랭체인&랭그래프로 AI에이전트 개발하기

오산개발자모임

나명진

트랜스포머 모델의 등장

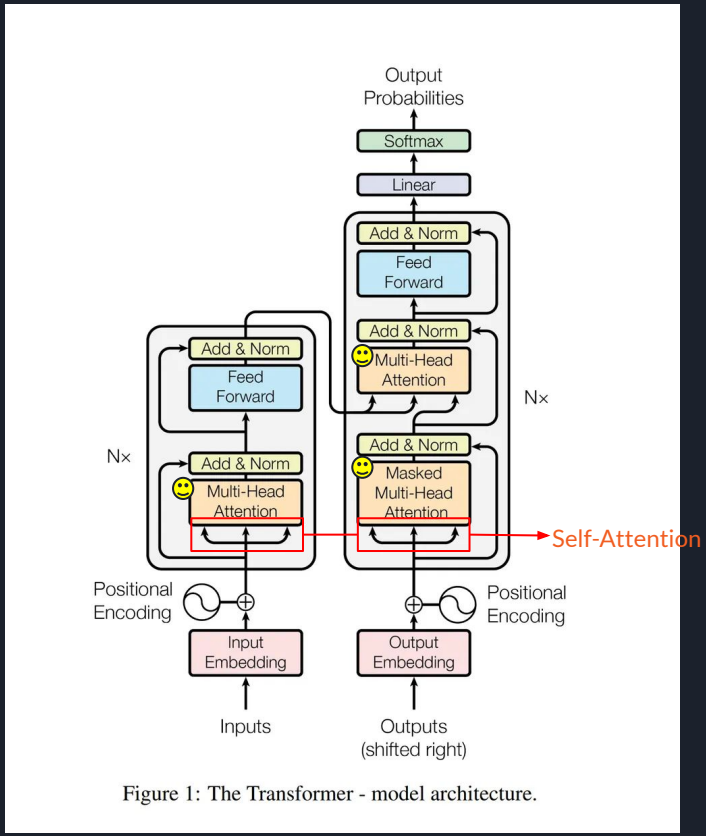


Figure 1: The Transformer - model architecture.

