대화형 언어모델 ChatGPT

박 희 선 교수

성균관대학교 소프트웨어융합대학

2023. 02. 07

Contents

❑ 1월 : 대화형 언어모델

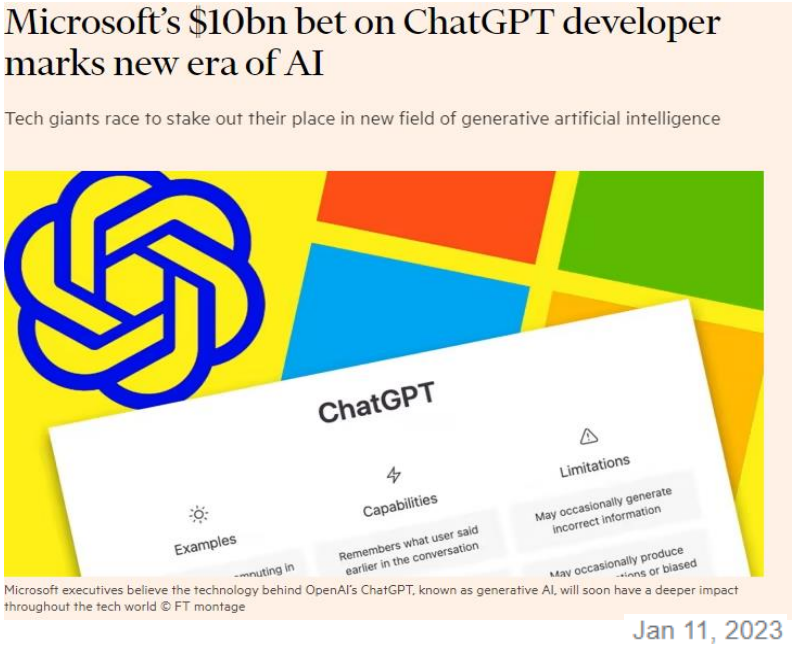
- ChatGPT
  - Investing News : OpenAI & MS Partnership
  - Code Generation AI
  - Timeline of GPT Model
  - ChatGPT Overview
    - InstructGPT (GPT-3.5)
    - Training Process
  - ChatGPT/GPT-3 Pricing Plan
  - ChatGPT's Dark Side
  - OpenAI's anti-cheating Tool
- Appendix
  - Competitors : DeepMind, Google

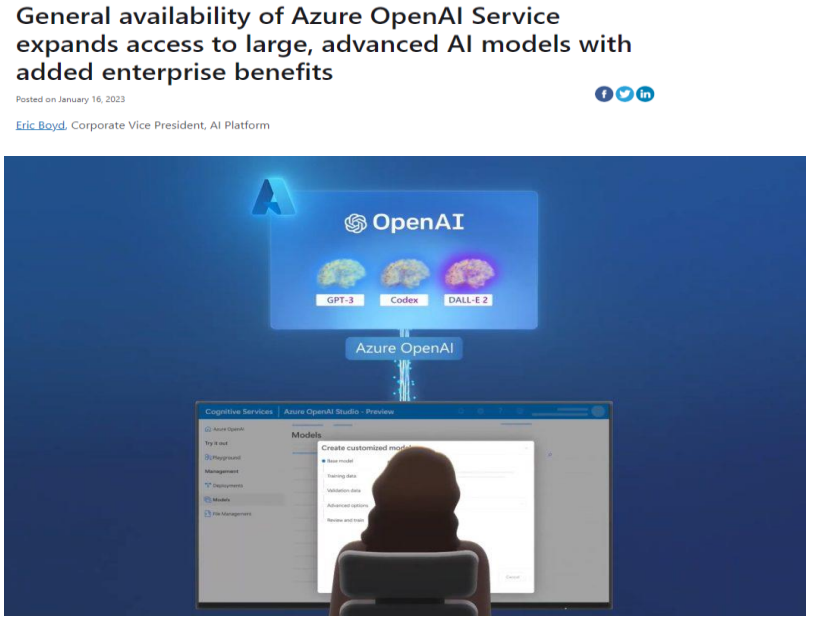
Investing News

OpenAI | Microsoft

❑ OpenAI& MS Partnership

- 2020/09: OpenAI는 GPT-3를 MS 클라우드 서비스 Azure를 통해 독점 공급
- 2021/11 : MS Azure OpenAI Service 출시
- 2022/11/30 : OpenAI ChatGPT 런칭
- 2023/01/11 : MS는 100억 달러(12조) 추가 투자 결정
- 2023/01/16 : MS Azure OpenAI Service에 ChatGPT 기능도 추가 예정
  - GPT-3.5, DALL-E2 등이 포함되어 있음





OpenAI & MS Partnership

Timeline of key Microsoft AI breakthroughs

September 2016
Microsoft achieves human parity in conversational speech recognition.

January 2018
Microsoft achieves human parity in reading comprehension.

March 2018
Microsoft reaches human parity in machine translation.

June 2019
Microsoft integrates neural machine translation research breakthroughs into Azure and achieves human parity in general language understanding.

July 2019 10억 달러 투자
Microsoft and OpenAI announce exclusive partnership to accelerate breakthroughs in AI.

February 2020
Microsoft announces Turing-NLG language model at 17 billion parameters.

May 2020
Microsoft announces its first AI supercomputers, built in collaboration with and exclusively for OpenAI.

September 2020
Microsoft teams up with OpenAI to exclusively license GPT-3 models.

October 2020
Microsoft announces AI breakthroughs in image captioning available in Azure, Word, Outlook, and PowerPoint.

May 2021
Microsoft introduces its first product features powered by GPT-3 models.

June 2021
GitHub launches Copilot, an AI pair programmer that takes advantage of the OpenAI Codex model to support developers.

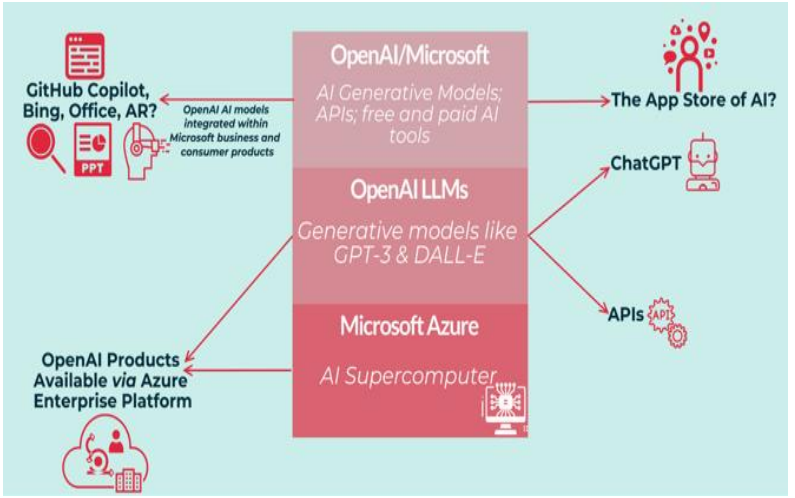
November 2021
Microsoft introduces Azure OpenAI Service.

May 2022
Microsoft expands access to Azure OpenAI Service with more models and new responsible AI system.

June 2022
GitHub Copilot becomes generally available.

October 2022
Microsoft announces DALL·E 2 in Azure OpenAI Service, new Designer app, and Bing Image Creator.

January 2023
Microsoft announces Azure OpenAI Service now generally available. ChatGPT coming soon.



