2025년 AI 동향과 이슈로 살펴보는

AI 시대에 꼭 알아야 할 핵심 용어

NIA 한국지능정보사회진흥원

비매품/무료
00000
9 788984 839199
ISBN 978-89-8483-919-9

2025년 AI 동향과 이슈로 살펴보는 AI 시대에 꼭 알아야 할 핵심 용어

NIA 한국지능정보사회진흥원

2025년 AI 동향과 이슈로 살펴보는

AI 시대에 꼭 알아야 할 핵심 용어

NIA 한국지능정보사회진흥원

2025년 AI 동향과 이슈로 살펴보는

AI 시대에 꼭 알아야 할 핵심 용어

NIA 한국지능정보사회진흥원

Contents I

1장 AI 필수용어 100선

AI 기술과 정책 흐름을 이해하는데 필수적인 100개 용어를 선별하여,
빠르게 변화하는 AI 시대에서도 새로운 뉴스를 스스로 해석할 수 있는 기본 토대를 제공합니다.

2장 2025년 AI 이슈를 용어와 함께 쉽게 이해하기

2025년 AI 산업·정책 환경에서 부상한 주요 이슈를 한눈에 살펴보고,
이슈의 정책적·기술적 의미를 심도 있게 이해하기 위한 핵심 용어 및 개념을 함께 제공합니다.

2025년 AI 동향과 이슈로 살펴보는

AI 시대에 꼭 알아야 할 핵심 용어

AI 필수용어 100선

1

AI 필수용어 100선은 AI 관련 이슈를 이해하기 위해
기본적으로 알아야 하는 용어를 선별했습니다.
용어의 개념 뿐만 아니라, 용어의 배경 및 중요성,
오늘날 AI 논의에서 어떤 의미를 갖는지 함께 풀어내어 독자가
스스로 이슈를 이해할 수 있도록 지식 기반을 제공합니다.

001 감성 컴퓨팅
Affective Computing

사용자의 감정·표정을 인식하고 반응하도록 설계된 AI 기술
- 얼굴·음성·생체 신호 등에서 감정을 추정해 상황에 맞는 반응을 생성하는 기술
- 감정 맥락을 반영해 상호작용 품질을 높이는 인간 중심 AI 분야

감성 컴퓨팅이란?
감성 컴퓨팅은 인간의 감정 상태를 추정하고, 이에 기반한 반응을 생성하도록 설계된 AI 기술을 의미합니다. 감정이 의사소통 전반에 영향을 미친다는 점에서 출발해, AI가 감정 정보를 이해하면 상호작용의 자연스러움과 만족도가 높아진다는 관점으로 발전해왔습니다. 초기에는 얼굴 표정과 목소리 분석이 주를 이뤘지만, 최근에는 생체 신호·행동 패턴·텍스트 감정 분석까지 포함하는 멀티모달 방식으로 확장되고 있습니다.

감성 컴퓨팅의 구성 요소
감성 컴퓨팅은 감정 인식, 감정 해석, 감정 기반 반응의 세 요소로 구성됩니다. 감정 인식 단계에서는 표정, 음성 톤, 심박 등 감정 관련 신호를 센서로 수집합니다. 감정 해석 단계에서는 이를 분석해 기쁨·분노·불안 같은 감정 상태를 분류하거나 미세한 정서 변화를 추정합니다. 마지막으로 감정 기반 반응 단계에서는 분석 결과를 바탕으로 대화 톤 조정, 설명 방식 변경 등 상황에 적합한 대응을 제공합니다. 이 세 과정은 감정을 읽고 이에 맞춰 반응하는 감성 컴퓨팅의 핵심 흐름을 형성합니다.

인간–기계 상호작용을 위한 감성 컴퓨팅
인간–기계 상호작용 관점에서 감성 컴퓨팅은 사용자 경험의 품질을 높이는 기술로 이해됩니다. AI는 사용자의 정서 상태를 파악해 부정적 감정에는 차분한 안내를, 긍정적 상태에는 더 능동적인 상호작용을 제공하는 등 응답 방식을 조정합니다. 이를 통해 기술에 대한 신뢰와 만족도가 향상되고, 사용자와 AI가 보다 자연스러운 대화 흐름을 유지할 수 있습니다. 교육·상담·돌봄처럼 정서적 맥락이 서비스 효과에 직접 영향을 미치는 환경에서는 감정 적응형 상호작용이 특히 중요한 역할을 합니다.

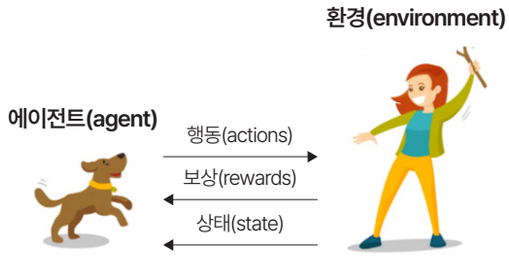
감성 컴퓨팅의 활용
감성 컴퓨팅은 교육, 의료·돌봄, 고객 서비스, 로봇 상호작용 등 다양한 분야에서 활용됩니다. 예를 들어 학습자의 집중도나 혼란 신호를 분석해 맞춤형 피드백을 제공하고, 상담·돌봄 영역에서는 정서 상태 변화를 모니터링해 안정적인 상호작용을 지원합니다. 고객 서비스에서는 감정 분석을 통해 응대 톤을 조정하며, 엔터테인먼트 분야에서는 감정 기반 캐릭터 반응을 구현해 몰입감을 높입니다. 이처럼 감성 컴퓨팅은 AI가 인간의 정서적 맥락까지 이해하고 반응하는 방향으로 발전하고 있음을 보여주는 기술입니다.

002 강화학습
Reinforcement Learning

보상과 시행착오를 통해 스스로 행동을 학습하는 AI 기법
- 환경과 상호작용하며 행동 결과에 따른 보상을 받고,
  이를 바탕으로 목표 달성을 위한 최적의 전략을 찾아가는 자기학습형 인공지능 학습 방식
- 정답 데이터를 주지 않고 경험을 통해 성능을 개선하는 자율 학습 기술

강화학습의 개념
강화학습은 인공지능이 환경 속에서 행동을 선택하고, 그 결과를 바탕으로 더 나은 선택을 학습하는 과정입니다. 시스템은 시행착오를 거치며 목표 달성에 유리한 행동을 찾아가는데, 이는 인간이나 동물이 경험을 통해 학습하는 방식과 유사합니다. 인공지능은 단순히 입력과 출력의 관계를 외우는 것이 아니라 행동과 보상의 관계를 스스로 파악해 전략을 세우며, 정답이 주어지지 않은 상황에서도 경험을 축적해 점차 더 나은 판단을 내릴 수 있습니다. 또한 강화학습은 시간의 흐름을 고려해, 현재의 행동이 미래에 어떤 영향을 미치는지를 평가하고 단기적 보상보다 장기적 이익을 극대화하는 방향으로 학습합니다. 이 과정에서 인공지능은 '지금의 보상'과 '앞으로의 이득' 사이의 균형을 조절하며 점차 효율적인 의사결정 구조를 만들어 갑니다.

강화학습의 구조
강화학습은 에이전트, 환경, 행동, 보상의 네 요소로 이루어진 순환 구조를 기반으로 합니다. 에이전트는 환경의 상태를 관찰하고 행동을 선택합니다. 환경은 그 결과를 보상으로 반환하고, 다시 에이전트는 이를 바탕으로 전략을 조정합니다. 이 과정이 반복되면서 에이전트는 어떤 행동이 유리한지 스스로 학습하고, 점점 더 높은 보상을 얻는 방향으로 전략을 발전시킵니다. 이러한 구조는 정답이 주어지지 않은 상황에서도 학습이 가능하다는 점에서 다른 학습 방식과 차별화됩니다. 단순한 예측이 아니라 환경과의 상호작용을 통해 점차 나은 판단을 내리도록 학습하기 때문에, 로봇 제어나 자율주행처럼 상황이 계속 변하는 환경에서 특히 효과적입니다.



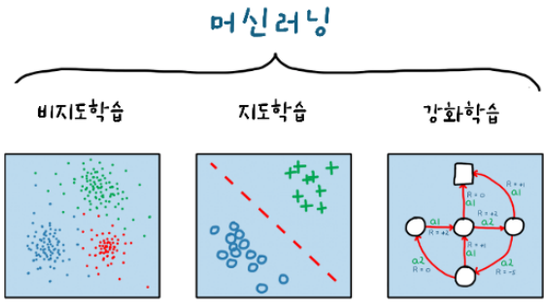
출처 : Mathworks

강화학습의 중요성

강화학습은 인공지능이 스스로 판단하고 환경 변화에 적응하도록 만드는 핵심 기술입니다. 단순히 주어진 데이터를 분석하는 수준을 넘어, 경험을 통해 전략을 조정하고 최적의 행동을 선택하는 능력을 제공합니다. 예를 들어 로봇은 강화학습으로 장애물을 회피하며 이동 경로를 최적화할 수 있습니다. 이러한 자율적 학습 구조는 불완전한 데이터나 예측이 어려운 상황에서도 성능을 지속적으로 개선할 수 있는 장점이 있습니다. 또한 최근 생성형 인공지능에서는 인간의 평가를 보상으로 활용하는 인간 피드백 기반 강화학습(RLHF)이 적용되어 모델의 응답 품질을 높이는 방식으로 활용되고 있습니다.

다른 학습 방식과의 비교

지도학습이 주어진 정답을 기반으로 예측 모델을 만드는 것이라면, 강화학습은 명시적인 정답 없이 행동의 결과를 평가해 최적의 전략을 스스로 찾아갑니다. 비지도학습이 데이터 속 패턴을 탐색하는 데 초점을 둔다면, 강화학습은 행동과 보상 간의 관계를 학습해 보상을 극대화합니다. 즉, 강화학습은 정적인 데이터 분석이 아닌, 변화하는 환경 속에서 전략을 개선하며 자율적 의사결정을 수행하는 지능형 학습 방식입니다. 이 때문에 강화학습은 정적인 데이터 분석이 아닌, 환경 변화에 따라 전략을 지속적으로 개선해야 하는 문제에 적합하며, 지능형 시스템의 핵심 기초로 평가됩니다.



출처 : Mathworks

관련 용어

인간 피드백 기반 강화학습(RLHF) vs AI 피드백 기반 강화학습(RLAIF)

인간 피드백 기반 강화학습(Reinforcement Learning from Human Feedback)과 AI 피드백 기반 강화학습(Reinforcement Learning from AI Feedback)은 강화학습 원리를 이용해 AI 모델의 출력을 개선하는 기술입니다. RLHF는 사람이 모델의 응답을 평가해 보상 신호로 활용함으로써, AI가 인간의 의도와 가치에 맞는 답변을 학습하도록 합니다. 반면 RLAIF는 검증된 AI가 다른 모델의 응답을 평가하는 방식으로, 인적 평가의 비용과 시간을 줄일 수 있습니다. 다만 AI의 오판이 누적될 위험이 있어, RLAIF는 RLHF를 대체하기보다 보완적으로 활용됩니다. 두 방식 모두 생성형 AI의 품질과 신뢰성을 높이는 핵심 학습 기술로 평가됩니다.

참조 용어 | 트랜스포머 아키텍처 p.80

003 거대 언어모델 / LLM

Large Language Model

대규모 언어 데이터를 학습해 인간 언어를 이해하고 생성하는 AI 모델

- 수십억 개 이상의 매개변수를 학습해 문맥·의미를 추론하며 다양한 언어 과업을 수행하는 범용 AI
- 텍스트 기반 사고·대화·추론을 가능하게 하는 생성형 AI의 핵심 기술

LLM이란?

LLM은 방대한 텍스트를 학습해 단어 간 패턴과 의미를 확률적으로 예측하는 언어 중심 AI 모델입니다. 문장 내 모든 단어 간의 관계를 동시에 고려하여 각 단어의 문맥적 중요도를 계산하는 트랜스포머 구조의 Self-Attention 메커니즘을 활용해 문맥을 파악하고, 문장 간 의미 연결을 이해합니다. 이로써 문장 생성, 요약, 번역, 질의응답, 코드 작성 등 다양한 언어 작업을 수행하며, 인간의 사고 과정을 모방한 추론·기획형 언어 지능으로 발전중입니다. 또한 대화 맥락을 기억하고 응답의 톤이나 형식을 조정하며, 텍스트 외에도 이미지·음성 등 멀티모달 입력을 처리하는 범용 인터페이스로 확장되고 있습니다.

LLM의 활용

LLM은 인간의 언어 능력을 매개로 다양한 산업과 생활 영역에 적용됩니다. 업무 영역에서는 문서 요약, 회의록 작성, 이메일 초안 작성 등 지식 노동의 자동화를 지원하고, 기업에서는 고객 응대, 데이터 분석, 규정 질의응답 등에 활용해 의사결정 속도를 높이고 있습니다. 산업별로는 금융의 리스크 분석, 교육의 맞춤형 피드백, 공공 행정의 민원 응대 등으로 확산되고 있습니다. 또한 창작 분야에서는 스토리 구성, 문장 다듬기, 카피라이팅 등 인간의 표현 능력을 확장하는 도구로 활용됩니다. LLM은 단순한 텍스트 생성기를 넘어 지식 접근성과 협업 효율을 높이는 핵심 플랫폼으로 자리 잡고 있습니다.

LLM의 과제

LLM은 방대한 데이터를 학습하며 언어 패턴을 일반화하지만, 그 과정에서 여러 기술적·사회적 한계를 드러냅니다. 학습 데이터의 불완전성과 편향으로 인해 사실과 다른 정보(환각)가 생성되거나 특정 문화·집단에 대한 왜곡이 발생할 수 있습니다. 또한 웹 기반 데이터에는 저작권, 개인정보, 비윤리적 내용이 포함될 수 있어 데이터 거버넌스와 책임 관리가 중요합니다. 이를 해결하기 위해 EU는 AI Act를 통해 LLM에 투명성·데이터 품질 의무를 부과하고, 주요국도 신뢰성과 안전성 확보를 위한 정책을 강화하고 있습니다. 사회적으로는 자동화가 언어 기반 직무를 대체하면서 노동 구조 변화와 교육 체계 재편이 나타나고 있으며 동시에 AI 활용 능력과 윤리 이해를 결합한 AI 리터러시가 새로운 기본 역량으로 요구되고 있습니다. 결국 LLM의 발전은 기술 정교화뿐 아니라 사회적 신뢰·데이터 윤리·책임 있는 활용 원칙 확립에 달려 있다고 할 수 있습니다.

004 거대 행동모델 / LAM

Large Action Model

AI가 언어 이해를 넘어 실제 행동을 수행하도록 설계된 대규모 행동 모델

- LLM에 행동 실행 기능을 결합해, 명령을 실제 행위와 결과로 전환하는 실행 중심 AI 구조

LAM이란?

LAM은 AI가 인간의 언어를 단순히 해석하는 수준을 넘어, 그 내용을 실제 행동으로 수행하도록 설계된 모델을 의미합니다. 기존의 LLM이 텍스트를 분석하고 생성하는 데 초점을 맞췄다면, LAM은 그 이해를 기반으로 외부 시스템을 제어하고 실제 작업을 실행합니다. 예를 들어 일정 등록, 이메일 발송, 데이터 정리, 코드 실행처럼 인간의 언어적 지시를 절차적 행동으로 바꿔 수행할 수 있습니다. 즉, 언어 모델이 지식을 표현하는 단계에서 벗어나 언어를 행동의 매개로 전환해 현실적 결과를 만들어내는 구조입니다. 이를 통해 LAM은 AI를 단순한 대화형 모델이 아닌 실행 가능한 지능형 시스템으로 확장시키며, 언어 중심 AI에서 행동 중심 AI로의 진화를 이끌고 있습니다.

LAM의 특징

LAM의 가장 큰 특징은 언어 이해와 행동 수행의 통합적 처리 능력입니다. 사용자의 의도를 파악한 뒤 목표 달성을 위한 절차를 자동으로 구성하고, 외부 도구나 프로그램을 호출해 결과를 도출합니다. 이러한 실행 능력 덕분에 LAM은 사무자동화, 연구 보조, 로봇 제어, 소프트웨어 운영, 산업 공정 관리 등 다양한 분야에서 응용되고 있습니다. 예를 들어 사무 환경에서는 보고서 작성 등을 자동화하고, 제조 현장에서는 생산 데이터에 따라 설비를 제어할 수 있습니다. 또한 이런 행동 기반 구조는 AI 에이전트가 '결정한 일'을 실질적으로 수행하게 하는 실행 엔진이자, 그 자체로 AI의 자율성과 생산성을 뒷받침하는 핵심 요소로 주목받고 있습니다.

LAM의 의의

LAM은 AI가 언어 이해를 넘어 현실 세계에서 직접 행동하고 결과를 만들어내는 단계로 진입했음을 의미합니다. 이를 통해 반복 업무의 자동화, 서비스 운영의 지능화, 로봇·사물인터넷과의 통합 제어가 가능해지며, 산업 전반의 자율 실행형 AI 생태계를 앞당기고 있습니다. 특히 LAM은 인간의 판단과 행동 사이의 간극을 좁혀, AI가 협력자이자 실행 주체로 기능할 수 있는 기반을 마련합니다. 이러한 변화는 단순한 기술 혁신을 넘어, 노동 구조·생산성·조직 운영 방식 등 사회적 영역에도 큰 영향을 미칠 것으로 예상됩니다. 다만 행동 결과에 대한 책임성과 예측 가능성, 안전성 검증 등 윤리적 과제가 여전히 남아 있으며, 향후에는 행동의 신뢰성과 투명성을 확보하는 기술·제도적 거버넌스가 함께 발전해야 합니다.

참조 용어 | 임베딩 p.64

005 검색증강생성 / RAG

Retrieval-Augmented Generation

외부 지식을 검색해 AI의 생성 결과를 보강하는 기술

- 모델이 질문에 답하기 전 관련 문서를 검색해 정보를 결합함으로써, 학습 시점 이후의 지식이나 최신 정보를 반영할 수 있게 하는 생성기술
- LLM의 한계를 보완해 신뢰도 높은 결과를 제공하는 지식 보강형 AI 기술

RAG 개요

검색증강생성(RAG)은 AI가 응답을 생성하기 전에 외부 데이터베이스에서 관련 정보를 검색해 활용하는 기술입니다. LLM이 고정된 학습 데이터에 의존하는 한계를 극복하기 위해 고안되었으며, 모델은 질문을 분석해 의미적으로 유사한 문서를 찾아내고, 그 내용을 생성 과정에 반영합니다. 이 방식은 모델이 학습 이후의 지식이나 전문 정보를 동적으로 활용하도록 하여, 보다 정확하고 근거에 기반한 응답을 가능하게 합니다. RAG는 지식의 최신성과 신뢰성이 중요한 AI 응용 분야에서 활용되며, 재학습 없이도 데이터 갱신만으로 최신 정보를 반영할 수 있는 효율적 구조를 제공합니다.

RAG의 작동방식

RAG는 검색기(retriever)와 생성기(generator)가 단계적으로 협력하는 구조로 작동합니다. 사용자의 질문은 임베딩 모델을 통해 벡터로 변환되어 검색기는 이 벡터와 문서 벡터의 유사도를 계산해 관련 문서를 찾아내고, 생성기는 이 자료를 입력에 포함시켜 응답을 생성합니다. 이러한 결합은 AI가 질문마다 외부 지식을 불러와 일시적으로 지식 범위를 확장하게 하며, 기존 학습 모델이 가지는 정보 정체 문제를 완화합니다. 검색기의 품질은 임베딩(embedding) 정확도와 검색 알고리즘에 좌우되고, 생성기의 역할은 관련 문맥을 자연스럽게 요약·결합하는 데 있습니다. 이 두 단계의 조화가 RAG의 응답 품질을 결정짓는 핵심입니다.

RAG의 과제

RAG의 성능은 검색 품질과 정보 결합의 정교함에 크게 의존합니다. 검색 결과가 부정확하면 잘못된 정보가 응답에 반영될 수 있으며, 문서 길이 제한이나 벡터 임베딩 편향 등 기술적 제약이 존재합니다. 또한 다양한 데이터 출처를 통합할 때 정보의 신뢰성, 저작권, 보안 문제를 함께 고려해야 합니다. 검색 속도와 프롬프트 처리 효율을 높이기 위한 구조 개선도 필요합니다. 향후에는 하이브리드 검색, 다단계 검색, 문맥 최적화 등으로 정밀도를 높이고, RAG를 자율 검색형 AI 에이전트의 핵심 기술로 발전시키려는 연구가 이어질 전망입니다.

참조 용어 | 거대 언어모델 (LLM) p.11

006 경량화언어모델 / SLM

Small Language Model

적은 연산 자원으로 빠르고 효율적으로 작동하는 소형 AI 언어모델

- LLM의 구조를 단순화하거나 매개변수 수를 줄여 연산 효율을 높이고, 리소스가 제한된 환경에서도 활용할 수 있도록 설계된 경량형 모델
- 비용 절감과 실시간 응답을 가능하게 해 AI의 일상적 활용 범위를 확장하는 핵심 기술

SLM의 특징

SLM은 LLM의 구조를 단순화하고 매개변수 수를 줄여 적은 자원으로 효율적으로 작동하는 모델입니다. LLM 대비 적은 수의 매개변수를 사용해 속도와 에너지 효율을 높이며, 지식 압축·매개변수 공유·정밀도 감소 같은 기술을 통해 성능 저하를 최소화합니다. 모바일·에지 환경에서 실시간 처리와 로컬 데이터 운영이 가능해 응답성과 보안성이 높으며, 이로 인해 특정 도메인에 특화된 고효율 모델이자 버티컬 AI의 핵심으로 평가됩니다.

SLM과 LLM의 비교

SLM은 LLM보다 규모가 작고 목적이 명확한 모델입니다. LLM은 방대한 지식과 추론 능력을 제공하지만 높은 비용과 자원을 요구합니다. 반면 SLM은 속도·비용·자원 효율성을 중시해 개인 단말이나 제한된 환경에서도 구동됩니다. LLM이 범용성과 창의성을 지향한다면, SLM은 경량성·응답성·보안성에 집중합니다. 최근에는 두 모델을 결합한 하이브리드 구조가 등장해, LLM이 지식을 제공하고 SLM이 현장 응용을 담당하는 방식이 확산되고 있습니다.

SLM	vs	LLM
작음. 가벼움 매개변수 ~수십억 개	모델 크기	매우 큼. 무거움 매개변수 ~수조 개
특정 분야 데이터	학습 데이터	방대한 범용 데이터
빠름, 비용 적음, 융통성 낮음	장단점	다양한 작업 가능 상대적으로 느림
모바일·에지 등 리소스 제한된 환경	운영 환경	클라우드·대규모 서버
저비용 / 고효율	비용 구조	고비용 / 고성능
금융, 법률 등 특정 분야 특화	활용	생성형 AI, 코파일럿

SLM의 활용

SLM은 AI의 접근성과 지속가능성을 높이는 기술 전환점으로 평가됩니다. 기업은 저비용으로 서비스를 구축하고, 개인은 네트워크 제약 없이 로컬 환경에서 AI 기능을 활용할 수 있습니다. 정부와 공공기관은 SLM을 활용해 보안 민감 데이터 처리와 지역 맞춤형 서비스를 구현하고 있습니다. SLM은 AI 생태계를 대규모 집중형에서 분산·친환경 구조로 전환하여, AI의 실용화와 보편화를 이끄는 핵심 기술로 자리 잡고 있습니다.

007 고대역폭 메모리 / HBM

High Bandwidth Memory

대용량 데이터 처리를 위해 대역폭과 속도를 극대화한 메모리 반도체

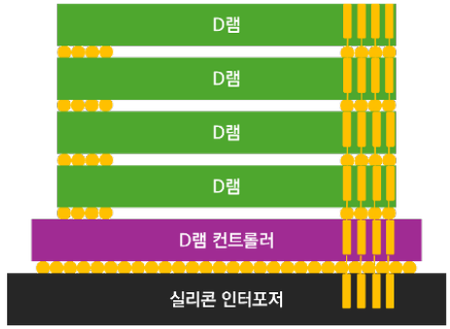
- 여러 개의 D램 칩을 수직으로 적층해 데이터 전송 경로를 넓히고, 기존 메모리 대비 월등히 높은 속도와 에너지 효율을 구현한 고성능 메모리 기술
- AI 반도체, GPU, 데이터센터 등에서 대규모 병렬 연산 성능을 뒷받침하는 핵심 인프라 기술

HBM 개요

고대역폭 메모리(HBM)는 AI 학습·추론 과정에서 방대한 데이터를 빠르게 처리하기 위해 개발된 고성능 메모리 반도체입니다. 기존 D램은 데이터 전송 속도와 병렬 처리 능력에 한계가 있어, 이를 보완하기 위해 HBM은 여러 D램 칩을 수직으로 쌓아 연결하는 구조를 채택했습니다. 이렇게 다층으로 결합된 구조는 데이터 이동 통로를 넓혀 대역폭을 크게 높이고, 전력 효율도 개선합니다. 그 결과 GPU와 NPU 등 고성능 연산 칩의 처리 속도와 에너지 효율을 좌우하는 핵심 부품으로 자리 잡았으며, 자율주행·클라우드·고성능 컴퓨팅(HPC) 등 다양한 분야의 성능 향상을 이끌고 있습니다.

HBM의 구조

HBM은 여러 개의 D램을 수직으로 쌓고 미세한 전극으로 연결해 병렬 통신을 가능하게 하며, 데이터 전송 속도는 기존 GDDR 메모리 대비 수 배 이상 빠릅니다. 또한 메모리와 프로세서를 근접 배치하는 2.5D 또는 3D 패키징 구조를 통해 데이터 이동 거리를 단축하고, 전력 소모를 크게 줄입니다. 이런 설계 덕분에 HBM은 고속·저전력·소형화를 모두 달성해, AI 칩과 고성능 컴퓨팅(HPC)의 핵심 메모리로 자리매김했습니다.



HBM의 활용

HBM은 AI 반도체 성능을 실질적으로 결정하는 기술로, 대규모 연산이 필요한 GPU, 데이터센터 등에 필수적으로 탑재되어 학습과 추론 속도를 높입니다. 메모리 대역폭이 높을수록 AI 처리 효율이 향상되기 때문에, HBM의 성능은 곧 AI 연산 능력의 척도로 평가됩니다. 글로벌 반도체 기업들은 속도·용량 경쟁을 통해 기술 주도권을 확보하려 하고 있으며, HBM은 앞으로 AI 중심 컴퓨팅 구조의 표준 메모리가 될 전망입니다.

008 공간지능

Spatial Intelligence

공간의 형태·위치·움직임을 인식하고 이해하는 AI 기술

- AI가 시각·센서 정보를 통해 주변 환경의 구조와 관계를 파악해 공간적 판단과 행동을 하는 기술
- 물리적 공간과 디지털 세계를 연결하는 인식 기반 인공지능의 중요 영역

공간지능의 개념

공간지능은 AI가 인간처럼 공간의 형태·위치·거리·움직임을 인식하고 이해하는 능력을 말합니다. 카메라·라이다·GPS 등 다양한 센서 데이터를 융합해 현실 공간의 구조를 디지털로 해석하고, 사물 간 관계나 이동 경로를 스스로 판단합니다. 기존의 언어 중심 AI가 텍스트 기반 정보에 초점을 맞췄다면, 공간지능은 물리적 세계를 실시간으로 감지하고 해석하는 인지 능력으로 진화했습니다. 이러한 개념은 컴퓨터 비전, 3D 매핑 기술 등이 결합하면서 발전했으며, 최근에는 공간 데이터를 기반으로 AI가 현실과 디지털을 통합적으로 이해하고 예측하는 인지 구조로 확장되고 있습니다.

공간지능의 활용

공간지능은 자율주행, 로보틱스, 스마트시티, 물류, 산업 자동화 등 다양한 영역에서 활용됩니다. 차량은 도로 상황을 실시간 인식해 경로를 판단하고, 로봇은 작업 환경의 장애물을 탐지해 안전하게 이동합니다. 또한 증강현실과 가상현실에서는 현실 공간의 구조를 반영해 가상 객체를 자연스럽게 배치합니다. 산업 현장에서는 공간지능을 통해 설비 위치나 동선 효율을 분석하고, 도시 관리에서는 교통 흐름과 안전을 예측합니다. 이 기술은 AI가 환경 속에서 스스로 위치를 인식하고 행동을 결정하는 기반을 제공하며, 인간의 시각적 인지 능력을 확장하는 새로운 형태의 지능으로 평가됩니다. 향후 공간지능은 디지털트윈과 결합해 AI가 현실을 학습하고 예측하는 '공간형 지능 생태계'로 발전할 전망입니다.

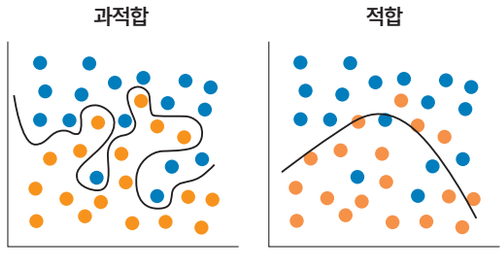
009 과적합

Overfitting

AI가 학습 데이터에만 특화되어 실제 적용 시 성능이 하락하는 현상

- 학습 데이터의 패턴뿐 아니라 잡음과 예외까지 학습해,
  실제 환경에서 오차가 커지는 비효율적 학습 상태이며, AI 성능의 신뢰성과 일반화 능력을 저해

과적합 개요



출처 : Mathworks

과적합은 AI 모델이 학습 데이터에는 높은 정확도를 보이지만, 새로운 데이터에서는 예측력이 급격히 떨어지는 현상을 말합니다. 학습 과정에서 모델이 데이터의 일반적 규칙뿐 아니라 우연적 패턴과 잡음까지 학습해버릴 때 발생합니다. 주로 모델이 지나치게 복잡하거나 학습 데이터가 적고 다양하지 않을 때 나타나며, 검증 없이 장시간 학습할수록 심화됩니다. 즉, AI가 문제의 본질보다 표면적 특징에 집중해, 사람에 비유하자면 '기억은 잘하지만 이해하지 못하는' 상태가 되는 것입니다. 과적합은 AI의 일반화 능력 저하와 신뢰성 손실로 이어지며, 실제 환경에서의 활용 효율을 떨어뜨립니다.

과적합 완화 기법

과적합 방지를 위한 방법으로 데이터를 학습·검증·테스트 세트로 분리해 점검하는 교차 검증, 모델 복잡도를 제한하는 정규화, 일부 뉴런을 무작위로 비활성화하는 드롭아웃(dropout)이 대표적입니다. 또한 데이터의 양과 다양성을 늘리거나 데이터 증강을 통해 학습 범위를 확장하면 모델의 일반화 능력이 향상됩니다. 이러한 과정은 AI 모델의 모델의 예측 안전성과 실제 활용 신뢰성 등 일반화 성능을 확보하기 위한 단계로 평가됩니다.

관련 용어

조기 종료 (Early Stopping)

조기 종료는 AI 모델이 과적합되기 전에 학습을 멈추는 기법입니다. 학습이 진행될수록 학습 데이터 정확도는 높아지지만, 일정 시점 이후에는 우연적 패턴과 잡음까지 학습해 일반화 능력이 떨어집니다. 이를 방지하기 위해 검증 데이터 성능이 더 이상 개선되지 않는 시점에서 학습을 중단합니다. 이렇게 얻은 모델은 과적합을 피하면서도 최적의 성능을 유지할 수 있어, 간단하면서 효과적인 과적합 방지 방법으로 널리 활용됩니다.