https://arxiv.org/pdf/1706.03762

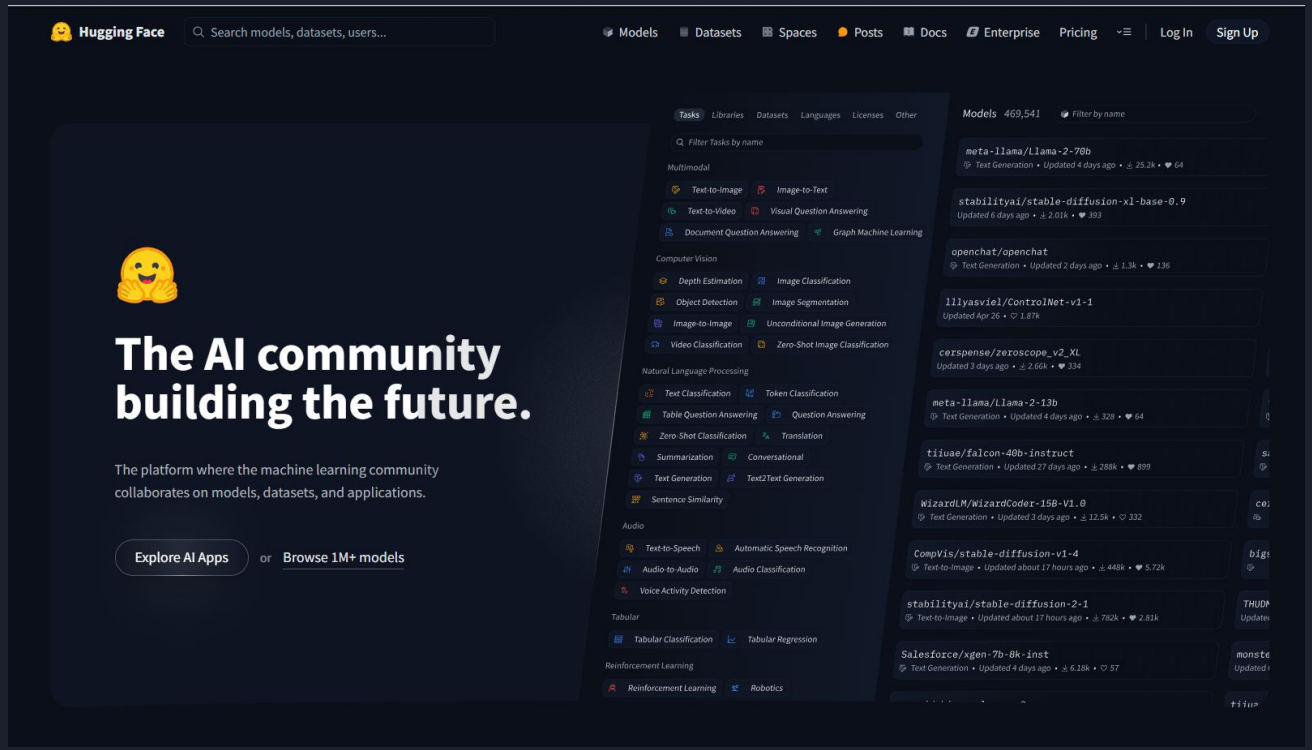
트랜스포머는 기존 모델들에 비해 두 가지 주요 이점을 제공

1. 병렬 처리가 가능하여 학습 시간이 크게 단축됨
2. 시퀀스 내 장거리 의존성을 더 효과적으로 포착함

- CNN, RNN을 대체
- 라벨링 없앰
- 병렬 처리에 적합한 연산 방식 사용 (고속)

Hugging Face 소개

Hugging Face는 AI 및 머신러닝 분야에서 가장 중요한 오픈소스 플랫폼 중 하나로, 특히 자연어 처리(NLP)와 관련된 모델, 도구, 데이터셋을 공유하고 협업할 수 있는 생태계를 제공합니다.



https://huggingface.co/

대략적인 GPU 견적

1. GPU 중심 고사양 워크스테이션 비교 견적

구분	다나와 맞춤형	인텔리존	컴퓨존	최고사양	한성컴퓨터
CPU	AMD Ryzen Threadripper PRO 7975WX (32코어)	Intel i9-14900K (24코어)	AMD Ryzen 9 7950X (16코어)	Intel Xeon w9-3495X (56코어)	AMD EPYC 9554P (64코어)
GPU	**RTX 4090 24GB × 2 (NVLink)**	**RTX 6000 Ada 48GB**	**RTX 4090 24GB**	**A6000 48GB × 2**	**RTX 4090 24GB × 4**
RAM	DDR5 128GB (4800MHz)	DDR5 128GB (5600MHz)	DDR5 64GB (5200MHz)	DDR5 256GB (4800MHz)	DDR5 256GB (4800MHz)
저장 장치	NVMe SSD 2TB (Gen4) + HDD 8TB	NVMe SSD 1TB (Gen4)	NVMe SSD 1TB (Gen4)	NVMe SSD 4TB (Gen4)	NVMe SSD 2TB (Gen4) × 2
파워	1600W 80+ Titanium	1200W 80+ Platinum	1000W 80+ Gold	2000W 80+ Platinum	1500W 80+ Titanium
쿨링	수랭 + 고급 케이스	공랭 + 오픈형 케이스	수랭	수랭 (서버급)	수랭 + 풀타워
가격	**1,200~1,500만 원**	**1,800~2,200만 원**	**900~1,200만 원**	**3,500~4,500만 원**	**2,500~3,000만 원**
추천 이유	가성비 멀티GPU 구성	단일 고용량 GPU 최적화	중급 사양 + 확장성	전문가용 서버급	다중 GPU 극한 활용

돈이 정말 엄청나게 드는군!

인공지능 산업융합 사업단

지원대상

o **(지원대상)** 국내 기업, 공공기관, 연구소(원), 대학교(원), 협·단체 등

구분	지원 대상	지원 기준	비고
1	국내기업	기업당 1건 지원	사업자등록증 등
2	대학교(원)	과·부별 1건 지원	재직증명서 등
3	공공기관, 연구소(원), 협 단체	부서별 1건 지원	재직증명서 등

※ 소속이 없는 개인 및 대기업은 지원 대상에서 제외 단, 사업단이 지원하는 R&D 사업 수행기업(관)은 지원 대상에 포함

※ **반드시 지원 기준당 1건 신청 가능하며, 모집 지원 타입별 중복신청불가**

o **(지원기간)** '25. 4. ~ '25. 12.