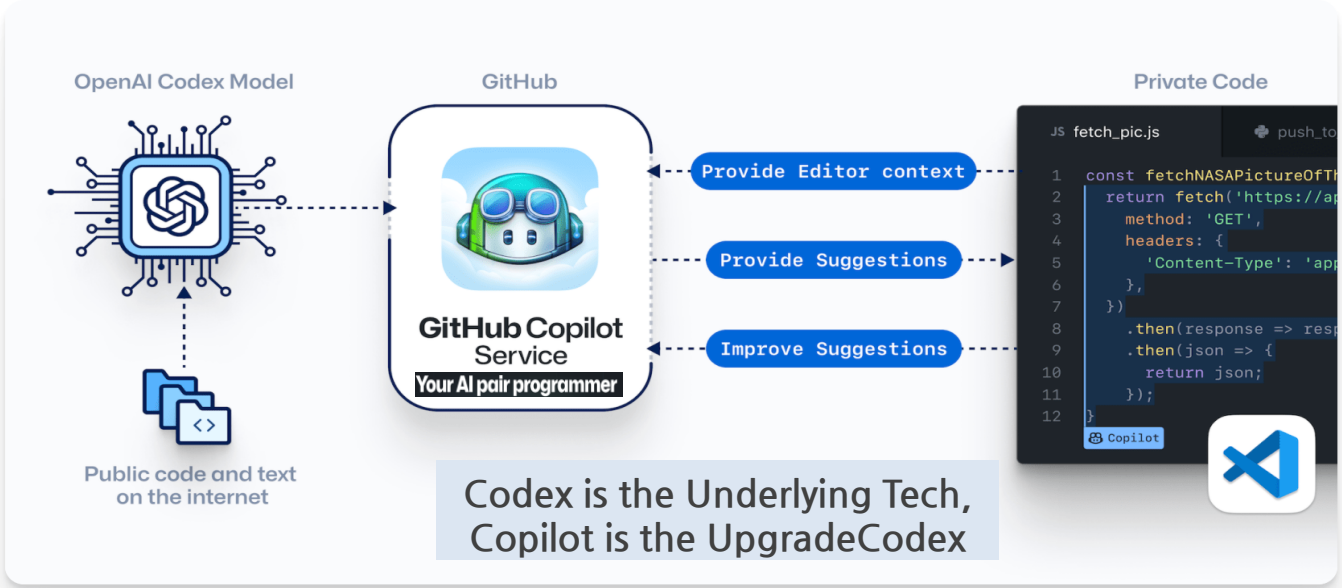
Code Generation AI

GitHub Copilot
GitHub
★★★★☆

OpenAI Codex
OpenAI

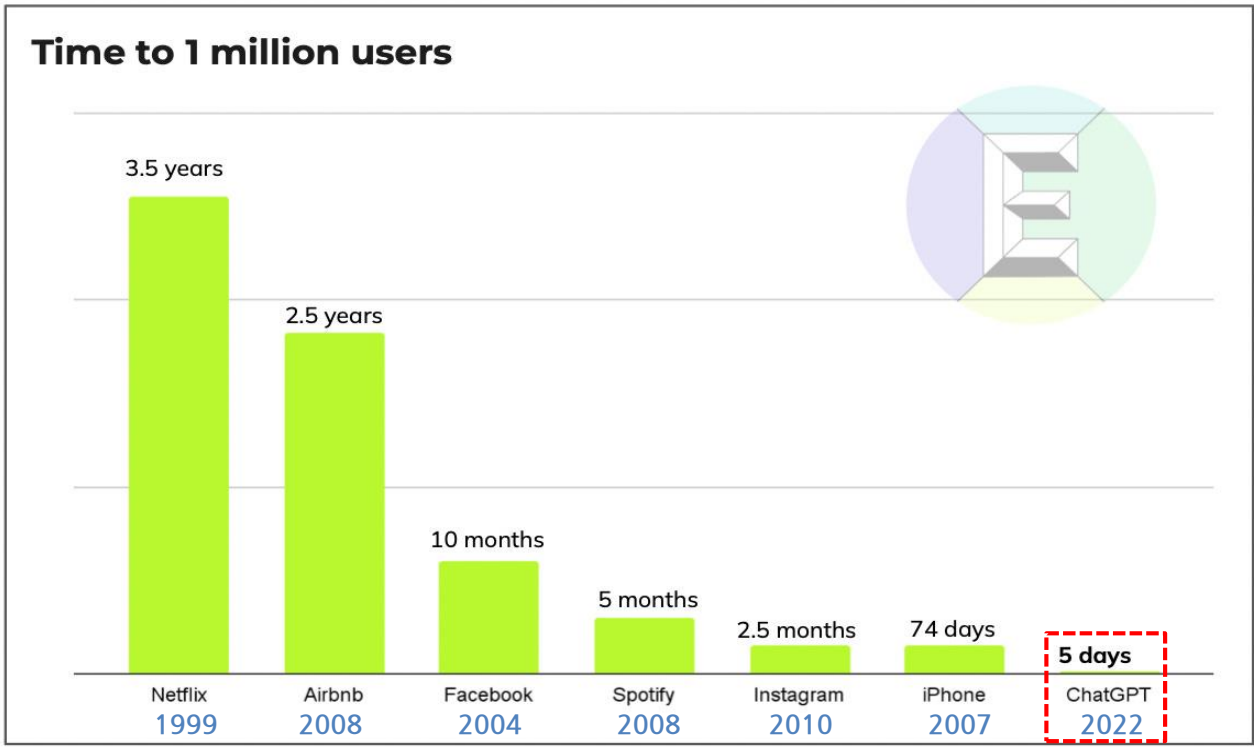
- MS : GitHub 인수 - 75억 달러(8조) (2018.06)
- OpenAI Codex : 자연어를 코드로 변환하는 코드 생성에 특화된 언어 모델
  - Training : GPT-3를 기반으로 5,400만개 GitHub repository에서 159GB의 Python code에 대해 추가 훈련
  - Codex가 Prompt 요청의 약 37%를 완료
  - **지원 언어 : Python**, JavaScript, Go, Perl, PHP, Ruby, Swift, TypeScript, Shell 등 12개 이상

➔ GitHub Copilot : Codex를 활용한 자동코드완성 AI 서비스 런칭 (2022.06)



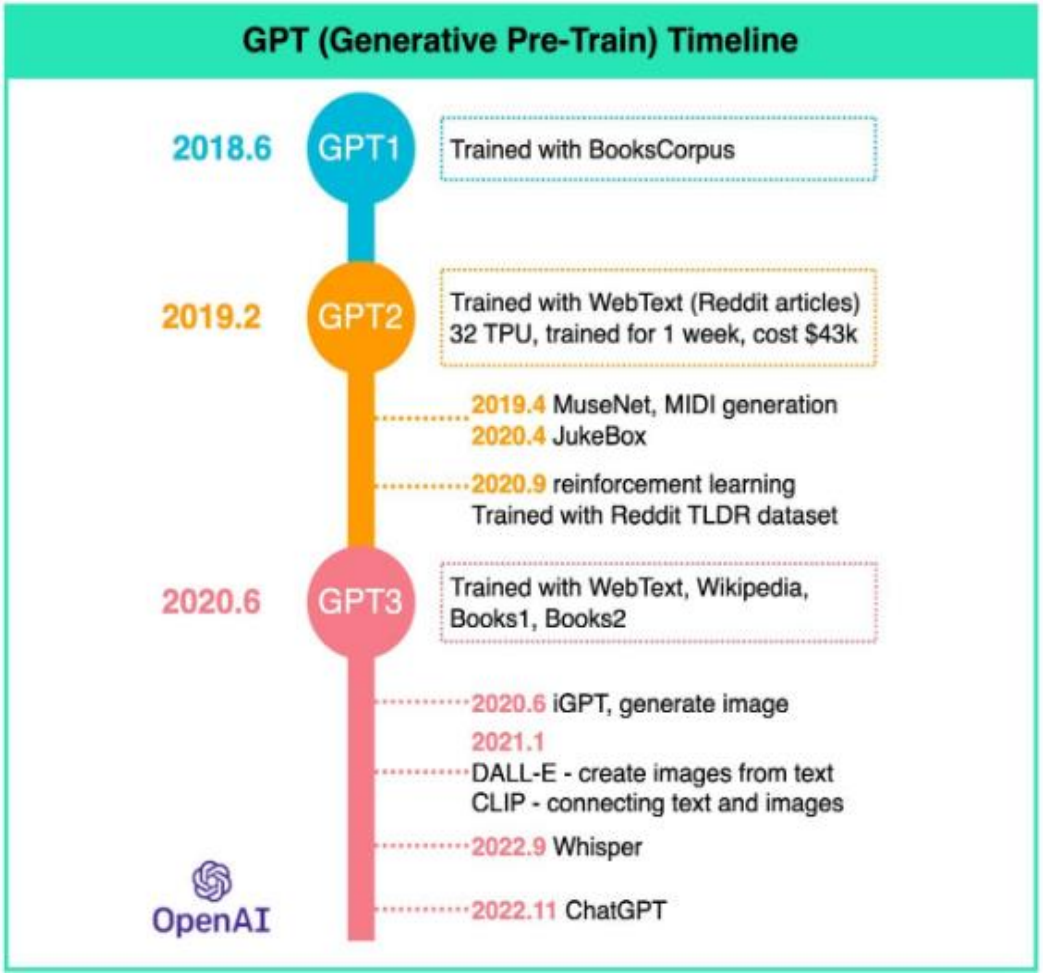
Is ChatGPT a game changer for AI?

- ChatGPT가 특별한 이유 : 공개 서비스
  - 높은 접근성, 민감한 주제까지 Bot이 다루도록 허용한 대담함
- ChatGPT 런칭: 2022/11/30
  - 출시 5일만에 100만명이 사용
  - 출시 2달만에 1,000만명 사용