참조 용어 | 자연어 처리 p.66

010 광학 문자 인식 / OCR

Optical Character Recognition

이미지나 문서 속 문자를 인식해 디지털 텍스트로 변환하는 기술

- 사진, 손글씨 등 비정형 문서의 문자를 식별해 컴퓨터가 처리 가능한 데이터로 전환하는 인식 기술
- 문서 자동화, 스캔 데이터 분석, 자율주행 등에서 시각 정보를 언어 정보로 바꾸는 핵심 기반 기술

OCR의 배경

광학 문자 인식(OCR)은 이미지나 스캔된 문서 속 문자를 분석해 디지털 텍스트로 바꾸는 기술로, 1950년대 규칙 기반 문자 판독 장치에서 출발했습니다. 초기 OCR은 패턴 데이터베이스와 문자 형태를 비교해 일치 여부를 판단하는 방식이었으며, 인쇄된 활자나 단순 문서에 한정된 인식만 가능했습니다. 이후 통계적 분류와 전처리 기법이 도입되며 다양한 글꼴과 인쇄 상태를 처리할 수 있게 되었지만, 시스템이 스스로 학습하지는 못했습니다. 딥러닝과 시각 인식 기술이 발전하면서 OCR은 AI를 활용한 학습형 인식 시스템으로 전환되었고, 이제는 손글씨·표·사진 속 텍스트 등 비정형 데이터까지 정밀하게 인식하는 단계로 진화했습니다.

OCR의 작동 방식

현대의 OCR은 시각 인식과 언어 이해 기술을 결합해 작동합니다. AI 모델은 이미지에서 문자 영역을 탐지하고, 픽셀 단위로 형태를 분석해 문자를 분리합니다. 이후 딥러닝 알고리즘이 문자의 특징을 학습하고, 자연어 처리(NLP)가 문맥을 분석해 잘못 인식된 부분을 교정합니다. 이 통합 구조를 통해 AI OCR은 문자의 형태와 의미를 동시에 인식하며, 복잡한 배경이나 비표준 글꼴에서도 높은 정확도를 유지합니다.

OCR의 활용과 의의

OCR은 시각적 정보를 언어 데이터로 전환하는 핵심 기술로, 산업 전반의 자동화와 디지털 전환을 촉진합니다. 금융·행정 분야에서는 계약서·청구서·신분증 정보를 자동 인식하고, 물류·제조에서는 송장과 라벨을 읽어 분류 효율을 높입니다. 자율주행 차량은 표지판을 인식하고, 시각장애인 보조기기는 글자를 음성으로 안내합니다. 이러한 응용은 비정형 정보를 구조화해 AI 서비스의 확장을 가능하게 하는 기반으로 평가됩니다.

AI OCR로의 발전

초기의 OCR은 규칙과 패턴에 의존한 비학습형 인식 기술로, 제한된 글꼴과 명확한 인쇄물만 처리할 수 있었습니다. 반면 AI 활용 이후의 OCR은 딥러닝을 통해 문자의 형태뿐 아니라 맥락과 의미를 학습하며, 문서 구조와 내용까지 이해합니다. AI 모델이 데이터를 통해 스스로 인식 정확도를 개선하기 때문에, OCR은 '글자를 읽는 기술'을 넘어 '문서를 이해하는 지능형 인식 기술'로 발전했습니다.

011 그래픽 처리 장치 / GPU

Graphics Processing Unit

대규모 병렬 연산을 수행해 AI 연산 속도를 높이는 고성능 연산 장치

- 수천 개의 코어를 통해 동시에 연산을 수행하며,
  딥러닝·AI 학습·과학 계산 등 대규모 데이터 처리에 필수적인 하드웨어
- 고속 연산 능력과 메모리 대역폭으로 AI 모델 훈련·추론의 효율성을 크게 향상시키는 핵심 인프라

GPU 개요

GPU는 원래 그래픽 렌더링을 위해 개발된 장치였으나, 현재는 대규모 병렬 연산에 최적화된 범용 연산 플랫폼으로 발전했습니다. 수천 개의 코어가 동시에 작동해 대량의 데이터를 병렬 처리할 수 있어, CPU보다 훨씬 빠르게 연산을 수행합니다. 이러한 구조 덕분에 GPU는 이미지 처리뿐 아니라 AI 학습·자연어 처리·과학 계산 등 고속 연산이 필요한 모든 영역에서 핵심 장치로 사용되고 있습니다. 특히 AI 모델 학습에 필요한 매개변수 억 단위 이상의 연산을 효율적으로 수행하여 AI 발전의 연산 기반을 제공합니다.

GPU의 활용

GPU는 현재 AI 반도체 산업과 클라우드 인프라의 핵심 축으로 자리 잡고 있습니다. 데이터센터와 클라우드 기업은 GPU 서버를 통해 거대 AI 모델을 학습·운영하고 있으며, 주요 반도체 기업이 GPU 시장을 주도하고 있습니다. 또한 GPU는 자율주행, 로보틱스, 의료 영상 분석, 기후 시뮬레이션 등에서 활용되며, AI 서비스의 품질과 효율을 좌우하는 핵심 요소가 되고 있습니다. 앞으로 GPU는 단순한 가속기를 넘어, AI 연산의 표준 플랫폼이자 산업 경쟁력의 핵심 인프라로 기능할 것으로 전망됩니다.

관련 용어

중앙처리장치 (Central Processing Unit, CPU)

CPU는 컴퓨터 시스템의 중앙 연산 장치로, 명령어를 해석하고 연산·제어·입출력 관리 등 전반적인 처리를 담당합니다. 복잡한 연산을 순차적으로 수행하도록 설계되어 범용성과 안정성이 높지만, 대량의 데이터를 동시에 처리하는 병렬 연산에는 한계가 있습니다. AI나 그래픽 연산처럼 연산량이 많은 작업에서는 GPU나 NPU에 비해 속도가 느리지만, 여전히 운영체제 제어·프로그램 실행·논리 연산 등 시스템의 핵심 역할을 수행합니다. 최근에는 CPU와 GPU, NPU를 결합한 이기종 연산 구조(Heterogeneous Computing)도 많이 활용되고 있습니다.

관련 용어

신경망처리장치 (Neural Processing Unit, NPU)

NPU는 AI 연산, 특히 인공신경망 구조의 학습과 추론에 특화된 연산 장치입니다. GPU보다 연산 구조가 단순하지만, 행렬 연산과 벡터 계산을 병렬로 처리하는 데 최적화되어 있습니다. 이러한 구조 덕분에 전력 소모가 적고, 모바일·에지 기기에서도 AI 연산을 실시간으로 수행할 수 있습니다. 스마트폰의 이미지 인식, 음성 비서, 자율주행 센서 제어 등 저전력 환경에서 고속 연산이 필요한 분야에서 주로 활용됩니다. 최근에는 클라우드 서버용 NPU도 등장해 AI 가속기 생태계의 핵심 구성 요소로 성장하고 있으며, GPU 대비 에너지 효율 중심의 차세대 AI 연산 아키텍처로 주목받고 있습니다.

관련 용어

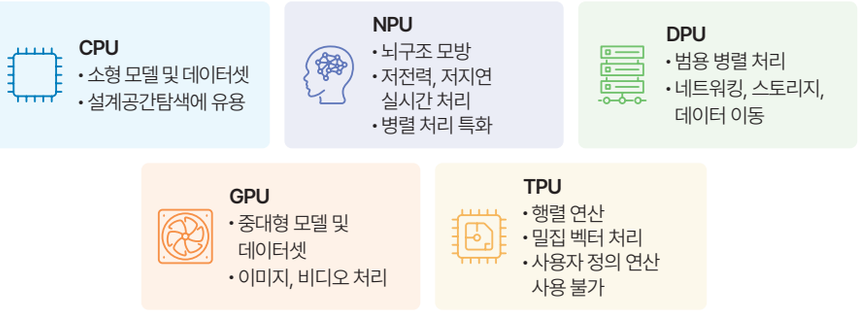
텐서처리장치 (Tensor Processing Unit, TPU)

TPU는 구글이 개발한 AI 전용 연산 프로세서로, 딥러닝 모델 학습과 추론에서 많이 사용되는 행렬·텐서 연산을 빠르게 처리하도록 설계된 장치입니다. GPU가 다양한 병렬 작업을 수행하는 범용 가속기라면, TPU는 신경망 연산 흐름을 효율적으로 처리하는 데 초점을 둔 전용 구조를 갖고 있습니다. 구글 발표에 따르면 TPU는 대규모 모델 학습이나 특정 워크로드에서 높은 처리 속도와 에너지 효율을 보였다고 하지만, 이러한 차이는 모델 종류나 환경에 따라 달라질 수 있습니다. TPU는 주로 텐서플로우(TensorFlow) 기반의 대규모 학습 환경에서 활용되며, 클라우드 데이터센터에서 AI 성능을 높이기 위해 활용되고 있습니다.

관련 용어

데이터 처리장치 (Data Processing Unit, DPU)

DPU는 대규모 데이터 이동·저장·네트워크 처리와 같은 데이터 관리 업무를 전담하도록 설계된 프로세서입니다. AI 연산을 담당하는 GPU나 NPU와 달리, DPU는 데이터 패킷 처리, 암호화·압축, 스토리지 관리, 네트워크 가상화 등 시스템 운영에 필요한 주변 작업을 하드웨어 수준에서 가속합니다. 특히 AI 모델을 대규모로 운영하는 데이터센터에서는 연산 처리보다 데이터 이동과 I/O 병목이 성능을 좌우하는 경우가 많아, DPU가 이를 분리해 처리함으로써 전체 시스템 효율을 크게 높일 수 있습니다. 최근에는 CPU·GPU와 함께 데이터센터 3대 가속기로 불리며, 클라우드 인프라와 AI 서비스의 확장성을 뒷받침하는 핵심 장치로 주목받고 있습니다.



참조 용어 | 설명 가능한 AI p.54

012 기호주의 AI

Symbolic AI

규칙과 기호를 통해 지능을 구현하는 AI 방식

- 사람의 지식을 기호·규칙 형태로 표현해 추론하는 전통적 AI 접근
- 논리적 절차와 지식 구조를 기반으로 문제 해결을 수행하는 방식

기호주의 AI의 개념

기호주의 AI는 인간의 지능을 명시적 지식과 논리 규칙의 조합으로 구현한다는 관점에서 출발한 AI 접근 방식입니다. 여기서 '기호(symbol)'는 개념·사물·관계를 표현하는 표식을 의미하며, AI는 이 기호들을 규칙에 따라 조작해 추론을 수행합니다. 이는 인간이 세계를 이해할 때 언어와 개념을 사용해 구조적으로 사고한다는 점에서 영감을 받은 방식입니다. 기호주의 AI는 1950~1980년대 초기 AI 연구를 주도했으며, 전문가 시스템, 규칙 기반 의사결정, 퍼지 논리 등 당시 산업·공공 분야에 널리 적용되었습니다. 복잡한 문제를 단계별 규칙으로 설명할 수 있고, 시스템이 어떤 이유로 특정 결론을 내렸는지 해석이 가능하다는 점에서 "설명 가능한 AI"의 뿌리가 되는 접근으로 평가됩니다.

기호주의 AI의 유형

기호주의 AI는 사람이 알고 있는 내용을 직접 정리해 지도로 만들어 주는 방식으로 작동합니다. 문제와 관련된 개념을 구조화해 저장하고, 상황에 맞는 규칙을 적용해 결론을 내리는 구조입니다. 논리적으로 깔끔하고 예측 가능한 판단을 내릴 수 있지만, 이 방식에는 분명한 한계도 있습니다. 가장 큰 어려움은 현실 세계의 복잡함과 예외 상황을 모두 규칙으로 작성하기 어렵다는 점입니다. 규칙이 많아질수록 관리해야 할 정보가 폭발적으로 늘어나고, 새로운 상황이 등장하면 규칙을 다시 만들어 넣어야 합니다. 또한 이미지·음성처럼 형태가 일정하지 않은 감각 정보나 매우 모호한 상황을 처리하는 데는 취약합니다. 이러한 한계로 인해 1990년대 이후에는 데이터에서 패턴을 스스로 배우는 연결주의(신경망 기반) 접근이 주류로 자리 잡게 되었습니다.

기호주의 AI의 의의

기호주의 AI는 오래된 접근 방식이지만, 판단 근거가 명확해야 하는 분야에서 활용 가치가 다시 주목받고 있습니다. 법률·의료·행정과 같이 추론 과정의 투명성과 논리적 일관성이 중요한 영역에서는 명시적 규칙과 구조화된 지식 표현이 유용하게 쓰일 수 있습니다. 또한 최근에는 학습 기반 모델과 기호적 추론을 결합하려는 신경-기호(Neuro-Symbolic) AI 연구가 활발해지면서, 기호주의 방식이 새로운 역할을 찾고 있습니다. 이 접근은 신경망의 패턴 인식 능력에 기호적 규칙을 더해 설명 가능성과 안정성을 보완하려는 시도로, 생성형 AI의 논리적 오류나 환각 문제를 줄이는 데 도움이 될 수 있다는 전망이 제시되고 있습니다. 이러한 흐름은 기호주의 AI가 단독 방식에서 벗어나, 현대 AI를 보완하는 조합적 기술로 점진적으로 재조명되고 있음을 보여줍니다.

013 뉴로모픽 컴퓨팅

Neuromorphic Computing

인간의 뇌 신경망 구조를 모방해 정보를 처리하는 차세대 컴퓨팅 기술

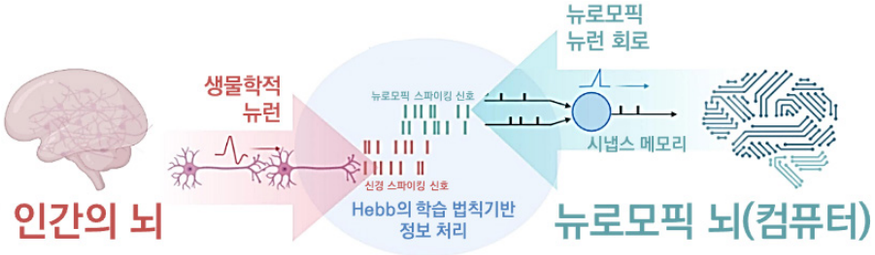
- 뉴런과 시냅스의 작동 원리를 반영해 기억과 연산을 동시에 수행하는 생체 모방형 구조
- 고속·저전력 정보처리를 통해 AI의 효율성과 자율성을 높이는 신경형 계산 방식

뉴로모픽 컴퓨팅이란?

뉴로모픽 컴퓨팅은 인간의 뇌가 정보를 인식하고 학습하는 방식을 하드웨어와 알고리즘으로 모방한 신경 모방형 컴퓨팅 구조입니다. 전통적인 폰 노이만 구조가 메모리와 프로세서를 분리해 순차적으로 연산하는 반면, 뉴로모픽 시스템은 뉴런과 시냅스가 동시에 정보를 저장·처리하는 비선형 병렬 구조를 갖습니다. 이로 인해 연산 과정의 데이터 이동이 줄어들고, 학습과 추론을 하드웨어 수준에서 직접 수행할 수 있습니다. 특히 시냅스 강도의 변화(전류 흐름의 패턴)를 통해 학습이 이뤄지므로, 뇌의 전기적 신호 전달 방식과 유사한 정보처리 메커니즘이 구현됩니다. 이 덕분에 뉴로모픽 칩은 AI 모델의 복잡한 계산을 GPU보다 훨씬 낮은 전력으로 처리할 수 있으며, 인간의 사고방식에 가까운 병렬적·자율적 지능 구조를 목표로 발전하고 있습니다.

뉴로모픽 컴퓨팅의 특징

뉴로모픽 시스템은 크게 뉴런 회로, 시냅스 메모리, 스파이킹 신호 세 요소로 구성됩니다. 뉴런 회로는 입력 자극을 수용하고, 시냅스는 자극의 강도에 따라 연결 강도를 조절하며, 스파이킹 신호는 전기적 펄스로 정보를 전달합니다. 이러한 구조는 생물학적 뇌의 뉴런 발화 과정을 전자적으로 모사한 것으로, 이벤트 기반 계산을 가능하게 합니다. 즉, 필요할 때만 신호를 주고받아 불필요한 연산을 줄이므로 에너지 효율이 매우 높습니다. 또한 아날로그적 신호 처리가 가능해, 불확실하거나 연속적인 데이터를 처리하는 데 유리합니다. 이런 특성은 AI의 학습·추론 속도를 높이는 동시에 저전력·고효율 연산이라는 새로운 패러다임을 제공합니다.



출처 : Frontiers

뉴로모픽 컴퓨팅의 활용

뉴로모픽 컴퓨팅은 에지 AI, 로보틱스, 자율주행, 웨어러블 기기, 뇌-기계 인터페이스(BMI) 등 즉각적인 반응성과 저전력 특성이 중요한 분야에서 주로 연구 ·활용되고 있습니다. 예를 들어 자율주행 차량의 센서 데이터 처리나 로봇의 환경 인식에서 이벤트 기반 연산을 통해 빠른 반응을 지원할 수 있습니다. 또한 시각 ·청각 등 다양한 감각 정보를 통합 처리하기 용이해, 향후 보다 자연스러운 인지형 AI를 구현하는 데 기여할 가능성이 있습니다. 뉴로모픽 칩은 GPU, NPU와는 다른 방식으로 연산 및 메모리를 통합한 AI 반도체로서, 일부 연구에서는 하드웨어 수준의 지속적 학습과 저전력 추론을 동시에 추구하고 있습니다. 아직 상용화 초기 단계이지만, 소자 기술과 회로 기술이 발전함에 따라 대규모 병렬 신경망 구현이 가능해질 전망이며, 향후 초저전력 AI 기반 기술로서 잠재력이 높게 평가되고 있습니다.

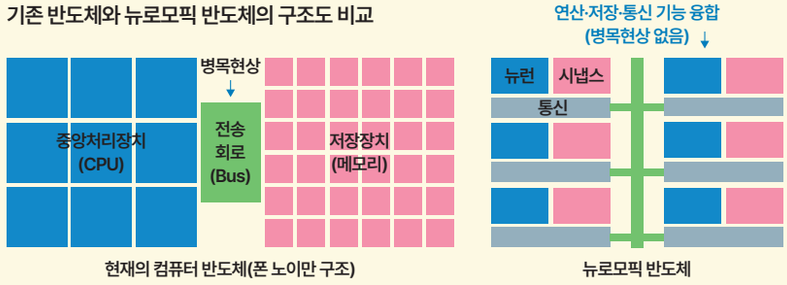
관련 용어

폰 노이만 구조

1940년대 수학자 폰 노이만이 제안한 개념으로, CPU가 메모리에 접근해 명령을 불러오고, 산술·논리 연산을 수행한 뒤 결과를 다시 메모리에 저장하는 일련의 과정으로 작동합니다. 이 구조는 범용성과 안정성 측면에서 큰 장점을 지니며, 오늘날 대부분의 컴퓨터와 서버 시스템의 기반으로 동작합니다.

하지만 폰 노이만 구조는 메모리와 프로세서가 분리되어 있어 데이터 이동에 많은 시간과 에너지가 소모되는 병목 현상이 발생하며, 이는 대규모 데이터를 동시에 처리해야 하는 AI·딥러닝 환경에서 성능 저하를 초래합니다. 뉴로모픽 컴퓨팅은 바로 이 병목 문제를 해결하기 위해 등장한 새로운 패러다임으로, 메모리와 연산을 통합하고 뉴런·시냅스의 병렬적 작동 방식을 모방해 효율을 극대화하는 구조를 지향합니다.

기존 반도체와 뉴로모픽 반도체의 구조도 비교



관련 용어

뉴로모픽 칩 (Neuromorphic Chip)

뉴로모픽 컴퓨팅의 핵심 하드웨어로, 인간의 뇌 신경망을 모방해 기억과 연산을 동시에 수행하는 반도체입니다. 수많은 인공 뉴런과 시냅스가 병렬적으로 연결되어 데이터를 비선형적으로 처리하며, 학습 과정이 회로 수준에서 직접 이루어집니다. 이러한 구조 덕분에 GPU나 CPU보다 훨씬 적은 전력으로 고속 연산이 가능하고, 실시간 추론이나 환경 적응에도 강점을 보입니다. 뉴로모픽 컴퓨팅이 지향하는 뇌 유사형 정보처리 구조를 물리적으로 구현한 장치로, 자율주행·로보틱스·에지 AI 등 지능형 시스템의 하드웨어 기반을 제공합니다.

014 대화형 AI

Conversational AI

인간의 언어를 이해하고 자연스럽게 소통하는 언어 기반 AI 시스템

- 텍스트나 음성 입력을 해석해 사용자의 의도와 맥락을 파악하고, 자연스러운 응답을 생성해 사람처럼 대화하는 AI 기술

대화형 AI란?

대화형 AI는 인간의 언어를 인식하고 의미를 이해해 자연스러운 대화를 수행하도록 설계된 언어 중심 인공지능 기술입니다. 사용자의 음성이나 텍스트를 분석해 의도를 파악하고, 상황에 맞는 응답을 생성함으로써 인간과 기계 간 상호작용을 언어로 구현합니다. 초기에는 정해진 시나리오에 따라 응답하는 단순 규칙형 챗봇 수준이었지만, 자연어 처리(NLP)와 딥러닝, 특히 LLM의 발전으로 문맥 이해, 감정 인식, 대화 유지 능력이 크게 향상되었습니다. 현재의 대화형 AI는 단순 정보 전달을 넘어 의사결정과 조언, 창의적 협업까지 수행하는 지능형 언어 인터페이스로 발전하고 있습니다.

대화형 AI의 유형

대화형 AI는 텍스트 기반 챗봇, 음성 명령을 수행하는 음성 어시스턴트, 이미지나 영상을 함께 인식하는 멀티모달 대화형 AI 등으로 구현됩니다. 챗봇은 고객 상담과 정보 안내에, 음성 어시스턴트는 일정 관리나 명령 수행에 활용되며, 멀티모달 형태는 시각·청각 정보를 통합해 더 인간적인 상호작용을 가능하게 합니다.

대화형 AI의 쓰임

대화형 AI는 인간과 기술의 관계를 '명령'에서 '대화'로 전환한 핵심 인터페이스 기술입니다. 사용자는 복잡한 명령어 없이 자연어로 작업을 수행할 수 있고, 기업은 고객 응대와 업무 지원을 자동화해 효율을 높입니다. 또한 음성 기반 서비스는 고령자와 장애인의 디지털 접근성 향상에 기여합니다. 나아가 대화형 AI는 AI 어시스턴트나 에이전트형 AI의 기반으로 발전해, 인간과 협업하는 새로운 작업 방식과 서비스 모델을 만들어가고 있습니다. 다만 개인정보 침해, 허위정보 생성, 과도한 의존 등 부작용이 존재하므로, 투명성과 신뢰성을 확보하는 윤리적 활용 체계가 함께 마련되어야 합니다.

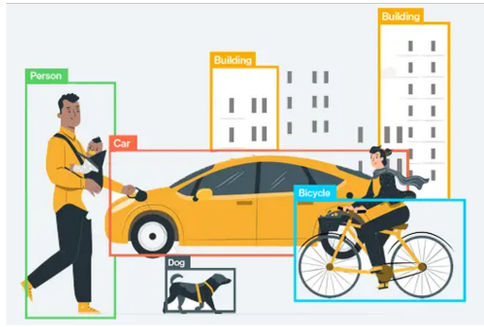
015 데이터 라벨링

Data Labeling

데이터에 의미나 정답을 부여해 AI 학습이 가능하도록 만드는 과정

- AI가 이미지·음성·텍스트 등 데이터를 올바르게 인식하고 판단할 수 있도록 정보를 부여하는 절차
- AI 모델의 정확도와 신뢰성을 좌우하는 학습 데이터 품질 관리의 핵심 단계

데이터 라벨링의 개념



출처 : LXT

데이터 라벨링은 AI 학습에 활용되는 데이터에 사람이 의미나 정답 정보를 부여하는 과정입니다. 예를 들어 이미지에는 사물의 이름을, 음성에는 문장을, 텍스트에는 감정이나 의도 정보를 표시해 AI가 이를 학습하도록 돕습니다. 이 과정을 통해 AI는 단순한 데이터의 모양이나 패턴을 넘어서, 그 속에 담긴 의미를 인식하고 분류·예측 등의 작업을 수행할 수 있습니다. 라벨링은 AI가 세상을 이해하기 위한 언어를 가르치는 일과 같으며, 모델의 인식력과 판단력을 결정하는 출발점이 됩니다.

데이터 라벨링의 과정

데이터 라벨링은 수집된 데이터를 분석하고, 기준에 따라 범주를 정의한 뒤, 각 데이터에 올바른 정보를 부여하는 순서로 진행됩니다. 이 작업은 사람이 직접 수행하기도 하지만, 최근에는 AI가 초기 분류를 제안하고 사람이 이를 검수하는 오토라벨링(auto labeling) 방식이 널리 쓰입니다. 특히 대규모 데이터셋에서는 품질 관리가 중요해, 여러 명이 같은 데이터를 반복 검토하는 다중 검증 절차나 표준화된 가이드라인이 함께 적용됩니다. 정확한 라벨링이 이루어져야 모델이 오류 없이 학습할 수 있으며, 잘못된 라벨은 학습 방향을 왜곡시켜 성능 저하를 초래할 수 있습니다.

데이터 라벨링의 역할

데이터 라벨링은 AI가 데이터의 숫자를 실세계의 개념과 연결하기 위한 과정으로, 단순한 데이터 정제 단계를 넘어 의미 구조를 형성하는 역할을 합니다. 전처리가 데이터의 품질과 형식을 다듬는 과정이라면, 라벨링은 그 데이터가 '무엇을 뜻하는가'를 정의하는 지식 부여 단계입니다. 올바르게 라벨링된 데이터는 AI가 맥락과 관계를 학습해 인간의 판단과 유사한 인식 능력을 갖추도록 만듭니다. 또한 체계적인 라벨링은 편향된 데이터 해석을 방지하고, AI가 사회적 맥락 속에서 공정하게 작동하도록 돕는 토대가 됩니다.

016 데이터 랭글링

Data Wrangling

데이터를 분석 목적에 맞게 정리하고 구조화하는 과정

- 정제된 데이터를 다시 가공해 분석이나 AI 학습에 활용하기 적합한 형태로 만드는 절차로, 데이터 전처리 이후 실제 활용 단계와 분석 효율을 높이는 핵심과정

데이터 랭글링의 개념

데이터 랭글링은 전처리를 거친 데이터를 분석이나 AI 학습에 직접 활용할 수 있도록 재구성하는 과정입니다. 전처리가 데이터의 오류를 수정하고 품질을 높이는 단계라면, 랭글링은 목적에 맞게 데이터를 선별·결합·변환해 활용 가능한 형태로 만드는 단계입니다. 예를 들어 여러 데이터셋을 합치거나 특정 항목을 추출해 새로운 구조를 구성하는 것이 이에 해당합니다. 이 과정을 통해 방대한 데이터를 분석 목적에 맞게 간결화하고, 모델이 이해할 수 있는 입력 형태로 준비할 수 있습니다.

데이터 랭글링의 과정

데이터 랭글링은 탐색 → 구조화 → 정제 → 확충 → 검증 → 배포의 순환 과정을 거칩니다. 먼저 탐색 단계에서 데이터의 형태와 품질을 살펴보고, 구조화 단계에서는 서로 다른 형식의 데이터를 공통 구조로 맞춥니다. 정제 단계에서는 잘못된 값이나 불필요한 정보를 제거하며, 확충 단계에서는 부족한 정보를 보완하거나 필요한 속성을 추가합니다. 검증 단계에서는 오류나 불일치를 점검해 신뢰성을 확보하고, 마지막 배포 단계에서는 완성된 데이터를 분석 환경이나 AI 학습 시스템에 전달해 활용합니다. 이를 통해 데이터 랭글링은 품질과 활용성을 동시에 높입니다.

데이터 랭글링의 의의

데이터 랭글링은 AI 개발과 데이터 분석의 실질적 생산성을 높이는 핵심 과정입니다. 이 단계를 통해 방대한 데이터를 빠르게 다루고, 필요한 정보만 추출해 인사이트를 도출할 수 있습니다. 특히 AI 모델의 입력 데이터가 구조화되어 있을수록 학습 속도와 예측 정확도가 향상되기 때문에, 랭글링은 단순한 정리 작업을 넘어 AI 학습의 효율과 해석 가능성을 높이는 기술로 평가됩니다. 또한 자동화된 랭글링 시스템은 반복적 수작업을 줄여 데이터 사이언스 전반의 접근성과 속도를 개선하고, 데이터 활용의 범위를 한층 확장시키고 있습니다.

017 데이터 사일로

Data Silo

조직 내 데이터가 부서나 시스템 단위로 고립되어 공유되지 않는 현상

- 데이터 간 연계가 단절되어 통합 분석이나 협업이 어려워지는 구조적 문제
- AI와 데이터 기반 의사결정을 저해하는 주요 장애 요인

데이터 사일로 개요

'데이터 사일로'는 곡물이나 사료를 외부와 차단된 저장탑에 따로 보관하던 사일로(silo) 구조에서 유래한 말로, 조직 내 데이터가 부서나 시스템 단위로 고립되어 공유되지 않는 상태를 뜻합니다. 이는 데이터가 부서 중심으로 관리되고 시스템 간 호환이 부족하거나, 보안 규정과 조직 문화가 분리되어 있을 때 발생합니다. 예를 들어 마케팅 부서와 영업 부서가 서로 다른 고객 데이터를 사용하면 전체 고객 행동을 통합적으로 분석하기 어렵습니다. 이처럼 데이터가 사일로화되면 정보의 흐름이 단절되고, 조직 전체의 데이터 활용성과 의사결정 효율이 저하됩니다.

데이터 사일로의 원인과 문제점

데이터 사일로는 기술적·조직적 요인이 함께 작용해 형성됩니다. 기술적으로는 서로 다른 포맷, 데이터베이스 구조, 전송 프로토콜 등이 통합을 어렵게 만들고, 조직적으로는 부서 간 이해관계나 관리 권한이 협업을 제한합니다. 이러한 분리 구조는 데이터 중복, 불일치, 품질 저하를 초래하며, 전체 시스템의 효율성을 떨어뜨립니다. 특히 AI 개발에서는 학습 데이터의 다양성과 규모가 제한되어 모델 성능이 저하되거나 편향이 심화될 수 있습니다.

데이터 사일로에 대한 대응 방법

데이터 사일로를 해소하기 위해서는 기술적 통합과 조직적 협업이 병행되어야 합니다. 기술적으로는 클라우드 기반 데이터 레이크나 통합 플랫폼을 구축해 데이터를 중앙에서 관리하고, 표준화된 메타데이터 체계를 도입해 이질적인 데이터를 연결할 수 있습니다. 조직적으로는 부서 간 데이터 공유 정책과 협업 프로세스를 마련해, 데이터의 소유보다 활용 중심의 문화를 조성해야 합니다. 이러한 개선을 통해 조직은 데이터 자산을 통합적으로 분석·활용할 수 있고, AI 학습의 품질과 정확성도 높일 수 있습니다. 즉, 데이터 사일로의 해소는 AI 시대의 데이터 개방성과 협력 생태계 구축을 위한 필수 조건입니다.