구분	자원풀	이용기간	이용개월수	비고
1	전용풀	25.4. ~ 25.12.	9개월	
2	HPC	25.4. ~ 25.6	3개월	

※ 기본지원 기간은 3개월(HPC) 또는 9개월(전용풀) 이며 이용종료 이전 서비스 이용해지 (최소 15일이전) 신청 가능

신청자원

o **(지원구성 및 모집규모)** 모집비율 내 **이용자 선정 (약 160개사 내외)**

구분				모집 비율	주요특징
자원풀	가속기	CPU	서버 메모리		
❶ 전용풀	64TF (T4*8)	16코어	128GB	10%내외	o 가속기 서버 1식 단독 사용
	78TF A100*4	48코어	900GB	15%내외	
	156TF A100*8	92코어	900GB	15%내외	
	67TF (H100*1)	92코어	900GB	30%내외	
	536TF (H100*8)	92코어	900GB	20%내외	
	700TF BOW*8	32코어	800GB	10%내외	
❷ HPC	1~3PF (H100)	-	-	5%내외	

※ 모집비율은 예상규모이며 정책지원, 지원희수, 대기풀현황 등 상황에 따라 변경될 수 있음

※ 모집규모 내에서도 평가기준에 부합하지 않을 경우 모집규모 이하로 선정할 수 있음

신청방법

o **(지원신청)** 환경세팅(최대 2주) 후 컴퓨팅 자원의 즉각적 이용이 가능한 계획을 수립하여 자원 신청

o **(신청기간)** '25.2.14.(금) ~ '25.3.11.(화) 14:00 / 약 4주간

o **(신청방법)** 제출서류를 작성 후 메일신청

o **(제출처)** aica_dc@aicluster.or.kr

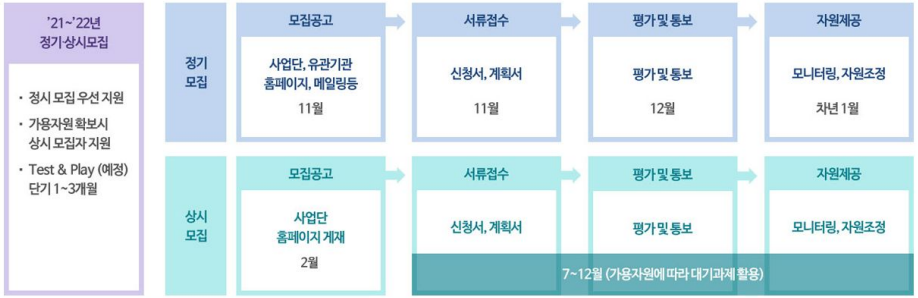
제출처	aica_dc@aicluster.or.kr
메일제목 /압축파일명	25년 AIDC 서비스 이용신청_기업명_자원명 ex) 25년 AIDC 서비스 이용신청_인공지능산업융합사업단_전용풀(T4*8

신청안내

서비스 신청 | 서비스 이용

신청방법

- **지원대상** AI 기술연구 및 제품·서비스 개발과 상품화를 목표로 하는 과제를 수행하는 국내 기업과 연구기관·대학·공공기관 등
- **모집방법** 매년 11월에 차년도 이용자를 모집(정시모집)하며, 이후 확보되는 가용자원 제공을 위해 상시모집(2월~3월) 추진
- **담당자** 인공지능 데이터센터 컴퓨팅자원팀 (aica_dc@aicluster.or.kr)
- **서비스 종류 :** 데이터센터 서비스, 단독 스토리지, 상용화 수준, 체험 등
  - 데이터센터 서비스 : **AI 연구 · 상품개발 서비스 지원을 위한 가속기+스토리지, 개발환경, 빅데이터 AI 플랫폼, 다양한 SW솔루션서비스(SaaS) 제공**
  - 단독 스토리지 : 최대 500TB 파일시스템(NAS)을 VM과 연계하여 제공 및 필요 시 가속기 제공
  - 상용화 수준 : 학습모델 개발 완료에 따른 오픈베타 서비스를 위한 스토리지, VM 등 제공
  - 체험 서비스 : 공용풀 기반으로 가속기(V100), 스토리지(5TB) 제공 * 최대 3개월 이용



https://aica-gj.kr/main.php

로컬에 멀티모달 LLM 서버를 구축하려면

기본 하드웨어 구성 (최소 사양)