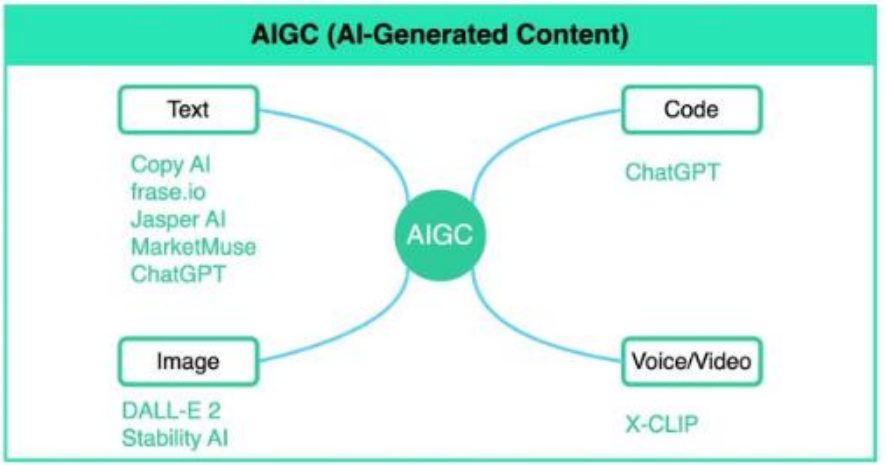
Exponential View via Linas Beliūnas

Timeline of GPT Model

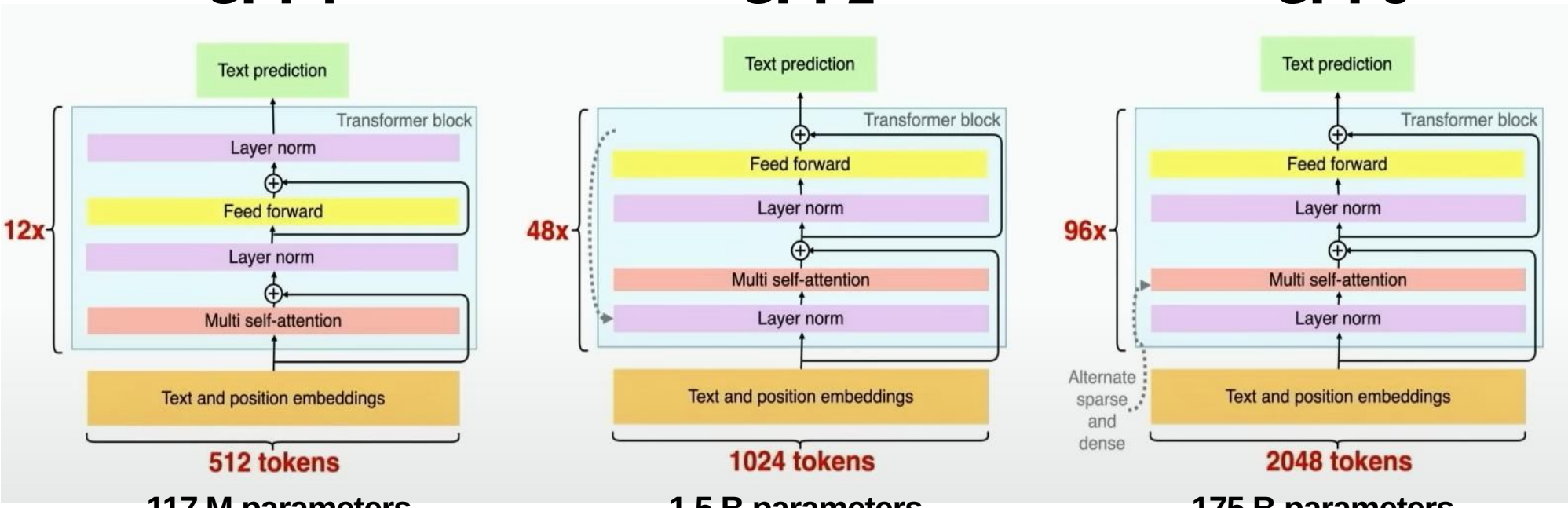


GPT-3 기반 Applications

- **DALL-E**: creating images from text
- **CLIP**: connecting text and images
- **Whisper**: multi-lingual voice to text
- **ChatGPT**: chatbot, article writer, code writer



GPT-1 vs GPT-2 vs GPT-3



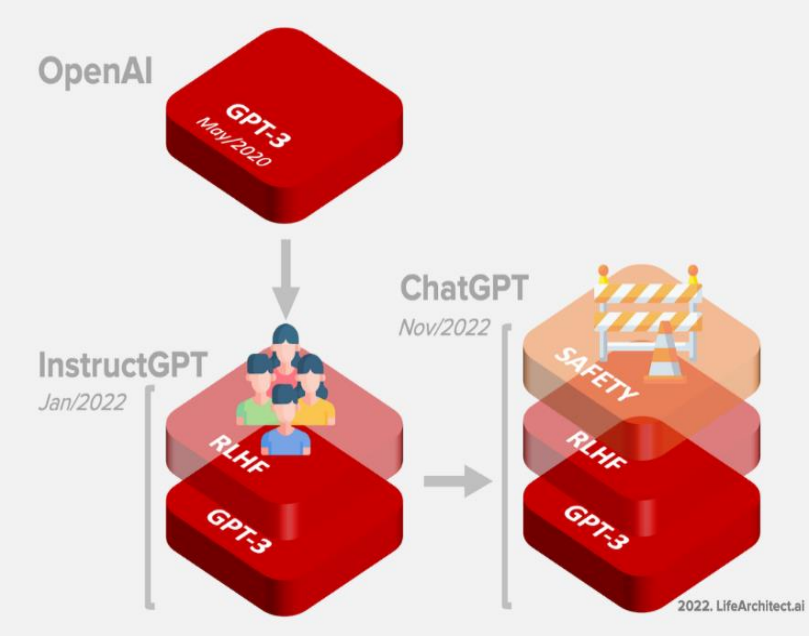
117 M parameters **1.5 B parameters** **175 B parameters**

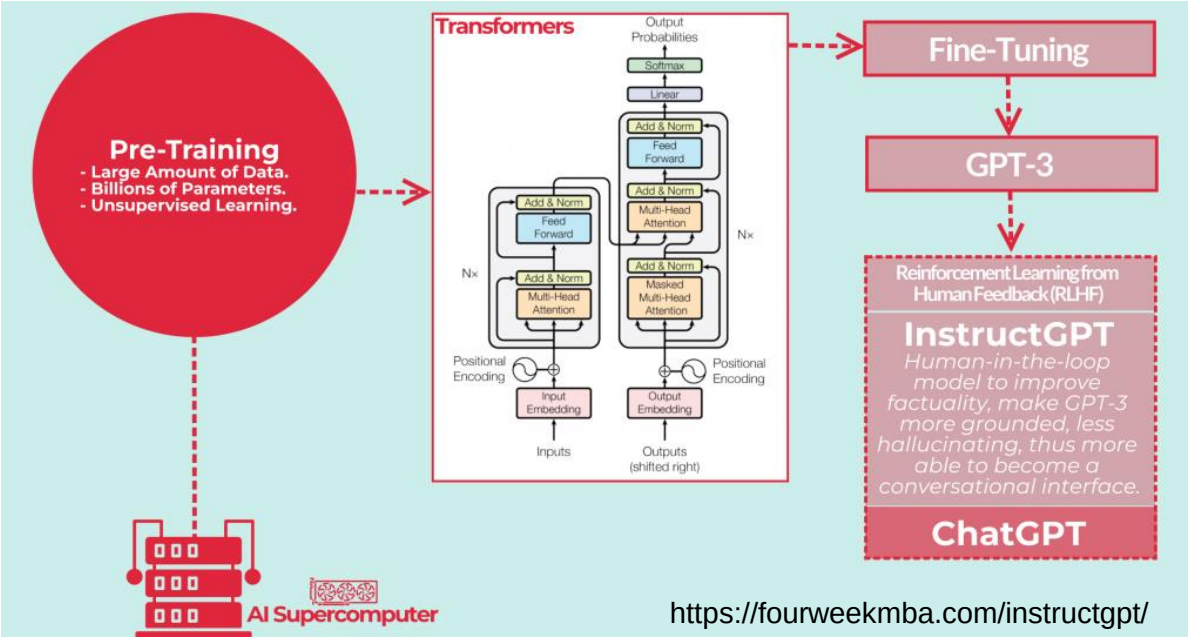
	GPT-1	GPT-2	GPT-3
Parameters	117 Million	1.5 Billion	175 Billion
Decoder Layers	12	48	96
Context Token Size	512	1024	2048
Hidden Layer	768	1600	12288
Batch Size	64	512	3.2M

ChatGPT Overview

❑ Fine-tuned version of GPT-3.5

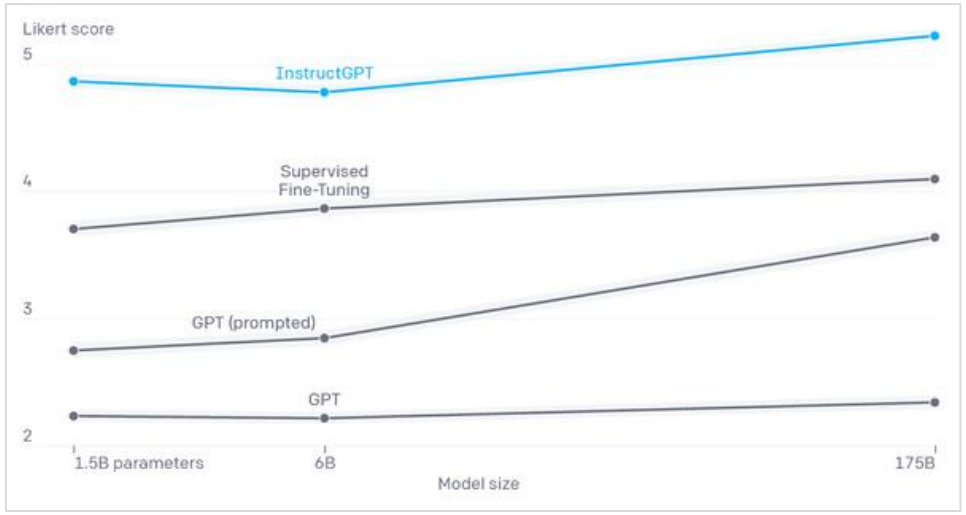
- GPT-3.5 = InstructGPT
  - GPT-3.5는 GPT-3을 기반으로 하지만 정책을 준수하도록 강제함으로써 AI가 인간의 가치와 align하도록 가드레일 내에서 작동함
  - InstructGPT = GPT3 + RLHF(Reinforcement Learning with Human Feedback)
    - InstructGPT : 13 Billion Parameters (vs GPT-3 : 175 Billion Parameters)



