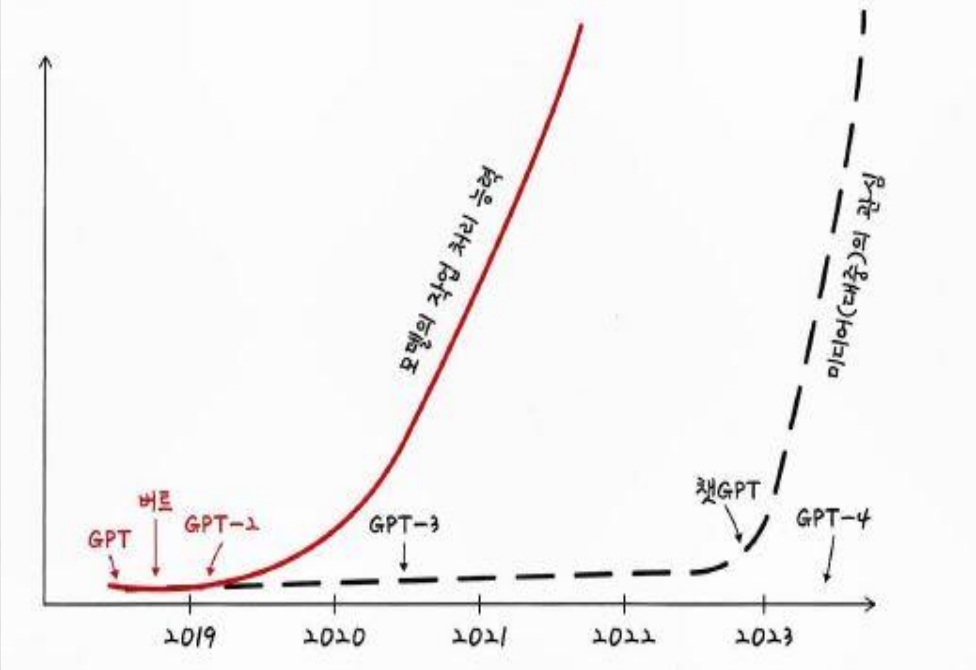
# LLM(Large Language Model)

# LLM 개념

## AI와 LLM

- LLM은 인공지능의 한 분야로, 대규모 데이터로 학습한 결과를 이용해서 인간의 언어를 처리하고, 생성하며, 맥락을 이해하는 데 사용



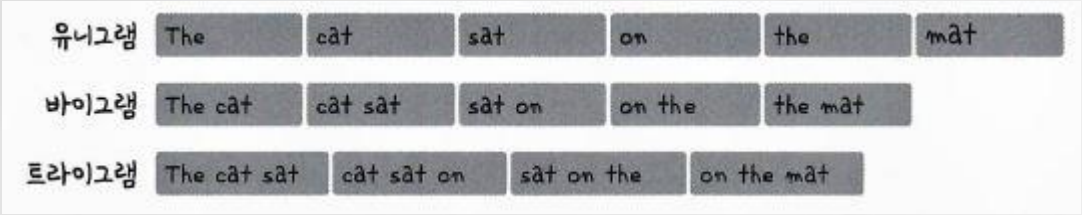
# 언어 모델

## 통계적 언어 모델

- 통계적 방법은 컴퓨터가 문장이나 단어를 얼마나 자연스럽게 표현할지를 수학적으로 계산
- 언어 모델에서 사용하는 확률/통계적 방법으로 'n-gram'이 있음
- N-gram은 일련의 단어나 문자가 얼마나 자주 함께 나타나는지를 살펴보는 방법
- 이때 'n'은 연속적으로 고려되는 단어의 수를 의미

# 언어 모델

## 통계적 언어 모델

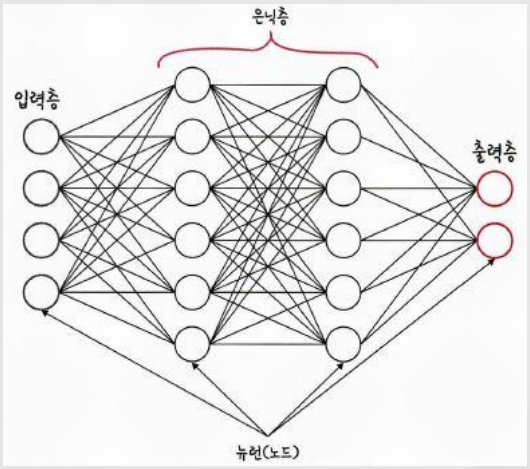



- "The cat sat on the mat"
- 1-gram(유니그램): 전체 문장을 각각의 단어로 나눔 따라서 유니그램(unigram)은 'The', 'cat', 'sat', 'on', 'the', 'mat'
- • 2-gram(바이그램): 전체 문장을 두 단어씩 나눕니다. 따라서 바이그램(bigram)은 "The cat', 'cat sat', 'sat on', 'on the', 'the mat'이 됨 참고로 단어가 겹쳐야 함
- • 3-gram(트라이그램): 전체 문장을 세 단어씩 나눔 따라서 트라이그램(trigram)은 'The cat sat', 'cat sat on', 'sat on the', 'on the mat'이 됨. 트라이그램 역시 단어가 중복되어 분할

# 언어 모델

## 신경망 언어 모델

- 패턴을 분석하는 것이 신경망 언어 모델
- 신경망은 일반적으로 입력층(Input Layer), 하나 이상의 은닉층(Hidden Layer), 출력층 (Output Layer)으로 구성
  - 입력층에서 외부로부터 데이터를 받아들이고
  - 은닉층에서 데이터를 처리 하여 다양한 특성과 패턴을 학습
  - 출력층에서 최종 결과를 생성