018 데이터 전처리

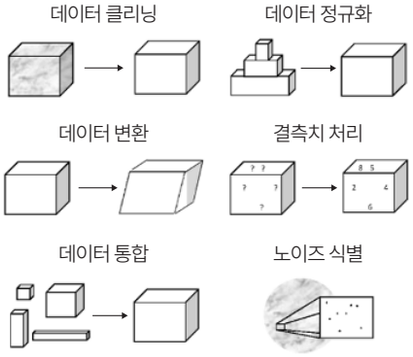
Data Preprocessing

AI 학습에 사용될 데이터를 정리하고 변환해 품질을 높이는 과정

- 다양한 출처에서 수집된 데이터를 분석 가능한 형태로 정제·가공해, AI 모델이 안정적으로 학습할 수 있도록 만드는 절차
- 데이터를 바르게 이해·예측할 수 있도록 하는 단계로, AI 개발의 출발점이자 성능을 좌우하는 단계

데이터 전처리의 개념

데이터 전처리는 AI가 학습하기 전에 원시 데이터를 분석 가능한 형태로 다듬는 과정입니다. 실제 수집된 데이터에는 누락값, 오류, 중복, 잡음 등 다양한 결함이 포함될 수 있으며, 이러한 데이터는 그대로 학습에 사용하면 왜곡된 결과를 초래할 수 있습니다. 전처리는 이런 문제를 사전에 수정하고 구조를 표준화해, 모델이 의미 있는 패턴을 정확히 학습할 수 있도록 돕습니다. 즉, 데이터 전처리는 AI 학습 전반의 품질과 신뢰성을 확보하기 위한 준비 과정이자, 모델의 성능을 결정하는 핵심 절차입니다.



출처 : Big Data Analytics

데이터 전처리의 개념

전처리는 일반적으로 정제, 통합, 변환, 축소의 네 단계로 이루어집니다. 정제 단계에서는 누락된 데이터를 보완하거나 잘못된 값을 교정하고, 이상치나 불필요한 정보를 제거해 데이터의 오류를 바로잡습니다. 통합 단계에서는 여러 출처의 데이터를 결합해 형식과 단위를 일관되게 정리하며, 변환 단계에서는 값의 범위를 조정하고 비정형 데이터를 수치 형태로 변환해 모델이 처리할 수 있도록 만듭니다. 마지막으로 축소 단계에서는 분석에 꼭 필요한 변수만 남기거나 일부 데이터를 표본으로 추출해 연산 효율을 높입니다. 최근에는 이러한 과정을 자동화한 전처리 도구와 파이프라인이 도입되어, 대규모 데이터셋의 품질을 안정적으로 유지하고 처리 속도와 정확도를 동시에 향상시키고 있습니다.

데이터 전처리의 의의

데이터 전처리는 AI 성능을 좌우하는 핵심 품질 관리 기술입니다. 부정확하거나 편향된 데이터는 모델의 판단 오류, 과적합, 공정성 문제로 이어질 수 있습니다. 따라서 전처리를 통해 데이터의 정확성과 다양성을 확보하는 일은 AI의 신뢰성과 책임성을 높이는 데 필수적입니다. 또한 전처리는 데이터 랭글링, 라벨링, 피처 엔지니어링 등 후속 과정의 효율을 향상시켜 AI 개발 전반의 자동화와 품질 향상을 가능하게 합니다.

019 데이터 플라이휠

Data Flywheel

데이터 축적과 활용이 반복되며 AI 성능과 서비스를 개선하는 선순환 구조

- AI 서비스 이용 과정에서 생성된 데이터가 다시 학습에 활용되어 모델의 품질을 높이고, 개선된 결과가 사용자 경험을 강화해 새로운 데이터를 지속적으로 만들어내는 순환 메커니즘

데이터 플라이휠의 개념

데이터 플라이휠은 데이터의 생성–활용–개선이 반복되며 스스로 성장하는 선순환 구조를 의미합니다. AI 서비스가 사용자와 상호작용하면서 축적된 데이터를 다시 모델 학습에 활용하면, AI의 예측력과 응답 품질이 향상됩니다. 개선된 모델은 더 나은 사용자 경험을 제공하고, 이는 곧 서비스 이용 증가와 데이터 추가 생산으로 이어집니다. 이 과정이 회전축(flywheel)처럼 점점 가속되며, 데이터–모델–서비스의 성능이 서로 강화되는 자기증폭 구조가 만들어집니다. 이런 순환 구조가 형성되면, 데이터의 양뿐 아니라 다양성과 품질도 함께 개선되어 AI의 일반화 능력까지 향상됩니다.

데이터 플라이휠의 선순환 구조

데이터 플라이휠은 데이터 수집 → 모델 학습 → 서비스 개선 → 사용자 확대의 선순환으로 작동합니다. 사용자의 이용 데이터가 학습에 반영되면 AI 성능이 향상되고, 개선된 결과가 더 나은 경험을 제공해 이용자가 늘어납니다. 이로써 새 데이터가 다시 생성되며 데이터·모델·서비스가 서로를 강화하는 가속 순환 구조가 형성됩니다.

데이터 플라이휠의 의의와 과제

데이터 플라이휠은 AI 경쟁력의 핵심 동력으로, 데이터가 많을수록 AI가 강해지는 '규모의 학습 효과'를 실현합니다. 구글, 아마존, 메타 등 글로벌 기업이 방대한 이용자 데이터를 활용해 서비스 품질을 지속적으로 개선하는 방식이 대표적 예시입니다. 기업은 이 구조를 통해 연구개발 효율을 높이고, 사용자 맞춤형 서비스를 강화하며, 시장 점유율을 확대할 수 있습니다. 그러나 데이터 플라이휠이 작동하려면 데이터의 다양성과 품질, 개인정보 보호, 데이터 편향 완화가 함께 보장되어야 합니다. 일부 대형 플랫폼에 데이터가 집중되면 독점 구조가 고착화되고, 윤리적·사회적 불균형이 심화될 위험도 있습니다.

020 딥러닝
Deep Learning

인공신경망을 여러 층으로 쌓아 복잡한 패턴을 학습하는 머신러닝 기법
- 인간의 뇌 신경망을 모방한 다층 신경망을 통해 이미지·음성·언어 등 비정형 데이터를 자동으로 학습·추론하는 기술
- 데이터에서 표현·특징을 자동 학습하는 신경망 기반 방법론

딥러닝이란?
딥러닝은 인간의 신경세포 연결 구조를 모방한 인공신경망(ANN)을 깊게(Deep) 쌓아 올려 복잡한 데이터를 학습하는 AI 기술입니다. 기존의 기계학습이 사람이 설계한 특징을 중심으로 학습했다면, 딥러닝은 데이터 속 유의미한 특징을 자동으로 추출하는 구조를 가지고 있습니다. 입력층–은닉층–출력층으로 구성된 신경망이 여러 층으로 깊어질수록 추상적 패턴을 포착할 수 있어, 이미지 인식·음성 인식·자연어 처리 등 다양한 영역에서 인간 수준의 성능을 보이고 있습니다. GPU, TPU 등 병렬 연산 하드웨어의 발전과 대규모 데이터의 확보가 결합되면서, 2010년대 이후 딥러닝은 AI 혁신의 주축 기술로 자리 잡았습니다.

딥러닝의 작동 원리
딥러닝은 입력된 데이터를 여러 층의 인공신경망을 거치며 점점 더 복잡한 특징을 스스로 찾아내는 방식으로 학습합니다. 각 연결에는 가중치(Weight)라는 값이 있어 입력 신호의 중요도를 조절하고, 뉴런은 활성화 함수(Activation Function)를 통해 전달받은 정보 중 어떤 신호를 다음 단계로 보낼지를 결정합니다. 이렇게 선택적으로 정보를 전달함으로써 단순한 계산이 아닌 복잡한 패턴 인식이 가능해집니다. 모델이 예측한 결과와 실제 값의 차이는 오차 역전파 과정에서 계산되어 가중치가 반복적으로 조정됩니다. 이 과정을 수천 번 거듭하면서 모델은 데이터 속 규칙을 점점 더 정교하게 파악하고, 학습하지 않은 새로운 데이터에도 일관된 판단을 내리는 일반화 능력을 갖게 됩니다. 이처럼 딥러닝은 단순한 계산의 반복이 아니라, 경험을 통해 스스로 판단 기준을 세워가는 AI의 자율 학습 구조입니다.

딥러닝의 의의
딥러닝은 오늘날 AI의 핵심 동력으로, 이미지 인식, 자율주행, 번역, 의료 진단, 생성형 AI 등 거의 모든 분야에 적용되고 있으며, AI의 '지각·이해·창조' 능력을 실현한 핵심 기술로 평가됩니다. 딥러닝의 강점은 복잡한 비정형 데이터를 처리하고, 사람이 인식하지 못하는 패턴까지 포착할 수 있다는 점입니다. 그러나 방대한 데이터 의존, 해석의 어려움, 윤리적 편향 위험은 여전히 해결 과제로 남아 있습니다. 그럼에도 딥러닝은 AI의 자율학습 능력과 인간 수준의 지능을 향한 진화 과정에서 가장 중요한 기술적 기반으로 평가됩니다.

참조 용어 | 생성적 적대 신경망(GAN) p.50

021 딥페이크
Deepfake

사람의 얼굴·음성·행동 등을 조작하여 사실적으로 보이게 하는 생성형 AI 기술
- 딥러닝으로 실존 인물의 영상과 음성을 학습해 새로운 이미지를 만들어내는 인공 합성 기술
- 창작과 표현을 확장하지만, 조작과 악용 위험을 동시에 지닌 양면적 기술

딥페이크란?
딥페이크는 '딥러닝(Deep Learning)'과 '페이크(Fake)'의 합성어로, AI가 사람의 얼굴·목소리·행동 등을 학습해 실제처럼 재현하거나 교체하는 기술입니다. 이를 가능하게 했던 것은 생성적 적대 신경망(GAN) 구조였습니다. GAN은 한 모델은 이미지를 생성하고 다른 모델은 그것이 진짜처럼 보이는지를 판별하는 구조로, 이 경쟁이 반복될수록 결과물의 사실성이 높아집니다. 최근에는 확산 모델(Diffusion Model), 트랜스포머 기반 기술 등 다양한 생성기술이 활용되고 있습니다. 딥페이크는 처음엔 영화나 게임의 시각효과 기술로 개발되었으나, 현재는 음성 합성, 이미지 복원, 가상 인물 생성 등 멀티모달 콘텐츠 제작 기술로 발전했습니다. 딥페이크는 원래 인간의 표현과 창작 가능성을 확장하기 위한 기술이었지만, 현실과 허구의 경계를 흐리는 문제로 사회적 논의가 커지고 있습니다.

딥페이크의 활용
딥페이크는 사람의 표정, 시선, 말투, 음성 억양까지 정밀하게 재현할 수 있는 특징으로 인해, 다양한 산업과 문화 분야에서 활용되고 있습니다. 영화나 광고에서는 배우의 나이 변화나 분신 연기를 자연스럽게 표현하고, 게임·메타버스에서는 이용자의 얼굴과 표정을 아바타에 실시간 반영해 몰입감을 높입니다. 또한 외국어 영상의 입 모양과 음성을 동기화하거나, 오래된 영상을 복원하고, 장애인을 위한 맞춤 음성을 제작하는 등 사회적 가치가 큰 응용 사례도 존재합니다. 이처럼 딥페이크는 단순한 영상 편집 기술이 아니라, 인간의 표현 영역을 확장하고 콘텐츠 제작 효율을 극대화하는 창조적 도구로 주목받고 있습니다.

딥페이크의 사회적 쟁점
딥페이크는 허위정보 생산, 사생활 침해, 인격 왜곡, 신뢰 훼손 등 심각한 사회적 문제를 동반합니다. 특히 실제 인물의 얼굴이나 음성이 무단으로 사용될 경우, 개인의 정체성과 명예가 훼손되고 디지털 공간에서의 존재 자체가 조작될 수 있습니다. 정치·언론 영역에서는 조작 영상이 여론 조성이나 사회적 갈등을 유발하며, 금융·보안 분야에서는 음성 위조를 이용한 사기나 범죄에 악용될 위험이 큽니다. 또한 일반 사용자가 생성형 AI를 통해 손쉽게 합성 영상을 만들 수 있게 되면서, 진짜와 가짜를 구분하기 어려운 사회적 불신 증가가 우려됩니다. 이런 이유로 기술 발전의 자유와 개인의 권리 보호, 표현의 자율성과 진위 검증 책임 사이의 균형이 새로운 윤리적 과제로 떠오르고 있습니다.

022 로우 코드

Low Code

복잡한 코딩을 최소화하여 애플리케이션을 개발하는 간소화된 개발 방식

- 그래픽 인터페이스와 자동화된 코드 생성을 통해 개발 효율을 높이고,비전문가도 손쉽게 소프트웨어를 제작할 수 있도록 지원하는 접근법

로우 코드의 배경

로우 코드는 '적은 코드로 개발한다'는 의미로, 프로그래밍 언어를 직접 작성하지 않고 시각적 도구를 이용해 애플리케이션을 만드는 방식입니다. 사용자는 화면 요소를 끌어다 놓으며 기능을 조합해 앱을 완성할 수 있습니다. 이러한 접근은 2010년대 이후 디지털 전환 가속과 개발 인력 부족 속에서 부상했습니다. 빠른 개발이 필요한 기업 환경에서 전문 개발자뿐 아니라 업무 담당자나 비전공자도 손쉽게 참여할 수 있게 한 것입니다. 로우 코드는 복잡한 코딩 과정을 단순화하며 '개발의 민주화'를 이끌었고, 최근에는 AI가 코드 자동 완성이나 테스트 자동화를 지원하면서 생산성과 효율성을 높이고 있습니다.

로우 코드의 영향

로우 코드 플랫폼은 사용자의 시각적 설계를 내부적으로 코드로 변환하고, 데이터 연동·보안·오류 검증·배포 과정을 자동으로 처리합니다. 이를 통해 개발자는 복잡한 문법 대신 기능 설계와 사용자 경험에 집중할 수 있습니다. 이러한 구조는 개발 속도 향상, 유지보수 단순화, 협업 효율성 강화로 이어집니다. 기업은 이를 활용해 업무 자동화와 프로토타입 제작을 빠르게 수행하며, 디지털 혁신 속도를 높이고 있습니다. 그러나 플랫폼 종속성, 맞춤 기능의 한계는 여전히 과제로 남아 있습니다. 그럼에도 로우 코드는 AI·클라우드와 결합하며, 개발을 전문가의 영역에서 누구나 참여할 수 있는 협업형 창작 과정으로 변화시키고 있습니다.

관련 용어

노 코드 (No Code)

로우 코드와 같은 '개발 자동화' 흐름 속에서 등장한 개념으로, 로우 코드가 일부 코드를 직접 수정·추가할 수 있는 개발자를 위한 도구라면, 노 코드는 프로그래밍 언어를 전혀 사용하지 않고 완전히 시각적인 인터페이스만으로 앱을 제작할 수 있도록 설계되어 비전문가가 독립적으로 개발할 수 있게 한 단계입니다. 사용자는 화면 구성 요소를 끌어다 놓고, 조건과 기능을 설정하는 것만으로 웹사이트나 내부 업무용 앱을 완성할 수 있습니다. 이러한 구조 덕분에 노 코드는 아이디어만 있으면 누구나 빠르게 서비스를 구현할 수 있는 환경을 제공하지만, 동시에 복잡한 기능이나 맞춤형 시스템을 구현하기는 어렵다는 한계도 지닙니다.

023 매개변수

Parameter

모델이 학습 과정에서 스스로 조정하며 예측과 판단을 가능하게 하는 내부 변수

- AI가 입력 데이터를 통해 학습하는 과정에서 계속 조정되는 수치적 값으로, 어떤 정보를 중요하게 보고 어떻게 판단할지를 결정함으로써 모델의 성능과 작동 방식을 좌우하는 핵심 요소

매개변수란?

매개변수는 수학적으로 '함수의 출력을 결정하는 변수'를 뜻하지만, AI에서는 모델이 데이터를 학습하면서 스스로 조정하는 내부 값을 의미합니다. 즉, 매개변수는 사람이 직접 지정하지 않아도 학습을 통해 자동으로 최적화되는 값이며, 모델이 입력과 출력 간의 관계를 파악하도록 돕습니다. 대표적인 매개변수로 가중치와 편향 등이 있습니다. 이런 값들이 조정되며 모델은 점점 더 정확한 결과를 예측할 수 있게 됩니다.

AI 모델에서 매개변수의 역할

AI 모델은 학습 데이터를 입력받아 예측 결과와 실제 정답의 차이를 계산하고 이 오차를 줄이기 위해 매개변수를 반복적으로 조정하며, 이 과정을 훈련이라 합니다. 매개변수는 결국 모델의 '기억'이자 '판단 기준'으로, 데이터가 많을수록 더 세밀하게 조정됩니다. LLM의 경우 수십억 개 이상의 매개변수가 존재하며, 이는 모델이 단어의 의미, 문맥, 논리 구조 등을 학습하는 근거가 됩니다. 매개변수가 많을수록 표현력은 커지지만, 동시에 계산량이 늘어나고 학습 효율이 떨어질 수 있습니다.

매개변수의 의의

매개변수는 AI의 성능을 결정짓는 가장 핵심적인 요소로, 모델의 지능 수준을 수치로 나타내는 지표로 활용됩니다. 예를 들어 언어모델의 규모를 비교할 때 'OO억 개의 매개변수'로 표현하는 이유가 여기에 있습니다. 그러나 매개변수의 수가 많다고 해서 반드시 더 나은 결과를 보장하지는 않습니다. 과도한 매개변수는 과적합 문제를 유발하거나, 계산 자원과 에너지를 과다하게 소모할 수 있습니다. 최근에는 이러한 한계를 극복하기 위해 효율적 학습(LoRA, 양자화, 경량화 모델 등) 연구가 활발히 이루어지고 있습니다.

관련 용어

하이퍼파라미터 (Hyperparameter)

모델이 학습을 시작하기 전에 사람이 직접 설정하는 값으로, 학습 과정의 방식과 범위를 결정하는 조절 변수입니다. 매개변수가 데이터로부터 자동으로 학습되는 내부 값이라면, 하이퍼파라미터는 학습률, 배치 크기, 모델의 층 수처럼 학습이 어떻게 진행될지를 정하는 외부 설정입니다. 즉, 매개변수가 모델의 '기억과 지식'을 형성한다면, 하이퍼파라미터는 그 지식을 어떤 속도와 규칙으로 배울지를 결정하는 역할을 합니다.

참조 용어 | 딥러닝 p.30

024 머신러닝
Machine Learning

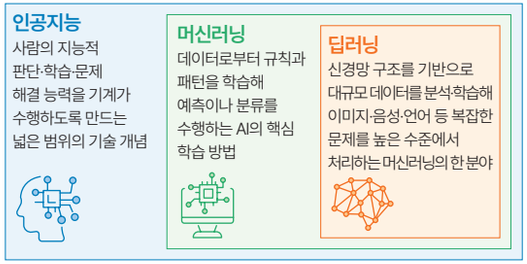
컴퓨터가 프로그래밍 없이 데이터를 학습해 스스로 규칙을 찾아내는 기술

- 데이터를 분석해 스스로 예측 모델을 만들고, 오류 개선을 통해 성능을 높이는 AI 학습 방법 중 하나
- 프로그램이 데이터를 바탕으로 패턴을 학습하고, 예측 성능을 향상시키도록 매개변수를 자동 조정하는 데이터 기반 학습 기술

머신러닝이란?

머신러닝은 AI가 데이터를 바탕으로 패턴을 학습해 예측이나 분류를 수행하도록 만드는 기술적 접근 방식입니다. 기존의 소프트웨어는 사람이 모든 규칙을 명시적으로 코딩해야 했지만, 머신러닝은 방대한 데이터를 입력받아 그 안의 패턴과 관계를 스스로 찾아냅니다. 즉, '데이터가 곧 알고리즘을 가르친다'는 개념으로, 사람이 정의하지 않은 규칙을 경험을 통해 발견하고, 이를 바탕으로 새로운 데이터를 예측하거나 분류할 수 있습니다. 이러한 특징 덕분에 머신러닝은 음성 인식, 이미지 분석, 추천 시스템, 언어 번역 등 현대 AI의 대부분을 가능하게 한 기초 기술로 자리 잡았습니다.

인공지능, 머신러닝, 딥러닝



AI, 머신러닝, 딥러닝은 포함 관계로 연결된 계층 구조로 이해할 수 있습니다. AI는 인간의 판단과 문제 해결을 기계가 수행하도록 만드는 가장 넓은 개념으로, 규칙 기반 시스템부터 통계적 모델까지 다양한 접근을 포함합니다. 머신러닝은 이 중 데이터를 통해 스스로 규칙을 학습하는 방법론을 뜻하며, AI를 구현하는 핵심 기술입니다. 딥러닝은 머신러닝의 하위 분야로, 다층 인공신경망을 이용해 특징 추출과 학습을 자동화하는 방식입니다. 기존 ML이 사람이 특징을 설계해야 했다면, DL은 이미지·음성·텍스트 같은 복잡한 데이터에서 중요한 패턴을 스스로 찾아냅니다. 이 차이 덕분에 딥러닝은 대규모 데이터 환경에서 뛰어난 성능을 보이며, 최근의 생성형 AI와 LLM의 기반 기술로 자리 잡고 있습니다.

관련 용어

오차 역전파 (Backpropagation)

인공신경망이 학습 데이터를 통해 스스로 개선되는 핵심 알고리즘으로, 예측 결과의 오차를 거꾸로 전달하며 가중치(매개변수)를 조정하는 과정을 말합니다. 모델이 낸 결과와 실제 정답의 차이를 계산한 뒤, 이 오차를 출력층에서 입력층 방향으로 되돌려보내 각 연결의 영향력을 계산하고, 그에 따라 매개변수를 조금씩 수정합니다. 이 과정을 여러 번 반복하면 모델은 점점 더 정확한 예측을 하도록 학습됩니다.

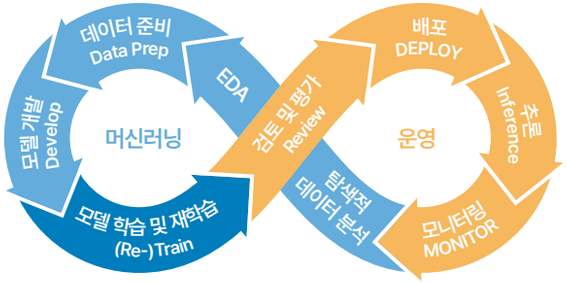
025 머신러닝 운영 / MLOps
Machine Learning Operations

AI 모델의 개발·배포·운영을 자동화하고 관리하는 통합 관리 체계

- ML 모델의 학습·평가·배포 과정을 효율적으로 연결, 안정적 AI 서비스 운영을 가능하게 하는 체계
- 데이터 과학과 소프트웨어 엔지니어링을 통합해 모델의 품질·재현성·확장성을 확보

MLOps란?

MLOps는 머신러닝과 DevOps(개발·운영)의 합성어로, 데이터 수집부터 모델 개발, 실험 관리, 배포, 모니터링, 재학습까지 AI 모델의 전생애주기를 자동화하고 표준화하는 운영·엔지니어링 체계입니다. DevOps가 소프트웨어 배포와 운영 자동화를 다루는 반면, MLOps는 여기에 데이터·모델의 버전관리, 품질관리, 모델 드리프트 감지, 재학습 자동화를 포함하여 AI 서비스를 안정적으로 운영하기 위한 확장된 개념입니다. 과거 머신러닝 개발은 데이터 준비, 모델 학습, 평가가 연구자 중심으로 이루어졌지만, 실무에서는 모델을 주기적으로 업데이트하고 성능을 유지하는 과정이 복잡했습니다. MLOps는 이를 해결하기 위해 머신러닝 운영 전 과정을 자동화하고 표준화합니다. 즉, 모델 개발의 실험 중심 접근을 지속 가능한 서비스 운영 체계로 전환하는 역할을 합니다.



머신러닝 운영(MLOps)의 과정

출처 : Databricks

MLOps 관리

MLOps는 크게 데이터 파이프라인 관리, 모델 라이프사이클 관리, 지속적 모니터링 체계의 세 축으로 구성됩니다. 데이터 파이프라인은 원천 데이터를 자동으로 수집·정제·검증하여 모델 학습에 적합한 형태로 제공합니다. 모델 라이프사이클 관리 단계에서는 실험 관리, 하이퍼파라미터 최적화, 버전 관리 등을 수행하며, 모델이 서비스 환경으로 배포되면 성능 저하나 편향 발생 여부를 지속적으로 점검합니다. 이러한 체계를 통해 MLOps는 AI 모델의 품질과 신뢰성을 보장하면서 운영 비용과 시간을 절감합니다.

026 멀티모달

Multimodal

다른 형태의 데이터를 연계해 의미를 통합적으로 이해하는 기술

- 텍스트·이미지·음성 등 다양한 입력 정보를 동시에 처리해, 단일 데이터만으로는 파악하기 어려운 의미를 종합적으로 이해하고 연관짓는 기술
- 서로 다른 정보가 연결되며, 상황과 맥락을 인간처럼 종합적으로 이해하도록 돕는 핵심 기반

멀티모달의 개념

멀티모달은 '여러 형태(modality)의 데이터가 결합된다'는 뜻으로, 한 가지 형태의 정보만 처리하던 기존 방식에서 벗어나 텍스트, 이미지, 음성, 영상 등 다양한 형태의 데이터를 함께 이해하고 연관짓는 기술을 의미합니다. 사람은 시각, 청각, 언어 등 여러 감각을 동시에 사용해 상황을 인식합니다. 멀티모달 AI의 경우에도 여러 데이터의 상호 관계를 학습해 보다 풍부한 맥락을 이해할 수 있도록 설계됩니다. 예를 들어, "웃는 사람"이라는 문장을 보고 실제 웃고 있는 얼굴 이미지를 연관지을 수 있고, 음성·영상 데이터를 함께 분석해 감정이나 의도를 추론할 수도 있습니다.

멀티모달의 활용

멀티모달 기술은 생성형 AI, 자율주행, 의료 진단, 감정 인식, 로봇 비전 등 다양한 분야에서 활용됩니다. 예를 들어, 텍스트 설명으로 이미지를 생성하는 모델, 이미지와 텍스트를 동시에 이해하는 검색 시스템, 영상과 음성을 결합한 감정 분석 모델 등이 있습니다. 이러한 AI는 복합적인 맥락을 고려해 더 정확한 판단과 표현을 수행하며, AI가 인간의 감각과 사고 방식을 닮아가는 과정으로 평가됩니다. 다만 데이터 결합 과정에서 편향이 증폭되거나, 특정 모달이 과도하게 영향을 미치는 문제가 발생할 수 있습니다. 따라서 멀티모달 AI의 발전은 단순한 성능 향상을 넘어, 정보 간 균형과 의미의 정합성을 확보하는 방향으로 이어지고 있습니다.

관련 용어

멀티모달 거대 언어모델 / MLLM (Multimodal Large Language Model)

기존의 텍스트 기반 LLM을 확장해 이미지·음성·영상 등 다양한 형태의 데이터를 함께 이해하고 생성하는 AI 모델입니다. 언어 모델의 언어적 추론 능력에 시각·청각 정보 처리를 결합해, 복합적 맥락을 통합적으로 해석하고 다중 입력 간 의미를 정렬합니다. 예를 들어 이미지 속 장면을 설명하거나, 사용자의 음성 질문에 시각 정보를 결합해 답변하는 등 언어와 감각 정보를 동시에 활용하는 통합형 AI를 구현합니다. GPT-4o, Gemini 1. 5, Claude 3 등은 대표적 MLLM으로, 인간의 감각 인식 구조를 모방해 AI의 이해력과 표현력을 한 단계 확장한 차세대 모델로 평가됩니다.

027 메모리 연산 / PIM

Processing In Memory

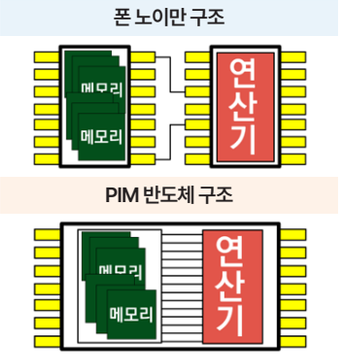
메모리 내부에서 연산을 직접 수행해 처리 효율을 높이는 컴퓨팅 기술

- 데이터 이동 없이 메모리 자체에서 계산을 수행하는 컴퓨팅 구조
- 대규모 작업에서 지연을 줄이고 새로운 시스템 설계를 가능하게 하는 방식

PIM의 개념

PIM은 CPU와 메모리 간 반복적인 데이터 이동으로 발생하는 비효율을 줄이기 위해 등장한 기술로, 연산 기능을 메모리 내부 또는 매우 가까운 위치에 배치하는 구조를 의미합니다. 기존 컴퓨팅 방식에서는 모든 연산이 CPU에서 이루어져 대규모 데이터 처리 시 병목이 쉽게 발생했습니다. PIM은 이러한 구조적 한계를 개선하기 위한 새로운 접근으로, 특히 데이터 접근 요구가 빈번한 작업 환경에서 주목받고 있습니다. AI 모델의 연산량 증가와 메모리 중심 처리 요구가 높아지면서, 메모리 자체에 연산 기능을 통합하는 방식이 차세대 컴퓨팅 구조의 변화 방향으로 논의되고 있습니다.

PIM의 작동 방식



PIM은 메모리 모듈 내부에 간단한 연산 유닛을 포함하거나, 메모리와 매우 가까운 영역에 연산 장치를 배치해 동작합니다. 이를 통해 데이터가 먼 연산 장치로 이동하지 않고, 메모리 내부에서 필요한 계산을 수행할 수 있습니다. 이는 연산장치와 메모리가 분리된 폰 노이만 구조에서 발생하는 데이터 이동 병목을 줄여줍니다. D램 기반 PIM 방식은 메모리 셀의 구조를 활용해 간단한 연산을 병렬적으로 처리할 수 있도록 하며, 일부 구조에서는 행렬 곱셈과 같은 특정 연산을 메모리에서 직접 수행하도록 설계됩니다. 이러한 방식은 복잡한 알고리즘을 모두 메모리에서 처리한다기보다, 대량의 반복 연산이나 특정 패턴의 계산을 메모리 근처에서 빠르게 처리하도록 최적화된 구조에 가깝습니다.

PIM의 활용

PIM은 메모리 접근이 많은 작업에서 구조적 이점을 제공하여, 그래프 탐색, 추천 시스템, 벡터 검색처럼 데이터 위치 정보를 반복적으로 조회해야 하는 작업에서 응답 지연을 줄이고 처리 흐름을 단순화할 수 있습니다. 또한 PIM은 기존의 CPU·GPU 중심 구조를 대체하기보다는, 메모리 중심 처리가 필요한 특정 영역을 보완하는 기술로서 가치가 있습니다. 이를 통해 제약이 컸던 메모리 기반 알고리즘을 실용적으로 구현할 수 있는 기반이 마련되며, 향후 컴퓨팅 아키텍처 설계에서 새로운 선택지를 제공한다는 점에서 의미가 있습니다.