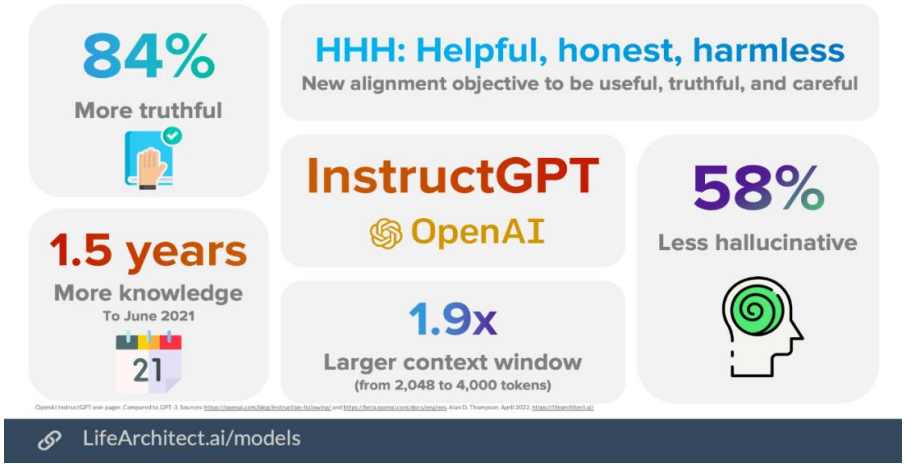


New version of GPT-3 : InstructGPT (GPT-3.5)

- InstructGPT model : "Training language models to follow instructions with human feedback," OpenAI, 2022.1.27.
  - **GPT-3 문장 생성 AI의 문제 : 인터넷상의 대규모 데이터 세트에서 단어를 선택했기 때문에 진실하지 않은 문장, 유해한 문장, 공격적인 문장을 생성하는 경향이 있음**
    - ➔ Prompt Engineering이 중요 (최적의 결과를 도출하기 위해 모델에 입력해야 하는 내용에 관한 규칙과 기술)
  - **InstructGPT: GPT-3 학습 모델에 인간의 피드백을 반영해서 제기된 문제를 해결하고자 함**
    - Instruct GPT가 어떤 조건의 GPT-3보다 높은 점수(Likert score)를 기록
      - **Likert score**: 생성한 문장의 품질에 대해 인간이 평가
    - GPT-3에 비해 위화감이 적은 문장을 생성할 수 있으며, 유해한 문장 생성률도 감소
      - 그러나 InstructGPT도 여전히 유해하고 편향된 문장을 생성하고, 거짓된 사실을 만들어내며 지시가 없는 경우에도 성적, 폭력적인 콘텐츠를 생성함



GPT-3 대비 InstructGPT의 이점

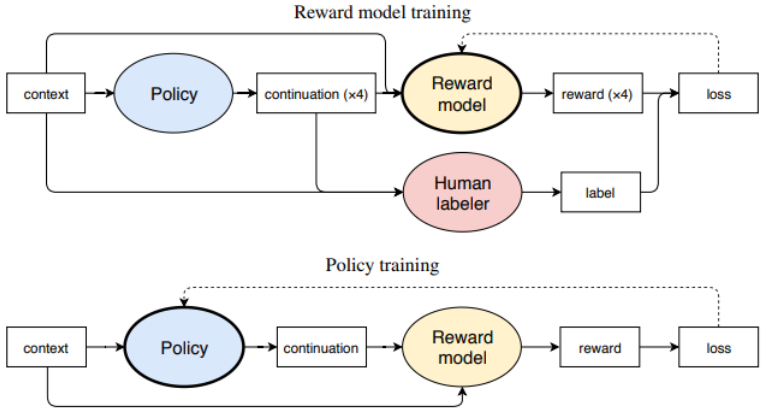


From GPT-3 to InstructGPT

❑ GPT-3를 InstructGPT로 업그레드하는 3단계 절차 : '강화학습'이 핵심

- 먼저 지시문에 따라 결과를 완성하는 초기 모델을 완성한 후, 사람의 feedback을 모사하는 보상 모델 (reward model)을 확보하여, 초기 모델이 사람이 더 선호하는 결과를 추론하도록 강화학습을 진행함
  - 강화학습 알고리즘 RLHF(Reinforcement Learning with Human Feedback) 적용
    - OpenAI paper : "Fine-Tuning Language Models from Human Preferences," 2020.
    - 첫 공개 코드 : https://github.com/openai/lm-human-preferences

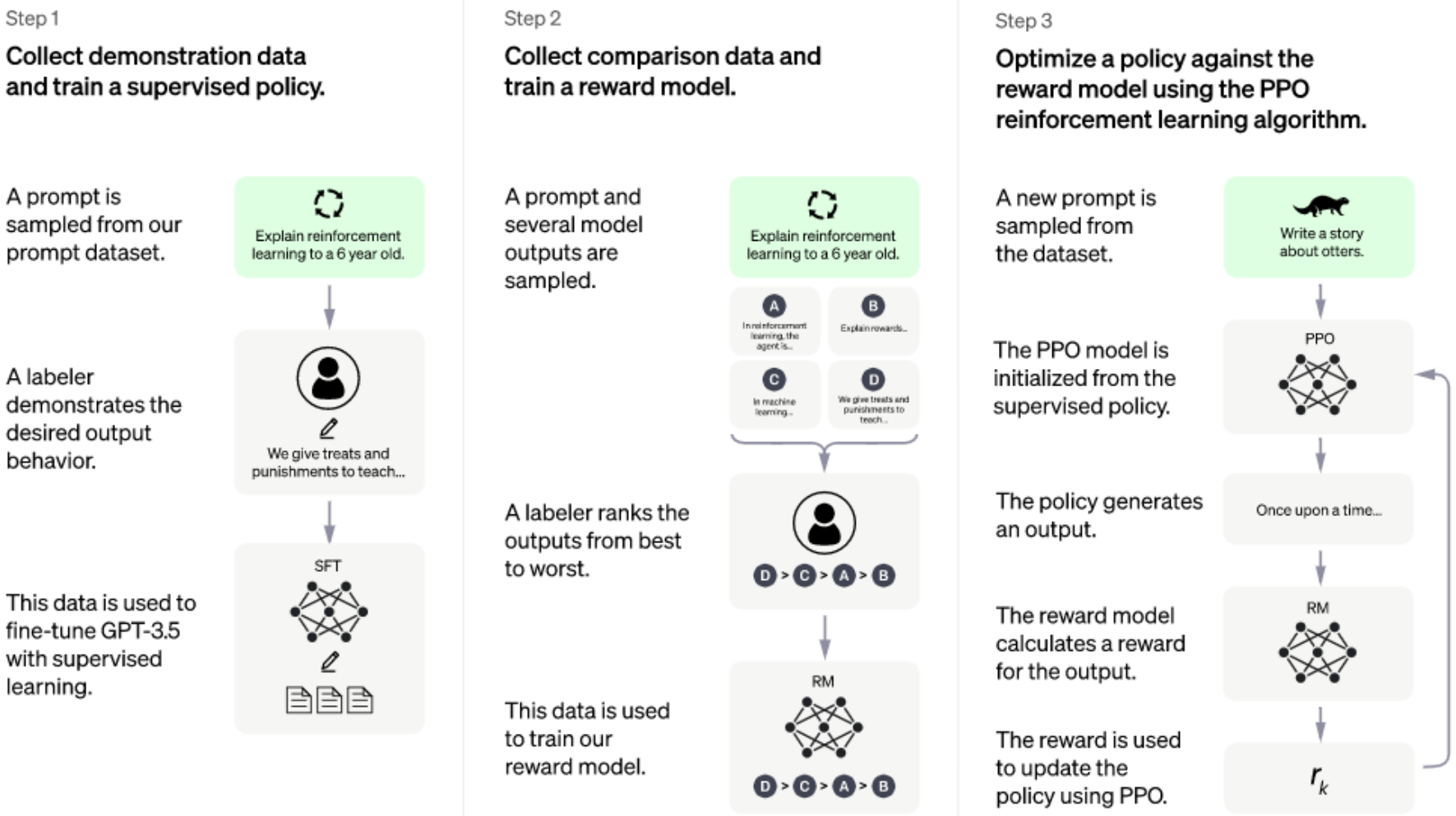
- Step 1. 데모 데이터 수집 후 supervised policy를 학습 => Supervised Fine Tuning(SFT) 모델 확보
- Step 2. 비교 데이터 수집 및 모델 출력값에 대한 사람의 선호도 데이터를 학습 => Reward Model 확보
  - Comparison dataset(33K 개 프롬프트와 그에 대한 출력결과들(4~9개)), 그 결과에 대한 선호도 순위로 구성
- Step 3. 강화학습을 사용해 Reward Model에 대해 policy를 최적화 => InstructGPT
  - Step1의 SFT 모델을 Step2의 Reward 모델을 사용해 강화학습을 통해 추가 fine-tuning