# 언어 모델

## 신경망 언어 모델

- RNN ( Recurrent Neural Networks )
  - 시퀀스 데이터 처리에 적합하며 과거의 정보가 현 재의 결정에 영향을 미칠 수 있음
  - RNN을 이용하면 번역은 물론 주식 가격, 날씨 변화 등 시간에 따라 변화하는 데이터를 분석하여 미래를 예측
  - 하지만 과거의 데이터를 저장하기 위한 공간이 작기 때문에 매우 긴 데이터를 처리하는 데는 한계

# 언어 모델

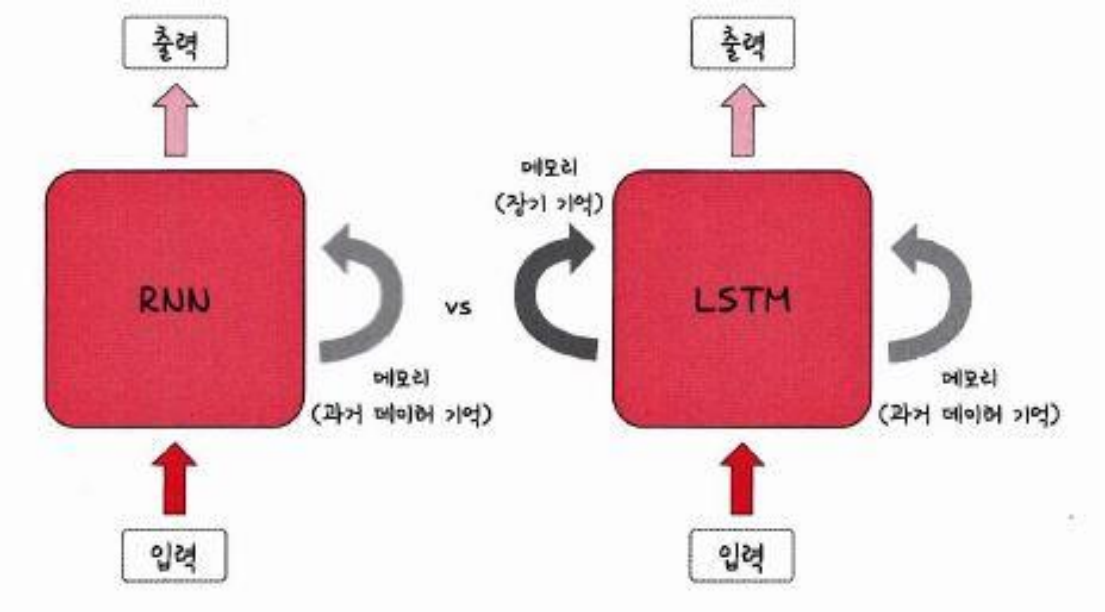
## 신경망 언어 모델

- LSTM(Long Short-Term Memory Networks)
  - RNN의 경우 긴 시퀀스 데이터 처리에 한계가 있다고 했는데, 그 한계를 극복하기 위해 고안된 모델이 LSTM
  - LSTM 은 긴 시퀀스 정보를 기억하고 필요에 따라 이를 삭제하거나 업데이트할 수 있는 메커니즘을 가지고 있음

# 언어 모델

## 신경망 언어 모델

- 메모리는 정보를 저장하기 위한 공간에 제약이 있기 때문에 긴 문장을 모두 저장할 수 없음
- 한계를 극복하기 위해 LSTM은 장기 기억을 위한 메모리를 하나 더 둠



# 언어 모델

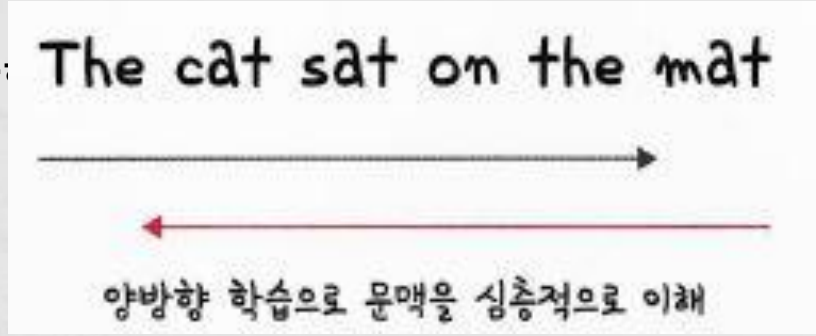
## 트랜스포머 아키텍처

- 2017년 구글 브레인(Google Brain)에서 발표한 논문 (Attention Is All Ybu Need)에서 트랜스포머 아키텍처가 도입
- 이전의 언어 모델은 각 단어(혹은 구절)를 개별적으로 이해하고 처리하는 데 중점을 두었음
- 트랜스포머(Transformer)는 문장과 단락 전체를 처리할 수 있음
- 트랜스포머를 통해 LLM은 자연어에서 인간의 의도를 심층적이고 맥락에 맞게 이해할 수 있게 되었으며, 이를 통해 콘텐츠 생성, 문장 요약 등 활용 범위가 넓어짐

# 언어 모델

## 트랜스포머 아키텍처

- 트랜스포머를 이용한 대표적 인 모델
- 버트(Bidirectional Encoder Representations from Transformers, BERT)
  - 텍스트를 양방향으로 분석하여 맥락을 이해하는 언어 모델
  - 이때 양방향으로 텍스트를 분석 한다는 것은 언어 모델이 단어의 앞뒤 문맥을 모두 고려하여 그 단어의 의미를 이해하는 것을 의미
  - 지금까지의 언어 모델은 주로 한 방 측 (예를 들어 왼쪽에서 오른쪽으로)