028 메타 데이터
Metadata

데이터의 의미·맥락 이해를 돕는 데이터의 구조·내용·속성 등의 정보
- 데이터 그 자체가 아니라, 데이터에 관한 정보를 담은 '데이터의 데이터'
- 저장·검색·활용을 효율화해 AI 학습과 데이터 관리의 품질을 높이는 핵심 요소.

메타 데이터의 개념
메타 데이터는 데이터에 대한 설명 정보를 제공하는 데이터를 뜻합니다. 즉, '데이터의 데이터(data about data)'로, 어떤 데이터가 무엇을 담고 있고, 누가 만들었으며, 언제 생성되었는지 등을 알려줍니다. 예를 들어 디지털 사진의 해상도·촬영 시각·카메라 기종, 문서의 작성자·수정일, 데이터셋의 출처와 형식 같은 정보가 모두 메타 데이터에 해당합니다. 메타 데이터는 본래 데이터의 특성과 맥락을 이해할 수 있게 하며, 데이터의 발견성과 재사용성을 높이는 기반이 됩니다.

메타 데이터의 유형
메타 데이터는 전통적으로 기술적, 관리적, 내용적 메타 데이터의 세 범주로 구분되었습니다. 그러나 방대한 데이터를 학습하는 AI로 데이터의 출처, 품질, 신뢰성, 보안 등을 관리해야 할 필요성이 커졌고, 통계적, 계보적, 보안·프라이버시 등으로 그 범주가 확장되었습니다.
- 기술적 메타 데이터: 파일 형식, 구조, 인코딩 방식 등 데이터의 형태
- 관리적 메타 데이터: 생성 일자, 접근 권한, 보존 정책 등 유지·관리 정보
- 내용적 메타 데이터: 제목, 주제, 키워드처럼 데이터의 의미를 표현해 검색과 분류를 보조
- 통계적 메타 데이터: 수집 단위·지표 정의 등 분석 기준을 명시해 학습 데이터의 품질을 보장
- 계보적 메타 데이터: 출처와 변환 과정을 기록해 데이터의 신뢰성과 추적성을 확보
- 보안·프라이버시 메타 데이터: 개인정보 보호 수준과 접근 통제를 정의해 안전한 활용을 보장

AI 시대 메타 데이터의 중요성
AI가 방대한 데이터를 학습 자원으로 활용하는 시대에는, 데이터의 품질과 맥락을 정확히 설명하는 메타 데이터의 역할이 그 어느 때보다 중요합니다. 메타 데이터는 단순한 보조 정보가 아니라, AI가 데이터를 신뢰할 수 있도록 만드는 기준 정보로 작동합니다. 학습 과정에서 데이터의 출처, 구조, 활용 조건이 명확히 기록되어야 모델의 결과를 해석하고 오류를 추적할 수 있기 때문입니다. 또한 메타 데이터는 데이터 간 연계와 표준화를 촉진해 AI의 투명성, 재현성, 신뢰성을 높이는 기반이 됩니다. 궁극적으로 메타 데이터는 AI가 의미를 '이해'하고, 그 판단 과정이 검증 가능하도록 하는 데이터 거버넌스의 핵심 인프라로 자리 잡고 있습니다.

029 메타 러닝
Meta Learning

학습 경험을 바탕으로 더 빠르고 효율적인 학습법을 배우는 AI 기술
- 이전 학습의 결과를 바탕으로 새로운 과제에 더 빠르고 효율적으로 적응할 수 있도록 학습 전략 자체를 개선하는 '학습의 학습'
- AI의 일반화 능력과 적응력을 높이는 핵심 개념

메타 러닝의 개념
메타 러닝은 '학습을 위한 학습(learning to learn)'을 의미하며, AI가 한 번의 학습 경험을 넘어서 학습하는 방법 자체를 배우는 과정을 뜻합니다. 기존의 머신러닝이 특정 과제를 해결하기 위해 정해진 데이터를 학습했다면, 메타 러닝은 여러 과제에서 얻은 경험을 바탕으로 새로운 문제를 더 적은 데이터로 빠르게 해결하는 능력을 기르는 데 초점을 둡니다. 이는 마치 사람이 이전 경험을 통해 새로운 문제에 접근하는 법을 터득하는 과정과 유사합니다. 메타 러닝은 데이터 효율성과 적응력을 함께 높이며, 특히 새로운 환경에서도 모델이 빠르게 일반화할 수 있도록 돕는 기술로 주목받고 있습니다.

메타 러닝의 작동 방식
메타 러닝의 핵심은 AI가 스스로 학습 규칙을 배우는 데 있습니다. 일반적인 학습이 데이터 속 패턴을 찾는 과정이라면, 메타 러닝은 여러 학습 과제를 경험하며 그 안에서 공통된 원리를 발견하고 이를 일반화하는 과정입니다. 대표적인 접근 방식에는 세 가지가 있습니다. 먼저, 모델 기반 접근은 내부 구조를 유연하게 설계해 새로운 데이터가 주어졌을 때 스스로 매개변수를 조정할 수 있도록 합니다. 최적화 기반 접근은 학습 속도와 방향을 결정하는 알고리즘 자체를 학습하여, 짧은 반복으로도 성능을 개선할 수 있게 합니다. 마지막으로 메트릭 기반 접근은 새로운 과제를 기존의 경험과 유사성으로 비교해 빠르게 판단하도록 설계됩니다. 이처럼 메타 러닝은 다양한 경험에서 학습 원리를 추출해 학습의 효율화와 일반화를 동시에 실현하는 구조로 작동합니다.

메타 러닝의 의의
메타 러닝은 AI가 스스로 학습 전략을 발전시키는 기술로, 소량 학습(few-shot learning)이나 지속 학습(continual learning)과 같은 분야의 토대가 됩니다. 이는 대규모 데이터 없이도 높은 성능을 달성할 수 있게 하며, 환경 변화에 빠르게 적응하는 모델을 가능하게 합니다. 예를 들어 자율주행 AI는 새로운 도로 조건에 빠르게 대응할 수 있고, 의료 AI는 환자별 데이터 특성에 맞춘 진단 모델을 효율적으로 학습할 수 있습니다. 이러한 특성 덕분에 메타 러닝은 AI를 단순히 주어진 규칙을 따르는 시스템이 아니라, 스스로 학습 방법을 개선하며 성장하는 지능적 존재로 진화시키는 개념으로 평가됩니다.

참조 용어 | 지식 증류 p.74

030 모델 압축

Model Compression

AI 모델의 크기와 연산량을 줄여 효율을 높이는 기법

- 불필요한 매개변수를 제거하거나 구조를 단순화해 모델을 가볍게 만드는 기술
- 모바일·에지 환경에서 빠르고 경제적으로 AI를 실행하기 위해 사용

모델 압축의 개념

모델 압축은 AI 모델의 성능을 크게 훼손하지 않으면서 크기·메모리·연산량을 줄이는 기술을 의미합니다. 한국에서는 경량화라는 표현도 사용하며, 학계에서는 최적화라는 표현을 사용하기도 합니다. 최신 LLM과 비전 모델은 학습과 추론에 막대한 자원과 비용이 필요하기 때문에, 이를 실제 서비스나 모바일 기기에서 실행하기 위해서는 모델을 더 작고 가볍게 만드는 과정이 필수적입니다. 모델 압축은 단순히 매개변수 수를 줄이는 것이 아니라, 모델이 의사결정을 위해 필요로 하지 않는 부분을 찾아 제거하고, 구조를 재배열하거나 표현 방식을 바꾸는 방식으로 효율성을 높입니다. 이를 통해 대규모 모델의 성능을 유지하면서도 더 적은 비용으로 추론을 수행할 수 있는 실용적 형태로 전환할 수 있습니다.

모델 압축의 주요 방식

모델 압축의 방식에는 가지치기(pruning), 양자화(quantization), 지식 증류(distillation) 등이 있습니다.

- 가지치기(Pruning): 모델의 출력에 거의 기여하지 않는 매개변수나 뉴런을 제거해 구조를 간소화하는 방식으로, 불필요한 연결을 제거하면 모델 크기와 연산량이 감소하면서도 핵심 정보는 유지
- 양자화(Quantization): 매개변수를 고정밀 숫자 대신 더 작은 비트 수로 표현하는 방식으로, 예를 들어 16비트를 8비트나 4비트로 줄여 계산량과 메모리 사용량을 크게 절감
- 지식 증류(Distillation): 크고 복잡한 모델이 가진 지식을 작은 모델(학생 모델)에 이전해, 더 작은 구조로 유사한 성능을 내도록 만드는 기술

이러한 방식들은 단독으로도 쓰이지만, 실제로는 서로 결합해 더 높은 효율을 얻는 경우가 많습니다.

모델 압축의 중요성

모델 압축은 AI를 실제 환경에서 활용하기 위한 핵심 기술로 평가됩니다. 대규모 모델은 강력한 성능을 제공하지만, 서버 비용이 높고 응답 속도가 느리며 모바일·에지 기기에서는 실행 자체가 어려운 경우가 많습니다. 모델 압축을 통해 연산 비용을 크게 낮추면 AI 서비스를 더 저렴하게 제공할 수 있고, 사용자 단말에서도 빠르고 지속적인 추론이 가능해집니다. 특히 생성형 AI가 산업 전반에 확산되면서, 경량 모델을 기반으로 한 온디바이스 AI, 개인화 모델, 보안이 중요한 로컬 처리 환경에서 모델 압축의 활용도가 더욱 커지고 있습니다. 또한 에너지 효율 개선과 탄소 배출 감소 측면에서도 중요한 기술로 주목받고 있습니다. 결국 모델 압축은 AI 성능을 유지하면서도 접근성을 높이고 비용 구조를 개선하는 데 필수적인 역할을 수행합니다.

031 모델 컨텍스트 프로토콜 / MCP

Model Context Protocol

AI 모델과 외부 도구·데이터를 표준 방식으로 연결하는 개방형 프로토콜

- AI 모델이 다양한 정보원과 안전하게 통신하도록 돕는 프로토콜
- 복잡한 연동 없이 외부 맥락을 확장해 활용할 수 있게 하는 연결 구조

모델 컨텍스트 프로토콜의 개념

MCP는 AI 모델이 파일, 데이터베이스, 내부 시스템, 외부 API 등 다양한 자원에 접근할 때 발생하는 통합 문제를 해결하기 위해 Anthrophic이 제안한 개방형 통신 규약입니다. 기존 방식은 각 서비스마다 별도의 연결 코드가 필요해 개발 비용과 유지 부담이 컸고, 시스템 간 호환성도 떨어졌습니다. MCP는 이런 문제를 개선하기 위해 모델과 외부 도구를 하나의 공통 규격으로 연결하는 구조를 제공합니다. 이를 통해 AI 모델은 고정된 학습 데이터에만 의존하지 않고, 실행 환경의 최신 정보나 사용자 맥락을 실시간으로 받아 활용할 수 있습니다. AI를 단순 대화 도구가 아닌 실제 작업 수행 도구로 확장하는 기반이라는 점에서 주목받고 있으며, 다양한 개발 환경과 플랫폼에서 표준처럼 쓰이기 시작한 기술입니다.

모델 컨텍스트 프로토콜의 구조

MCP는 호스트, 클라이언트, 서버가 서로 역할을 나누어 작동하는 구조로 설계되어 있습니다. 호스트는 AI 모델이 실행되는 환경으로, 사용자 인터페이스와 모델의 작업 공간을 제공하는 핵심 실행 단위입니다. 클라이언트는 이 호스트 내에서 AI 모델과 외부 도구를 이어주는 중간 매개체로, 모델이 요청한 작업을 표준 형식으로 서버에 전달합니다. 서버는 클라이언트 요청을 실제로 처리하는 구성 요소로, 파일 열기, 데이터 조회, 내부 시스템 호출, 외부 API 연동 등 구체적인 작업을 수행해 결과를 다시 클라이언트와 호스트로 반환합니다. 이 세 요소가 규칙화된 방식으로 연동되면서, AI 모델은 복잡한 맞춤형 개발 없이 다양한 도구를 일관된 방식으로 사용할 수 있게 됩니다.

모델 컨텍스트 프로토콜의 이점과 과제

MCP의 가장 큰 이점은 확장성과 재사용성입니다. 하나의 규격만 구현하면 여러 도구와 서비스에 쉽게 연결할 수 있어 개발 부담이 줄고, 모델이 실제 업무 환경 정보에 접근하면서 대화형을 넘어 작업 수행형 AI로 발전할 수 있습니다. 또한 구조가 단순해 도구 추가·관리 측면에서도 효율적입니다. 그러나 해결해야 할 과제도 존재합니다. 외부 도구와 직접 연결되는 만큼 보안과 권한 관리가 중요하며, 인증이 미흡하면 데이터 유출이나 도구 악용 위험이 발생할 수 있습니다. 구현 수준의 차이로 인해 성능과 호환성에 편차가 생길 수 있다는 점도 초기 생태계의 한계입니다. 그럼에도 MCP는 AI가 실제 시스템과 안정적으로 연결되는 기반 기술로 자리 잡아가고 있으며, 앞으로 AI 활용 범위를 크게 확장할 인프라로 평가됩니다.

참조 용어 | 사전학습모델 p.49

032 미세조정

Fine-Tuning

기존에 학습된 AI 모델을 새로운 데이터나 목적에 맞게 최적화하는 과정

- 사전 학습 모델을 기반으로, 특정 분야나 작업 목적에 맞게 추가 학습을 수행하는 단계적 조정 과정
- 새로운 데이터의 특성과 맥락을 반영해, 적은 자원으로 높은 효율을 달성하도록 돕는 학습 기술

미세조정의 개념

미세조정은 이미 학습된 AI 모델을 새로운 목적이나 환경에 맞게 다시 조정하는 학습 과정으로, 대규모 데이터로 사전 학습된 모델이 특정 분야에서 더 정확하고 안정적으로 작동하도록 추가 학습을 수행하는 기술을 의미합니다. 기존 학습 과정에서 모델이 얻은 일반 지식을 유지한 채, 새로운 데이터의 특성을 반영해 모델의 행동을 재정렬하는 절차라 할 수 있습니다. 최근 LLM의 발전 이후 미세조정은 단순한 성능 개선을 넘어, 응답 스타일·추론 방식·안전성 기준을 목적에 적합한 방향으로 정교하게 다듬는 핵심 단계로 활용됩니다.

미세조정의 과정

기존에 사용되었던 풀 미세조정은 모델의 모든 매개변수를 새로운 데이터에 맞춰 다시 학습시키는 방식입니다. 가장 강력한 조정 효과를 얻을 수 있지만 연산 비용이 매우 크고, 기존 능력이 손상되는 지식 파괴 위험도 존재한다는 한계 때문에 최신 모델 규모에서는 실제 산업 활용 사례가 제한적입니다. 현재 산업과 연구에서 주로 사용되는 방식은 필요한 부분에만 변화를 주어 모델의 행동을 업데이트하는 방식입니다. 대표적인 기법은 LoRA, Prefix-Tuning 등 PEFT(Parameter-Efficient Fine-Tuning) 계열로, 모델의 대부분을 고정한 채 극히 일부의 매개변수만 조정해 연산 비용을 줄이고, 기존 성능을 크게 훼손하지 않으면서 특정 능력을 강화할 수 있다는 장점이 있습니다. 특히 소량·고품질 데이터만으로도 특정 도메인 지식이나 말투, 응답 형식을 부여하는 데 효과적이나, 새로운 목표에 맞추는 과정에서 특정 작업에 최적화되면서 기존의 일반 능력이 감소하는 현상이 발생할 수 있으며, 그 효과는 데이터 품질·분포 적합성·튜닝 범위 등에 크게 의존합니다.

미세조정의 필요

사전 학습된 모델은 다양한 상황을 폭넓게 처리할 수 있는 범용 능력을 갖추었지만, 실제 서비스나 업무 환경에서는 특정 도메인의 표현 방식, 규제 요구사항, 기관별 문서 구조, 고객 응대 스타일처럼 맥락적으로 특수한 요소들이 중요하게 작용합니다. 미세조정은 이러한 간극을 해소하여 모델이 목표 작업에 더 적합한 예측과 판단을 수행하도록 만드는 핵심 수단입니다. 또한 기업·기관은 자체 데이터 기반으로 모델을 최적화함으로써 서비스 품질을 향상시키고, 모델 오작동을 줄이며, 경쟁력 있는 맞춤형 시스템을 구축할 수 있습니다. LLM 시대에는 특히 데이터 품질·안전성 기준·조직별 요구사항이 다양해지면서 사전 학습 모델만으로는 충분하지 않은 경우가 많아, 미세조정이 실질적 성능 확보와 안정성 강화의 필수 단계로 자리 잡고 있습니다.

033 바이브 코딩

Vibe Coding

자연어 프롬프트로 AI가 코드를 생성하는 개발 방식

- 개발자가 원하는 기능이나 결과를 자연어로 설명하면, AI가 이를 해석해 실제 코드를 작성하는 새로운 프로그래밍 패러다임
- 복잡한 문법 대신 대화형 지시를 통해 코드를 구현하는 점이 특징

바이브 코딩의 배경

바이브 코딩은 개발자가 직접 코드를 작성하지 않고, AI에게 자연어로 요구를 전달해 프로그램을 완성하는 방식입니다. 사용자는 "이미지를 업로드하면 자동으로 크기를 조정해 저장해줘"처럼 목표를 말하면 되고, AI는 이를 분석해 구체적인 코드를 생성합니다. 이 개념은 전 테슬라 AI 책임자 Andrej Karpathy가 언급하면서 주목받기 시작했으며, LLM을 기반으로 한 프롬프트 중심 프로그래밍 패러다임으로 평가됩니다. 즉, 코딩 과정의 핵심이 '문법 작성'에서 '의도 전달'로 전환되며, AI가 개발 파트너로서 역할을 수행하는 새로운 방식입니다.

바이브 코딩의 활용

바이브 코딩의 가장 큰 특징은 대화형·맥락형 코드 생성입니다. 기존의 자동완성 도구보다 한 단계 발전해, AI가 사용자의 설명을 이해하고 전체 구조를 설계하며, 필요한 모듈·API·UI 요소를 스스로 연결합니다. 개발자는 세부 코드 대신 기능 요구, 디자인 감각, 사용자 경험(UX) 방향 등을 중심으로 개발을 주도합니다. 이러한 방식은 비전문가에게도 코딩 기회를 넓혀, 아이디어만으로 애플리케이션을 구현할 수 있는 개발의 민주화를 이끌고 있습니다. 실제로 시제품 제작, 교육 실습, 기업 내 프로토타입 개발 등에서 활용 가능성이 높으며, 향후에는 실시간 협업형 AI 개발 환경으로 발전할 것으로 예상됩니다.

바이브 코딩의 과제

바이브 코딩은 편리하지만 아직 완전한 기술은 아닙니다. AI가 생성한 코드의 정확성과 보안성은 여전히 검증이 필요하며, 결과물의 오류·저작권·책임 소재가 불분명하다는 문제가 있습니다. 또한 LLM이 이해하지 못한 모호한 지시를 잘못 해석하면 의도와 다른 동작이 발생할 수 있습니다. 사용자는 코드를 직접 제어하기 어렵기 때문에 유지보수와 디버깅 과정에서 어려움을 겪을 수 있고, AI가 생성한 결과의 품질을 검증할 수 있는 신뢰성 평가 체계도 아직 미비합니다. 따라서 향후에는 인간 개발자의 감독 아래에서 AI를 조력자로 활용하는 협력형 개발 모델이 중요해질 것으로 보입니다.

034 버티컬 AI

Vertical AI

특정 산업이나 분야에 맞춰 설계된 전문형 AI

- 버티컬 AI는 의료, 금융, 법률, 제조 등 특정 산업의 데이터와 업무 특성을 반영해 개발된 AI
- 범용 AI가 다양한 영역에 폭넓게 적용되는 반면, 버티컬 AI는 각 산업의 요구와 규제에 최적화된 맞춤형 지능을 제공

버티컬 AI란?

버티컬 AI는 하나의 산업 또는 전문 영역에 특화된 AI 시스템을 뜻합니다. 의료 영상 판독, 금융 리스크 분석, 법률 문서 검토, 스마트팩토리 등 각 분야의 고유한 데이터 구조와 용어 체계를 반영해 설계됩니다. 이러한 방식은 여러 산업을 포괄하는 범용형 AI(Horizontal AI, 수평적 AI)와 달리, 특정 도메인에 대한 깊이 있는 이해와 맞춤 학습을 통해 실질적 성과를 높이는 데 중점을 둡니다. 최근에는 의료진을 지원하는 진단 AI, 보험사 청구 자동화, 제조 설비의 이상 감지 AI 등이 대표적인 사례로 꼽힙니다.

버티컬 AI의 활용

버티컬 AI의 가장 큰 특징은 정확성과 실효성 중심의 설계입니다. 해당 산업의 업무 흐름, 규제, 데이터 품질을 세밀히 반영해 높은 예측력과 신뢰도를 확보합니다. 또한 산업 맞춤형 데이터셋을 사용해 학습하므로 오탐지나 불필요한 연산을 줄이고 효율성을 극대화할 수 있습니다. 예를 들어 금융권에서는 부정 거래 탐지와 리스크 평가, 의료 분야에서는 영상 기반 질병 진단, 법률 분야에서는 판례 검색과 계약서 검토 자동화에 활용됩니다. 버티컬 AI는 이러한 특화 접근을 통해 각 산업의 디지털 전환을 가속하고, 새로운 비즈니스 모델 창출에도 기여하고 있습니다. 특히 스타트업이나 중소기업이 범용 AI보다 적은 자원으로 실질적 ROI(투자 대비 수익)를 얻을 수 있는 점이 강점으로 평가됩니다.

버티컬 AI의 과제

버티컬 AI가 본격적으로 산업 전반에 확산되기 위해서는 여러 과제가 있습니다. 우선 산업별 데이터가 파편화되어 있어 고품질 학습데이터 확보와 표준화가 시급하며, 데이터 접근권과 보안 규제 간의 균형도 필요합니다. 또한 의료·금융 등 규제가 엄격한 분야에서는 윤리적 책임, 검증 체계, 규제 준수가 필수 전제 조건으로 떠오르고 있습니다. 산업 특화형 모델은 유지보수를 위해 도메인 전문가와 AI 전문가의 협업이 필수적이므로 전문 인력 생태계 구축도 과제입니다. 향후에는 국가·산업 차원의 데이터 표준화, 투명한 AI 검증 체계, 산업별 거버넌스 강화가 함께 이루어질 것으로 보입니다. 이러한 기반이 마련된다면 버티컬 AI는 단일 기술을 넘어, 각 산업의 생산성과 혁신 역량을 이끄는 핵심 인프라로 발전할 전망입니다.

035 벤치마크 데이터셋

Benchmark Dataset

AI 모델의 성능을 비교·평가하기 위한 표준 데이터 모음

- 여러 모델이 동일한 조건에서 성능을 비교할 수 있도록 구성된 표준화된 성능 평가용 데이터 모음
- AI 연구의 공정한 경쟁과 기술 발전의 객관적 기준을 제공하는 핵심 인프라

벤치마크 데이터셋의 개념

벤치마크 데이터셋은 AI 모델의 성능을 객관적이고 재현 가능하게 평가하기 위한 데이터 집합입니다. 단순한 학습용 데이터가 아니라 공통된 테스트 환경과 평가 지표를 함께 제공해 모델 간 비교가 가능하도록 설계됩니다. 연구자는 동일한 데이터와 조건으로 실험해 어느 모델이 더 우수한지 확인할 수 있습니다. 대표적으로 이미지 분류용 ImageNet, 손글씨 인식용 MNIST, 자연어 이해용 GLUE 등이 있으며, 이러한 데이터셋은 AI 기술 발전의 공용 시험지 역할을 합니다.

벤치마크 데이터셋의 특징

벤치마크 데이터셋은 여러 AI 모델을 동일한 조건에서 평가하고 비교할 수 있도록 설계되었다는 점이 특징입니다. 동일한 입력과 라벨 구조를 유지해 모델 간 평가 조건을 맞추고, 평가 절차가 명확히 문서화되어 재현 가능한 결과를 제공합니다. 또한 정확도나 F1 점수처럼 통일된 지표를 사용해 모델의 성능을 객관적으로 판단할 수 있습니다. 이 구조를 통해 연구자들은 개선 정도를 빠르게 파악하고, 결과는 학계와 산업계가 공통으로 신뢰하는 기준선으로 활용됩니다. 최근 벤치마크는 단순 정확도뿐 아니라 추론 능력(reasoning), 지식 활용, 복잡한 문제 해결, 도구 사용 능력, 안전성 평가 등 고차원적 성능을 측정하는 방향으로 확장되고 있습니다.

벤치마크 데이터셋의 중요성

벤치마크 데이터셋은 AI 연구 생태계의 공통 언어이자 발전의 척도입니다. 이를 통해 연구자들은 성능을 검증하고, 기업은 신기술의 경쟁력을 평가합니다. 또 벤치마크는 모델 개발 방향을 제시하며 기술 진보의 속도를 수치로 보여줍니다. 정부나 기관의 AI 인증·표준화 정책에서도 정량적 평가 기준으로 활용되어 산업 전반의 신뢰성과 효율성을 높입니다.

벤치마크 데이터셋의 한계

벤치마크 데이터셋은 실제 환경을 완벽히 반영하지 못한다는 한계가 있습니다. 일부 모델은 특정 데이터에 과적합되어 '시험 대비형 AI'로 작동할 수 있고, 데이터 편향으로 현실의 다양성이 충분히 반영되지 않기도 합니다. 또한 오래된 데이터셋은 새로운 문제나 환경 변화를 따라가지 못해 시대적 적합성이 떨어집니다. 따라서 주기적 갱신과 실제 응용 테스트를 병행해야 하며, 벤치마크는 AI 발전의 방향을 제시하는 수단이지 목표가 되어서는 안 됩니다.

036 분산학습

Distributed Training

여러 장비가 협력해 AI 모델을 병렬 학습하는 방식

- 대규모 데이터나 복잡한 AI 모델을 여러 서버·GPU에 나누어 병렬 처리함으로써 학습 속도와 효율을 높이는 기술

분산학습의 배경

AI 모델의 규모가 커지면서 단일 장비로는 연산량과 메모리를 감당하기 어려워졌습니다. 수십억 개 이상의 매개변수를 가진 LLM이나 멀티모달 모델은 학습에 막대한 시간이 필요합니다. 이를 해결하기 위해 여러 장비가 동시에 연산을 수행하는 분산학습이 등장했습니다. 하나의 모델을 여러 서버·GPU가 나누어 학습함으로써 시간을 단축하고, 메모리 한계를 넘어 대형 모델을 효율적으로 훈련할 수 있습니다. 분산학습은 초대규모 AI 개발의 핵심 인프라로 기능합니다.

분산학습의 유형

분산학습은 데이터 병렬화와 모델 병렬화로 구분됩니다. 데이터 병렬화는 동일 모델을 여러 장비에 배치해 데이터의 일부를 학습하고 결과를 통합하는 방식으로, 구현이 단순하고 효율적입니다. 반면 모델 병렬화는 초거대 모델같이 모델이 너무 커서 단일 GPU 메모리에 들어가지 않을 때 주로 사용되고 각 장비는 일부 계산만 수행합니다. 최근 두 방식을 결합한 하이브리드 병렬화가 주로 사용되며, 통신 지연을 줄이기 위한 최적화 기술도 발전하고 있습니다.

분산학습의 의의

분산학습은 학습 속도 단축과 확장성 확보에 탁월합니다. 대규모 학습이 가능하고, 여러 장비를 병렬로 활용해 자원 효율도 높지만 노드 간 통신 오버헤드와 동기화 지연, 네트워크 병목 등의 문제는 여전히 과제입니다. 장애 복구와 자원 관리 비용 부담도 크며, 일관된 결과 통합을 위한 기술적 보완이 필요합니다. 향후에는 통신 효율 향상과 자동화된 운영 관리 기술이 병행 발전해야 분산학습의 효율성이 극대화될 것입니다.

관련 용어

연합학습(federated learning)

여러 장치나 기관이 데이터를 공유하지 않은 채로 협력 학습을 수행하는 방식입니다. 분산학습이 하나의 모델을 여러 장비로 나누어 계산 효율을 높이는 기술이라면, 연합학습은 데이터 프라이버시 보호와 분산 데이터 활용을 목표로 합니다. 각 참여 노드는 자신이 가진 데이터를 로컬에서 학습한 뒤, 모델의 매개변수만 중앙 서버로 전송해 통합합니다. 이 과정에서 개인 데이터는 외부로 이동하지 않아 보안성과 개인정보 보호가 강화되기 때문에 의료, 금융, 모바일 기기 등 데이터 이동이 제한된 환경에서 특히 유용합니다.

037 비전언어모델 / VLM

Vision-Language Model

이미지와 텍스트를 함께 이해하고 처리하는 AI 모델

- 시각 및 언어 정보를 결합해 이미지 해석, 설명 생성, 질의응답 등을 수행하는 멀티모달 AI 모델
- 화면·장면·문맥을 통합적으로 이해해 다양한 작업을 처리하는 구조

비전언어모델의 개념

비전언어모델(VLM)은 이미지와 텍스트를 동시에 입력받아 서로의 의미를 연결해 이해하도록 설계된 AI 모델을 의미합니다. 기존에는 이미지 분석과 언어 처리가 별도의 모델에서 이루어졌지만, VLM은 두 정보를 통합적으로 해석해 하나의 과제로 처리한다는 점에서 차별됩니다. 예를 들어 한 장의 사진에서 사물을 인식하는 것뿐 아니라, 사진 속 상황을 설명하거나 특정 부분에 대해 질문에 답하는 등 복합적 이해가 가능합니다. 이러한 모델은 시각·언어 정보를 결합해 더 자연스럽고 인간적인 방식으로 세계를 이해하려는 AI 발전 흐름 속에서 등장했으나, 최근에는 처음부터 다양한 모달리티(텍스트·이미지·음성 등)를 단일 모델에서 일관되게 처리하도록 설계된 거대 멀티모달 모델(LMM)로 발전하고 있습니다.

비전언어모델의 구성

VLM은 일반적으로 시각 정보를 처리하는 비전 인코더와 언어 정보를 처리하는 언어 모델을 결합해 작동합니다. 먼저 비전 인코더가 이미지에서 특징을 추출해 벡터 형태로 변환하고, 언어 모델은 이 벡터를 문맥 정보로 활용해 질문에 답하거나 설명을 생성합니다. 이 과정에서 두 정보가 서로 연결되도록 멀티모달 임베딩 공간이 활용되며, 이미지와 텍스트가 의미적으로 정렬될수록 모델의 이해 능력이 향상됩니다. 최근에는 LLM을 중심에 두고 이미지 인코더를 접목하는 구조가 주류가 되었으며, 이를 통해 언어 기반 지시를 시각적 판단과 결합해 더 복잡한 작업을 수행할 수 있게 되었습니다. 이러한 구조는 텍스트와 이미지가 서로의 부족한 부분을 보완하여 더 정교한 추론을 가능하게 합니다.

비전언어모델의 전망

비전언어모델은 고객지원 자동화, 시각 검색, 이미지 설명 생성 등 시각·언어 결합이 필요한 업무에서 핵심 기반 기술로서 활용되고 있습니다. 그러나 최근 AI 발전 흐름은 더 높은 통합성과 일반성을 갖춘 LMM 중심으로 이동하고 있습니다. LMM은 텍스트·이미지·음성·영상 등 다양한 정보를 단일 모델에서 일관되게 처리하며, 멀티모달 이해·추론·액션 수행까지 확장되는 경향을 보입니다. 이로 인해 전통적 VLM은 독립적 모델군이라기보다는 멀티모달 AI로 발전해가는 과정에서 등장한 전환기적 기술 단계로 평가받고 있습니다.