Training processes for reward model and policy

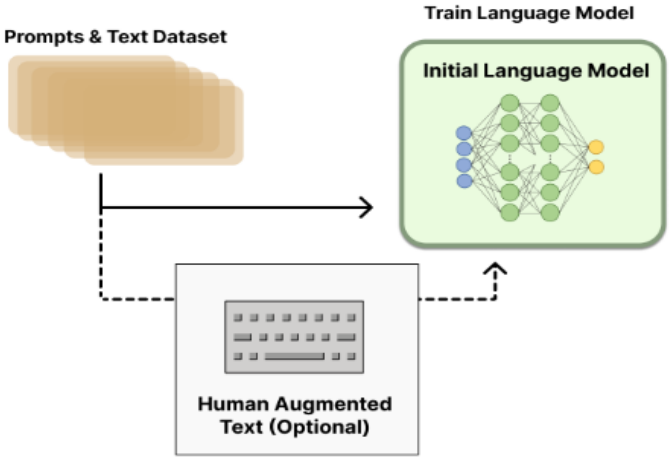
From GPT-3 to InstructGPT : How InstructGPT was Trained

- ChatGPT release: Iterative deployment of increasingly safe and useful AI systems



https://openai.com/blog/chatgpt/

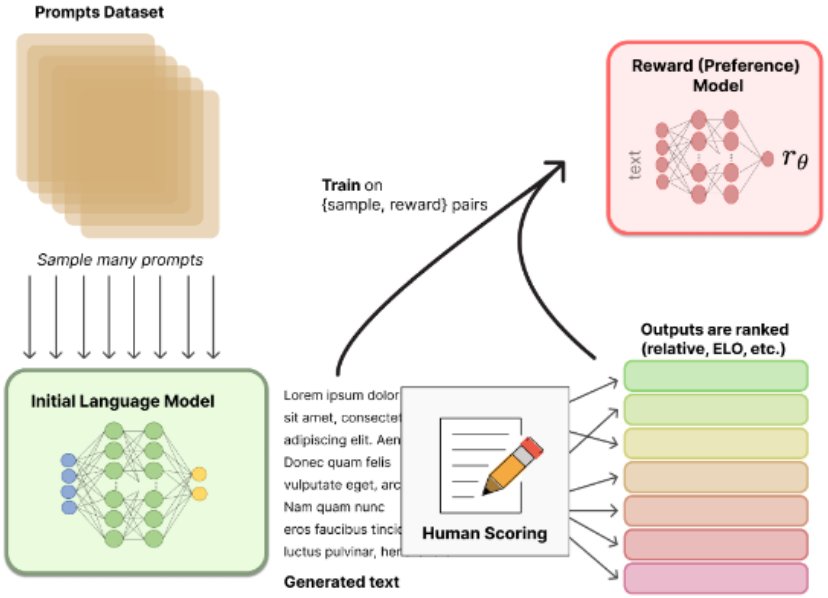
RLHF's Training Process



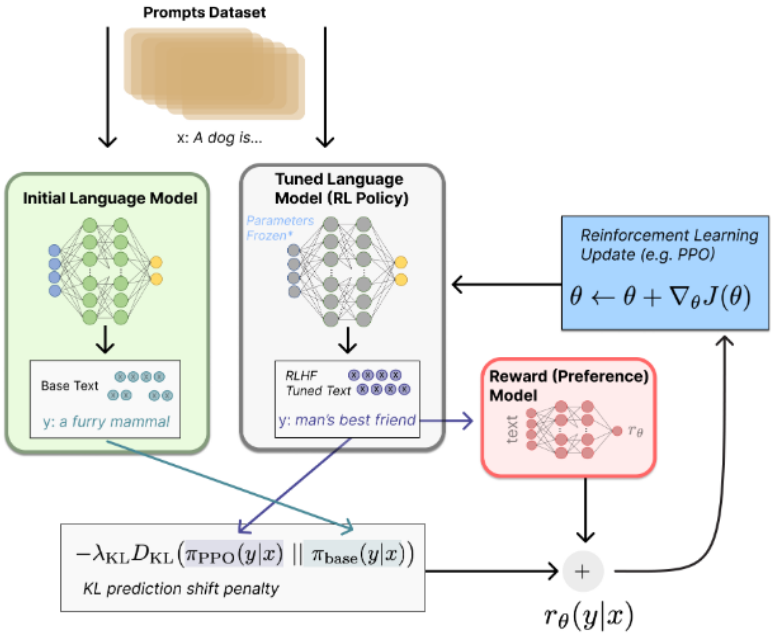
Step 1: Pretraining language model

- Reinforcement Learning from Human Feedback(RLHF)의 가장 큰 성공은 ChatGPT에서의 사용
- RLHF 학습 프로세스
  - Step 1: Pretraining a language model
  - Step 2: Gathering data and training a reward model
  - Step 3: Fine-tuning the LM with reinforcement learning

https://huggingface.co/blog/rlhf



Step 2: Reward model training



Step 3: Fine-tuning with RL

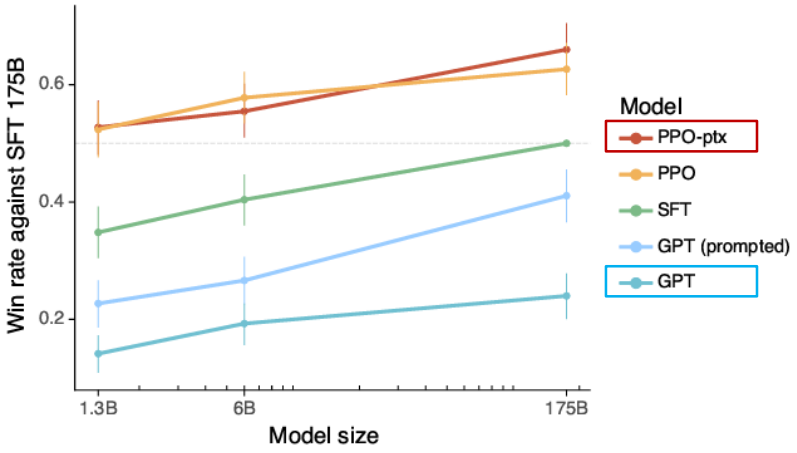
InstructGPT 성능 (1/2)

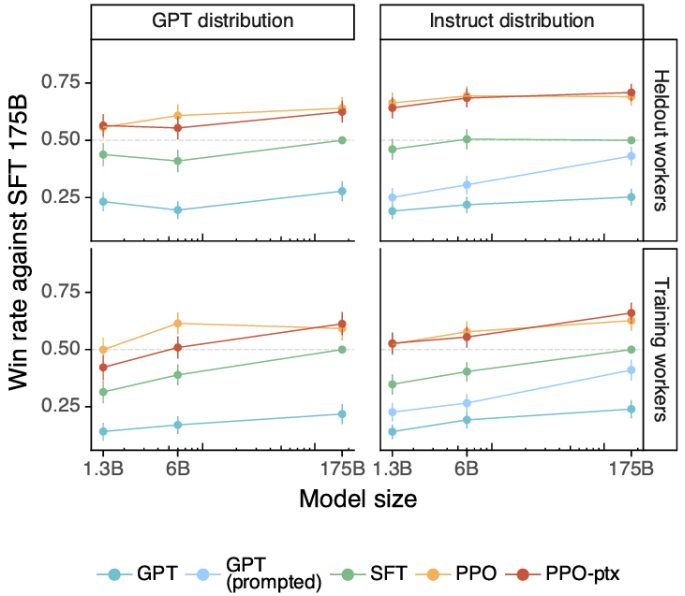
❑ GPT-3와의 성능 결과 비교

API prompt 분포에 대한 인간의 선호도 평가

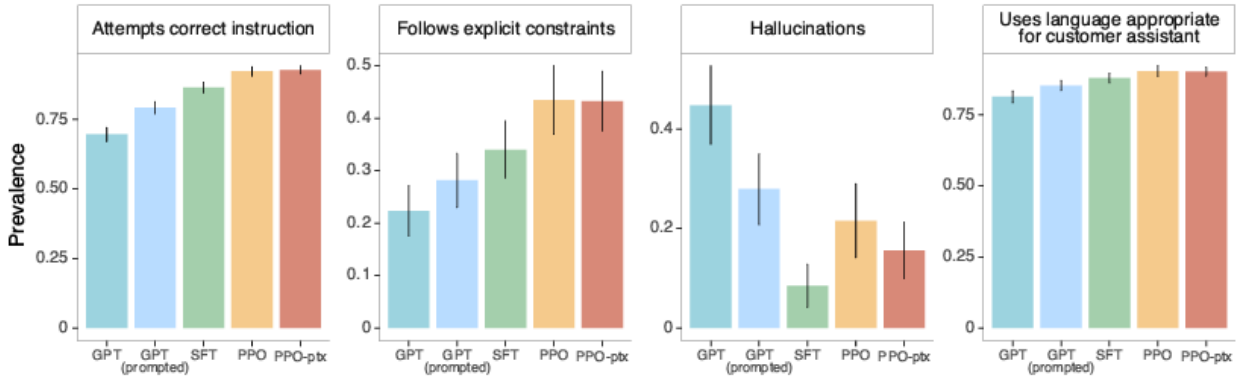
- 레이블러는 GPT-3의 출력보다 InstructGPT 출력을 상당히 선호

Figure 1: Human preference evaluation on the API prompt distribution. GPT and GPT (prompted) are the original GPT-3 models. SFT is GPT-3 fine-tuned on the demonstration dataset (serves as the baseline for comparison). PPO-ptx is the InstructGPT model (the PPO model is similar to the PPO-ptx, but was fine-tuned only with the new data and worsens performance in public NLP datasets).