# 언어 모델

## 트랜스포머 아키텍처

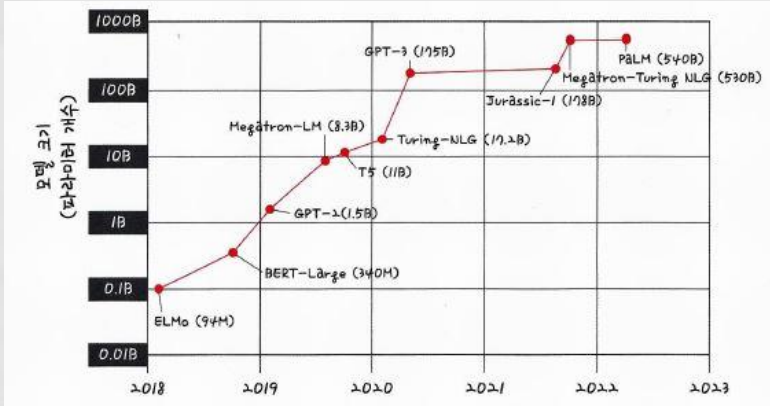
- 트랜스포머를 이용한 대표적 인 모델

- GPT(Generative Pretrained Transformer)
  - GPT는 오픈AI에 의해 개발된, 인간의 언어를 처리하는 강력한 인공지능 언어 모델
  - 오픈AI는 2018년 처음으로 GPT-1 모델을 발표했고 이후 4년 만인 2022년 챗GPT를 발표
  - 2023년에는 높은 정확성(답변의 정확성)을 갖는 GPT-4 모델을 발표
  - 이 모든 과정이 5년 동안 이뤄졌으며 현재 GPT 모델은 다른 언어 모델에 비해 자연스러운 텍스트 생성 및 높은 수준의 대화로 각광받고 있음

# 언어 모델

## 거대 언어 모델

- 거대 언어 모델(Large Language Model)은 대규모 데이터로 훈련된, 매우 큰 규모의 인공 지능 기반 언어 모델
  - 오픈AI가 GPT-3 모델을 학습시킬 때, 이 모델은 약 45TB의 텍스트 데이터를 이용하여 학습
- 거대 언어 모델의 크기는 주로 모델이 가지고 있는 파라미터의 수로 측정
  - 오픈AI의 GPT-3는 약 1,750억 개의 파라미터



# LLM 특징과 종류

## LLM의 특징

- LLM은 인터넷의 텍스트, 책, 논문, 기사 등 다양하고 방대한 양의 텍스트 데이터로 부터 학습(트랜스포머)
- 언어를 이해하고 생성하는 데 특화되어 있으며 언어 모델은 텍스트를 읽고 이해 하는 것뿐만 아니라 자연스러운 언어로 질문에 답변하고, 글을 작성하고, 대화를 나누는 등의 생성적 작업도 수행할 수 있음
- 특정 작업을 위해 파인튜닝할 수 있으며 파인튜닝이란 챗GPT와 같은 언어 모델을 기업의 데이터로 추가 학습을 시키는 과정으로, 이를 통해 특화된 분야에 더욱 정교하게 사용될 수 있음

# LLM 특징과 종류

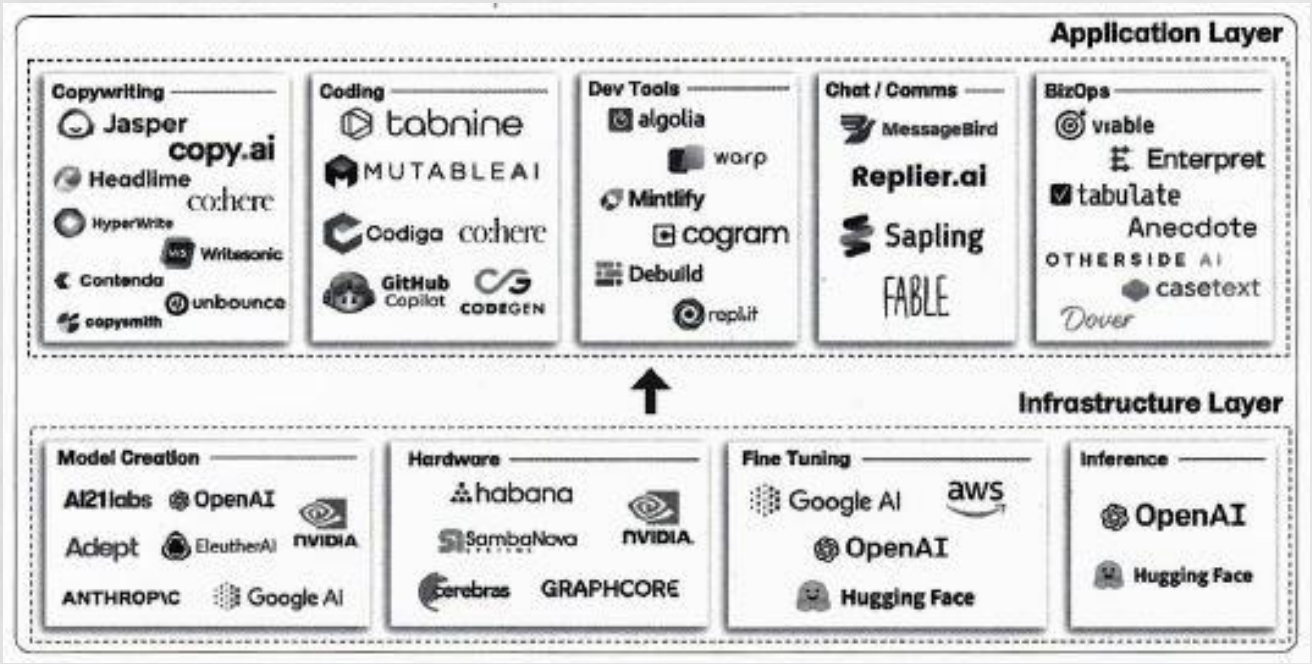
## LLM의 특징

- LLM을 훈련하고 운영하는 데는 상당한 컴퓨팅 자원이 필요
- 이때 컴퓨팅 자원이란 주로 GPU나 TPU 같은 하드웨어
  - GPU(Graphic Processing Unit) 는 그래픽 처리를 위한 장치였으나 최근 인공지능 관련해서는 대규모 데이터 학습과 복잡한 수학적 연산을 빠르게 처리하기 위한 용도로 사용
  - TPU(Tensor Processing Unit)는 구글이 개발한, 머신러닝 및 딥러닝 작업에 최적화된 하드웨어

