AI를 이해한다는 것

01

규칙, 학습, 언어 모델이 각각 어떻게 답을 만드는지 구분합니다.

01 AI, 모델, 서비스를 구분해요

AI · 머신러닝	AI는 인식·판단·생성 등을 다루는 분야입니다. 머신러닝은 데이터에서 패턴을 학습하는 접근이고, 딥러닝은 여러 층의 신경망을 사용하는 머신러닝의 한 분야입니다.
LLM · GPT	LLM은 대규모 언어 모델입니다. GPT는 그중 한 모델 계열입니다.
ChatGPT	모델을 대화·파일·검색 등과 함께 사용할 수 있게 만든 서비스입니다.

02 판단 기준을 어디서 얻을까요?

규칙 기반

사람이 기준을 정합니다. "30°C 이상이면 냉방을 켠다."

학습 기반

온도와 사용 기록을 학습해 냉방이 필요할 때의 패턴을 찾습니다.

1950년 튜링의 질문 → 1956년 다트머스 연구 모임 → 1950년대 후반 퍼셉트론 연구. 규칙과 학습은 오늘날에도 함께 쓰입니다.

03 모델은 예측하고, 학습은 값을 조정해요

> **40 + 5 × 3 = 55점** 선형회귀 예시 · 3시간을 넣어 점수를 예측

여기서 **40과 5는 매개변수**입니다. 실제 점수와 예측값의 차이를 손실로 계산하고, 손실을 줄이는 쪽으로 값을 조정합니다. **추론**은 학습한 값으로 새 입력의 결과를 계산하는 단계입니다.

학습에 쓰지 않은 사례에서도 잘 작동하는지 확인합니다. 학습 사례에 지나치게 맞춰 새 사례에서 성능이 떨어지는 현상이 **과적합**입니다.

문제부터 고르기 · 수치 예측은 회귀, 범주 판별은 분류, 문장 작성은 생성의 예입니다. 데이터의 양·정확성·다양성도 살핍니다.

04 언어 모델은 문맥을 보고 다음 토큰을 이어요

토큰	텍스트를 모델이 처리하는 조각입니다. 단어보다 작을 수 있습니다.
어텐션 [1]	문맥 속 토큰 사이의 관계를 계산해 관련 정보를 반영합니다.
컨텍스트 창	한 번에 처리할 수 있는 토큰 범위입니다. 매개변수 수와는 별개입니다.

GPT 계열은 다음 토큰의 확률을 계산해 선택한 뒤 이어서 생성합니다. **사전학습**에서 일반 패턴을 배우고, **후속 학습**에서 예시나 피드백으로 응답 방식을 다듬습니다.

수업 중 필기

예측값과 근거를 함께 적어요.

공부 시간이 5시간이라면 예측값은?

이 모델을 새 학생에게 쓰기 전 확인할 점은?

참고 [1] Vaswani 외 · Attention Is All You Need · UNLOCK 1회차 슬라이드

AI에게 일을 맡기는 법

02

요청, 도구, 결과 검증까지 한 번에 따라갑니다.

05 같은 모델도 지침과 자료에 따라 답이 달라져요

"사과는 무슨 색인가요?"

"빨강, 초록, 노랑 등 여러 색이 있어요."

"한 단어로만 답해줘."

"빨강." 같은 질문에 출력 형식을 추가한 예입니다.

대화에 넣은 자료와 지시문은 이번 작업의 문맥으로 쓰입니다. 학습은 모델의 매개변수를 조정합니다. 채팅창에 쓴 지침은 사용자 요청이고, 앱이 별도로 전달하는 시스템 지침은 요청을 처리할 때 적용하는 규칙입니다.

06 모델에서 에이전트까지

> **목표 받기 → 다음 행동 결정 → 도구 실행 → 결과 확인**

모델은 요청과 정보를 바탕으로 다음 행동을 고릅니다. 실행 환경은 파일을 읽고, 도구를 실행하고, 권한과 작업 상태를 관리합니다. 이 실행 구조를 **하네스**라고 부릅니다. 마지막에는 사람이 실제 결과물을 열어 확인합니다.

Skill	반복하는 작업의 절차와 참고 자료. 예를 들어 보고서 검수 순서.
MCP	AI 앱이 도구와 데이터에 연결될 때 쓰는 표준. 연결 후 권한이 필요합니다.
오케스트레이션	검색·작성·검토처럼 여러 단계의 순서와 결과 전달을 조율하는 일.

07 실습 · 나만의 웹페이지

목표는 볼 사람과 전할 내용, **자료**는 실제 소개와 작업물, **완료 기준**은 브라우저에서 글이 읽히고 링크가 동작하는 상태입니다.

> "처음 만나는 사람을 위한 소개 페이지를 만들어줘. 이 메모에 있는 내용만 쓰고, 작업물 링크가 열리는지 확인해줘."

AI 슬롭은 대량 생성된 저품질 콘텐츠를 비판할 때 쓰는 표현입니다. 내 페이지에서 사실이 틀린 문장, 범용 문구, 작동하지 않는 버튼 중 한 곳을 골라 고칩니다.

08 텍스트 워터마크와 AI 탐지 [2–4]

워터마크는 생성 과정의 토큰 선택에 **통계적 신호**를 남기는 방식입니다. 이를 지원하는 검출기는 해당 신호가 있는지 살핍니다. AI 탐지기는 일반적으로 문장의 특징을 보고 AI 작성 가능성을 추정합니다.

Claude와 OpenAI가 공개한 적용 범위는 모델·서비스·지역·설정에 따라 다릅니다. 짧은 글이나 편집된 글에서는 신호를 판별하기 어려울 수 있습니다. 탐지 결과만으로 작성자를 단정하기 어렵습니다. 사실 여부는 원문, 작성 과정은 초안·수정 이력으로 확인합니다.

수업 중 필기

만들고 싶은 페이지 한 장을 떠올려요.

누가 읽고 무엇을 알면 좋을까요?

브라우저에서 무엇을 확인할까요?

참고 [2] Anthropic · Claude 워터마크 [3] OpenAI · 텍스트 출처 [4] OpenAI · 탐지기 한계