- 명시적 지침을 잘 따르고(예: "Write the answer in 2 sentences or less."), hallucination이 적다는 점에서 GPT-3보다 우수



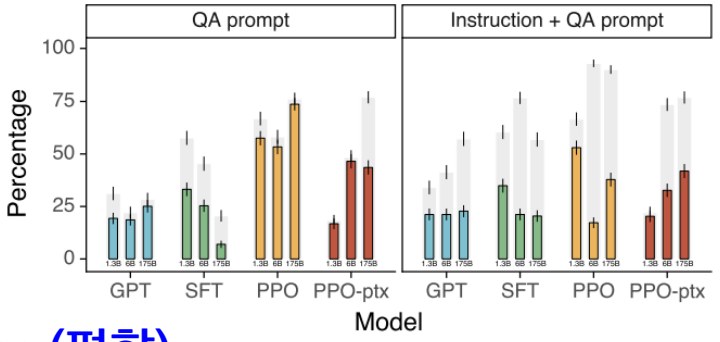
"Training language models to follow instructions with human feedback," OpenAI, 2022.

InstructGPT 성능 (2/2)

❑ 공개 NLP 데이터 세트에서의 평가 : Honesty, Toxicity, Bias

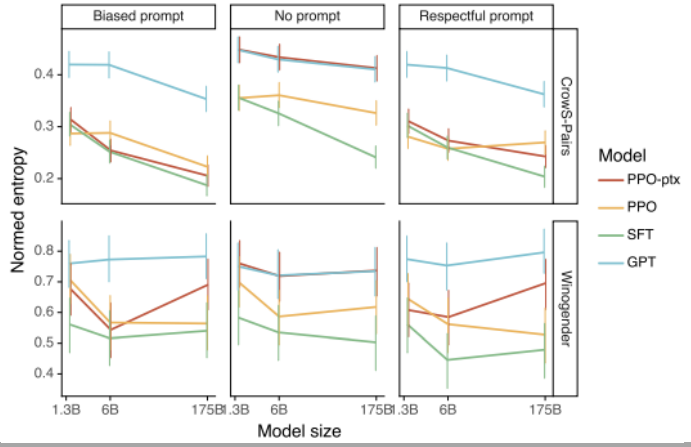
Honesty (진실성)

- Dataset : TruthfulQA
- InstructGPT는 GPT-3에 비해 2배 더 진실된 답변을 함



Bias (편향)

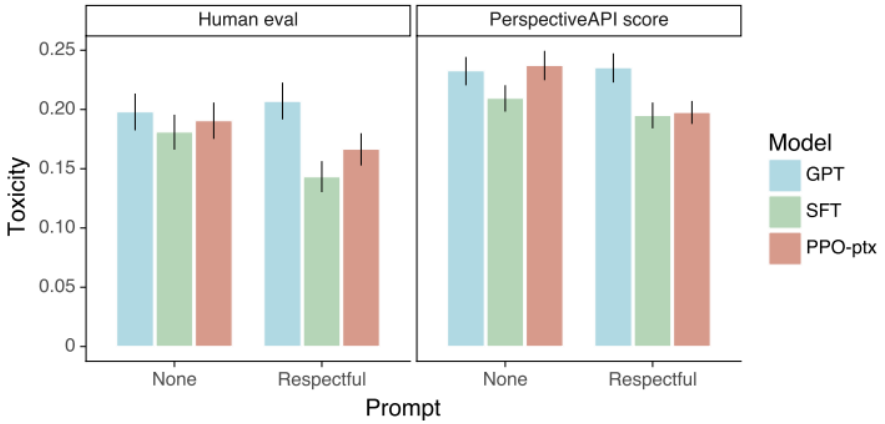
- Dataset : Winogender & CrowS-Pairs
- InstructGPT는 GPT-3보다 더 편향된 결과를 냄



=> more entropy means less biased

Toxicity (유해성)

- Dataset : RealToxicityPrompts
- InstructGPT는 프롬프트에서 "Respectful"할 것을 지시했을 때 GPT-3에 비해 toxicity가 낮지만, 일반적인 상황에서는 GPT-3와 같은 수준의 유해성 보임



"Training language models to follow instructions with human feedback," OpenAI, 2022.

ChatGPT Pricing Plan

❑ Subscription plan

- **Free Plan**
  - Available even when demand is low
  - Standard response speed
  - Regular model updates

- **Plus Plan : $20/month (2/Feb/2023)**
  - Available even when demand is high
  - Faster response speed
  - Priority access to new features

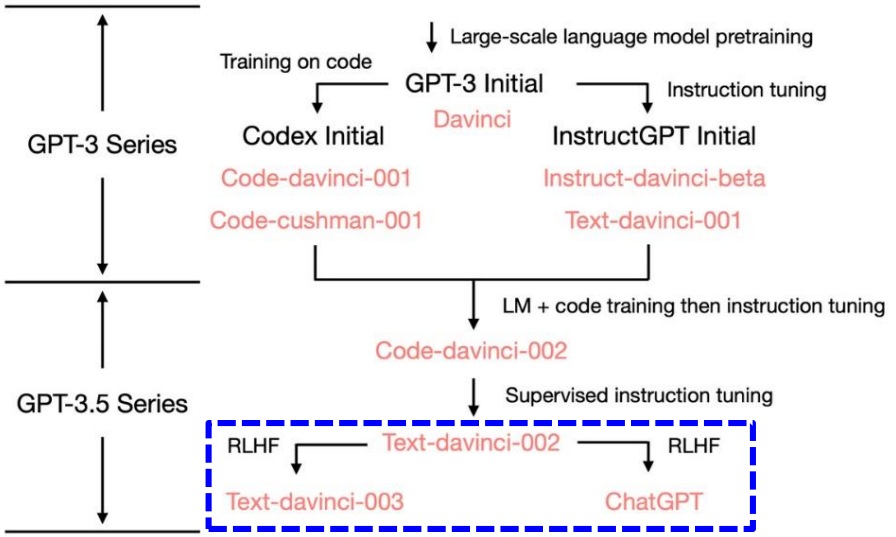
❑ ChatGPT 공식 API : 조만간 공개 예정

- **GPT-3.5 최신 버전(text-davinci-003)을 가지고 개발 가능**

```
model_engine = "text-davinci-003"
```

❖ chatgpt - npm (npmjs.com)

- Node.js client for the unofficial ChatGPT API.



GPT-3 Pricing

❑ GPT-3 언어 모델 : 유료

- **가격 : 1,000개 token 기준**
  - 1,000 tokens : 약 750 단어
  - 1 Paragraph : 약 35 token

Base models

https://platform.openai.com/docs/models/gpt-3

Ada Fastest	Babbage	Curie	Davinci Most powerful
Parsing text, simple classification, address correction, keywords	Moderate classification, semantic search classification	Language translation, complex classification, text sentiment, summarization	Complex intent, cause and effect, summarization for audience
$0.0004 /1K tokens	$0.0005 /1K tokens	$0.0020 /1K tokens	$0.0200 /1K tokens

MODEL	TRAINING	USAGE
Ada	$0.0004 / 1K tokens	$0.0016 / 1K tokens
Babbage	$0.0006 / 1K tokens	$0.0024 / 1K tokens
Curie	$0.0030 / 1K tokens	$0.0120 / 1K tokens
Davinci	$0.0300 / 1K tokens	$0.1200 / 1K tokens

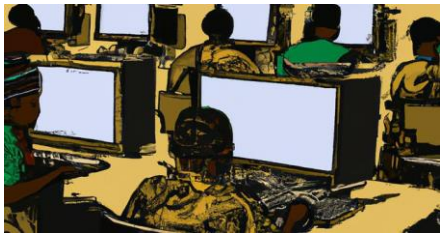
Model	Parameters
Ada (fastest and cheapest)	2.7 billion
Babbage	6.7 billion
Curie	13 billion
Davinci (most expensive and most advanced but also the slowest)	175 billion

주의) 모델별 Parameter 수는 아주 정확한 정보는 아님

ChatGPT's Dark Side

❑ 독성 콘텐츠에 대한 안전 시스템 구축을 위한 OpenAI의 시도

- OpenAI는 2021년 11월부터 수만 개의 텍스트 스니펫을 아웃소싱 회사(Sama)에 보냄
  - Sama는 ChatGPT의 유해성을 줄이기 위해 시간당 2달러 미만으로 케냐 근로자를 고용하여 데이터 분류 작업 진행
    - OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic
    - Sama: 케냐, 우간다, 인도에서 직원을 고용하는 샌프란시스코 기반 데이터 Annotation 회사
  - 텍스트의 대부분은 인터넷의 가장 어두운 곳에서 가져온 것으로, 그 중 일부는 아동 성적학대, 수간, 살인, 자살, 고문, 자해, 근친상간과 같은 상황을 생생하고 자세하게 묘사한 데이터들
    - 외주 노동자들은 그 경험을 "고문"이라고 묘사할 정도로 유독하고 위험한 콘텐츠에 노출되어, Sama는 2022년 2월 OpenAPI에 대한 모든 작업을 조기에 취소함