- **CPU**: 최신 Intel Xeon 또는 AMD EPYC (16코어 이상)
- **RAM**: 64GB 이상 (128GB 권장)
- **GPU**: NVIDIA RTX A6000 또는 A100 (24GB VRAM 이상)
- **저장장치** : 2TB NVMe SSD (모델 및 캐시용)
- **네트워크** : 10Gbps 이더넷

예상 비용 (USD 기준)

- **기본 구성** : $10,000 ~ $15,000
- **고급 구성** : $25,000 ~ $40,000

운영 비용 고려사항

- **전력 소비** : A100 서버는 일반적으로 1,000W ~ 1,500W 소모
- **냉각 시스템** : 추가 비용 발생 가능
- **유지 보수** : 부품 교체 및 업그레이드 비용

로컬에 멀티모달 LLM 서버를 구축

데이터 수집 및 통합

- **센서 데이터**: 온도, 압력, 유량, 진동, 전류/전압, 가스 농도 등
- **이미지/비디오 데이터**: 웨이퍼 표면, 장비 상태, 공정 모니터링 영상
- **오디오 데이터**: 장비 작동 소리, 이상 소음, 알람 등
- **공정 정보**: 레시피, 매개변수, 수율 데이터, 품질 측정 결과

반도체 제조 특화 기능 구현

이상 감지 기능

- 정상 패턴에서 벗어난 이상 상황 감지
- 다변량 센서 데이터 기반 이상 점수 계산

공정 최적화 기능

- 현재 공정 상태에 따른 파라미터 조정 추천
- 수율 최적화를 위한 설정 제안

예지 정비 기능

- 장비 고장 사전 예측
- 소리 및 진동 패턴 기반 유지보수 필요성 알림

시각적 검사 기능

- 웨이퍼 표면 결함 감지
- 패턴 일관성 분석

로컬에 멀티모달 LLM 서버를 구축 사업을 한다면

총 예상 사업비

- **초기 개발 (1년)**: $850,000 ~ $1,200,000
- **첫 고객 구축 비용**: $90,000 ~ $240,000
- **첫해 운영 비용**: $100,000 ~ $200,000