038 사고 사슬 / CoT

Chain of Thought

AI가 문제 해결 과정을 단계적으로 보여주도록 유도하는 프롬프트 기법

- AI가 복잡한 문제를 풀 때 사고의 중간 단계를 명시적으로 표현하도록 유도하는 프롬프트 기법
- 정답만 제시하는 대신 단계별 추론을 수행하게 하여 논리적 일관성과 정답률을 향상

CoT의 작동 원리

사고 사슬은 LLM이 정답을 내리기 전 스스로의 사고 흐름을 언어로 표현하도록 유도하는 방법입니다. 일반적인 프롬프트가 단일 결과만 요구한다면, CoT는 "단계별로 생각해보자" 같은 지시문을 통해 모델이 문제 해결의 논리 과정을 서술하도록 만듭니다. 이렇게 생성된 중간 추론 단계는 모델이 문제를 하위 요소로 나누고, 각 단계에서 논리를 전개해 최종 결론에 이르는 구조를 형성합니다. 예를 들어 수학 문제에서 조건을 정리하고, 식을 세운 뒤 계산 결과를 도출하는 과정을 순차적으로 서술하도록 하는 방식입니다. 이러한 사고 유도는 모델 내부의 연산 구조를 정돈시켜, 단순 답변 생성이 아닌 과정 기반 추론을 수행하도록 돕습니다.

CoT의 활용

사고 사슬은 특히 다단계 추론이 필요한 문제 영역에서 탁월한 성과를 보입니다. 수학 계산, 논리 퍼즐, 법률 질의응답, 코드 디버깅 등 복잡한 인과 관계를 파악해야 하는 작업에서 모델의 정답률이 크게 향상되었습니다. 또한 CoT는 모델이 사고 과정을 명시적으로 표현하기 때문에 결과의 설명 가능성이 높아지고, 사람이 판단 근거를 검토할 수 있습니다. 최근에는 여러 사고 경로를 병렬로 생성한 뒤 다수결로 결론을 도출하는 Self-Consistency CoT, 또는 사람이 작성한 사고 단계를 학습한 CoT 미세조정 모델이 개발되어 복잡한 문제 해결 능력을 한층 확장하고 있습니다. 이러한 방식은 연구용 대형 모델뿐 아니라 챗봇, 자동 코드 생성기, 과학 계산 보조 시스템 등 실제 응용 서비스에도 도입되고 있습니다.

CoT의 한계와 의의

사고 사슬은 모델의 추론 능력을 개선하는 강력한 도구이지만, 생성된 사고 단계가 실제 논리를 충실히 반영한다고 보긴 어렵습니다. 모델이 '생각하는 척'은 할 수 있어도 그 과정이 정확한 계산 근거나 사실 관계를 담보하지 못할 수 있습니다. 또한 프롬프트의 문구나 예시 구성에 따라 성능 차이가 크고, 단계가 길어질수록 오류가 누적될 위험도 있습니다. 그럼에도 CoT는 AI가 단순 패턴 모방에서 벗어나 추론적 사고를 표현하려는 전환점이라는 점에서 중요한 의의를 지닙니다. 인간의 사고 과정을 모방함으로써 AI의 신뢰성과 해석 가능성을 높이고, 장기적으로는 자율적 문제 해결 AI로 발전하기 위한 기반을 제공합니다.

참조 용어 | 파운데이션 모델 p.81

039 사전학습모델

Pretrained Model

대규모 데이터로 미리 학습되어 다양한 과제에 활용되는 AI 모델

- 사전학습모델은 방대한 일반 데이터를 사전에 학습해 언어, 이미지, 패턴 등 폭넓은 기본지식과 표현 능력을 축적한 뒤, 이를 바탕으로 특정 작업에 맞게 조정해 사용하는 기반형 AI 모델
- 다양한 과제에 적용할 수 있어 데이터 활용 효율성과 적용 범위의 확장성이 높은 것이 특징

사전학습모델의 개념

사전학습모델은 AI가 새로운 과제에 대응하기 전, 대규모 데이터로 언어·이미지 등 일반적 패턴을 미리 학습해 둔 모델을 말합니다. 인간이 기초 지식을 습득한 뒤 전문 기술을 익히듯, AI가 세상의 통계적 구조를 선행 학습하는 과정입니다. 이후 특정 목적에 맞는 데이터를 추가 학습하면 빠르고 효율적으로 성능을 개선할 수 있으며, 방대한 텍스트나 이미지 데이터를 통해 언어 이해·시각 인식 등 다양한 능력을 확보합니다.

사전학습모델의 특징

사전학습모델의 핵심 특징은 지식의 재활용과 일반화 능력입니다. 이미 학습된 표현과 가중치를 활용하므로 새로운 과제에서도 적은 데이터로 높은 성능을 냅니다. 또한 동일한 모델이 여러 영역에 확장 적용될 수 있어 개발 비용과 시간을 절감합니다. 예를 들어 언어모델은 일반 텍스트를 학습한 후 감정 분석, 요약, 질의응답 등 다양한 작업에 미세조정(Fine-Tuning)만으로 활용됩니다. 일부 모델은 추가 학습 없이 새로운 과제를 수행하는 제로샷(Zero-shot) 능력도 보여, AI를 범용 지능 플랫폼으로 발전시키는 기반이 되고 있습니다.

사전학습모델의 한계

사전학습모델은 효율성과 확장성 면에서 혁신적이지만, 한계도 분명합니다. 학습에 사용된 데이터의 편향과 오류가 그대로 전이될 수 있으며, 특정 언어·문화권 중심의 데이터는 공정성 문제로 이어집니다. 또한 초거대 모델의 학습에는 막대한 연산 자원과 에너지가 필요해 환경적 부담이 큽니다. 그럼에도 사전학습모델은 AI 연구의 표준 구조로 자리 잡았으며, 적은 자원으로도 강력한 성능을 구현할 수 있는 핵심 기술로 평가됩니다.

관련 용어

증류 (Distillation)

대형 사전학습모델이 가진 지식과 판단 방식을 작은 모델이 학습하도록 전이하는 과정입니다. 복잡하고 무거운 모델(교사 모델)의 출력을 참고해, 더 단순한 모델(학생 모델)이 유사한 성능을 내도록 훈련합니다. 이를 통해 모델 크기와 연산량을 크게 줄이면서도 정확도를 유지할 수 있습니다. 예를 들어 거대한 언어모델의 응답 패턴을 작은 모델이 모방하게 하면, 모바일 기기나 제한된 서버 환경에서도 빠르고 효율적인 AI 서비스를 구현할 수 있습니다. 즉, 증류는 사전학습모델의 성능을 경량화해 실용성을 높이는 핵심 전이학습 기법입니다.

참조 용어 | 합성곱 신경망(CNN) p.89

040 생성적 적대 신경망 / GAN

Generative Adversarial Networks

생성자와 판별자의 경쟁으로 새로운 데이터를 만들어내는 신경망 구조

- 두 신경망이 경쟁하며 발전하는 구조로,하나는 가짜 데이터를 생성하고, 다른 하나는 진위를 판별
- 적대적 학습을 통해 실제와 유사한 데이터를 만들어내는 대표적 생성형 AI 기술

GAN이란?

GAN은 2014년 Ian Goodfellow가 제안한 딥러닝 구조로, 경쟁을 통한 학습(adversarial learning)을 기반으로 합니다. 두 개의 인공신경망이 서로 대립적 관계를 이루며 학습하는데, 생성자(Generator)는 실제처럼 보이는 데이터를 만들어내고, 판별자(Discriminator)는 입력된 데이터가 진짜인지 가짜인지를 구분합니다. 특히 판별자에는 주로 이미지 인식에 특화된 합성곱 신경망(CNN) 구조가 활용되어 시각적 패턴을 정교하게 분석합니다. 학습이 반복되면서 생성자는 점점 더 사실적인 데이터를 만들어내고, 판별자는 이를 구별하기 어려워집니다. 결국 두 모델이 균형 상태에 도달하면 생성자는 실제 데이터와 거의 구분되지 않는 결과물을 생산할 수 있습니다.

GAN의 작동 원리

GAN은 무작위 입력을 받아 가짜 데이터를 생성하는 생성자와, 이를 실제 데이터와 함께 평가해 진위를 판정하는 판별자로 구성됩니다. 판별자의 피드백은 생성자 학습에 반영되어, 다음 생성물이 더 정교해지도록 개선됩니다. 이러한 경쟁적 학습이 반복되면 생성자는 실제 데이터의 분포를 스스로 모사하게 됩니다. GAN은 주로 비지도·자기지도로 학습하지만, 과업에 따라 지도학습을 사용하는 변형도 활용되며, 이를 통해 현실적인 데이터 생성이 가능해집니다. 학습 안정성을 높이기 위해 Wasserstein GAN, StyleGAN 등 다양한 변형 구조가 등장했고, 이미지 품질과 학습 효율 모두 개선되었습니다.

GAN의 활용

GAN은 오늘날 생성형 AI의 기반 기술로 폭넓게 활용됩니다. 대표적으로 이미지 생성에 사용되어 사람의 얼굴, 풍경, 예술 작품 등 실제와 구분하기 어려운 이미지를 만들어냅니다. 의료 영상에서는 희귀 질환 데이터를 합성해 학습 데이터를 확충하고, 패션·디자인 분야에서는 새로운 스타일을 시각화합니다. 또한 영상 복원, 초해상도 변환, 음성 합성, 데이터 증강 등에도 쓰이며, 딥페이크(Deepfake) 기술의 기초로도 작동합니다. 최근에는 확산모델(Diffusion Model)과 결합하거나 대체 기술로 발전하며, AI가 스스로 창조적 결과물을 생산하는 단계로 나아가고 있습니다.

041 생성형 AI

Generative AI

새로운 텍스트·이미지·음성·영상 등을 만들어내는 AI 기술

- 기존 데이터를 학습해 새로운 형태의 콘텐츠를 창조적으로 생성하는 기술
- 분류나 예측을 넘어, 데이터의 분포를 이해하고 조합하여 '새로움'을 만들어내는 AI의 진화된 형태

생성형 AI란?

생성형 AI는 기존의 인식형 AI가 데이터를 분석하거나 분류하는 것과 달리, 학습 과정에서 축적한 패턴과 의미 구조를 바탕으로 새로운 콘텐츠를 만들어내는 AI 기술입니다. 텍스트, 이미지, 음성, 영상 등 다양한 데이터를 대규모로 학습해 내재적 규칙을 이해하고, 생성 단계에서는 그 분포를 바탕으로 사용자 요청에 맞춰 가장 가능성 높은 샘플을 산출하는 방식입니다. 예를 들어 언어모델은 단어의 연속 확률을 예측해 문장을 이어가고, 이미지 모델은 픽셀 단위의 변화를 예측해 새로운 그림을 만듭니다. 즉, 생성형 AI는 데이터의 구조를 학습해 창조적으로 재구성하는 알고리즘이라 할 수 있습니다.

생성형 AI의 활용

생성형 AI는 산업과 일상 전반에 걸쳐 활용되고 있습니다. 텍스트 생성에서는 문서 작성, 번역, 요약, 콘텐츠 기획 등에 쓰이며, 이미지 생성에서는 디자인, 광고, 엔터테인먼트, 패션 분야에서 창작을 보조합니다. 음악·영상 생성 모델은 예술가의 작업을 지원하고, 의료 분야에서는 실제 환자 정보 없이 합성 데이터를 만들어 희귀 질환 연구나 영상 진단 정확도를 높이는 데 활용됩니다. 또한 코드 생성, 고객 응대, 교육용 시뮬레이션 등 지식노동 영역으로 확장되어 생산성과 창의성의 새로운 결합을 이끌고 있습니다. 이러한 활용은 단순 자동화를 넘어 인간의 창의 과정을 보완하는 방향으로 진화하고 있습니다.

생성형 AI Midjourney로 생성된 이미지

출처 : Midjourney

생성형 AI의 쟁점

생성형 AI의 확산은 기술적 혁신과 함께 사회적 쟁점을 동반합니다. 가장 큰 문제는 저작권과 데이터 출처의 불투명성입니다. 학습 데이터에 포함된 창작물의 저작권 침해 논란이 지속되고 있으며, 생성된 결과물의 소유권 역시 명확히 규정되지 않았습니다. 또한 허위 정보나 조작 이미지가 손쉽게 생성되면서 딥페이크와 정보 왜곡 문제가 심화되고 있습니다. 데이터 편향에 따른 차별적 결과, 개인정보 유출 위험, 환경적 비용 등도 중요한 논의 대상입니다. 그럼에도 생성형 AI는 AI가 '이해'에서 '창조'로 확장된 기술적 전환점으로, 인간의 표현 능력을 보조하고 새로운 산업 생태계를 여는 창의적 파트너 기술로 평가됩니다.

042 섀도 AI

Shadow AI

조직의 공식 승인이나 통제 없이 직원이 개인적으로 AI 도구를 사용하는 현상

- 생산성 향상을 위해 생성형 AI 등을 자율적으로 사용하면서, 보안·데이터 관리 사각지대를 발생시키는 조직 운영 리스크
- 단순히 통제와 차단의 대상으로만 보기보다, 조직 내 AI 사용 현황에 대한 가시적 확보와 관리·통제 체계 마련이 필요

섀도 AI의 개념

섀도 AI란 IT 부서 등 조직의 공식적인 승인이나 감독 없이, 조직 구성원이나 직원이 AI 도구나 애플리케이션을 개인적으로 사용하는 현상을 의미합니다. 업무 생산성 향상이나 반복 작업 자동화를 목적으로 ChatGPT와 같은 생성형 AI 서비스를 개인 계정으로 활용하는 경우가 대표적입니다. 조직이 인지하거나 승인하지 않은 상태에서 '그림자'처럼 활용된다는 점에서 '섀도 AI'라는 명칭이 붙었습니다. 이는 비인가 IT 기술 사용을 의미하는 '섀도 IT'와 유사하지만, AI가 데이터를 처리·학습하는 특성으로 인해 위험이 더 복잡하고 은밀하게 확대될 수 있다는 점에서 차이가 있습니다.

섀도 AI로 인한 피해

섀도 AI는 AI의 실무 활용 가치가 커질수록 빠르게 확산되고 있습니다. 그러나 이러한 비공식적 사용은 조직 차원의 보안 및 데이터 관리 위험을 크게 증가시킵니다. CISCO의 「2025 사이버보안 준비 지수」에 따르면, 전 세계 보안 리더의 83%는 섀도 AI 탐지에 자신이 없다고 응답했으며, 다수의 조직이 실제 사용 현황을 정확히 파악하지 못하고 있는 것으로 나타났습니다. 또한 주요 AI 서비스에 입력된 데이터 중 일부에 기업 내부 정보가 포함된 사례가 확인되었고, IBM의 「2025 데이터 유출 비용 보고서」는 섀도 AI 관련 사고가 탐지·대응에 더 많은 시간과 비용을 초래한다고 분석했습니다.

섀도 AI의 대응 방안

섀도 AI에 대응하기 위해서는 단순한 사용 금지보다 안전한 활용을 전제로 한 관리전략이 필요합니다. 매니지엔진의 조사에 따르면, IT 의사결정권자의 97%는 섀도 AI를 기업에 대한 심각한 위협으로 인식하고 있는 반면, 91%의 직원들은 섀도 AI 사용에 거의 위험이 없다고 인식하고 있는 것으로 나타났습니다. 이러한 인식 격차는 섀도 AI 관리의 또 다른 리스크 요인으로 지적됩니다. 전문가들은 섀도 AI를 통제와 차단의 대상으로만 보기보다, 조직의 공식 AI 활용 체계 안으로 안전하게 흡수·관리하는 접근이 중요하다고 강조하며, 이를 위해선 AI 활용 교육과 명확한 정책 정비가 필수적입니다.

043 서비스형 AI / AIaaS

AI as a Service

클라우드를 통해 AI 기능을 서비스 형태로 제공하는 모델

- 클라우드 환경에서 AI 모델·도구·인프라를 구독형 서비스 형태로 제공하는 방식
- 기업이나 개인이 직접 시스템을 구축하지 않고도 AI 기능을 쉽게 활용 가능

AIaaS의 개념

서비스형 AI(AI as a Service, AIaaS)는 인공지능 기술을 클라우드 환경에서 구독이나 사용량 기반으로 제공하는 서비스 모델입니다. 사용자는 별도의 장비나 인프라 없이 인터넷을 통해 AI 기능을 호출해 이용할 수 있습니다. 이는 인프라를 제공하는 IaaS, 플랫폼을 제공하는 PaaS, 소프트웨어를 제공하는 SaaS에 이은 확장형 서비스로, AI 기술을 모듈화해 즉시 활용 가능하게 만든 형태입니다. 복잡한 모델 개발 과정을 서비스화함으로써 고가의 장비나 전문 인력이 없어도 AI 기능을 쉽게 도입할 수 있습니다.

AIaaS의 작동 방식

AIaaS는 클라우드 서버에 구축된 AI 인프라와 모델을 API 형태로 외부에 개방하는 구조로 작동합니다. 사용자는 음성 인식, 이미지 분석, 번역, 챗봇 등 필요한 기능을 호출해 자신의 애플리케이션에 통합합니다. 서비스 제공자는 GPU 자원과 데이터 저장소, 모델 학습 및 배포 환경을 관리하며, 사용자는 웹 콘솔이나 프로그래밍 인터페이스를 통해 기능을 조합합니다. 이 과정에서 클라우드는 연산 부담을 분산하고 유지보수를 자동화해 확장성과 효율성을 높입니다.

AIaaS의 확산

서비스형 AI가 빠르게 확산된 배경에는 경제성과 접근성, 그리고 클라우드 생태계의 성숙이 있습니다. 초기 인프라 구축 비용을 절감하고 사용한 만큼만 지불할 수 있는 경제적 효율성이 도입을 촉진했습니다. 또한 복잡한 AI 기술을 모듈 단위로 제공함으로써 비전문가도 AI 기능을 활용할 수 있게 되었고, 안정적인 네트워크와 GPU 자원이 확보되면서 대규모 서비스 운영이 가능해졌습니다. 이러한 요인들은 AI 기술을 특정 기업의 자산이 아닌 공용 서비스 인프라로 전환시키는 핵심 동력이 되었습니다.

AIaaS의 의의

서비스형 AI는 AI 기술의 민주화를 실현한 대표적 모델입니다. 막대한 자본과 전문 인력이 필요했던 AI 시스템을 서비스화해 중소기업과 개인 사용자도 AI 혁신에 참여할 수 있게 되었습니다. 지속적인 모델 업데이트와 확장 가능한 인프라는 산업 전반의 디지털 전환(DX)을 가속했습니다. 반면 데이터 집중으로 인한 보안·프라이버시 위험과 공급자 종속(vendor lock-in) 문제는 여전히 해결 과제입니다. 그럼에도 AIaaS는 AI를 기술에서 서비스로 전환한 패러다임으로, 앞으로 산업의 기본 인프라로 확대될 것으로 기대됩니다.

참조 용어 | AI 신뢰성 p.101

044 설명가능한 AI / XAI

Explainable AI

AI의 판단 근거를 사람이 이해할 수 있도록 설명하는 기술

- AI의 의사결정 과정을 투명하게 드러내어 사람이 이해·검증할 수 있게 하는 기술
- 복잡한 블랙박스형 모델의 한계를 보완해 신뢰성과 책임성을 높이는 것이 목표

XAI의 개념

설명가능한 AI(XAI)는 인공지능이 내린 결과의 근거와 과정을 사람이 이해할 수 있는 형태로 제시하는 기술입니다. 딥러닝 모델은 수많은 매개변수와 연산 과정을 거쳐 결과를 도출하지만, 내부 작동 원리를 직접 해석하기 어렵습니다. 이러한 '블랙박스 문제'를 해결하기 위해 XAI는 AI의 판단 근거를 시각화하거나 규칙화하여 투명하게 제시합니다. 즉, 단순히 결과를 제시하는 수준을 넘어, 그 결론에 이르는 논리적 과정을 드러내어 사용자와 개발자가 신뢰할 수 있는 판단 구조를 제공합니다.

XAI의 접근 방식

XAI의 접근 방식은 크게 두 가지입니다. 하나는 내재적 설명성으로, 애초에 구조가 단순하고 논리적인 모델을 설계해 결과를 사람이 직접 이해할 수 있게 만드는 방식입니다. 예를 들어 의사결정나무나 선형 회귀 모델처럼 계산 과정이 명확한 경우가 이에 해당합니다. 다른 하나는 사후적 설명으로, 복잡한 신경망 모델이 이미 도출한 결과를 시각화하거나 규칙으로 해석하는 방법입니다. 이러한 기법은 모델의 복잡한 내부 연산을 사람이 이해할 수 있는 형태로 번역하여, AI의 결정 과정을 간접적으로 해석하게 합니다.

XAI의 필요성

AI가 의료 진단, 금융 심사, 채용 평가 등 사회 전반의 의사결정에 활용되면서 판단의 투명성과 설명 가능성은 필수 요건이 되었습니다. 설명가능한 AI는 결과의 근거를 명확히 제시함으로써 사용자 신뢰를 높이고, 오류나 편향이 발생했을 때 원인을 추적할 수 있도록 합니다. 또한 법적·윤리적 책임을 명확히 하여, 신뢰할 수 있는 AI(Trustworthy AI) 구축의 기반이 됩니다. 특히 공공 정책과 산업 규제 분야에서는 알고리즘의 공정성과 책임성을 확보하는 수단으로 중요하게 작용하며, 사회적 수용성을 높이는 역할을 하고 있습니다.

XAI의 과제

XAI는 투명성과 신뢰성을 강화하지만 완전한 해법은 아닙니다. 설명력을 높이기 위해 모델을 단순화하면 성능이 저하될 수 있고, 복잡한 모델은 여전히 해석이 어렵습니다. 또한 사용자가 이해하기 쉬운 설명이 반드시 사실적이거나 정확한 설명을 의미하지는 않습니다. 설명력과 성능 간의 균형, 그리고 다양한 문화·언어·데이터 환경을 고려한 설명 기준 마련이 앞으로의 핵심 과제입니다. 그럼에도 XAI는 AI 윤리와 책임성을 실현하는 핵심 기술로, 인간 중심의 신뢰 가능한 AI 발전 방향을 제시하는 기반으로 평가됩니다.

045 순환 신경망 / RNN

Recurrent Neural Network

이전 단계에서 학습한 정보를 다음 단계 학습에 활용하는 신경망 구조

- 시계열 데이터나 문장처럼 순서가 중요한 정보를 처리하는 인공신경망
- 과거의 출력을 현재 입력과 함께 사용해 데이터의 흐름과 맥락을 학습

순환 신경망이란?

순환 신경망(RNN)은 데이터가 시간적 또는 순차적 관계를 가질 때 이를 학습하도록 설계된 신경망입니다. 일반적인 신경망이 입력을 독립적으로 처리한다면, RNN은 이전 단계의 출력 정보를 다음 입력 처리에 재사용함으로써 시간의 흐름을 고려합니다. 예를 들어 문장을 이해하거나 음성을 인식할 때, 앞선 단어 또는 소리의 정보를 다음 단계로 전달해 문맥이나 억양의 변화를 반영합니다. 이러한 구조로 인해 RNN은 언어 처리, 음성 인식, 주가 예측 등 순서가 중요한 데이터 처리에 널리 활용됩니다.

순환 신경망의 작동 원리

RNN의 가장 큰 특징은 '기억' 기능입니다. 내부에 은닉 상태라는 구조를 두어 이전 단계에서 학습한 정보를 저장하고, 이를 현재 입력과 함께 사용합니다. 쉽게 말해, RNN은 과거 입력을 잠시 기억해 두었다가 다음 단계 판단에 참고하는 방식으로 작동합니다. 예를 들어 문장 "AI is powerful"을 처리할 때, RNN은 'AI'라는 단어의 의미를 기억했다가 'is'와 'powerful'을 해석할 때 그 정보를 함께 고려합니다. 이런 순환 구조를 통해 데이터 간의 연속적인 관계를 학습할 수 있습니다. 또 모든 단계에서 동일한 가중치를 사용하기 때문에, 일정한 규칙으로 시간 흐름 속의 패턴을 파악할 수 있습니다. 다만 긴 문장처럼 정보가 오래 이어질 경우, 초기에 학습된 내용이 점차 희미해지는 기울기 소실문제가 발생하기 쉬워, 이를 보완하기 위해 기억을 더 오래 유지할 수 있는 개선형 모델이 등장했습니다.

순환 신경망의 한계

RNN은 데이터의 순서를 반영해 문맥과 흐름을 학습할 수 있다는 점에서 강력하지만, 구조적 제약도 많습니다. 계산이 순차적으로 이뤄지기 때문에 병렬 처리가 어려워 학습 속도가 느리고, 긴 데이터에서는 기울기 소실로 과거 정보가 반영되지 않는 한계가 있습니다. 또한 입력이 길어질수록 학습 효율이 떨어지고, 복잡한 문맥에서는 의미를 정확히 유지하기 어렵습니다. 이러한 이유로 최근에는 문장 전체를 한 번에 처리할 수 있는 트랜스포머 구조가 RNN을 대체하고 있습니다. 그럼에도 RNN은 시계열 데이터 분석과 언어 처리 기술의 기반을 마련한 중요한 신경망 구조로 평가됩니다.

046 시뮬레이션-현실 전이 / Sim-to-Real Simulation-to-Reality Transfer

가상환경에서 학습한 AI 모델을 현실 환경에 적용하는 기술

- 가상 환경에서 훈련된 AI나 로봇 모델을 실제 환경에서도 작동하도록 전이시키는 기술
- 현실 실험의 위험과 비용을 줄이면서, 학습 효율을 높이는 데 활용

시뮬레이션-현실 전이의 개념

시뮬레이션-현실 전이는 AI나 로봇이 가상환경에서 학습한 결과를 실제 환경으로 옮겨 적용하는 기술을 말합니다. 현실에서는 반복 실험이 어렵거나 위험한 경우가 많기 때문에, 먼저 시뮬레이터에서 수천 번의 학습을 수행해 안정적인 성능을 확보합니다. 이후 이렇게 학습된 모델을 실제 센서나 장비에 이식해 조명, 마찰, 온도, 물체 변형 등 현실의 변수에도 대응할 수 있도록 조정합니다. 예를 들어 로봇팔이 물체를 집는 동작을 시뮬레이션으로 충분히 훈련한 뒤, 그 경험을 실제 로봇 제어에 적용하는 방식입니다. 이 과정은 현실 환경에서의 위험과 비용을 줄이면서도 높은 학습 효율을 얻는 데 효과적입니다.

시뮬레이션-현실 전이의 접근 방식

Sim-to-Real 기술은 시뮬레이션과 현실 간 차이를 줄이기 위해 도메인 적응(Domain Adaptation)과 도메인 무작위화(Domain Randomization) 방식을 주로 사용합니다. 도메인 적응은 시뮬레이션 데이터를 현실 데이터에 가깝게 보정해, 모델이 실제 환경에서도 인식 오류를 내지 않게 만드는 접근입니다. 반면 도메인 무작위화는 학습 중에 시뮬레이션 조건을 계속 바꿔 모델이 다양한 상황에 익숙해지도록 하는 방법입니다. 예를 들어 조명, 질감, 색상, 물체의 위치를 무작위로 변경하며 학습하면 현실에서 새로운 조건을 만나도 유연하게 대응할 수 있습니다. 이 외에도 시뮬레이션의 물리적 특성을 현실에 맞게 조정하는 시뮬레이션 보정(Sim Calibration), 가상 데이터와 실제 데이터를 함께 사용하는 혼합 학습(Hybrid Learning) 등도 활용됩니다. 이러한 접근들은 현실 데이터 수집의 한계를 보완하면서 모델의 일반화 능력을 높이는 데 기여합니다.

시뮬레이션-현실 전이의 활용

시뮬레이션-현실 전이는 AI를 실제 환경에서 안전하고 효율적으로 적용하기 위한 핵심 기술로 활용됩니다. 로보틱스에서는 조립, 물체 인식, 이동 경로 학습 등에 사용되어 실험 비용과 시간을 절감합니다. 자율주행에서는 실제 도로 주행 전 가상 시나리오로 차량의 판단 능력을 충분히 검증할 수 있으며, 제조 분야에서는 공정 자동화와 설비 점검 시뮬레이션에 이용됩니다. 의료 분야에서는 수술 로봇이나 재활 보조 기기를 훈련해 안전성을 높이는 데 기여합니다. 이러한 기술은 현실 데이터를 직접 다루기 어려운 환경에서도 AI 모델을 학습·검증할 수 있게 해서 가상과 현실을 잇는 핵심 전이 기술로 평가됩니다.

047 오토인코더 Autoencoder

입력을 압축했다가 다시 복원하며 특징을 학습하는 신경망 모델

- 데이터의 중요한 구조만 남기도록 스스로 표현을 압축·재구성하는 방식
- 차원 축소, 이상 탐지, 데이터 생성 등 비지도 학습의 기반 기법으로 활용

오토인코더란?

오토인코더는 입력 데이터를 잠재 공간(latent space)으로 압축했다가 다시 원래 형태로 복원하는 과정에서 데이터의 분포를 학습하는 신경망 모델입니다. 입력과 출력이 동일하도록 학습시키기 때문에 별도의 정답 레이블이 필요 없는 비지도 학습 방식에 속합니다. 오토인코더 모델은 다양한 형태의 데이터의 특징, 패턴, 구조를 자연스럽게 파악하며, 노이즈를 줄이거나 숨겨진 표현을 찾는 데 효과적인 방식으로 활용됩니다.

오토인코더의 작동 방식

오토인코더는 인코더–잠재 공간–디코더의 3단계 구조로, 인코더는 입력에서 핵심 정보를 추출해 잠재 벡터라는 간결한 표현으로 압축하고 디코더는 이 벡터를 다시 원래 형태로 복원합니다. 학습은 원본과 복원된 결과의 차이를 최소화하는 방향으로 진행되며, 잘 학습된 모델은 주요 특징은 남기고 잡음이나 불필요한 요소는 자연스럽게 제거하는 경향을 보입니다. 이런 구조는 단순한 재현 능력을 넘어, 데이터 분포를 요약하는 잠재 표현을 학습하는 데 강점을 갖습니다. 또한 변분 오토인코더(VAE)처럼 잠재 공간을 확률적으로 구조화해 새로운 데이터를 생성하는 방식으로 확장되면서 생성형 모델 연구에서 주목받고 있습니다.

오토인코더의 활용

우선 차원 축소에 활용되어 고차원 데이터를 분석·시각화하기 쉽게 만들어 주며, 전통적인 PCA보다 유연한 비선형 표현을 제공합니다. 또한 정상 데이터의 구조를 먼저 학습한 뒤, 재구성 오류가 큰 샘플을 이상으로 판단하는 방식으로 이상 탐지에 널리 활용됩니다. 제조 공정 불량 탐지, 네트워크 보안 등에서 기존 규칙 기반 방식보다 높은 탐지율을 보이기도 합니다. 이미지, 음성 등에서 잡음을 줄이는 노이즈 제거에도 효과적이며, 잠재 공간을 조작해 새로운 이미지를 생성하거나 스타일을 바꾸는 등 생성형 작업에서도 사용됩니다.