# LLM 특징과 종류

## LLM의 종류

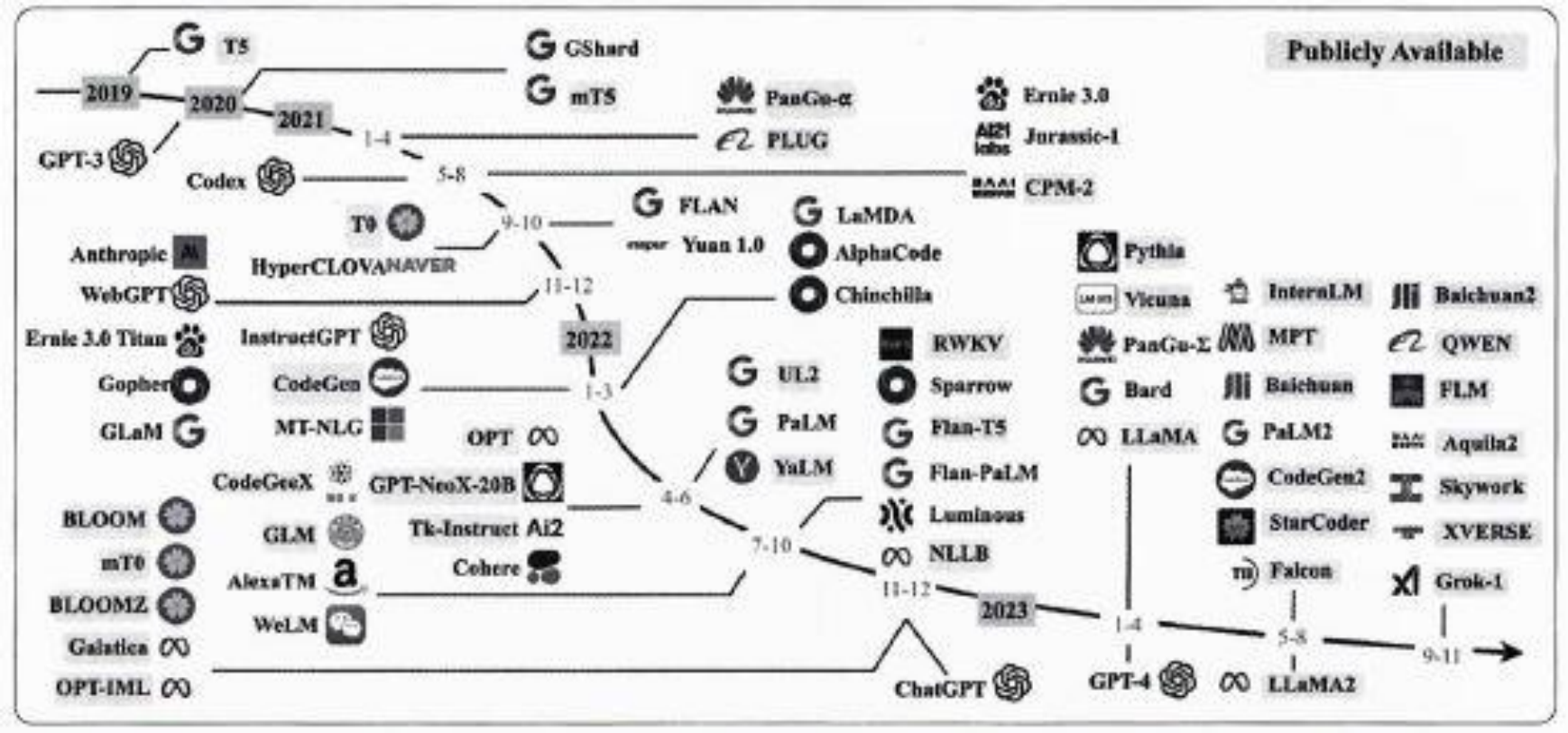
- LLM 관련해서 다음과 같은 생태계가 구축되어 있는 것
- LLM 생태계는 인프라스트럭처 레이어(Infrastructure Layer)와 애플리케이션 레이어(Application Layer)로 구분