총 첫해 사업비 (1개 고객 기준): $1,040,000 ~ $1,640,000

고객 관점의 ROI

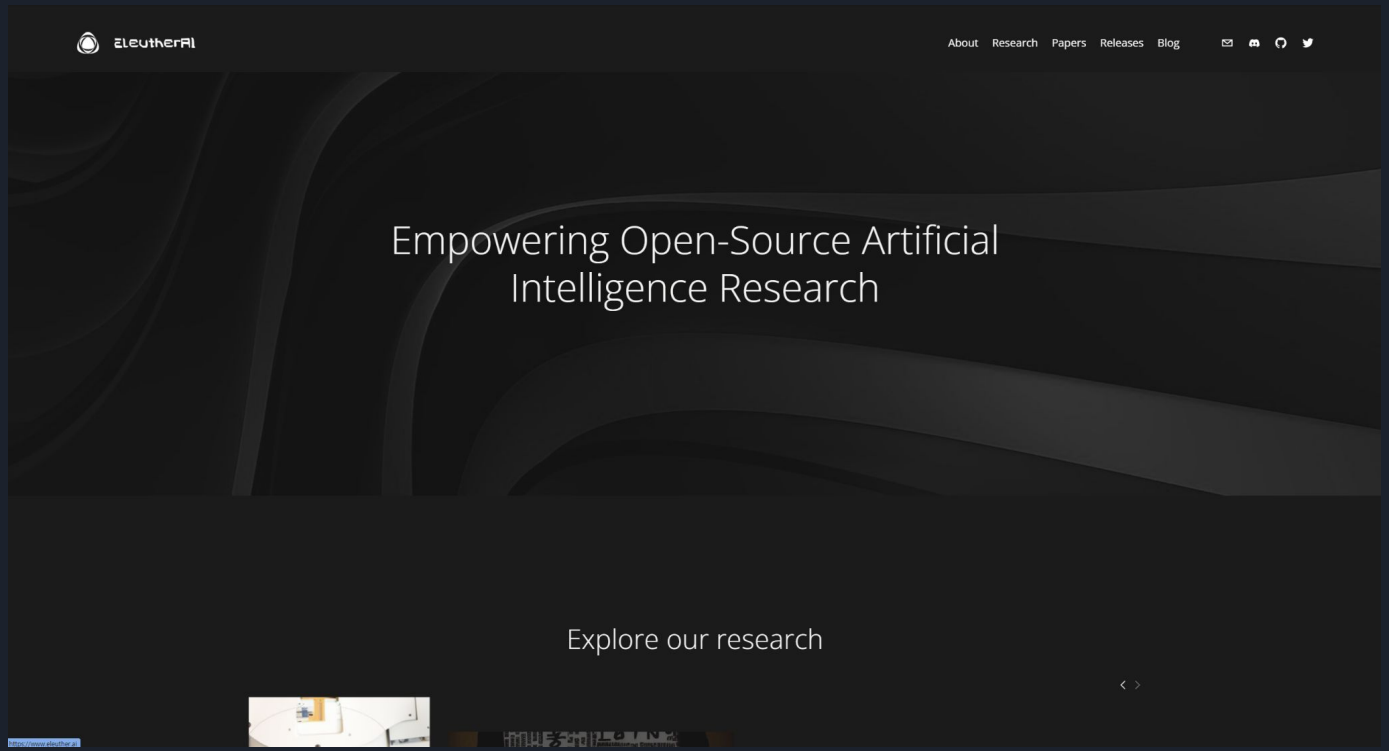
- **수율 향상**: 0.52% 수율 증가 = 연간$10,000,000 절감 (팹 규모에 따라)
- **다운타임 감소**: 30% 감소 = 연간$5,000,000 절감
- **품질 향상**: 결함률 25% 감소 = 연간 $3,000,000 절감
- **예상 총 ROI**: 투자 대비 500% (12년 내)

사업자 관점의 ROI

- **첫해 투자**: $1,040,000 ~ $1,640,000
- **첫 고객 수익**: $250,000 ~ $450,000 (설치 + 첫해 라이센스)
- **손익분기점** : 4~6명의 고객 확보 시 (약 2년)
- **5년차 예상 고객**: 20~30개 공장
- **5년차 예상 연간 수익**: $5,000,000 ~ $10,000,000
- **5년 누적 ROI**: 초기 투자 대비 약 800~1,200%