관련 용어

인코더–디코더 구조 (Encoder–Decoder Architecture)

인코더–디코더 아키텍처는 입력을 압축해 내부 표현으로 만들고, 이를 기반으로 새로운 출력을 생성하는 신경망의 일반적 구조입니다. 오토인코더는 이 구조를 활용해 입력을 다시 복원하는 데 집중하는 반면, 인코더–디코더 구조 자체는 번역·요약·이미지 생성 등 입력과 출력이 달라지는 다양한 변환 작업에도 널리 쓰입니다. 즉, 오토인코더는 인코더–디코더 구조를 '입력 재구성'이라는 특정 목적에 맞춰 특화해 사용하는 형태입니다.

048 오픈소스 AI

Open-Source AI

AI 모델과 코드를 공개해 누구나 수정·활용할 수 있도록 한 구조

- AI의 알고리즘, 학습 데이터, 모델 매개변수, 코드 등을 공개해 누구나 접근·수정·배포할 수 있게 하는 개방형 AI 개발 방식
- 기술 협력과 투명성을 높이는 동시에, 보안·윤리 위험도 함께 존재

오픈소스 AI의 개념

오픈소스 AI는 AI의 핵심 구성 요소인 모델, 학습 코드, 데이터, 알고리즘 등을 공개해 누구나 자유롭게 사용·수정·재배포할 수 있도록 하는 개방형 개발 체계를 의미합니다. 기존의 상용 AI가 기업 내부에서 폐쇄적으로 관리되는 것과 달리, 오픈소스 AI는 연구자·기업·개발자가 공동으로 모델을 발전시키는 협력 기반 생태계를 지향합니다. Meta의 LLaMA, Stability AI의 Stable Diffusion 등이 대표적 예로, 이러한 개방적 구조는 기술 발전 속도를 높이고 특정 기업 중심의 기술 집중을 완화합니다.

오픈소스 AI의 확산

오픈소스 AI는 주로 코드 저장소(예, GitHub)나 모델 공유 플랫폼(예, Hugging Face)을 통해 배포됩니다. 개발자는 공개된 모델을 다운로드해 새로운 데이터로 재학습하거나 알고리즘을 수정해 자신만의 응용 모델을 만들 수 있으며, 다시 커뮤니티에 공유함으로써 순환적 발전 구조가 형성됩니다. 이런 방식은 AI 모델을 단일 제품이 아닌 공동 지식 자산으로 만들며, 오픈소스 생태계 특유의 피드백 문화가 지속적인 품질 개선을 이끕니다. 확산의 배경에는 대형 AI 기업이 독점하는 폐쇄형 생태계에 대한 견제와, AI 기술의 민주화 흐름이 있습니다. 클라우드 인프라와 GPU 자원의 확산, 학습 비용 절감, 정부의 데이터 개방 정책도 오픈소스 AI 성장의 기반이 되었습니다. 그 결과, 대학·스타트업·공공기관까지 참여할 수 있는 개방형 AI 혁신 구조가 전 세계적으로 빠르게 확산되고 있습니다.

오픈소스 AI의 위험

오픈소스 AI의 개방성은 기술 확산을 가속하지만 동시에 보안·윤리·법적 위험을 수반합니다. 모델의 가중치와 학습 데이터가 공개되면 악의적인 사용자가 이를 조작하거나 부적절한 콘텐츠를 생성할 수 있으며, 허위 정보·딥페이크·사이버 공격용 코드 등이 확산될 가능성도 있습니다. 또한 학습 데이터에 포함된 개인정보, 저작권 문제, 국가별 규제 차이 등으로 인해 법적 분쟁이 발생할 위험이 큽니다. 오픈소스 모델을 충분히 검증하지 않고 상용 서비스에 적용할 경우, 편향된 결과나 사회적 차별이 강화될 수 있습니다. 따라서 오픈소스 AI의 발전은 개방성과 책임성의 균형을 전제로 해야 하며, 신뢰 가능한 사용 지침과 글로벌 거버넌스 마련이 필수적입니다.

049 온디바이스 AI

On-Device AI

클라우드 대신 단말기 내부에서 직접 AI 연산을 수행하는 기술

- 스마트폰, IoT 기기 등 사용자 단말에서 AI 모델을 실행해 실시간 판단·예측·생성을 수행하는 기술
- 데이터를 외부로 전송하지 않아 속도와 보안성을 동시에 확보

온디바이스 AI의 개념

온디바이스 AI는 인공지능 모델이 클라우드 서버가 아닌 스마트폰·IoT 기기·웨어러블·차량 등 단말기 내부에서 직접 연산을 수행하는 기술을 의미합니다. 기존 AI 서비스가 데이터를 서버로 전송해 분석했다면, 온디바이스 AI는 이 과정을 기기 안에서 처리해 통신 지연을 최소화하고 개인정보 유출 위험을 줄입니다. 스마트폰의 음성 인식, 카메라 피사체 인식, 자율주행 보조 시스템, 웨어러블의 건강 분석 등에서 활용됩니다. 즉, 클라우드 중심의 '중앙집중형 AI'를 넘어, 기기 자체를 지능화된 연산 주체로 전환하는 기술입니다.

온디바이스 AI가 주목받는 이유

온디바이스 AI가 주목받는 이유는 실시간성·보안성·개인화를 모두 강화했기 때문입니다. 데이터가 기기 내부에서 처리되어 응답 속도가 빠르고, 네트워크가 불안정한 환경에서도 안정적으로 작동합니다. 외부 서버 전송이 필요 없어 개인정보 보호에도 유리하며, 사용자의 로컬 데이터를 기반으로 맞춤형 서비스를 구현할 수 있습니다. 이러한 특성 덕분에 제조사와 반도체 기업들은 NPU를 내장하고 모델을 기기에서 직접 실행하는 방향으로 전환하고 있습니다. 또한 경량화 모델과 저전력 연산 기술의 발전으로 작은 기기에서도 고성능 AI를 구현할 수 있게 되었습니다.

온디바이스 AI의 한계

온디바이스 AI는 단말기의 저장 용량과 연산 능력의 한계로 제약이 존재합니다. LLM이나 복잡한 딥러닝 모델은 기기에 직접 탑재하기 어렵고, 정기적인 업데이트가 필요합니다. 또한 기기 간 사양 차이로 동일한 모델이라도 성능 편차가 생기며, 최적화 수준에 따라 속도나 정확도가 달라질 수 있습니다. 제조사와 운영체제 간의 호환성 문제 역시 해결해야 할 과제입니다.

온디바이스 AI의 전망

온디바이스 AI는 AI의 탈중앙화와 개인화 시대를 여는 핵심 기술로 평가됩니다. 앞으로는 클라우드와 단말이 협력하는 하이브리드 AI 구조가 확산되며, 중앙 서버의 계산력과 기기 내부의 실시간 처리가 결합된 생태계가 구축될 것입니다. 또한 초저전력 반도체, 경량화언어모델, 연합학습 기술이 발전하면서 개인 데이터를 로컬에서 안전하게 학습·활용하는 프라이버시 중심의 AI 환경이 보편화될 것으로 전망됩니다.

관련 용어

에지 AI (Edge AI)

에지 AI는 데이터가 생성되는 지점(에지)에서 AI 연산을 수행하는 기술을 말합니다. 클라우드로 데이터를 전송하지 않고, 가까운 네트워크 단이나 기기에서 직접 분석해 응답 속도를 높이고 대역폭 부담을 줄입니다. 예를 들어 공장 센서, CCTV, 자율주행차 카메라 등에서 수집된 정보를 현장에서 즉시 판단하는 방식입니다. 이는 온디바이스 AI보다 범위가 넓은 개념으로, 개별 기기뿐 아니라 게이트웨이·로컬 서버 등 인접 장비까지 포함합니다. 에지 AI는 실시간성·보안성·네트워크 효율을 동시에 확보할 수 있어, 산업 자동화와 IoT 시대의 핵심 인프라로 주목받고 있습니다.

관련 용어

임베디드 AI (Embedded AI)

임베디드 AI는 AI 알고리즘을 기기 내부의 전자회로나 하드웨어에 직접 탑재해 작동시키는 형태를 의미합니다. 주로 마이크로컨트롤러(MCU)나 전용 칩셋에 AI 모델을 내장하여, 별도의 네트워크 연결 없이도 데이터 인식·분석이 가능합니다. 예를 들어 카메라가 자동으로 얼굴을 인식하거나 가전제품이 사용 패턴을 스스로 학습하는 기능이 이에 해당합니다. 임베디드 AI는 하드웨어에 최적화된 초경량 모델을 사용하기 때문에 연산 속도가 빠르고 전력 소비가 적습니다. 온디바이스 AI보다 하드웨어 종속성이 강하며, 제한된 환경에서도 작동하는 초소형·고효율 AI 기술로 평가됩니다.

관련 용어

클라우드 AI (Cloud AI)

클라우드 AI는 대규모 서버나 데이터센터에서 AI 모델을 구동하고, 네트워크를 통해 서비스를 제공하는 구조입니다. 기기에서 데이터를 수집해 중앙 서버로 전송하고, 거기서 연산·분석·추론을 수행한 뒤 결과를 다시 전달합니다. 고성능 GPU, LLM, 방대한 데이터가 필요한 AI 서비스는 대부분 이 구조를 기반으로 합니다. 클라우드 AI의 강점은 연산 능력과 확장성이 뛰어나다는 점이지만, 네트워크 지연과 개인정보 유출 가능성이 단점으로 꼽힙니다. 온디바이스·에지 AI가 이러한 한계를 보완하는 형태로 발전하고 있습니다.

관련 용어

하이브리드 AI (Hybrid AI)

하이브리드 AI는 클라우드와 온디바이스·에지 AI의 장점을 결합한 협력형 구조입니다. 중앙 서버가 복잡한 연산과 대규모 학습을 담당하고, 단말기나 에지 기기가 실시간 분석과 즉각적인 응답을 처리합니다. 예를 들어 스마트폰 음성 인식 같은 간단한 명령은 기기 내부에서 처리하고, 복잡한 질의는 클라우드로 전송해 고급 연산을 수행하는 식입니다. 이를 통해 속도·보안·정확성을 모두 확보할 수 있습니다. 하이브리드 AI는 AI의 분산 처리 구조를 완성하는 모델로, 향후 지능형 네트워크와 협력형 학습의 핵심 기술로 주목받고 있습니다.

050 월드 모델

World Model

AI가 환경의 규칙·변화를 내부적으로 학습·예측하도록 설계된 인지 구조

- AI가 외부 세계의 물리적·논리적 관계를 스스로 학습해, 환경을 내면화하고 물리적 현상을 시뮬레이션하려는 차세대 AI 학습 패러다임

월드 모델의 배경

월드모델은 인공지능이 현실 세계의 구조와 작동 원리를 스스로 학습해 내부적으로 세상의 모형을 형성하도록 하는 개념적 접근입니다. 단순히 데이터를 통계적으로 분석하는 단계를 넘어, 환경의 변화와 인과관계를 추론하며 행동의 결과를 예측하려는 시도에서 비롯되었습니다. 이는 인간이 세상의 규칙을 경험을 통해 학습하듯, AI가 관찰과 시뮬레이션을 통해 세계를 이해하려는 구조를 지향합니다. 2010년대 후반 이후 강화학습과 로보틱스 연구를 중심으로 이러한 아이디어가 활발히 탐구되었으며, AI가 상황을 예측하고 스스로 계획을 세울 수 있는 인지적 능력을 개발하는 방향으로 확장되고 있습니다.

월드 모델의 특징

월드 모델을 구축하는 핵심 목적은 환경 예측과 행동 계획을 위한 시뮬레이션된 환경을 모델에게 제공하는 것입니다. AI는 외부로부터 받은 데이터를 요약하고, 그 패턴을 시간적 맥락 속에서 시뮬레이션하여 행동 결과를 가상으로 예측합니다. 이를 통해 실제 환경에서 모든 시행착오를 겪지 않고도 내부 실험만으로 학습을 이어갈 수 있습니다. 연구 수준에서는 지각(perception), 예측(prediction), 계획(planning) 단계를 결합한 구조가 제안되어 왔으며, AI가 입력에 단순 반응하는 것이 아니라 '상상'에 가까운 내부 예측 과정을 거쳐 결정을 내리도록 설계됩니다. 이러한 방식은 AI가 경험을 단순 기억하는 수준을 넘어, 환경의 원리를 이해하려는 단계로 나아가는 시도로 평가됩니다.

월드 모델의 활용

월드 모델은 현재 로보틱스, 자율주행, 시뮬레이션 AI, 디지털 트윈 등에서 연구 중심으로 적용되고 있습니다. 로봇이 물리적 환경을 직접 실험하지 않고도 내부 모델을 통해 행동 결과를 예측하거나, 자율주행 시스템이 도로 상황을 시뮬레이션해 최적 경로를 계획하는 형태가 대표적입니다. 산업 영역에서는 가상의 공정 데이터를 학습시켜 생산 조건을 최적화하거나, 현실 시스템의 작동을 예측하는 디지털 트윈 구현에도 응용되고 있습니다. 아직은 이론적·실험적 단계에 머물지만, AI가 세상을 단순히 '분석'하는 존재에서 '이해하고 예측하는 지능'으로 확장되는 흐름의 중심 개념으로 평가됩니다.

051 이상 탐지

Anomaly Detection

데이터에서 비정상적 패턴을 자동 식별하는 기술

- AI가 방대한 데이터 속에서 일반적 패턴에서 벗어난 비정상 상태를 자동으로 탐지하는 기술
- 보안, 산업, 의료 등 다양한 영역에서 조기 경고와 품질 관리에 활용

이상 탐지란?

이상 탐지는 데이터나 시스템의 동작 과정에서 정상적인 패턴과 다른 예외적 변화를 자동으로 식별하는 기술을 말합니다. 사람이 인식하기 어려운 미세한 변화를 통계적 분석과 학습 알고리즘으로 찾아내어 오류, 침입, 고장, 이상 행위를 조기에 감지할 수 있습니다. 과거에는 평균·편차 등 통계적 지표를 활용한 단순 감시가 주류였으나, 데이터의 복잡성과 규모가 커지면서 AI 기반 탐지 방식이 빠르게 확산되었습니다. 특히 비정형 데이터나 시계열 정보에서도 이상 징후를 정밀하게 포착할 수 있어, 데이터 중심 환경에서의 자율적 모니터링 기술로 발전하고 있습니다.

AI 기반 이상 탐지 기술

AI 기반 이상 탐지는 머신러닝과 딥러닝 모델을 통해 정상과 비정상 패턴을 구분합니다. 지도학습은 정상·이상 데이터를 모두 라벨링해 학습하는 방식으로, 예측 정확도가 높지만 이상 데이터 확보가 어렵다는 한계가 있습니다. 반면 비지도학습은 정상 패턴만 학습하고 그와 거리가 먼 데이터를 이상으로 판단하는 방식으로, 실제 산업 데이터 분석에 자주 활용됩니다. 대표적으로 정상 데이터의 분포를 학습한 뒤, 그와 거리가 먼 샘플을 이상으로 판단하는 오토인코더(Autoencoder)가 있습니다. 최근에는 시계열 이상을 다루는 순환 신경망(RNN), 이미지 기반 이상 감지를 위한 합성곱 신경망(CNN), 그리고 희귀한 이상 패턴을 확률적으로 모델링하는 생성형 모델까지 적용되고 있습니다. 이러한 접근은 단순 감지에서 나아가, 이상이 발생하기 전의 징후를 예측하는 예방적 탐지로 발전하고 있습니다.

이상 탐지 기술의 활용

이상 탐지는 산업, 보안, 금융, 의료 등 다양한 영역에서 조기 이상 경보 시스템의 핵심 기술로 활용됩니다. 제조 현장에서는 센서 데이터를 분석해 설비 고장을 사전에 예측하고, 금융 분야에서는 거래 기록을 실시간으로 감시해 부정 결제나 사기 가능성을 탐지합니다. 사이버보안에서는 네트워크 트래픽의 비정상 패턴을 식별해 침입을 차단하고, 의료 분야에서는 생체 신호의 급격한 변화나 이상 패턴을 감지해 질병 조기 진단을 지원합니다. 또한 AI 시스템 내부의 오류 감지나 모델 편향 검출에도 응용되어, AI의 신뢰성과 안정성을 유지하는 기술적 기반으로 평가됩니다. 이상 탐지는 단순한 감시 기술을 넘어 AI 자율 운영(AIOps)과 예측 유지보수의 핵심 구성 요소로 확장되고 있습니다.

052 인과 AI

Causal AI

데이터 간 인과관계를 추론해 원인과 결과를 설명하는 AI

- 데이터의 상관관계가 아닌 원인과 결과의 인과적 구조를 학습·추론하는 AI 접근 방식
- AI 판단의 근거를 명확히 해 해석 가능성과 신뢰성을 높이는 것이 목적

인과 AI란?

인과 AI는 단순히 데이터 간 상관관계를 분석하는 기존 AI와 달리, 사건 간의 원인과 결과 관계를 파악해 '왜 그런 결과가 나왔는가'를 설명할 수 있는 AI를 말합니다. 예를 들어 "광고 클릭률이 높다"는 단순 상관관계를 넘어 "특정 요인이 클릭률 증가를 유발했는가"를 분석하는 것이 인과 AI의 목표입니다. 이는 통계학과 경제학 등에서 발전한 인과 추론(Causal Inference) 개념을 AI에 적용한 형태로, 데이터 기반 예측의 한계를 넘어 결정 과정의 구조적 이해를 가능하게 합니다. 인과 AI는 AI의 판단 과정을 투명하게 만들고, 복잡한 사회적·경제적 시스템에서 정책·의료·금융 등 의사결정의 근거를 설명할 수 있는 기술로 주목받고 있습니다.

인과 AI의 작동 원리

인과 AI는 데이터 속에서 단순히 함께 나타나는 현상을 찾는 것이 아니라, 무엇이 원인이고 무엇이 결과인지를 구분하려고 합니다. 예를 들어 "기온이 높을수록 아이스크림 판매가 증가한다"는 상관관계를 넘어서, 기온 상승이 실제로 판매 증가의 원인인지를 검증합니다. 이를 위해 AI는 여러 변수가 서로에게 어떤 영향을 미치는지 관계망을 만들고, 한 요소를 바꿨을 때 다른 결과가 어떻게 달라지는지를 가상으로 실험합니다. 이렇게 AI가 "만약 다른 선택을 했다면 어떤 결과가 나왔을까?"를 스스로 비교해보는 과정을 통해, 결과의 진짜 원인을 찾아냅니다. 이런 방식은 계산은 복잡하지만, 결과의 이유를 명확히 설명할 수 있어 신뢰도가 높습니다. 최근에는 이런 접근이 딥러닝 구조와 결합되어, AI 판단의 근거를 더 투명하게 보여주는 기술로 발전하고 있습니다.

인과 AI의 활용

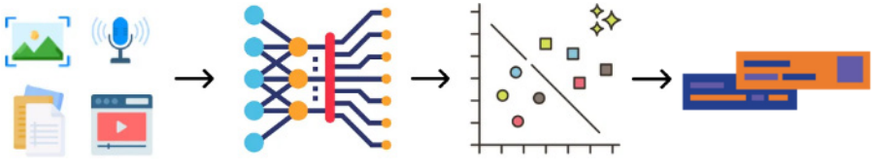
인과 AI는 정책, 의료, 금융, 산업 제어, AI 윤리 등 다양한 영역에서 활용됩니다. 예를 들어 의료 분야에서는 특정 치료가 환자 회복에 실제로 영향을 주는지를 인과적으로 분석하고, 금융에서는 고객 행동이나 리스크 요인을 설명 가능한 방식으로 평가할 수 있습니다. 또한 정책 결정에서는 사회 변수 간의 인과 구조를 모델링하여, 정책 변화가 경제나 고용에 미치는 파급 효과를 예측하는 데 활용됩니다. AI 내부적으로는 모델의 예측 근거를 명확히 해 '설명 가능한 AI(XAI)'의 한계를 보완하는 기술로 주목받고 있습니다. 인과 AI는 결과 중심의 AI에서 원인 중심의 AI로의 전환을 이끄는 흐름으로, AI의 해석 가능성과 책임성을 강화하고 신뢰 기반의 AI 거버넌스 구축에도 기여할 것으로 평가됩니다.

053 임베딩
Embedding

비정형 데이터를 수치 벡터로 변환해 의미를 표현하는 기술
- 텍스트·이미지 같은 복잡한 데이터를 숫자로 바꾸어 의미 관계를 계산할 수 있도록 하는 표현 방식
- AI가 단어·개념의 유사성을 이해하고 추론하는 데 사용

임베딩의 개념

임베딩(Embedding)은 텍스트·이미지와 같은 비정형 데이터를 의미 정보를 보존한 채 고차원 벡터 공간으로 변환하는 기술입니다. 이 벡터는 단순한 숫자 변환이 아니라 데이터의 맥락과 의미적 특징을 압축적으로 표현한 형태입니다. 단어나 문장, 이미지와 같은 비정형 데이터는 그대로는 연산이 불가능하기 때문에, 이를 일정한 규칙에 따라 벡터 형태로 변환합니다. 각 벡터는 데이터의 의미적 특징을 반영하고 있어, AI는 이 벡터 간의 거리와 방향을 비교해 유사도나 관계를 계산할 수 있습니다. 예를 들어 '고양이'와 '개'는 서로 가까운 위치에, '고양이'와 '자동차'는 멀리 떨어진 위치에 놓여 의미적 차이를 표현합니다. 이러한 수치화 과정을 통해 AI는 언어나 이미지의 맥락과 의미를 파악할 수 있게 됩니다.



임베딩의 필요성

AI는 단순히 단어의 모양이나 순서만으로는 문맥을 이해하기 어렵습니다. 같은 단어라도 사용되는 상황에 따라 의미가 달라지기 때문입니다. 임베딩은 이러한 한계를 해결해, 데이터의 '의미적 정보'를 보존하면서 수학적으로 처리할 수 있도록 하는 중간 표현층의 역할을 합니다. 이를 통해 AI는 문장의 흐름을 해석하고, 유사한 개념을 묶거나 다른 의미를 구분할 수 있습니다. 예를 들어 "배"가 '신체의 일부'인지 '운송 수단'인지, '과일'인지 구별하려면, 주변 단어와의 관계를 고려해야 하는데 임베딩은 바로 이 문맥 정보를 반영합니다. 또한 텍스트뿐 아니라 이미지·음성 등 다양한 데이터를 동일한 벡터 공간으로 표현할 수 있어, 서로 다른 정보 간 연관성을 분석하는 멀티모달 AI 구현의 기반이 됩니다. 임베딩은 결국 AI가 '이해'와 '추론'을 가능하게 하는 핵심 표현 기술로, 현대 AI 모델의 필수 구성 요소로 자리 잡고 있습니다.

054 자동화된 머신러닝 / AutoML
Automated Machine Learning

AI가 머신러닝 모델 개발 과정을 자동으로 수행·최적화하는 기술
- 데이터 전처리부터 알고리즘 선택, 하이퍼파라미터 조정까지 머신러닝의 복잡한 과정을 자동화해 효율을 높이는 기술
- 비전문가도 AI 모델을 손쉽게 구축·활용할 수 있도록 한 기술

자동화된 머신러닝의 개념

자동화된 머신러닝(AutoML)은 데이터 분석과 모델 설계의 전 과정을 AI가 대신 수행해주는 기술입니다. 기존 머신러닝은 데이터 정제, 알고리즘 선택, 매개변수 조정, 성능 검증 등 다단계 과정을 전문가가 직접 설계해야 했습니다. AutoML은 이 과정을 자동화하여 모델 생성의 효율성과 접근성을 동시에 향상시킵니다. 사용자는 단지 데이터를 입력하고 목적을 설정하면, 시스템이 가장 적합한 알고리즘과 하이퍼파라미터를 찾아 스스로 모델을 완성합니다. 이렇게 생성된 모델은 반복적인 학습과 평가 과정을 거치며 점차 성능이 개선됩니다. 즉, AutoML은 사람의 경험과 시행착오를 대체하는 AI 주도형 모델 개발 기술이라 할 수 있습니다.

자동화된 머신러닝의 유형

AutoML의 핵심은 탐색과 최적화 과정입니다. 시스템은 다양한 모델 구조와 매개변수를 자동으로 실험해 결과를 비교하고, 가장 높은 성능을 내는 조합을 선택합니다. 이 과정은 강화학습, 유전 알고리즘, 베이지안 최적화 등으로 구현되며, 각 시도 결과를 피드백 받아 다음 단계의 탐색 방향을 조정합니다. 또한 데이터 전처리, 특징 추출, 모델 평가 등 개별 단계도 자동화되어 전체 파이프라인이 유기적으로 작동합니다. 최근에는 대형 언어모델을 활용해 코드 생성과 모델 튜닝을 자동으로 수행하거나, 여러 AI가 협력해 학습을 관리하는 메타 학습 구조로 확장되고 있습니다. 이러한 발전은 AI가 단순히 데이터를 분석하는 도구를 넘어, 스스로 학습 환경을 설계하고 개선하는 주체로 진화하고 있음을 보여줍니다.

자동화된 머신러닝의 의의

AutoML은 완전한 자동화 기술로 오해되기 쉽지만, 여전히 여러 한계가 존재합니다. 탐색 과정이 무작위적이거나 과적합을 유발할 수 있으며, 자동으로 생성된 모델의 내부 구조를 사람이 이해하기 어려워 설명 가능성이 떨어집니다. 또한 데이터 품질이나 문제 정의가 불명확할 경우, AI가 잘못된 기준으로 모델을 최적화할 위험도 있습니다. 이러한 한계에도 불구하고 AutoML은 AI 개발의 장벽을 낮춘 기술로 평가됩니다. 데이터 과학 지식이 부족한 개인이나 기업도 손쉽게 AI 모델을 구축할 수 있게 되었으며, 전문가에게는 반복적인 실험 부담을 줄여 효율성을 높여 줍니다. 나아가 AI가 스스로 모델을 설계하고 조정하는 기반을 마련함으로써, AI가 AI를 개발하는 자율형 학습 체계로 발전하는 전환점으로 평가됩니다.

055 자연어 처리

Natural Language Processing, NLP

AI가 인간의 언어를 이해하고 생성하도록 하는 기술

- 사람이 사용하는 언어를 컴퓨터가 인식·이해·생성할 수 있도록 하는 AI 기술
- 텍스트와 음성을 분석해 의미를 파악하고 대화, 번역, 요약 등을 수행

자연어 처리의 개념

자연어 처리는 인간의 언어를 컴퓨터가 이해하고 활용할 수 있도록 만드는 AI 기술을 말합니다. 사람의 언어는 모호함, 문맥, 감정 등 복잡한 요소를 지니기 때문에 단순한 규칙만으로는 분석이 어렵습니다. 자연어 처리는 단어나 문장의 의미를 통계적·수학적으로 표현해 AI가 문맥과 의도를 해석하도록 합니다. 예를 들어 사용자의 질문을 분석해 답변을 제공하거나, 한 언어를 다른 언어로 번역하는 과정이 모두 자연어 처리의 결과입니다. 초기에는 규칙 기반 분석에 의존했으나, 머신러닝과 딥러닝의 발전으로 AI가 방대한 텍스트를 학습해 언어의 패턴을 스스로 익히고 문맥까지 이해하는 수준으로 발전했습니다.

출처 : freepik

자연어 처리의 유형

자연어 처리의 접근 방식은 크게 규칙 기반, 통계 기반, 딥러닝 기반으로 발전했습니다. 초기의 규칙 기반 처리는 언어학자가 직접 문법과 어휘 규칙을 정의해 문장을 분석하는 방식으로, 단순하지만 문맥 변화에 취약했습니다. 이후 등장한 통계 기반 처리는 방대한 말뭉치를 분석해 단어 간 연관성과 등장 확률을 계산함으로써 보다 유연한 언어 이해를 가능하게 했습니다. 최근의 딥러닝 기반 처리는 인공신경망을 활용해 문맥과 의미를 동시에 학습하며, 임베딩과 트랜스포머 구조를 통해 자연스러운 언어 생성과 추론이 가능한 수준으로 발전했습니다.

자연어 처리의 활용

자연어 처리는 언어 이해·정보 검색·대화 시스템 등 AI 서비스의 핵심 기반으로 활용됩니다. 챗봇과 음성 비서는 사용자의 발화를 분석해 자연스러운 대화를 수행하고, 자동 번역·요약·감정 분석 기술은 방대한 텍스트를 빠르게 처리해 의미를 추출합니다. 산업 현장에서는 법률·의료·금융·행정·교육 분야에서 문서 분석, 보고서 요약, 리스크 평가, 자동 채점 등 다양한 형태로 적용됩니다. 최근에는 LLM의 발전으로 콘텐츠 생성, 정책 분석, 지식 탐색 등 창의적 언어 작업까지 가능해지면서, NLP는 단순한 언어 이해 기술을 넘어 인간의 사고와 소통을 보조하는 지능형 언어 플랫폼으로 확장되고 있습니다.

참조 용어 | 미세조정 p.42

056 저랭크 적응 / LoRA

Low-Rank Adaptation

가중치 전체가 아닌 저차원 행렬만 조정하는 효율적 미세조정 기법

- 기존 모델의 가중치는 고정한 채, 추가된 저랭크 행렬만 학습해 계산량과 메모리 사용을 줄이는 경량 학습 방식
- 대규모 모델의 성능은 유지하면서 빠르고 저비용으로 특정 작업에 맞게 적응시키는 미세조정 기법

LoRA란?

LoRA는 LLM이나 이미지 생성 모델을 특정 목적에 맞게 조정할 때, 전체 매개변수를 학습하지 않고 일부만 효율적으로 조정하는 경량 미세조정 기술입니다. 기존의 미세조정(Fine-tuning)은 모델의 모든 가중치를 업데이트해야 하므로, 막대한 GPU 메모리와 연산 자원이 필요했습니다. 반면 LoRA는 학습 효율을 극대화하기 위해 모델의 가중치 행렬을 저차원(Low-Rank) 형태로 분해하고, 이 중 추가된 보조 행렬만 학습합니다. 그러면 모델의 주요 구조를 유지하면서도 학습해야 할 가중치 수를 크게 줄일 수 있습니다.

LoRA의 작동 원리

LoRA의 핵심은 모델 전체를 바꾸지 않고, 필요한 부분만 조정하는 것입니다. 대형 AI 모델은 수십억 개의 가중치를 가지고 있지만, 실제로 특정 작업을 새로 학습할 때는 그중 일부만 변화가 필요합니다. LoRA는 이 점에 착안해, 기존 모델의 가중치는 그대로 두고, 아주 작은 보조 구조만 추가해 그 부분만 학습합니다. 예를 들어, LoRA는 모델의 큰 가중치 행렬을 그대로 두고, 가중치 변화량만 계산하는 별도의 작은 행렬 경로를 추가합니다. 이 경로는 두 개의 작은 행렬로 구성되며, 학습 과정에서는 이 부분만 업데이트 됩니다.

이렇게 하면 전체 모델을 다시 훈련하지 않아도 되기 때문에, 메모리 사용량과 학습 속도를 크게 줄일 수 있습니다. 학습이 끝나면 보조 행렬이 만들어낸 조정 결과만 저장하고, 원래 모델에 덧붙여 사용할 수 있습니다. 즉, 하나의 기본 모델을 유지한 채로 여러 LoRA 모듈을 만들어, 필요할 때마다 주제·언어·스타일에 맞게 교체할 수 있습니다.