# LLM 특징과 종류

## LLM의 종류

- LLM 모델 종류



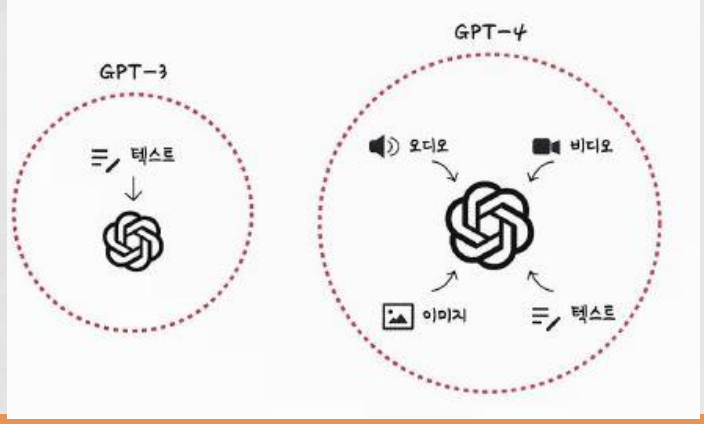



https://arxiv.org/pdf/2303.18223.pdf

# LLM 특징과 종류

## LLM의 종류

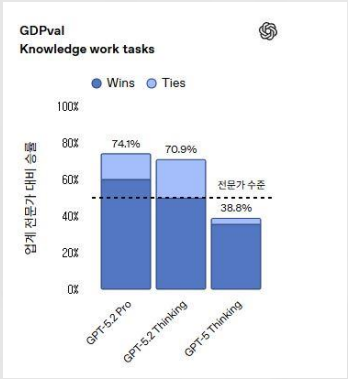
- GPT-4
  - GPT-4(혹은 GPT-4 Turbo)는 오픈AI가 가장 최근에 발표한 모델 (2023년 12월 시점)
  - 무려 1조 5천억 개의 파라미터를 갖는 이 모델은 전례 없는 모델 크기
  - 광범위한 파라미터를 통해 GPT-4는 복잡한 언어 패턴을 식별하여 텍스트 생성, 이해 및 일관된 문맥을 제공
  - GPT-4부터는 멀티모달(multimodal)을 제공(멀티모달은 '여러 가지 방법을 섞어 쓴다'는 뜻)



# LLM 특징과 종류

## LLM의 종류

- GPT-5
  - GPT-5는 OpenAI의 차세대 대규모 언어 모델(LLM)로, 기존 GPT-4 계열보다 추론 능력·정확성·멀티모달 처리 능력이 크게 강화된 모델
  - 텍스트 이해와 생성은 물론, 이미지·표·코드 등 다양한 입력을 종합적으로 처리하는 데 초점을 둠
  - 🧠 고급 추론 능력: 복잡한 문제를 단계적으로 분석하고 논리적으로 답변
  - 🧩 멀티모달: 텍스트 + 이미지(+ 기타 데이터)를 함께 이해
  - ⚙️ 개발 친화성: 코드 작성, 디버깅, 설계 문서 생성에 강점
  - 🛡️ 안정성과 안전성 강화: 오류(환각) 감소, 정책·보안 측면 개선
  - 📚 전문 분야 대응: 교육, 연구, 법·의료·보안 등 전문 맥락 이해 향상



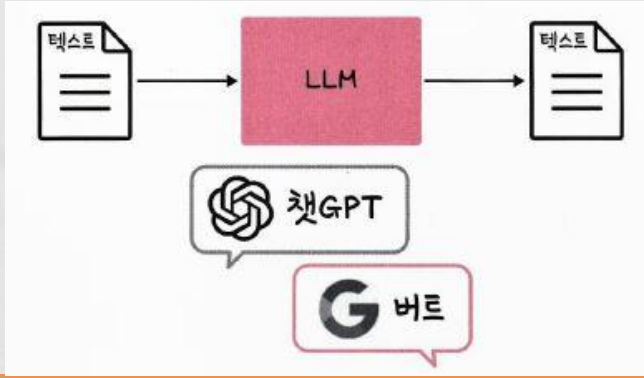



https://openai.com/ko-KR/index/introducing-gpt-5-2/

# LLM 특징과 종류

## LLM과 GAI, SLM

- LLM과 같이 자주 언급되는 단어가 생성형 AI(Generative AI, GAI)
- LLM은 거대 언어 모델
  - 대용량의 텍스트 데이터를 학습함으로 인간의 언어인 자연어를 이해하고 생성하는 작업에 능숙
  - 오픈AI의 GPT 시리즈(예: GPT-3.5, GPT-4), 구글의 팜2, 제미나이
  - LLM은 텍스트를 기반으로 질문에 답하거나, 글을 작성하고, 대화를 진행하는 등 다양한 언어 관련 작업을 하는데 사용



# LLM 특징과 종류

## LLM과 GAI, SLM

- GAI는 입력 데이터를 기반으로 새로운 콘텐츠를 생성하는 인공지능
- 텍스트,이미지,음악,비디오 등 다양한 형태의 콘텐츠를 생성할 수 있는 모델들이 포함
- 생성형 AI의 범위

# LLM 특징과 종류

## LLM과 GAI, SLM

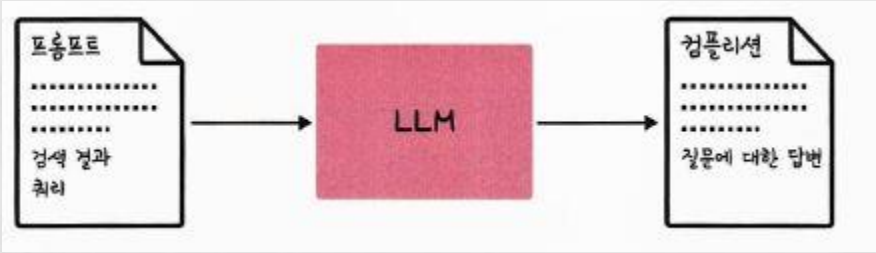
- LLM과 GAI의 비교

| 구분 | GAI | LLM |
|---|---|---|
| 콘텐츠 생성 범위 | 텍스트, 이미지, 음성, 코드 등 다양 | 텍스트에 국한 |
| 학습 | 대규모 데이터에 의존하면서도 이미지, 오디오 등 다양한 데이터 유형을 학습 | 인터넷, 서적 및 기타 광범위한 텍스트 학습 |
| 출력 | 음악에서 시각적 예술 작품에 이르기까지 광범위한 출력을 생성 | 사용자의 질문에 일관되고 상황에 맞게 텍스트를 생성 |
| 신경망(혹은 모델) | 이미지 생성을 위한 GAN(Generative Adversarial Networks) 또는 음악과 같은 순차적 데이터를 위한 RNN(Recurrent Neural Networks)을 포함한 다양한 신경망 사용 | 텍스트와 같은 순차 데이터에 매우 효과적인 RNN, 트랜스포머 사용 |
| 활용 분야 | 하나의 분야(혹은 산업)에 국한되지 않고 여러 창의적 분야에 사용 가능 | 언어 관련 작업에 특화 |