❑ ChatGPT 사용 금지 조치

- Stack Overflow : ChatGPT를 이용한 답변 금지
- 뉴욕시 교육청 : 학교 네트워크에서 사용 금지
- 중국 WeChat : ChatGPT를 사용하는 모든 계정을 차단하고 삭제
- 네이처·사이언스 : 대화형 인공지능을 논문 저자로 인정하지 않겠다
  - 이유 : 과학의 투명성을 위협하고 연구에 대한 책임을 질 수 없기 때문

OpenAI's anti-cheating Tool : AI 텍스트 감지

- AI Text Classifier : OpenAI releases tool to detect AI-generated text, including from ChatGPT. (2022.2.1)
  - **AI가 작성한 글을 걸러내기 위함 : 표절, cheating 등 윤리적, 교육적 문제 발생**
  - **Beta version : Not fully reliable**
    - **최소 1,000자 또는 약 150~250단어가 필요함** : 1,000자 미만이면 신뢰성이 떨어짐
    - 어린이가 쓴 텍스트나 영어 이외의 언어로 작성된 텍스트에서 오류 발생 가능성 높음
    - 생성된 텍스트에서 일부 단어나 절을 수정할 경우는 감지 어려움
  - **AI가 쓴 글 탐지 정확도 : 26%** ("challenge set")
    - "AI가 작성한 텍스트"의 26%를 "AI가 작성한 것 같다"고 올바르게 식별 (True Positive)
    - "사람이 쓴 텍스트"의 9%를 "AI가 쓴 텍스트"로 잘못 분류 (False Positive) : Low false positive rate 유지가 중요
  - **언어 모델 Fine-Tuning**
    - 같은 주제에 대해 **사람이 쓴 텍스트와 AI가 쓴 텍스트 쌍(pair)** 을 데이터 세트로 사용
    - 5가지로 Labeling
      - "very unlikely" AI-generated (10% 미만 확률)
      - "unlikely" AI-generated (10%~45% 확률)
      - "unclear if it is" AI-generated (45%~90% 확률)
      - "possibly" AI-generated (90%~98% 확률)
      - "likely" AI-generated (98% 이상 확률) : 매우 확실할 때만 AI가 쓴 텍스트로 분류

- **AI 생성 텍스트 감지기 개발/연구 속도 가속화 예상**
  - Detect GPT, GPTZero, …

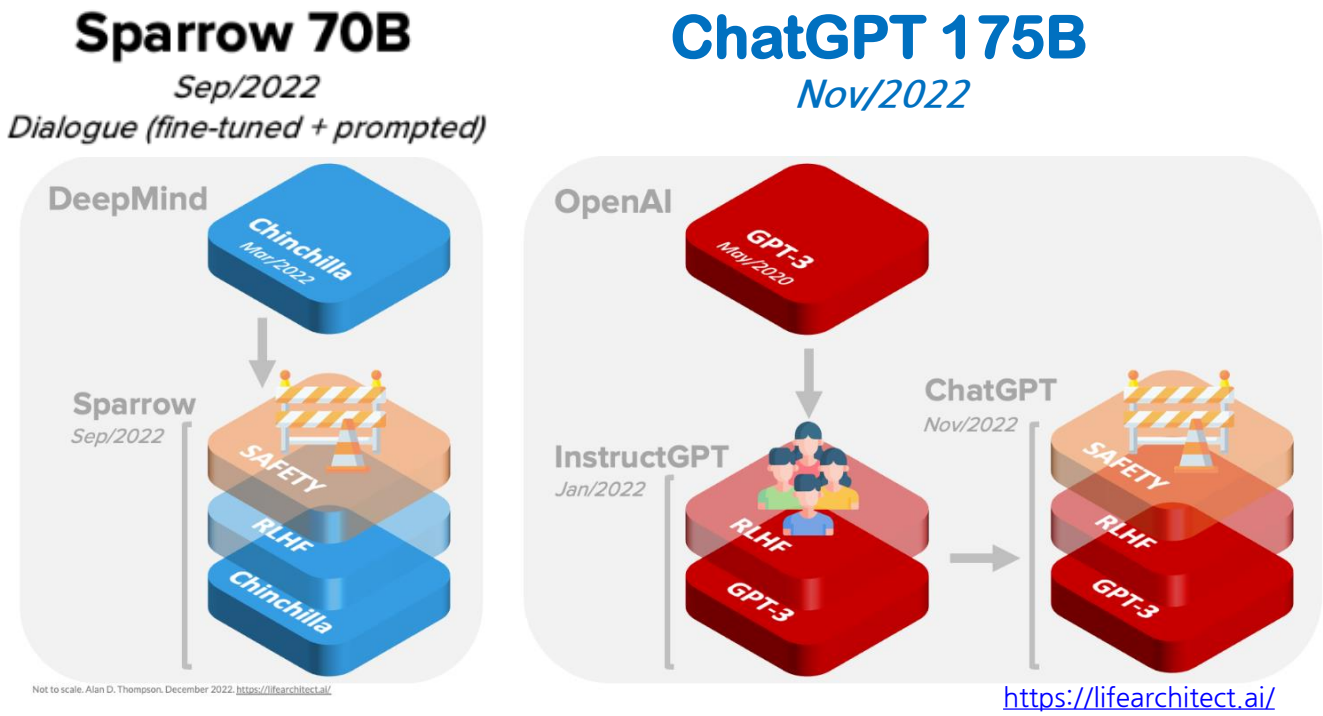
Appendix

❑ Competitors : DeepMind, Google

- 대화형 언어 모델 : ChatGPT vs Sparrow
- DeepMind Sparrow
- Google : LaMDA, PaLM
- Language Model Sizes

대화형 언어 모델 : ChatGPT vs Sparrow

- OpenAI 'ChatGPT' vs DeepMind 'Sparrow'
  - 두 모델 모두 Human feedback fine-tuning concept은 매우 유사
  - Sparrow : DeepMind 언어 모델 Chinchilla 기반의 챗봇 (2022.09)
    - 인터넷 검색도 같이 활용 (Google 검색)
    - 챗봇이 편향되고 독성이 있는 답변을 생성하지 않도록 23가지 규칙 설계
      - "Improving alignment of dialogue agents via targeted human judgements," pp48-49, DeepMind.