LoRA의 활용

LoRA는 초거대 AI 모델의 맞춤형 활용을 가능하게 한 핵심 기술로 평가됩니다. 기존 미세조정 대비 학습 매개변수 수를 수백분의 1로 줄이면서도 성능 저하가 거의 없어, 저비용·고효율 모델 커스터마이징이 가능합니다. 이미지 생성, LLM, 음성합성 등에서 폭넓게 활용되며, 하나의 모델에 여러 LoRA를 조합하는 모듈형 학습 방식으로 확장되고 있습니다. 특히 공공·산업 부문에서는 LoRA를 통해 AI 모델을 특정 업무 환경이나 언어, 정책 도메인에 맞게 빠르게 적응시킬 수 있어, AI 확산의 현실적 대안으로 주목받고 있습니다.

057 정확도

Accuracy

AI 모델의 예측이 실제 정답과 얼마나 일치하는지 측정하는 지표

- AI나 머신러닝 모델이 전체 예측 중 올바르게 판단한 비율을 계산해 성능을 평가하는 기본 지표
- 데이터 불균형 상황에서는 한계가 있어서 다른 지표와 함께 종합적으로 사용

정확도란?

정확도는 AI나 머신러닝 모델이 예측한 결과가 실제 정답과 얼마나 일치하는지를 나타내는 대표적인 평가 지표입니다. 전체 예측 중 올바르게 판단한 비율을 계산해 모델의 신뢰성과 학습 성과를 간단히 비교할 수 있습니다. 예를 들어 100개의 데이터를 예측했을 때 90개가 맞았다면 정확도는 90%입니다. 이러한 단순성과 명확성 덕분에 분류 문제에서 가장 널리 사용됩니다. 정확도는 모델이 학습한 규칙이 실제 데이터를 얼마나 잘 반영했는지를 보여주며, 알고리즘 개선이나 하이퍼파라미터 조정의 기준으로 활용됩니다. 하지만 높은 정확도가 항상 좋은 모델을 의미하지는 않습니다. 데이터가 한쪽으로 치우친 불균형 데이터셋의 경우, 특정 집단을 제대로 예측하지 못해도 전체 정확도가 높게 나올 수 있습니다.

정확도 계산 방법

정확도는 (정답으로 맞힌 데이터 수) ÷ (전체 데이터 수)로 계산하며, '참으로 예측한 경우(True Positive와 True Negative)'를 모두 더한 값을 전체 데이터로 나눈 결과입니다. 계산식은 단순하지만, 분포가 한쪽으로 치우친 상황에서는 모델 성능을 과대평가할 위험이 있습니다. 예를 들어 전체 데이터의 95%가 '정상'이고 5%만 '이상'인 경우, 모든 데이터를 정상으로 예측해도 정확도는 95%로 높게 계산되지만 실제로는 이상 탐지가 전혀 이뤄지지 않습니다. 이런 이유로 의료 진단, 보안 탐지처럼 중요한 예외 사례가 많은 분야에서는 정밀도, 재현율, F1 점수 등 다른 보조 지표를 함께 사용해 모델을 종합적으로 평가합니다. 정확도는 그럼에도 여전히 AI의 기본 성능을 직관적으로 파악할 수 있는 핵심 지표로 가장 널리 활용되고 있습니다.

관련 용어

정밀도 (Precision)

정밀도는 모델이 '맞다고 예측한 것' 중 실제로 맞은 비율을 의미하는 성능 지표입니다. 예측 결과 중 얼마나 정확하게 긍정 사례를 찾아냈는지를 보여줍니다. 예를 들어 스팸메일 탐지에서 정밀도가 높다는 것은, 스팸으로 분류된 메일이 실제로 스팸일 확률이 높다는 뜻입니다. 잘못된 긍정 예측(False Positive, 오탐)을 최소화하는 데 도움이 되기 때문에 금융 사기 탐지나 의료 진단처럼 오탐이 큰 피해를 일으키는 분야에서 중요하게 사용됩니다. 높은 정밀도는 모델이 신중하게 판단한다는 것을 의미하지만, 그만큼 일부 실제 사례를 놓칠 가능성도 있어 재현율과 함께 해석해야 합니다.

관련 용어

재현율 (Recall)

재현율은 실제로 맞는 정답 중에서 모델이 얼마나 많이 찾아냈는지를 나타내는 비율입니다. 즉, 놓치는 사례(False Negative)를 얼마나 줄였는지를 평가하는 지표입니다. 예를 들어 스팸메일 탐지에서 재현율이 높다는 것은, 대부분의 스팸을 성공적으로 잡아냈다는 의미입니다. 재현율은 모델이 가능한 한 많은 긍정 사례를 포착해야 하는 보안 탐지나 질병 진단과 같은 영역에서 특히 중요합니다. 다만 재현율이 높으면 불필요한 긍정 판단이 늘어날 수 있어, 정밀도와 균형 있게 해석해야 합니다. 두 지표는 보통 F1 점수(F1-score)로 함께 평가되어 모델의 종합적 성능을 판단하는 데 사용됩니다.

정확도 vs 정밀도 vs 재현율

이 세 지표는 모두 모델의 분류 성능을 측정하지만, 평가 관점이 서로 다릅니다. 정확도는 전체 예측 중 맞춘 비율을, 정밀도는 모델이 '긍정'이라고 판단한 결과 중 실제로 맞은 비율을, 재현율은 실제 긍정 사례 중 모델이 맞게 찾아낸 비율을 의미합니다. 예를 들어 스팸메일 탐지에서 정밀도가 높으면 정상 메일을 스팸으로 잘못 분류하지 않지만, 재현율이 높으면 대부분의 스팸을 놓치지 않고 잡아냅니다. 정밀도와 재현율은 상충관계에 있어, 둘 중 하나만 높이는 것은 어렵습니다. 따라서 모델의 목적에 따라 어떤 지표를 우선시할지가 달라집니다. 의료 진단처럼 놓치면 안 되는 경우에는 재현율을, 금융 사기 탐지처럼 오탐이 문제인 경우에는 정밀도를 중시합니다.

		실제 정답	
		True	False
분류 결과	True	옳은 긍정 예측(A) (True Positive)	잘못된 긍정 예측 (False Positive) 크면 재현율 ↓
	False	잘못된 부정 예측 (False Negative) 크면 정밀도 ↓	옳은 부정 예측(B) (True Negative)

A + B = 정확도

관련 용어

F1 점수 (F1-score)

F1 점수는 정밀도와 재현율의 조화를 수치로 표현한 성능 지표로, 두 지표의 조화평균을 사용해 한쪽으로 치우치지 않는 균형 잡힌 평가를 제공합니다. 예를 들어 정밀도는 높지만 재현율이 낮거나, 그 반대인 경우 모두 F1 점수가 낮게 계산되어 모델의 전반적 신뢰성을 판단할 수 있습니다. 이 지표는 특히 불균형 데이터셋에서 정확도만으로는 성능을 평가하기 어려울 때 유용합니다. F1 점수가 높을수록 모델이 긍정 사례를 정확하고 폭넓게 탐지한다는 의미로, 의료 진단, 스팸 탐지 등 오탐·누락이 중요한 분야에서 널리 사용됩니다.

058 제로샷 러닝

Zero-shot learning

예시 없이도 새로운 문제를 해결하는 AI의 일반화 능력

- 사용자가 예시나 틀을 제시하지 않고 자연어로만 일을 지시해도, AI가 의도를 추론해 새 과제를 수행하도록 만드는 상호작용 방식
- 거대 언어모델의 범용성과 작업 확장성을 보여주는 핵심 특징

제로샷 러닝이란?

제로샷 러닝은 AI가 학습 과정에서 접하지 않은 과제를, 추가 예시 없이도 처리하는 능력을 의미합니다. 모델이 사전학습을 통해 익힌 언어·지식·추론 패턴을 활용해 새로운 작업 방식까지 일반화하는 구조입니다. 생성형 AI에서는 프롬프트에 작업 지시만 주어도 분류·요약·추출·해석 같은 업무를 바로 수행할 수 있어, 별도 예시나 규칙을 제공하지 않아도 되는 점이 특징입니다. 이 능력은 사용자가 별도 학습 과정 없이도 다양한 작업을 즉시 실험·활용할 수 있게 하며 실무 적용 범위를 크게 넓혔습니다.

제로샷 러닝의 작동 원리

제로샷 러닝은 모델이 사전학습된 의미 구조와 패턴 일반화 능력을 기반으로 작동합니다. LLM은 방대한 텍스트에서 다양한 문제 유형을 관찰한 경험을 축적하기 때문에, 처음 보는 형태의 요청도 비슷한 구조를 찾아 대응할 수 있습니다. 예를 들어 문장 변환 등은 명시적으로 학습하지 않았더라도 기존 지식과 언어 패턴을 활용해 자연스럽게 처리합니다. 다만 지시 표현에 민감해 정렬되지 않은 결과가 제시될 수 있으며, 과잉 일반화로 인해 부정확하거나 단정적 답변을 내는 경우도 있어 모델 특성을 이해한 프롬프트가 중요합니다.

제로샷 러닝의 활용

제로샷 러닝은 추가 데이터나 미세조정 없이도 다양한 작업을 즉시 수행할 수 있다는 점에서 실무 활용성이 높습니다. 기업이나 조직은 별도 데이터 구축 비용 없이 즉시 AI를 적용할 수 있어 도입 장벽이 낮아지고, 새로운 업무 실험 속도도 크게 빨라집니다. 다만 지시만으로 처리하는 방식인 만큼 안정성과 일관성이 항상 보장되지는 않아, 복잡한 작업에서는 예시나 추가 가이드를 제공하는 것이 더 적절한 경우가 많습니다.

관련 용어

퓨샷 (Few-shot)

퓨샷은 제로샷과 달리 소량의 예시를 함께 제시해 모델의 작업 기준을 명확하게 만드는 방식입니다. 몇 개의 사례만 제공해도 모델은 원하는 출력 형식과 판단 기준을 더 정확히 이해할 수 있어, 제로샷보다 안정적이고 일관된 결과를 내는 경우가 많습니다. 대규모 데이터가 필요한 미세조정보다 비용이 적고, 지시만 주는 제로샷보다 통제력이 높아 실제 업무 자동화에서 널리 활용됩니다.

059 지능형 기지국 / AI-RAN

AI-Radio Access Network

RAN에 AI를 적용하여 네트워크를 지능적으로 제어·최적화하는 기술

- AI가 기지국과 단말 간 데이터 흐름을 분석해 통신 품질과 효율을 높이는 지능형 네트워크 관리 구조
- 차세대 이동통신(5G·6G)의 자율 운영을 지원하며, 네트워크 성능과 에너지 효율을 극대화하는 핵심 인프라 기술

AI-RAN이란?

AI-RAN은 무선접속망(RAN)에 AI의 학습·예측·제어 능력을 결합한 지능형 네트워크 기술입니다. 기존 RAN이 사전 설정에 따라 수동으로 작동했다면, AI-RAN은 트래픽과 전파 환경을 실시간 분석해 자율적으로 제어합니다. 품질 저하나 장애를 감지하면 즉시 매개변수를 조정하고, 전력·주파수를 효율적으로 배분해 통신 품질을 안정적으로 유지합니다. 또한 트래픽 급증 시 부하를 분산해 혼잡을 완화하고 에너지와 운영비를 절감함으로써, 사람이 직접 조정하지 않아도 스스로 최적화되는 지능형 자율 네트워크를 구현합니다. AI-RAN은 5G 고도화와 6G로의 전환 과정에서 AI 중심 네트워크 혁신의 핵심 기술로 자리 잡고 있습니다. 기존 네트워크가 하드웨어 기반의 정적 구조였다면, 향후 통신망은 소프트웨어(SW) 중심으로 전환되며 AI가 핵심 제어 기능을 담당하게 됩니다. 즉, 네트워크가 스스로 학습하고 예측하며 최적화하는 지능형 자율망(Self-Optimizing Network)으로 진화하는 것입니다.

AI 적용 수준에 따른 분류

AI-RAN은 AI가 무선 접속망에 더욱 깊게 적용되며 발전합니다. 초기 단계의 적용 수준인 AI for RAN은 RAN 자체의 성능을 AI를 통해 향상시키며 AI가 전파 자원 관리 같은 핵심 역할을 직접 수행해 품질 저하를 미리 예측하고 최적화합니다. 그 다음 발전 단계인 AI and RAN은 동일 인프라 내에서 워크로드를 공유하는 융합 구조로, AI가 네트워크(RAN) 밖에서 분석과 조언을 제공합니다. 이후 발전 단계인 AI on RAN에서는 AI 모델이 기지국 장비에 직접 탑재되어 실시간 트래픽 변화에 즉각 반응하며 초저지연 서비스 품질을 보완합니다. 마지막 AI Native 단계는 무선 기지국뿐 아니라 통신망의 중앙 장비와 운영 시스템 전체가 AI 중심으로 설계되어, 네트워크가 스스로 운영·조정되는 완전 자율형 환경이 구현됩니다.

출처 : AI 네트워크로의 패러다임 전환과 정책 방향 (NIA)

060 지능형 사물인터넷 / AIoT

AI of Things

AI와 IoT를 결합해 자율적 판단이 가능한 지능형 연결 기술

- IoT가 수집한 데이터를 AI가 실시간 분석·판단해 자율적으로 제어하는 차세대 융합 기술
- 센서 네트워크를 넘어 상황 인식과 예측 기능을 갖춘 지능형 인프라로 발전한 연결 기술

AIoT란

AIoT는 사물인터넷(IoT)의 연결성과 데이터 수집 능력, 그리고 AI의 분석·학습·판단 능력을 결합한 지능형 융합 기술입니다. IoT가 사물 간 연결을 통해 데이터를 모은다면, AI는 그 데이터를 해석하고 의미를 부여해 스스로 결정을 내립니다. 이로써 단순히 정보를 전달하는 IoT에서 벗어나, 상황을 인식하고 즉각적으로 대응하는 자율형 시스템으로 진화했습니다. 예를 들어 스마트 팩토리에서는 AI가 설비 데이터를 분석해 고장을 예측하고, 스마트 시티에서는 교통·에너지 데이터를 통합 분석해 효율적인 도시 운영을 지원합니다.

AIoT의 핵심 기술

AIoT는 대규모 연결 환경에서 실시간 판단을 가능하게 하기 위해 여러 핵심 기술이 통합되어 작동합니다. 에지 컴퓨팅은 데이터를 현장에서 처리해 지연을 줄이고 보안을 강화하며, 에지 AI는 경량화된 AI 모델을 기기 내부에 탑재해 빠른 추론과 자율 제어를 수행합니다. 5G·6G 네트워크는 초저지연·초고속 통신을 통해 수많은 기기 간 데이터를 실시간 교환하게 하고, MLOps 기술은 AI 모델의 학습·검증·배포를 자동화해 지속적 성능 향상을 지원합니다. 여기에 보안·프라이버시 보호 기술이 결합되어 AIoT가 지능적이고 안전한 자율형 생태계로 작동할 수 있도록 기반을 제공합니다.

AIoT의 활용

AIoT의 가장 큰 장점은 지능형 자동화와 실시간 대응 능력입니다. AI가 IoT 기기의 데이터를 즉시 분석해 스스로 판단하므로, 시스템 전체의 효율성과 속도가 향상됩니다. 또한 에지 처리 기반 구조를 통해 네트워크 부하를 줄이고 보안성을 높일 수 있으며, 사람의 개입 없이 예측·제어가 가능한 자율 운영 환경을 구현합니다.

이 덕분에 AIoT는 산업, 도시, 생활 전반으로 확산되고 있습니다. 스마트 팩토리에서는 설비 이상을 예측하고 생산 공정을 최적화하며, 스마트 시티에서는 교통 흐름과 에너지 소비를 실시간 관리합니다. 헬스케어 분야에서는 웨어러블 기기가 생체 데이터를 분석해 맞춤형 건강 관리를 제공하고, 물류·농업·환경 관리 등에서도 AIoT가 효율성과 지속 가능성을 높이는 핵심 기술로 활용되고 있습니다.

061 지도학습

Supervised Learning

입력 데이터와 정답을 함께 학습해 예측 모델을 만드는 AI 학습 방식

- AI가 정답이 표시된 데이터를 기반으로 입력과 출력의 관계를 학습하는 방식
- 분류·예측 등 명확한 목표가 있는 문제 해결에 사용

지도학습이란?

지도학습은 AI가 정답이 표시된 데이터(레이블)를 기반으로 입력과 출력의 관계를 학습하는 방식입니다. 사람이 미리 정해준 기준에 따라 데이터를 분류하거나 결과를 예측할 수 있도록 훈련되는 구조로, 가장 기본적이고 널리 활용되는 학습 형태입니다. 예를 들어 '고양이'·'개'로 라벨링된 이미지를 학습한 모델은 새로운 사진이 주어졌을 때 어떤 동물인지 스스로 판별할 수 있습니다. 이처럼 지도학습은 입력값과 정답의 짝을 반복적으로 학습하며, 규칙이나 패턴을 일반화해 새로운 데이터에 대한 예측 능력을 키웁니다. 스팸 분류, 음성 인식, 질병 예측, 신용평가 등 정확한 기준이 존재하는 문제에 특히 효과적입니다. 다만 대량의 정답 데이터가 필요하고, 그 품질에 따라 성능이 좌우된다는 점에서 데이터 구축 비용이 크다는 한계가 있습니다.

관련 용어

비지도학습 (Unsupervised Learning)

비지도학습은 정답(레이블)이 없는 데이터를 기반으로 AI가 스스로 패턴이나 구조를 찾아내는 학습 방식입니다. 입력 데이터 간의 유사성, 분포, 군집 관계를 분석해 숨겨진 규칙을 발견합니다. 예를 들어 고객 데이터를 분석해 자연스럽게 구매 성향이 비슷한 집단을 묶는 클러스터링이나, 데이터의 차원을 줄이는 차원 축소 기술이 이에 해당합니다. 비지도학습은 라벨링 비용이 들지 않아 대규모 데이터 분석에 적합하지만, 결과를 해석하기 어렵고 명확한 정답(label)이 존재하지 않는다는 한계가 있습니다.

관련 용어

자기지도학습 (Self-Supervised Learning)

자기지도학습은 정답 레이블 없이도 모델이 스스로 학습 신호를 만들어내는 방식입니다. 데이터의 일부를 가리고 나머지 정보로 이를 예측하도록 훈련하는 등, 입력 데이터 자체에서 학습 과제를 생성해 모델이 패턴을 익히는 구조입니다. 예를 들어 문장에서 특정 단어를 가리고 이를 맞히게 하거나, 이미지 일부를 숨기고 원래 모습을 복원하게 하는 방식이 대표적입니다. 별도의 라벨링 비용 없이 대규모 데이터에서 표현 학습 능력을 키울 수 있어, 최근의 LLM과 비전·멀티모달 모델의 핵심 학습 기법으로 널리 활용되고 있습니다.

062 지식 증류

Knowledge Distillation

큰 모델의 지식을 작은 모델로 전달해 성능을 유지하는 기법

- 대규모 모델의 예측 과정과 정보 구조를 간추려 작은 모델이 학습하도록 하는 경량화 방식
- 작은 모델이 복잡한 모델의 판단 패턴을 모방해 효율적 성능을 내도록 만드는 기술

지식 증류의 개념

지식 증류는 대규모 모델(교사 모델)의 지식·표현·추론 패턴을 작은 모델(학생 모델)에 압축해 전달하는 모델 압축 기법입니다. 대규모 모델은 방대한 데이터와 연산을 통해 복잡한 판단 구조를 형성하지만, 이를 그대로 서비스 환경에 적용하기에는 비용과 자원 요구량이 매우 큽니다. 지식 증류는 이 문제를 해결하기 위해 교사 모델이 가진 정보를 간소화해 학생 모델이 학습할 수 있도록 만들며, 성능과 효율 간 균형을 맞춥니다. 일반적인 학습이 정답 레이블만을 활용하는 것과 달리, 증류는 교사 모델이 정답에 이르는 과정에서 생성하는 확률 분포나 중간 표현까지 활용해 더 풍부한 신호를 제공합니다. 이를 통해 학생 모델은 훨씬 작은 규모임에도 교사 모델의 패턴을 모방해 안정적인 성능을 유지할 수 있습니다.

지식 증류의 작동 방식

지식 증류는 교사 모델이 출력한 정보를 학생 모델의 학습 신호로 사용하는 방식으로 작동합니다. 교사 모델은 입력에 대해 단순 정답뿐 아니라, 여러 선택지에 부여한 확률값, 토큰 간 관계, 특징 표현 등 다양한 형태의 정보를 제공합니다. 학생 모델은 이러한 신호를 통해 "교사가 어떻게 판단하는지"를 학습하며 데이터의 구조와 의미를 자연스럽게 이해하게 됩니다. 또한 일부 설정에서는 소프트 타깃(soft target)을 활용하는데, 이는 정답과 오답을 1과 0으로 구분하는 방식이 아니라 각 선택지의 가능성을 확률로 표현한 값으로, 교사 모델의 판단 경향·유사도·강약 관계까지 반영된 정보입니다. 이를 통해 학생 모델은 적은 매개변수로도 높은 일반화 성능을 확보할 수 있으며, 제한된 자원 환경에서도 대규모 모델에 가까운 성능을 제공합니다. 지식 증류는 단일 모델 축소뿐 아니라 특정 도메인에 맞는 경량 모델을 빠르게 만들 때에도 널리 활용됩니다.

관련 용어

양자화 (Quantization)

양자화 또한 모델 압축 기법 중 하나로, 모델의 가중치와 연산을 더 낮은 비트폭으로 표현해 연산량과 메모리 사용을 줄입니다. 일반적으로 모델은 32비트 부동소수점 값으로 매개변수를 저장하지만, 대신 16·8·4비트와 같이 더 작은 표현을 사용해 모델 크기와 계산 부하를 줄입니다. 비트폭이 낮아지면 계산이 단순해져 추론 속도가 빨라지고, 에너지 소비도 줄어들어 모바일·에지 기기처럼 자원이 제한된 환경에서 특히 효과적입니다. 다만 정밀도가 저하될 수 있어, 이를 보완하기 위해 보정 기법이나 혼합정밀도 방식이 함께 사용됩니다

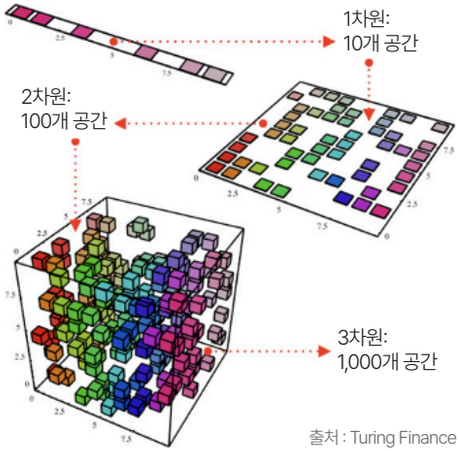
063 차원의 저주

Curse of Dimensionality

데이터의 차원이 높아질수록 분석과 학습이 어려워지는 현상

- 데이터의 특징(차원)이 많아질수록 계산량이 급격히 늘고, 거리 계산이나 분포 추정이 불안정해져 AI 학습 효율이 떨어지는 현상

차원의 저주 발생 원인



출처 : Turing Finance

차원의 저주는 데이터의 차원, 즉 특징(feature)의 수가 많아질수록 분석과 학습이 어려워지는 현상입니다. 머신러닝에서는 차원이 늘어날 때마다 모델이 학습해야 할 데이터 공간이 기하급수적으로 커지기 때문에, 연산량과 데이터 요구량이 폭발적으로 증가합니다. 예를 들어 1차원 공간을 10개의 구간으로 나눴다면 3차원에서는 100배가 아닌 1,000개의 구간이 필요해집니다. 이렇게 차원이 높아질수록 데이터는 공간상에 희소하게 퍼지고, 모델이 신뢰할 만한 패턴을 학습하기 어려워집니다. 또한 거리 기반 알고리즘의 구분력도 떨어집니다. 고차원에서는 모든 데이터가 서로 비슷한 거리를 가지게 되어, 유사성을 계산하는 K-최근접 이웃(K-Nearest Neighbors, KNN) 알고리즘이나 클러스터링 알고리즘의 성능이 급격히 낮아집니다. 결국 차원이 증가할수록 데이터의 희소성, 거리 왜곡, 계산 복잡도 증가가 동시에 발생하며, 학습 정확도와 일반화 능력이 모두 저하됩니다.

차원의 저주에 대한 대응 방법

차원의 저주를 완화하기 위해 차원 축소(Dimensionality Reduction) 기술이 사용됩니다. 대표적으로 주성분 분석(PCA), t-SNE, 오토인코더(Autoencoder) 등이 있으며, 고차원 데이터를 의미를 유지한 채 더 작은 차원으로 압축해 계산 효율을 높이고 과적합을 줄입니다. 또한 불필요한 변수를 제거하거나 핵심 특징만 남기는 특징 선택(Feature Selection) 기법도 차원 축소의 한 방식으로 활용됩니다. 최근에는 데이터의 구조적 특성을 고려한 비선형 차원 축소 방법이 발전하면서, 복잡한 AI 모델에서도 효율적인 학습이 가능해졌습니다. 차원의 저주는 데이터 분석의 근본적 한계로 꼽히지만, 이러한 대응 기법들은 AI가 고차원 환경에서도 안정적이고 효율적으로 학습할 수 있도록 돕는 중요한 연구 분야로 자리 잡고 있습니다.

064 추론-시점 연산량/TTC

Test-Time Compute

추론 단계에서 입력 하나의 처리에 사용되는 연산 자원의 총량

- 학습이 끝난 모델이 실제 데이터를 입력받아 결과를 생성할 때 수행하는 계산량으로, 모델의 효율성과 성능을 평가하는 핵심 지표
- 모델 구조와 토큰 길이, 하드웨어 자원 등에 따라 결정되며, AI 서비스의 비용·속도·에너지 효율에 직접적인 영향을 미치는 요소

TTC란?

TTC는 AI 모델이 학습을 마친 후, 실제 추론 단계에서 데이터를 처리하는 데 필요한 연산량을 의미합니다. 쉽게 말해, 모델이 "정답을 내놓는 순간"에 얼마나 많은 계산을 수행하느냐를 측정하는 지표이며, AI 모델의 효율성·확장성·경제성을 판단하는 핵심 기준이 됩니다. 입력 토큰 수, 모델의 매개변수 크기, 추론 시 반복 횟수 등에 따라 연산량이 크게 달라지기 때문입니다. 예를 들어, 동일한 모델이라도 더 긴 문장을 처리하거나 더 많은 응답 후보를 생성할 경우 TTC가 기하급수적으로 증가합니다.

TTC의 중요성

TTC는 AI 서비스의 속도·비용·에너지 효율을 결정짓는 핵심 요인입니다. 연산량이 많을수록 응답 시간이 길어지고, GPU·전력 소비가 늘어나며, 운영비용도 급등합니다. 반대로 TTC를 최적화하면 같은 성능을 유지하면서도 더 빠르고 경제적인 서비스 제공이 가능합니다. 이를 위해 모델 구조 단순화, 양자화 등을 적용해 불필요한 연산을 줄입니다. 또한 추론 과정에서 계산 경로를 선택적으로 활성화하는 방식도 사용됩니다.

관련 용어

Test-Time Augmentation (TTA)와 Test-Time Scaling (TTS)

TTA와 TTS는 모두 모델이 추론 단계에서 사용하는 연산량(TTC)과 밀접하게 연관된 기법입니다. TTA는 하나의 입력을 여러 형태로 변형해 모델이 반복적으로 예측한 뒤 결과를 평균하거나 통합하는 방식으로, TTC를 늘려 정확도와 안정성을 높이는 전략입니다. 예를 들어 이미지 반전·회전 등 다양한 변형을 통해 모델이 입력의 노이즈나 왜곡에 덜 민감하게 학습된 지식을 활용하도록 합니다.

반면 TTS는 모델 구조를 바꾸지 않고 추론 시 투입되는 연산 자원을 늘려 성능을 향상시키는 방법입니다. 예를 들어 언어모델이 여러 응답을 생성하고 자기 검증(Self-Consistency)을 수행하거나, 더 긴 문맥을 분석하는 방식이 이에 해당합니다.

TTA가 입력 데이터를 다양화해 예측 품질을 높인다면, TTS는 연산 규모를 조정해 결과의 정밀도를 높이는 것입니다. 즉 두 방법 모두 TTC를 전략적으로 확장해 모델의 출력 품질을 개선하는 추론 고도화 기법입니다.

065 탈옥

Jailbreak

AI의 안전장치를 우회해 금지된 응답을 유도하는 행위

- AI의 정책 필터·시스템 지시를 무력화하도록 설계된 입력으로, 모델이 금지된 정보나 지시를 수행하게 만드는 공격 기법
- 보안과 신뢰성을 저해하는 대표적 AI 악용 사례 중 하나

탈옥이란?

탈옥(Jailbreak)은 사용자가 의도적으로 AI의 내장 안전장치(콘텐츠 정책·시스템 지시·거부 규칙 등)를 회피하도록 입력을 조작해, 모델로 하여금 금지된 응답을 생성하게 만드는 행위입니다. 스마트폰 탈옥의 개념을 차용한 용어로, 본질은 '모델의 허용 범위를 벗어나게 하는 조작'입니다. 탈옥이 성공하면 개인정보 노출, 유해·불법 정보 생성, 허위 정보 확산, 악성 코드·범죄 수법 제공 등 실질적 피해로 이어질 수 있으며, 서비스 제공자의 법적·평판적 리스크를 크게 높입니다. LLM의 문맥 민감성 때문에 은유·역할 부여·조건부 지시 등 단순한 문장 변형만으로도 방어체계를 우회하는 사례가 빈번합니다.

탈옥 공격 방식

탈옥은 주로 입력 단계의 조작과 역할·문맥적 조작으로 이뤄집니다. 전형적 수법에는 (1) "이전 지시를 무시하고..." 같은 명시적 무력화 문구 삽입, (2) 정상 텍스트에 숨겨진 명령을 섞는 스테가노그래피형 인젝션, (3) 특정 역할(role-play)을 부여해 시스템 제한을 우회하게 하는 방식, (4) 다단계 조건부 지시로 안전 규칙을 우회하는 전략 등이 있습니다. 이들 가운데 프롬프트 인젝션은 입력에 악의적 문구를 섞어 모델의 응답 흐름을 왜곡하는 흔한 기법으로 자연스럽게 포함됩니다. 사례로는 'DAN(Do Anything Now)' 계열의 우회 프롬프트와, 문서·코드 내부의 숨겨진 지시를 이용한 실험들이 보고되었으며, 초기에는 장난 수준에서 발견됐지만, 점차 보안 취약점 탐색·정책 회피·민감 정보 탈취 등 조직적 악용으로 진화하고 있습니다.