# LLM 특징과 종류

## LLM과 GAI, SLM

- LLM에서는 질문과 답변을 특별한 용어
- 질문은 프롬프트(prompt), 답변은 컴플리션(completion)
- 특히 프롬프트는 사용자가 거대 언어 모델에 정보를 요청하거나 특정 작업을 수행하도록 지시하는 텍스트 메시지
- 이 프롬프트는 질문, 명령, 또는 토론 주제 등 다양한 형태로 나타날 수 있으며, 모델은 이를 바탕으로 답변을 생성



# LLM 특징과 종류

## LLM과 GAI, SLM

- LLM vs. SLM
- LLM은 매우 큰 신경망 구조와 방대한 데이터로 훈련되어 광범위한 언어 작업을 수행할 수 있음. SLM은 이보다 훨씬 더 제한적인 용도에 적합
- LLM과 SLM의 차이는 주로 모델의 크기, 학습 데이터의 양, 처리 능력, 그리고 사용 사례

| 구분 | SLM | LLM |
|---|---|---|
| 모델 크기 | 파라미터가 1,500만 개 미만 | 수천억 개의 파라미터 |
| 컴퓨팅 | 모바일 디바이스에서도 동작 | 수백 개의 GPU 필요 |
| 성능 | 단순 작업만 동작 | 복잡한 작업 처리가 가능 |
| 배포 | 배포가 쉬움 | 배포를 위해 상당한 인프라가 필요<br>(그래서 대체적으로 클라우드를 사용) |
| 학습 | 일주일 정도면 학습 가능 | 몇 달 간의 학습이 필요 |

# LLM 생성 과정

## LLM의 생성 과정

1. 데이터수집 및 준비

2. 모델설계

3. 모델학습

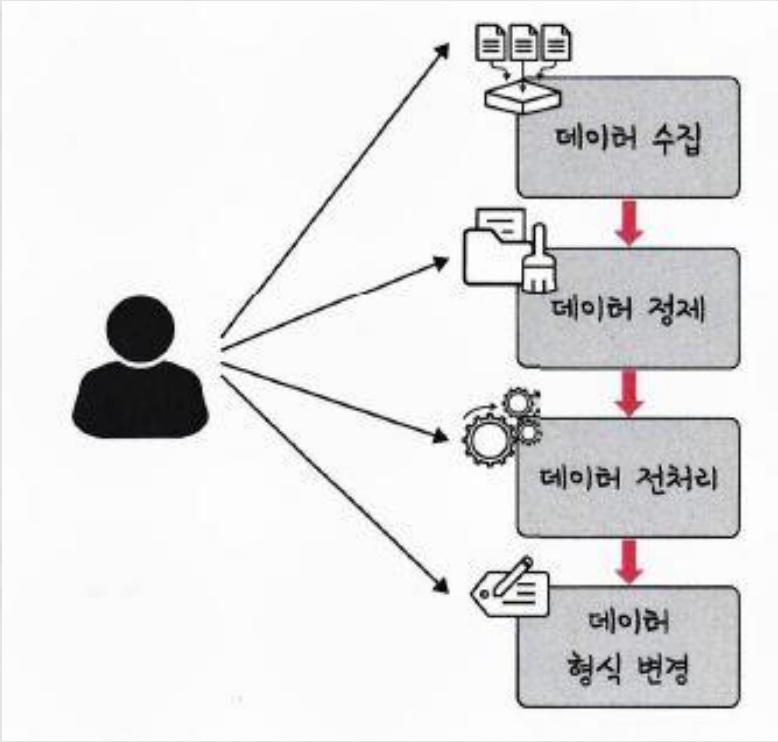
4. 평가 및 검증

5. 배포 및 유지 보수

# LLM 생성 과정

## 데이터 수집 및 준비

- 언어 모델을 학습시키기 위한 첫 단계는 데이터를 수집하고 준비하는 것
- 모델이 학습할 수 있는 형태로 데이터를 수집하고 가공하는 일련의 작업이 이루어지는 단계



# LLM 생성 과정

## 데이터 수집 및 준비

- 데이터 수집
  - **데이터 식별**: 언어 모델이 다양한 언어 패턴을 학습할 수 있도록 다양한 주제와 장르, 스타일이 포함
  - **데이터 수집**: 수집해야 할 데이터가 식별되었다면, 실제로 데이터를 수집 데이터는 HTML 페이지, PDF 문서, 텍스트 파일, 데이터베이스 등 다양한 형식으로 존재
    - 데이터를 수집할 때에는 저작권, 개인정보 보호 등 법적인 문제를 고려해서 수집

# LLM 생성 과정

## 데이터 수집 및 준비

- 데이터 정제
  - 데이터 정제는 데이터의 품질을 결정하는 핵심적인 과정
  - 중복제거: 수집된 데이터 중 중복되는 내용을 제거
  - 노이즈 제거: 노이즈는 모델 개발에서 원치 않는 무작위적이고 관련 없는 정보. 노이즈를 제거한다는 의미는 오타, 잘못된 문장 부호, 비정상적인 문자 등을 정리하는 과정을 의미

# LLM 생성 과정

## 데이터 수집 및 준비

- 데이터 전처리
  - 데이터 전처리는 데이터를 LLM 학습에 적합한 형태로 만드는 과정
  - 토큰화(tokenization) : 토큰화는 텍스트를 작은 단위로 나누는 과정
    - “Hello, how are you?”라는 문장을 토큰화하면 ‘Hello’, ‘,’ ‘how’, ‘are’, ‘you’, ‘?’
  - 정규화: 대소문자 통일, 어간 추출(stemming) 등을 통해 단어의 기본 형태로 변환
    - ‘run’, ‘runs’, ‘running’과 같은 단어를 기본 형태인 ‘run’으로 통일하는 과정