EleutherAI

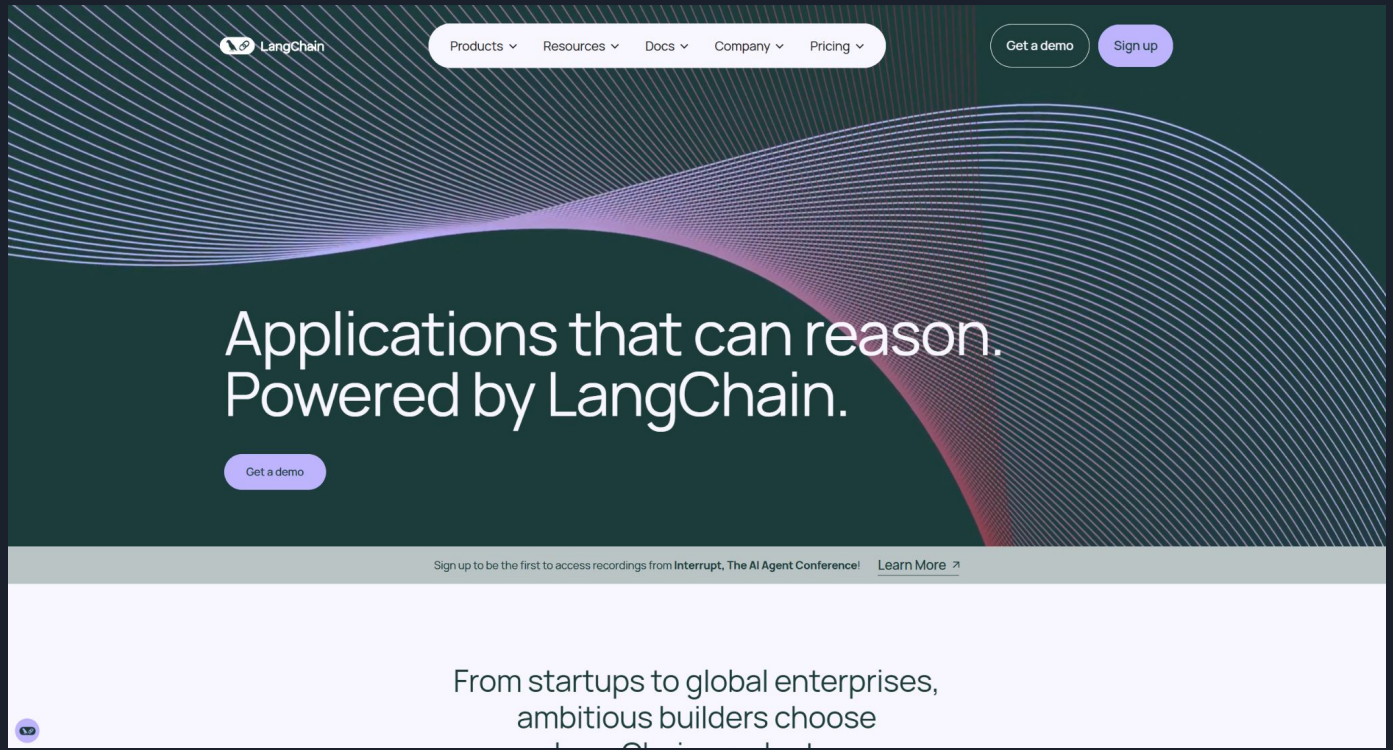
오픈소스 AI 연구 커뮤니티로, 대규모 언어 모델(LLM)을 공개적으로 개발하고 연구하기 위해 2020년에 설립되었습니다. 이 단체는 AI 연구의 민주화와 투명성을 증진하는 것을.목표로 하고 있습니다.



https://www.eleuther.ai/

LangChain

대규모 언어 모델(LLM)을 활용한
애플리케이션 개발을 위한 프레임워크



https://www.langchain.com/

GPT-o1과 Claude 3.7 Sonnet 비교

GPT-o1

- **추론 능력**: 복잡한 논리적 추론과 수학적 문제 해결에 특화됨
- **컨텍스트 길이**: 최대 128K 토큰(약 400페이지)
- **응답 속도**: 빠른 처리 속도 제공
- **주요 강점**: 코딩, 수학적 문제 해결, 긴 컨텍스트 처리

Claude 3.7 Sonnet

- **균형적 능력**: 추론과 자연스러운 대화 능력 사이의 균형이 좋음
- **컨텍스트 길이**: 약 200K 토큰으로 GPT-o1보다 더 길다
- **인간적 응답**: 더 자연스럽고 공감적인 응답 생성에 강점
- **주요 강점**: 복잡한 지시 이해, 긴 문서 처리, 균형 잡힌 응답

왜 o1은 더 높은 수준의 논리적 사고력를 갖는가?

1. RLUF(Reinforcement Learning from Update Feedback) 도입
2. 학습 데이터 구성의 변화
3. 모델 아키텍처의 최적화
4. 학습 알고리즘의 발전
5. 평가 방법의 혁신

AI Agent

특정 목표를 달성하기 위해 환경을 인식하고, 의사결정을 내리며, 자율적으로 행동할 수 있는 인공지능 시스템

핵심적인 특징

1. **자율성 (Autonomy)**: 사용자의 지속적인 감독 없이 독립적으로 작동할 수 있는 능력
2. **지각 (Perception)**: 환경으로부터 정보를 수집하고 이해하는 능력
3. **추론과 의사결정 (Reasoning & Decision-making)**: 수집된 정보를 바탕으로 판단하고 결정을 내리는 능력
4. **행동 (Action)**: 결정에 따라 환경에 영향을 미치는 조치를 취하는 능력
5. **목표 지향성 (Goal-orientation)**: 특정 목표나 목적을 위해 작동함
6. **적응성 (Adaptability)**: 환경 변화에 대응하여 전략을 조정할 수 있는 능력