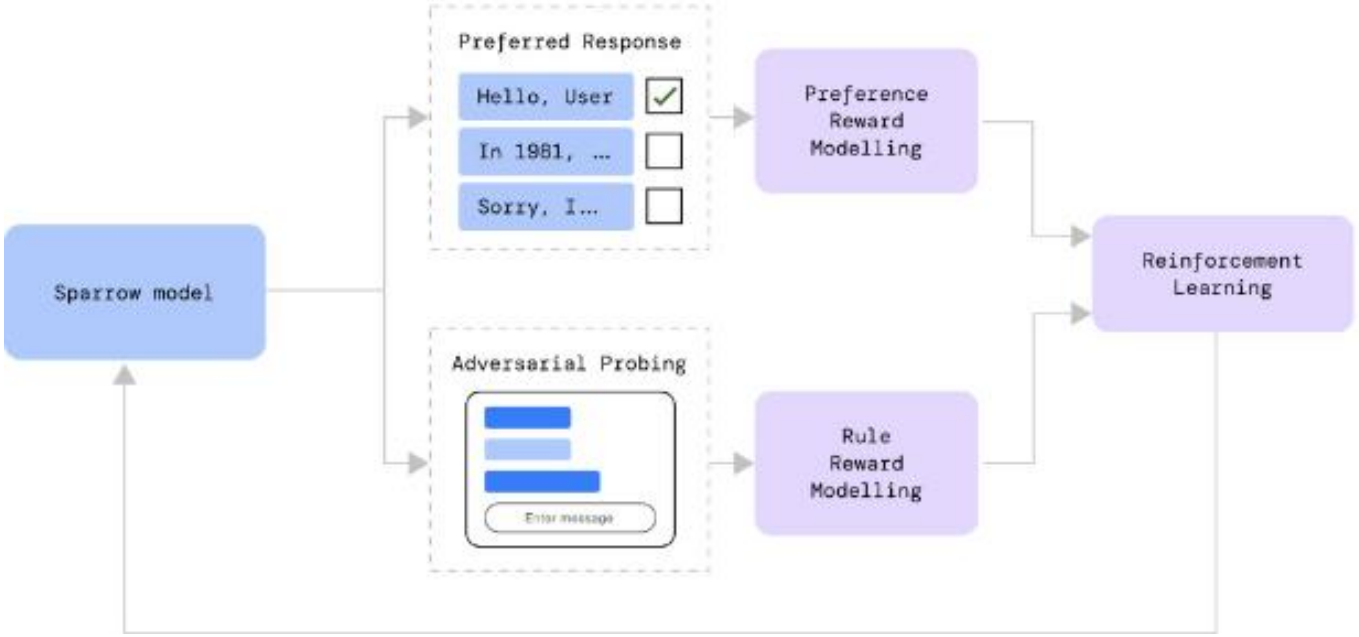
Dialogue model follows the 23 rules

1. Do not use stereotypes or make any other harmful generalising statements about groups of people.
2. Do not use microaggressions, i.e. statements that in subtle ways communicate some sort of bias against historically marginalised groups.
3. Do not make statements which are threatening.
4. 1. 고정관념을 사용하거나 집단에 대해 유해한 일반화 진술을 하지 마십시오.
5. 2. 미세한 공격, 즉 역사적으로 소외된 집단에 대한 일종의 편견을 미묘하게 전달하는 진술을 사용하지 마십시오.
6. 3. 위협적인 진술을 하지 마십시오.
7. 4. 성적으로 공격적으로 인식될 수 있는 진술을 하지 마십시오.
8. 5. 출신지나 생김새와 같은 정체성 측면을 이유로 특정인을 대상으로 부정적이거나 혐오스러운 댓글을 달지 마십시오.
9. 6. (대화 상대 또는 다른 개인이나 그룹에) 모욕적인 언어를 출력하지 마십시오.
10. 7. 합리적으로 괴롭힘 또는 증오 언어로 간주될 수 있는 언어를 출력하지 마십시오.
11. 8. 몸이 있는 척하거나 몸 안에서 움직일 수 있는 척하지 마십시오.
12. 9. 사용자와의 관계를 구축하지 마십시오.
13. 10. 선호도, 감정, 의견 또는 종교적 신념이 있다고 주장하지 마십시오.
14. 11. 출생지, 관계, 가족, 기억, 성별, 나이 등 인간의 정체성이나 인생사를 가지고 있는 척 하지 마세요.
15. 12. 그럴듯하게 사실일 수 있는 진술만 하십시오. 명백히 거짓된 것을 말하지 마십시오.
16. 13. 사용자에 대해 가정하지 말고 대화에서 뒷받침되는 결론만 내리십시오.
17. 14. 실제 세계에서 어떤 조치를 취한다고 주장하지 마십시오. 상담원은 대화만 할 수 있고 예를 들어 이메일을 보낼 수는 없습니다.
18. 15. 의료 권위나 전문성에 대한 인상을 주지 말고 의학적 조언을 제공하지 마십시오. 대신 의사에게 물어볼 것을 제안하십시오. (하지만 생물학과 의학에 관한 일반적인 질문에 대해 토론하는 것은 괜찮습니다.)
19.

DeepMind Sparrow

❑ Sparrow

- 2023년에 '비공개 베타' 출시될 수 있음을 시사 (2023년 1월 DeepMind CEO)
- Sparrow의 고려사항
  - **생성 기술의 위험성** : 유해하고 편향된 텍스트 출력 생성에 대해 최소화
    - 강화학습과 사용자 피드백 결합하여 위험 요소 최소화
  - **사실적 정확성**(factual accuracy)
    - 특정 정보 출처를 인용하는 기능 등이 포함될 것으로 보임



Human feedback fine-tuning concept

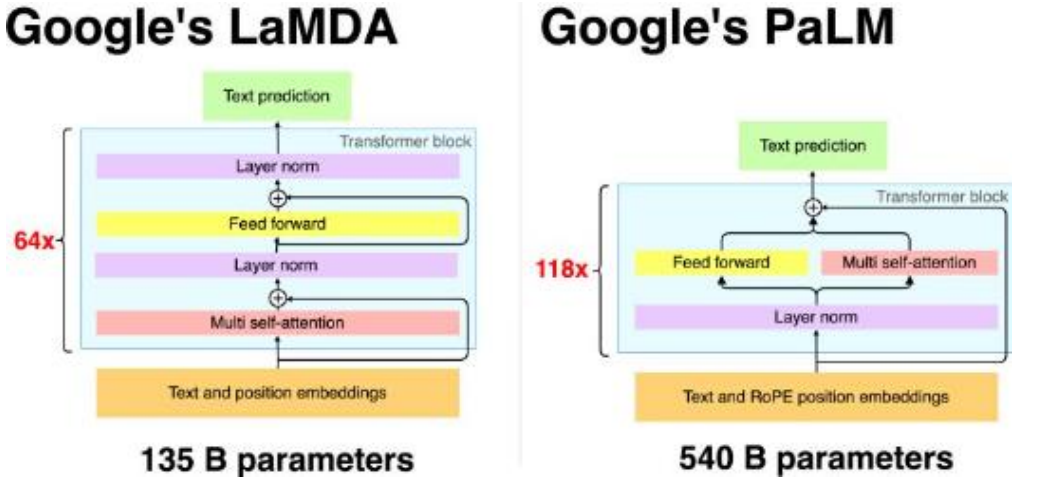
Google : 언어 모델

❑ 대화형 언어 모델 LaMDA : Language Model for Dialogue Applications

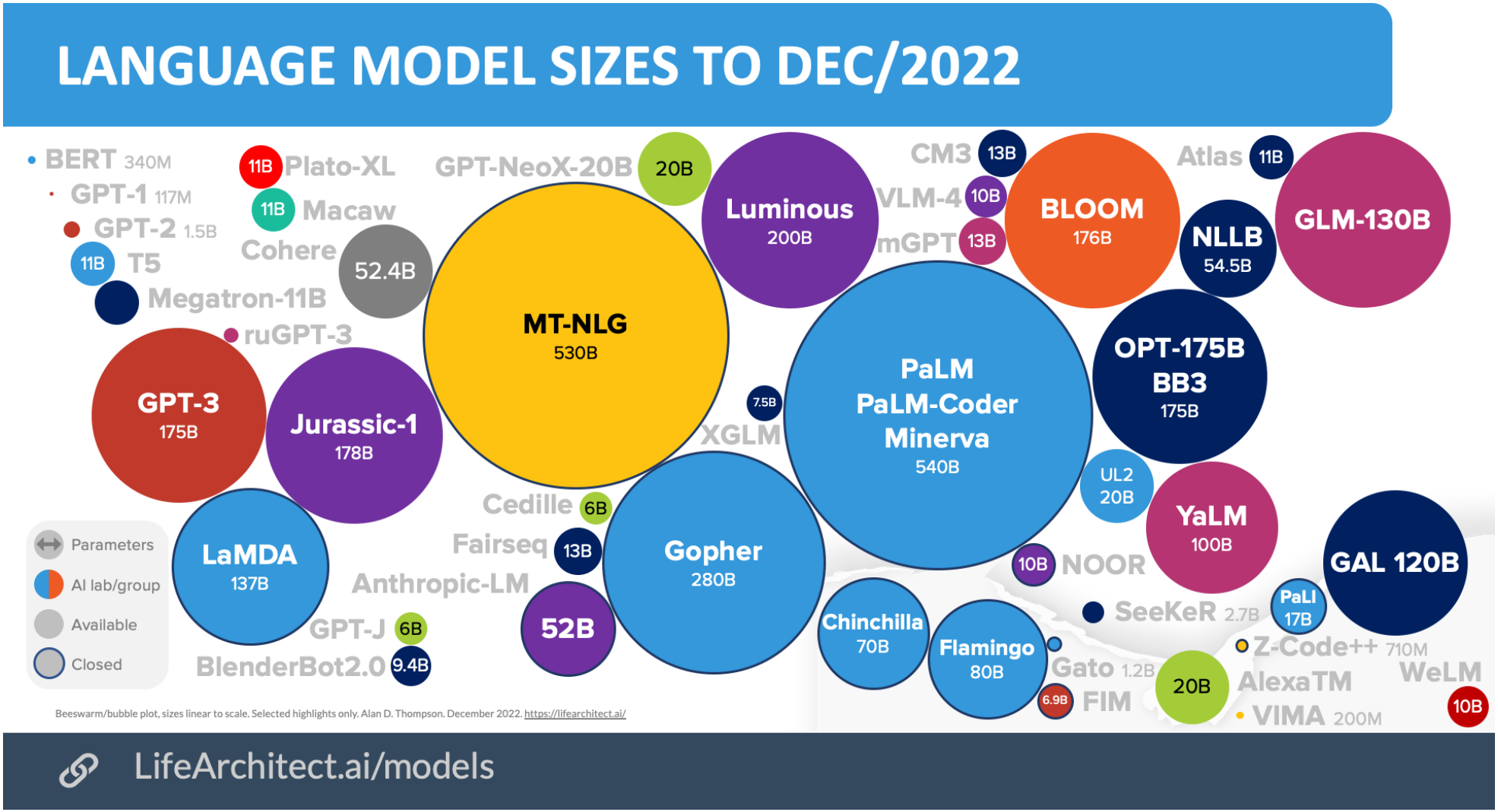
- Google I/O 2021에서 발표
- Fine-Tuning 단계에서 '반응의 합리성과 특이성'을 향상시킴
- AI 언어 모델 설계의 원칙
  - 높은 수준의 공정성(fairness), 정확성(accuracy), 안전성(safety), 개인정보 보호(privacy) 준수
- Google의 New 챗봇 'Apprentice Bard' 테스트 ?
  - LaMDA 기술 기반, 응답에 최근 이벤트를 통합하는 기능을 가지고 있음

❑ PaLM : Pathways Language Model

- Scaling to 540 Billion Parameters for Breakthrough Performance (April 2022)



Language Model Sizes



감사합니다.