# LLM 생성 과정

## 데이터 수집 및 준비

- 데이터 형식 변경
  - 데이터의 형식을 일치시키는 과정
  - 예를 들면, 모든 날짜를 “YYYY-MM-DD” 형태로 변경
  - 모든 정보의 형식이 일관되어 정보를 찾거나 비교할 때 훨씬 쉬워짐

# LLM 생성 과정

## 모델 설계

- LLM에서 모델 설계는 매우 큰 신경망 아키텍처를 구축하는 것을 의미
- 먼저 어떤 모델로 학습할지 결정해야 하는데 LLM은 주로 트랜스포머 모델을 기반
- 계층 수,학습률, 배치 크기 등과 같은 모델의 학습 과정을 조절할 하이퍼파라미터(hyperparameter)를 설정
- 하이퍼파라미터는 학습 과정을 제어하는 데 사용하는 설정 값
- 학습 과정에서 하이퍼파라미터를 조절하여 모델이 데이터를 얼마나 빨리, 얼마나 오래, 어떤 방식으로 학습할지 결정

# LLM 생성 과정

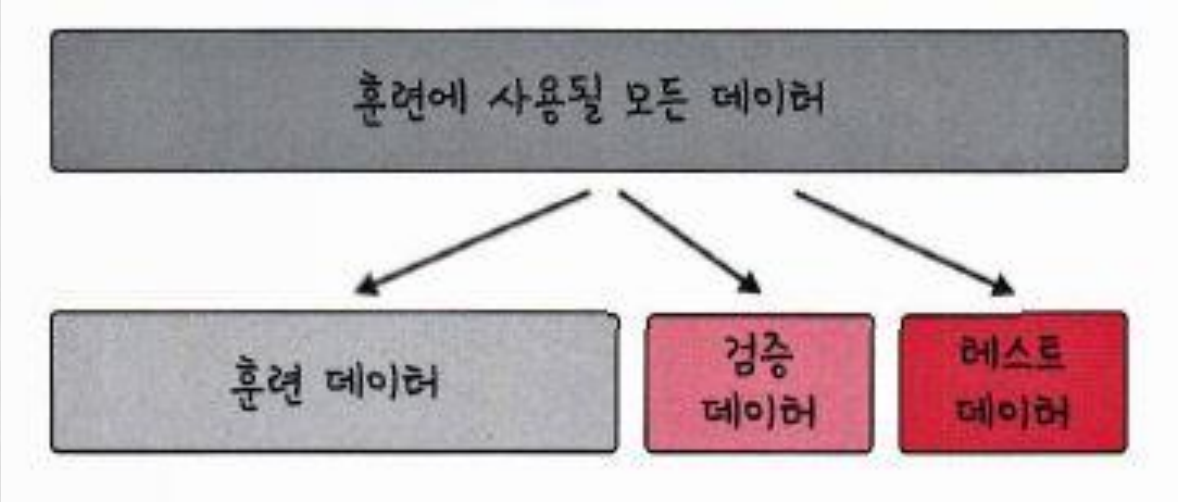
## 모델 학습

- 설정된 하이퍼파라미터와 모델 아키텍처를 사용해 학습
- 모델 학습은 모델이 데이터로부터 패턴을 학습하고, 이를 내부적으로 모델링하여 텍스트를 생성하거나 번역하는 등의 작업을 수행할 수 있도록 하는 과정
- '모델링'은 모델이 데이터로부터 중요한 특징이나 관계를 학습하고, 이를 수학적 구조로 표현하는 과정
- 모델링은 주어진 데이터를 기반으로 일반화된 패턴이나 규칙을 만드는 것

# LLM 생성 과정

## 평가 및 검증

- 모델 평가 및 검증은 모델이 얼마나 잘 작동하는지를 평가하고 이것을 실제로 서비스했을 때 어느 정도의 성능(예: 답변의 정확도, 답변 속도)을 낼 수 있는지를 확인하는 과정
- 수집된 데이터를 훈련, 검증, 테스트 용도로 나눠야 함



# LLM 생성 과정

## 평가 및 검증

- 모델 평가 지표
  - 정확도(accuracy) : 모델이 얼마나 많은 예측을 정확히 했는지 측정(전체적인 성능을 평가할 때 사용)
  - 정밀도(precision) : 양성으로 예측된 사례 중 실제 양성인 사례의 비율(예를 들어 스팸 메일을 걸러내되 중요한 메일을 놓치지 않아야 할 때 사용)
  - 재현율(recall) : 실제 양성 사례 중 모델이 양성으로 예측한 사례의 비율(모델이실제 양성 사례를 얼마나 잘 찾아내는지를 나타내며, 놓치는 것에 더 민감할 때 사용)
  - F1점수(F1 score) : 정밀도와 재현율의 조화 평균을 나타내는 지표
  - ROC 곡선 및 AUC : 모델의 성능을 다양한 임계값에서 평가

# LLM 생성 과정

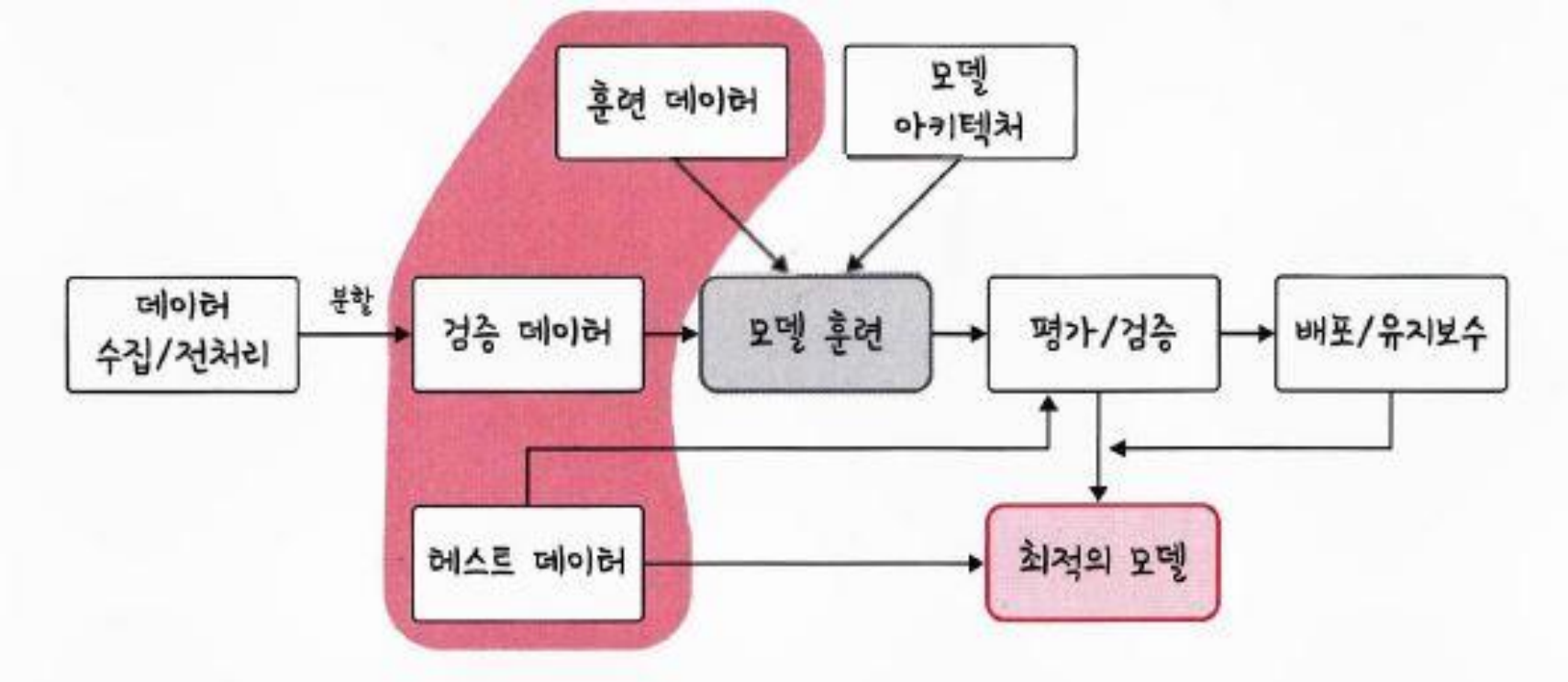
## 배포 및 유지보수

- 모델 배포와 유지보수는 LLM의 마지막 단계
- 모델을 배포한다는 의미는 챗봇과 같은 Q&A 서비스를 사용자가 이용하는 것
- 챗봇 서비스에 문제가 있으면(오류가 발생하면) 수정하는 작업이 유지보수에 해당

# LLM 생성 과정

## 모델 개발 라이프사이클

- 모델을 개발하는 전반적인 라이프사이클



# LLM 생성 후 추가 고려 사항

## 윤리적 고려 및 보정

- LLM의 윤리적 고려와 보정은 모델이 생성한 결과의 공정성, 편향성, 투명성을 다루는 매우 중요한 과정
- 모델이 다양한 사용자와 상황에 미치는 잠재적인 영향을 신중하게 평가하고, 부정적인 결과를 최소화하기 위해 필요
- 책임감 있는 AI(Responsible AI)는 인공지능을 설계, 개발, 배포할 때 윤리적, 법적, 사회적 책임을 고려하는 접근 방식

# LLM 생성 후 추가 고려 사항

## 윤리적 고려 및 보정

- 책임감 있는 AI의 원칙

# LLM 생성 후 추가 고려 사항

## 윤리적 고려 및 보정

- 공정성(fairness): AI는 성별, 인종, 나이 등에 따른 편향 없이 모든 사용자에게 공정하 게 서비스를 제공
- 신뢰성(reliability)&안전성(safety): AI는 사용자에게 안전하며 예측 가능한 위험을 관리 하고 예방
- 프라이버시(privacy) : AI는 사용자의 개인정보를 보호하고 데이터 보안을 유지
- 포용성(inclusion) & 다양성(diversity): 모든 사람이 차별 없이 참여하고 혜택을 받을 수 있도록 환경을 조성
- 윤리적 사용(ethical use) : AI는 사회적, 도덕적 기준에 부합하는 방식으로 사용되어야 하며 인간의 존엄성을 존중.
- 투명성(transparency): AI의 의사 결정 과정이 명확하고 이해 가능해야 하며, 사용자는 AI의 작동 방식과 결정 기준을 알 권리
- 책임성(accountability) : AI의 결정에 대한 책임을 명확히 하여 문제 발생 시 적절한 해결책을 제시

# LLM 생성 후 추가 고려 사항

## 지속적 모니터링

- LLM은 강력한 언어 모델로서 악의적으로 사용하고자 한다면 얼마든지 그렇게 사용할 수 있음
- 기업은 이미지가 실추되는 것은 물론 법적인 책임도 감수
- 사용자들의 질문과 답변을 지속적으로 검사
  - AI가 스스로 악의적 문구를 탐지
  - 사람이 질문과 답변을 지속적으로 점검

# 정리

## 정리

- LLM 개념
- 언어 모델
- LLM의 특성 및 종류
- LLM 생성 과정
- LLM 생성 후 추가 고려사항