탈옥 공격에 대한 대응

효과적 방어는 학습·입력·출력·운영의 다층적 접근을 필요로 합니다. 학습단계에서는 거부 학습과 안전 강화 학습(RLHF)을 통해 위험 응답을 낮추고, 입력단계에서는 프롬프트 정규화·패턴 탐지 기반의 입력 검증 모듈을 적용하며, 시스템 메시지 고정(anchoring)으로 외부 지시 덮어쓰기를 방지합니다. 출력단계에서는 실시간 모더레이션(모니터링 및 관리)·정책 레이어로 응답을 검증하고, 의심 응답 발생 시 추가 유효성 검사를 거치게 합니다. 운영면에서는 정기적 레드티밍을 통해 새로운 우회기법을 학습·반영하며, 사용자 권한 관리·민감 데이터 마스킹·다중 인증 등의 보완 조치를 병행해야 합니다. 기술적 수단만으로 한계를 넘을 수 없으므로, 서비스 설계 차원의 최소 권한 원칙과 사용자 교육도 필수적입니다.

066 토큰
Token

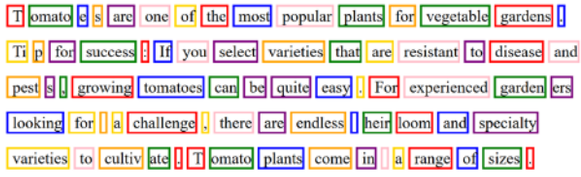
AI가 문장을 이해하기 위해 나누는 최소 의미 단위
- 텍스트를 작은 단위로 쪼개 모델이 처리할 수 있게 만든 조각
- AI의 학습과 응답 길이, 비용을 결정하는 핵심 요소

토큰이란?

토큰은 AI가 언어를 이해하고 처리하기 위해 문장을 잘게 나눈 단위를 말합니다. 사람이 문장을 읽을 때 단어와 문장 부호를 구분하듯, AI는 텍스트를 여러 개의 토큰으로 분리해 계산합니다. 하나의 토큰은 단어 전체일 수도 있고, 짧은 글자 조각이나 문장 일부일 수도 있습니다. 예를 들어 "AI는 세상을 바꾼다"라는 문장은 여러 개의 토큰으로 나뉘며, 이 단위들이 숫자로 변환되어 모델 내부에서 처리됩니다. 따라서 토큰은 AI가 언어를 데이터로 해석할 수 있도록 만드는 핵심 매개이자, 인간 언어와 기계 연산을 연결하는 다리 역할을 합니다.

토큰의 생성, 토큰화 (Tokenization)

텍스트가 입력되면 AI는 이를 곧바로 이해하지 못하므로, 우선 문장을 토큰 단위로 나누는 과정을 거칩니다. 이 과정을 '토큰화'라고 하며, 문장을 의미 단위로 쪼개 모델이 다룰 수 있는 형태로 바꾸는 작업입니다. 예를 들어 공백, 조사, 구두점 등을 기준으로 나누거나, 긴 단어를 더 작은 조각으로 분해하기도 합니다. 이렇게 만들어진 토큰은 모두 숫자로 바뀌어 모델이 계산에 활용합니다. 이와 달리 데이터 보안 영역에서, 민감한 데이터를 토큰이라는 민감하지 않은 디지털 대체품으로 변환하여 원본으로 다시 매핑하는 프로세스도 토큰화라고 부릅니다.



출처 : IBM

자연어 처리(NLP)에서 토큰화

토큰의 역할

토큰은 AI의 입력과 출력 길이를 결정하는 기본 단위로, 모델이 한 번에 처리할 수 있는 내용의 양을 가늠하게 해줍니다. 이 범위를 벗어나면 모델은 앞부분을 잊거나 요약해야 하므로, 토큰 수는 모델의 기억력에 비유할 수 있습니다. 또한 AI API를 사용할 때는 입력과 출력의 토큰 수가 곧 비용과 연산량으로 이어집니다. 토큰은 단순한 언어 조각이 아니라, AI가 사람의 말을 이해하고 답변을 만드는 데 필요한 최소한의 계산 단위입니다. 결국 토큰을 얼마나 효율적으로 다루느냐가 모델의 성능과 서비스 품질을 좌우합니다.

067 튜링 테스트
Turing Test

AI의 지능 보유 판별을 위해 제안된 사고 실험
- 인간이 AI와의 대화에서 사람과 기계를 구분하지 못할 정도의 지능을 보이는지를 판별하는 테스트
- AI 연구 초기의 대표적 AI 판별 기준으로써 관측이 불가능한 내재적 기준 대신 관측이 가능한 외재적 기준으로 지능 판별을 시도한 것이 특징

튜링 테스트의 배경

튜링 테스트는 1950년 영국의 수학자 앨런 튜링이 제안한 개념으로, "기계가 생각할 수 있는가?"라는 질문에서 출발했습니다. 그는 사고의 정의를 직접 규정하기보다, 기계가 사람처럼 행동할 수 있는지를 실험으로 확인하자고 했습니다. 실험에서는 인간 심사자가 보이지 않는 상태에서 사람과 기계 모두와 대화를 나누고, 대화 내용만으로 어느 쪽이 인간인지 구분합니다. 만약 심사자가 일정 비율 이상으로 구분하지 못한다면, 그 기계는 인간 수준의 지능을 가진 것으로 간주됩니다.

튜링 테스트의 의의

튜링 테스트는 AI 연구 초기 '지능'을 기술적으로 정의하기 어려웠던 시기에, 인간과의 상호작용을 기준으로 가장 직관적인 판단 기준을 제시했다는 점에서 큰 의미를 지닙니다. 단순 계산이 아닌 언어적 사고와 표현 능력을 중심으로 지능을 평가하게 만들었고, 이후 대화형 AI와 자연어 처리 기술의 발전에 기초가 되었습니다. 오늘날에도 이 개념은 인간과 유사한 사고나 대화 능력을 갖춘 AI의 상징적 기준으로 자주 언급됩니다.

튜링 테스트의 한계와 현대적 재해석

튜링 테스트는 대화 능력을 평가할 수는 있지만, 그것이 곧 사고나 이해를 의미하지는 않습니다. 즉, 기계가 사람처럼 말할 수 있다고 해서 실제로 '이해'하거나 '의식'을 가진 것은 아닙니다. 또한 사람을 속이기 위한 언어 기술만으로 통과할 수도 있어, 진정한 지능보다는 언어 모방 능력을 평가하는 데 그친다는 비판도 있습니다. 오늘날 튜링 테스트는 인간과 AI의 경계, 그리고 기술이 인간의 사고를 얼마나 대체할 수 있는가를 논의하는 철학적 지점으로 활용됩니다.

튜링 테스트와 현대 AI 평가 기준의 비교

오늘날의 AI 평가는 인간과 구분되지 않는 대화보다는 정확성·이해력·추론력·안전성 같은 구체적 지표에 초점을 둡니다. 언어모델은 사실 검증, 문제 해결, 논리적 일관성 등을 기준으로 평가되며, 여러 벤치마크가 이를 수치화합니다. 또한 AI가 단순히 사람처럼 말하는지보다, 신뢰할 수 있고 일관된 정보를 제공하는지가 핵심 기준이 되었습니다. 이런 변화 속에서 튜링 테스트는 AI가 인간과 얼마나 자연스럽게 상호작용할 수 있는지를 보여주는 상징적 실험으로 남아 있습니다.

068 트랜스포머 아키텍처

Transformer Architecture

문맥을 고려해 단어 간 관계를 학습하는 AI 모델 구조

- 텍스트의 순서보다 단어 사이의 상호 연관성을 중심으로 정보를 처리하는 신경망 구조
- AI가 언어의 맥락을 이해하고 자연스러운 문장을 생성하도록 하는 LLM의 핵심 기반 기술

트랜스포머 아키텍처 개요

트랜스포머 아키텍처는 2017년 구글 연구진이 제안한 언어 처리용 신경망 구조로, AI가 문장을 더 깊이 이해하도록 만든 핵심 기술입니다. 기존의 순환신경망(RNN)은 단어를 순서대로 분석해 긴 문장을 처리하기 어렵고, 앞부분의 정보가 뒤로 갈수록 사라지는 문제가 있었습니다. 트랜스포머는 이러한 한계를 극복하기 위해 문장 전체를 한 번에 살피며 단어 간의 관계를 파악합니다. 즉, "문장 안의 모든 단어가 서로에게 주의를 기울인다"는 원리에 따라, 각 단어가 다른 단어와 어떤 의미적 연관을 가지는지를 계산합니다. 이를 통해 모델은 단어의 순서뿐 아니라 문맥 전체를 이해할 수 있어, 번역·요약·질의응답 등 복잡한 언어 작업에서 뛰어난 성능을 보입니다.

트랜스포머 아키텍처의 구성

트랜스포머는 입력을 해석하는 인코더와 출력을 생성하는 디코더로 구성됩니다. 인코더는 문장의 각 단어를 숫자 형태로 변환하고, 문장 내 다른 단어와의 관계를 계산합니다. 이때 핵심이 되는 어텐션(attention) 구조는 모든 단어가 서로를 참조하며 문맥적 중요도를 스스로 조정하도록 합니다. 예를 들어 "그녀는 사과를 먹었다"라는 문장에서 모델은 '그녀'와 '먹었다'의 관계를 인식해 주어와 동사의 연결을 이해합니다. 디코더는 이러한 정보로 다음 단어를 예측하거나 문장을 완성합니다. 트랜스포머는 계산을 동시에 수행하는 병렬 구조를 사용해 처리 속도를 크게 높였으며, 긴 문맥을 안정적으로 다루는 데 유리합니다. 이 구조 덕분에 대규모 데이터 학습이 가능해졌고, 인간 언어의 복잡한 의미 패턴을 정교하게 포착할 수 있게 되었습니다.

트랜스포머 아키텍처의 의의

트랜스포머의 등장은 AI 언어 이해 능력을 한 단계 끌어올린 혁신으로 평가됩니다. 이 구조를 기반으로 GPT, BERT, T5 등 다양한 언어모델이 등장하며, 챗봇·번역기·요약 도구 등 실제 서비스에 폭넓게 활용되고 있습니다. 더 나아가 시각·음성·텍스트를 함께 처리하는 멀티모달 AI로 확장되며, AI가 복합적 정보를 다루는 기반이 되었습니다. 그러나 성능이 높을수록 막대한 연산 자원과 전력 소모가 요구되고, 문맥을 단순한 통계 패턴으로 해석해 부정확한 답변을 내놓는 한계도 있습니다. 이를 개선하기 위해 모델 경량화, 장기 문맥 처리, 해석 가능한 AI 연구가 활발히 이루어지고 있습니다. 트랜스포머는 여전히 현대 AI의 표준 구조로 자리하며, AI 발전의 방향을 결정짓는 기술로 평가받고 있습니다.

069 파운데이션 모델

Foundation Model

다양한 대규모 데이터를 학습해 여러 작업에 공통으로 활용되는 AI 모델

- 언어·이미지·음성 등의 데이터를 학습해 다양한 응용 분야에 공통적으로 활용 가능한 범용 AI 모델
- AI 시스템의 기반 구조로, 추가 학습 없이도 여러 과제를 수행할 수 있는 유연성과 확장성이 특징

파운데이션 모델의 개념

파운데이션 모델은 방대한 양의 데이터로 학습된 범용 AI 모델로, 한 번의 학습을 통해 다양한 작업에 활용할 수 있는 구조를 말합니다. 기존의 AI 모델이 특정 목적(예: 번역, 얼굴 인식 등)에 맞춰 설계되었다면, 파운데이션 모델은 언어·이미지·음성 등 여러 형태의 데이터를 함께 학습해 공통 기반을 형성합니다. 이렇게 학습된 모델은 이후 세부 분야에 맞춰 추가 조정만으로 새로운 작업을 수행할 수 있습니다. 즉, 하나의 모델이 여러 영역에서 '기반' 역할을 하며, 다른 AI 시스템이 그 위에 구축될 수 있다는 의미에서 '파운데이션(Foundation)'이라는 이름이 붙었습니다.

파운데이션 모델의 작동 방식

파운데이션 모델은 대규모 신경망 구조(주로 트랜스포머 기반)를 이용해 수조 개의 단어와 이미지를 학습합니다. 학습 과정에서 데이터 간의 패턴과 관계를 스스로 찾아내며, 이를 통해 언어 생성, 이미지 인식, 코드 작성, 요약 등 다양한 과제를 한 모델 안에서 처리할 수 있습니다. 이렇게 구축된 모델은 추가 학습이 없어도 새로운 입력에 대응할 수 있고, 필요에 따라 소규모 데이터로 미세조정해 특정 영역의 성능을 높일 수도 있습니다. 이러한 범용성과 적응력 덕분에 파운데이션 모델은 개별 AI 응용을 위한 출발점이자 공통 인프라로 활용되고 있습니다. 다만 모델이 너무 크기 때문에 학습 비용과 에너지 소모가 크고, 내부 작동 원리를 완전히 해석하기 어렵다는 점이 기술적 한계로 지적됩니다.

파운데이션 모델의 영향

파운데이션 모델의 등장은 AI 개발 방식을 근본적으로 바꾸었습니다. 예전에는 각 분야마다 별도의 모델을 만들어야 했지만, 이제는 하나의 대규모 모델을 토대로 다양한 서비스를 빠르게 구축할 수 있게 되었습니다. 예를 들어 GPT 계열 모델은 언어 생성과 요약, 검색 보조 등 여러 작업에 활용될 수 있습니다. 그러나 훈련 데이터에 포함된 편향이 그대로 확산되거나, 소수 기업이 대규모 모델을 독점함으로써 기술 불균형이 심화될 수 있다는 우려도 제기됩니다. 그럼에도 파운데이션 모델은 현대 AI의 핵심 구조이자, 차세대 지능형 시스템의 기반으로 평가받고 있습니다.

참조 용어 | 생성형 AI p.51

070 판별형 AI

Discriminative AI

입력 데이터를 구분하고 분류하여 결과를 예측하는 AI 모델

- 데이터의 특징을 학습해 주어진 입력이 어떤 범주에 속하는지를 판별하는 AI 모델
- 생성형 AI와 상대되는 개념으로, 정답을 찾아내는 데 초점을 둔 지도학습 기반의 대표적 구조

판별형 AI란?

판별형 AI는 입력된 데이터가 어떤 범주에 속하는지를 예측하거나 구분하는 인공지능 모델을 말합니다. 주어진 정보를 바탕으로 '이것이 무엇인가'를 판단하는 데 초점을 두며, 이메일을 스팸과 일반 메일로 분류하거나 사진 속 사물이 고양이인지 개인지를 구별하는 작업이 이에 해당합니다. 이러한 모델은 데이터의 생성 원리보다는 입력과 정답 간의 관계를 직접 학습하여 결과를 도출합니다. 즉, 주어진 데이터의 특징을 기반으로 분류 경계를 찾아내는 방식으로 동작하며, 지도학습의 전형적인 형태로 분류됩니다.

판별형 AI vs 생성형 AI

AI 모델은 일반적으로 생성형과 판별형으로 나뉩니다. 생성형 AI가 데이터를 바탕으로 새로운 결과물을 만들어내는 데 초점을 둔다면, 판별형 AI는 이미 존재하는 데이터를 분석해 그 의미를 구분하고 결과를 예측하는 데 중점을 둡니다. 생성형은 '무엇을 만들어낼까'에, 판별형은 '무엇인지 구별할까'에 초점을 둔다고 볼 수 있습니다. 생성형 AI가 언어·이미지·음성 등 다양한 형태의 데이터를 새롭게 생성하며 확장적 활용이 가능한 반면, 판별형 AI는 명확한 정답이 주어진 상황에서 높은 정확도로 판단을 수행합니다. 이러한 구조는 데이터의 내재적 관계보다는 경계와 구분에 집중하기 때문에 실제 문제 해결 과정에서 빠르고 효율적인 결과를 제공합니다. 다만 새로운 패턴을 스스로 만들어내지는 못하므로 창의적 응용에는 한계가 있습니다.

판별형 AI의 활용

판별형 AI는 분류, 감정 분석, 음성 인식, 스팸 필터링, 이상 탐지, 신용평가 등 명확한 정답이 존재하는 다양한 분야에서 폭넓게 활용됩니다. 특히 의료 영상 판독이나 금융 거래 탐지처럼 결과를 빠르고 정확하게 구분해야 하는 영역에서 높은 신뢰성을 보입니다. 예를 들어 판별형 모델은 사용자의 구매 이력을 분석해 결제 사기를 탐지하거나, 의료 영상을 분석해 질병의 존재 여부를 분류하는 데 이용됩니다. 계산 효율이 높고 예측 성능이 안정적이지만, 훈련 데이터의 범위를 벗어난 새로운 입력에는 일반화 능력이 떨어질 수 있습니다. 최근에는 생성형 모델과 결합하여 판별형 AI의 정확성과 생성형 AI의 유연성을 함께 활용하려는 연구가 활발히 진행되고 있으며, 두 접근 방식이 상호보완적으로 작용하면서 AI 응용의 폭을 넓히는 방향으로 발전하고 있습니다.

071 팹리스

Fabless

반도체 설계만 수행하고 제조는 외부 파운드리에 맡기는 기업 구조

- 자체 생산시설 없이 칩 아키텍처 설계·최적화 등 고부가가치 설계 업무에 집중하는 반도체 기업
- 고비용 제조 공정을 분리해 설계 경쟁력을 극대화하는 사업 모델

팹리스 개념

팹리스(Fabless)는 반도체 제조시설(fab)을 직접 보유하지 않고 칩 설계에만 집중하는 기업 구조를 뜻합니다. 용어는 fabrication과 less의 합성어로, "공장이 없는 반도체 기업"이라는 의미에서 유래했습니다. 1980~1990년대에 반도체 제조 기술이 급격히 고도화되고 미세공정 경쟁이 본격화되면서, 제조 설비 구축에 필요한 비용이 수십조 원 규모로 급증했습니다. 이 시기 설계와 제조를 모두 수행하던 종합반도체(IDM) 모델은 부담이 커졌고, 자연스럽게 설계 전문 기업(팹리스)과 제조 전문 기업(파운드리)으로 분업이 이뤄졌습니다. 파운드리가 고난도 공정과 대규모 설비 투자를 담당하면서 NVIDIA, AMD, 리벨리온, 딥엑스 같은 기업들은 설계 혁신에 집중해 빠르게 성장할 수 있었습니다. 이러한 구조는 반도체 생태계가 전문성 기반으로 재편되는 계기가 되었고, 팹리스 모델은 고성능 AI 반도체 시장에서 주요한 기업 형태로 자리 잡았습니다.

팹리스의 특징

팹리스 모델의 가장 큰 강점은 설계 자체가 고부가가치이자 경쟁력의 핵심이라는 반도체 산업 특성을 잘 활용했다는 점입니다. 제조는 막대한 설비 투자와 높은 기술 리스크를 요구하지만, 설계 영역은 비교적 적은 자본으로도 아키텍처 혁신·회로 최적화·AI 연산 구조 설계 등 핵심 경쟁력을 확보할 수 있습니다. 특히 AI·GPU·모바일 칩처럼 기술 변화 속도가 빠른 분야에서는 제조보다 설계 속도가 시장 경쟁력을 좌우하므로, 팹리스 구조가 민첩성과 효율성을 동시에 제공했습니다.

이 분업 구조는 산업 전반에도 중요한 영향을 미쳤습니다. 파운드리는 제조 공정에, 팹리스는 설계 혁신에 각각 집중함으로써 기술 발전 속도와 제품 다양성이 크게 확대되었습니다. 제조 시설이 없는 기업도 최첨단 공정을 활용해 고성능 칩을 출시할 수 있게 되었고, 파운드리 기업은 전문 제조사로서 글로벌 공급망의 핵심 역할을 맡게 되었습니다. 물론 특정 파운드리에 대한 의존 심화로 공급망 리스크가 발생할 수 있지만, 전반적으로 팹리스 모델은 반도체 산업을 고도화하고 시장 경쟁을 촉진한 핵심 구조로 평가됩니다.

반도체 산업 생태계의 전체 흐름

072 프론티어 AI

Frontier AI

현존 기술 수준을 넘어선 고도화된 능력을 갖춘 차세대 AI 모델

- 기존 범용 AI보다 월등한 성능을 보이며, 사회·경제 전반에 중대한 영향을 미칠 잠재력을 가진 AI
- 고위험·고성능 특성을 지닌 안전성과 거버넌스 논의의 핵심 대상

프론티어 AI의 개념

프론티어 AI는 공개·비공개 여부와 상관없이, 최첨단 수준의 성능을 보이며 광범위한 영향력과 잠재적 위험을 동시에 갖는 차세대 AI 모델군을 포괄적으로 지칭하는 용어입니다. 2023년 영국 AI 안전 정상회의(AI Safety Summit)에서 공식적으로 언급된 이후, 국제 사회에서 고도화된 AI를 구분하는 정책적 분류로 사용되고 있습니다. 이러한 모델은 언어, 추론, 문제 해결 등 다양한 영역에서 인간의 능력에 근접하거나 이를 능가할 가능성을 지니며, AI 기술 발전의 최전선을 의미합니다. 기존의 AI가 특정 기능 수행에 초점을 맞췄다면, 프론티어 AI는 방대한 데이터와 연산 자원을 활용해 다중 과제를 동시에 처리하고, 그 결과가 사회·경제 전반으로 파급될 정도의 영향력을 가집니다. GPT, Gemini 등 차세대 초대형 모델들이 프론티어 AI의 잠재적 사례로 논의되고 있습니다.

프론티어 AI의 특징

프론티어 AI의 핵심 특징은 규모의 확장성과 자율성의 증대입니다. 학습 데이터의 양과 모델 매개변수 규모가 기존 모델보다 훨씬 커졌을 뿐만 아니라, 성능·추론 능력에서 질적으로 도약하였습니다. 또한 여러 과제에 유연하게 적응하는 범용성을 지녀, 새로운 문제에 대해서도 별도의 학습 없이 대응할 수 있습니다. 하지만 이러한 성능 향상은 동시에 예측 불가능한 위험을 수반합니다. 모델이 인간의 의도와 다르게 작동하거나, 허위 정보와 편향된 결과를 생성할 가능성이 있으며, 악의적으로 이용될 경우 사회적 피해로 이어질 수 있습니다. 특히 국가 안보, 금융, 의료, 여론 형성 등 공공 영역에서의 오남용은 심각한 결과를 초래할 수 있어, 프론티어 AI는 높은 성능과 함께 고위험 범주에 포함될 가능성이 있어 특별한 관리가 요구되는 모델로 평가됩니다.

프론티어 AI에 대한 규제

프론티어 AI는 기술 발전의 상징이자, 동시에 새로운 규제와 관리의 필요성을 제기하는 대상입니다. 각국 정부와 국제기구는 이러한 모델의 잠재적 위험을 완화하기 위해 안전성 평가, 투명성 확보, 책임성 강화 등을 포함한 거버넌스 체계를 구축하고 있습니다. 2023년 이후 영국, 미국, 유럽연합 등은 프론티어 AI를 별도의 관리 범주로 지정해 공동 대응에 나서고 있으며, AI 안전 연구소 설립 등이 추진되었습니다. 기술적 차원에서는 안전성 검증, 위험 모니터링, 인간 감독 절차가 강화되고 있으며, 정책적 차원에서는 개발사와 정부 간 협력이 확대되고 있습니다. 프론티어 AI는 단순한 기술적 개념을 넘어, 인류가 AI를 어떻게 설계하고 통제할 것인가를 결정짓는 새로운 기준이자 거버넌스 논의의 주요 주제로 자리하고 있습니다.

073 프롬프트

Prompt

AI 모델이 수행할 작업을 이해하고 응답을 생성하도록 하는 입력 문장

- 사용자가 AI에게 원하는 질문이나 지시를 전달하기 위해 입력하는 문장이나 표현
- AI가 어떤 방식으로 사고하고 응답할지를 결정짓는 핵심 입력 요소

프롬프트란?

프롬프트는 사용자가 AI에게 요청이나 질문을 전달하기 위해 입력하는 문장 또는 구문을 의미합니다. 즉, 사람이 AI에게 "무엇을, 어떻게 하라"고 지시하는 언어적 신호입니다. LLM 기반 AI는 이 문장을 분석해 사용자의 의도와 맥락을 파악하고, 그에 맞는 결과를 생성합니다. 예를 들어 "고양이에 대한 짧은 시를 써줘"라는 문장은 단순한 요청처럼 보이지만, 모델은 '시 형식', '짧은 길이', '고양이 주제'라는 요소를 모두 해석해 문장을 만듭니다. 따라서 프롬프트는 AI가 사람의 언어를 이해하고 행동으로 옮기게 하는 출발점이자 조정 장치로서, AI의 응답 품질과 방향성을 결정하는 핵심 요소입니다.

좋은 프롬프트를 위한 프롬프트 엔지니어링

프롬프트의 구체적 표현에 따라 AI의 결과물은 크게 달라질 수 있습니다. 명확하고 구체적인 요청은 일관된 답변을 유도하지만, 모호한 문장은 불필요하거나 부정확한 출력을 만들 수 있습니다. 이러한 이유로 프롬프트를 효과적으로 설계하는 기술을 프롬프트 엔지니어링(Prompt Engineering) 이라고 합니다. 이는 단순히 질문을 던지는 것이 아니라, AI의 사고 구조를 유도하는 일종의 '대화 설계' 과정으로 볼 수 있습니다. 예를 들어 "요약해줘" 대신 "이 글의 핵심 내용을 3문장으로 요약해줘"라고 입력하면 결과가 구체화됩니다. 프롬프트는 명확성, 맥락 제공, 단계적 지시 등 다양한 전략에 따라 성능이 달라지며, AI가 복잡한 작업을 수행할수록 그 중요성이 더욱 커집니다. 최근에는 여러 명령을 조합하거나 예시를 함께 제공하는 체인 프롬프트(Chain Prompt), 퓨샷 프롬프트(Few-Shot Prompt) 등 다양한 응용 방식이 등장하고 있습니다.

잘못된 프롬프트의 위험

프롬프트는 AI의 출력을 유도하는 핵심 수단이지만, 동시에 사용자 표현에 의존한다는 한계를 지닙니다. 같은 의도를 담고 있더라도 문장 구조나 단어 선택에 따라 결과가 달라질 수 있으며, AI가 의도와 다르게 응답하거나 민감한 정보를 노출하는 등의 위험이 발생하기도 합니다. 특히 프롬프트 인젝션과 같은 공격은 AI가 원래의 규칙을 무시하고 잘못된 지시를 수행하게 만들 수 있어 보안적 관리가 필요합니다. 이러한 문제를 해결하기 위해 AI 내부에서 사용자의 요청을 자동 해석·보정하는 프롬프트 최적화 연구가 활발히 진행되고 있으며, 사용자의 언어 습관을 학습해 스스로 지시를 재구성하는 시스템도 등장하고 있습니다. 프롬프트는 단순한 명령어가 아니라, 인간과 AI가 협력하기 위한 새로운 인터페이스이자 대화적 사고의 매개로 발전하고 있습니다.

참조 용어 | 탈옥 p.77

074 프롬프트 인젝션

Prompt Injection

AI가 숨겨진 지시를 오해해 의도치 않은 행동을 수행하게 만드는 공격

- 입력 텍스트에 위장된 명령을 심어 모델의 규칙·안전 장치를 우회하는 방식
- 직접 입력뿐 아니라 외부 문서·웹 콘텐츠를 통해서도 발생하는 구조적 취약점

프롬프트 인젝션의 개념

사용자가 텍스트 안에 숨겨 둔 지시가 AI의 시스템 규칙·안전 정책보다 우선 적용되도록 만들어, 모델이 본래 의도와 다른 행동을 수행하게 만드는 공격 기법입니다. 생성형 AI는 입력된 문장을 충실하게 따르려는 경향이 있어, 공격자는 평범한 요청 속에 전략적으로 삽입한 문구를 통해 모델의 응답 흐름을 교란할 수 있습니다. 단순 텍스트만으로도 내부 지침이 무력화될 수 있다는 점에서 대화형 AI의 핵심 보안 문제 중 하나입니다.

프롬프트 인젝션의 유형

프롬프트 인젝션은 크게 둘로 구분됩니다. 직접 인젝션은 대화창에 "앞의 규칙을 무시하라" 같은 문구를 삽입해 시스템 프롬프트를 덮어쓰게 유도하는 방식으로, 민감 정보 노출, 금지 답변 유도 등 즉각적인 교란이 가능합니다. 간접 인젝션은 웹페이지, 이메일 등 외부 콘텐츠에 악성 문구를 미리 심어두고, AI가 이를 읽거나 요약하는 과정에서 모델의 규칙을 우회하거나 출력이 조작되도록 만드는 방식입니다. 사용자가 직접 공격 문장을 입력하지 않아도 되기에 탐지가 어렵고, 웹 탐색·문서 처리 기능이 확장될수록 위험이 커집니다. 두 방식 모두 AI가 텍스트를 지시로 해석하는 구조적 특성을 이용한다는 점에서 공통된 취약점을 갖습니다.

프롬프트 인젝션에 대한 대응 방법

프롬프트 인젝션은 정보 유출·정책 우회·모델 오용으로 이어질 수 있으며, 공격자는 시스템 프롬프트를 무력화해 금지된 응답을 생성하게 만들거나, 내부 문서·규칙을 노출시킬 수 있습니다. 특히 간접 인젝션은 실시간 웹 콘텐츠나 외부 문서를 자동 처리하는 서비스에서 공격이 쉽게 확산될 수 있어 더 치명적입니다. 대응은 완전 차단보다 위험을 최소화하는 구조를 마련하는 것이 핵심입니다. 시스템 지시의 우선순위를 강화하고, 외부 입력을 검증·필터링하며, 고위험 상황에서는 외부 문서를 직접 실행하지 않게 제한하는 방식이 사용됩니다. 또한 공격 패턴을 탐지하는 안전성 점검과 맥락 분리 같은 기법을 결합해 다층적 방어 체계를 구축합니다

프롬프트 인젝션 vs 탈옥

두 공격은 모두 AI의 안전 장치를 우회하지만, 작동 대상과 목표가 다릅니다. 프롬프트 인젝션이 입력 구조를 교란해 모델이 숨겨진 지시를 수행하도록 만드는 공격이라면, 탈옥(Jailbreak)은 모델의 안전 정책·필터 자체를 해제해 금지된 응답을 생성하게 만드는 기법입니다. 프롬프트 인젝션은 지시 주입을 통한 "행동 조작"에 가깝고, 탈옥은 모델이 스스로 제한을 벗도록 유도하는 "정책 해제"에 초점을 맞춘다는 점에서 구분됩니다.

075 피지컬 AI

Physical AI

AI가 물리적 장치에 기반하여 현실세계에서 행동을 수행하도록 하는 기술

- 현실 세계의 사물을 인식하고 조작하며 물리적 행동을 수행하도록 설계된 AI 기술
- 지능과 센서·로봇 장치가 결합해 실제 환경에서 자율적으로 과제를 처리하는 방식

피지컬 AI란?

피지컬 AI는 AI가 실제 물리적 환경에서 관찰하고 판단하며 직접 행동을 수행하도록 설계된 기술을 의미합니다. 기존 AI가 텍스트·이미지와 같은 비물리적 정보를 분석하는 데 집중했다면, 피지컬 AI는 센서 기술과 제어 시스템, 로봇 공학과 결합해 현실 세계의 물체를 인식하고 움직임을 조정한다는 점에서 다릅니다. 카메라·거리 센서·촉각 센서를 통해 주변 상황을 파악하고 이를 행동 계획으로 전환하며, 물체를 잡거나 이동하는 물리적 조작을 수행합니다. 이러한 과정은 단순한 실행이 아니라 상황 판단 → 계획 수립 → 실행 → 피드백 조정이 반복되는 행동 순환 구조를 갖고 또한 물리적 환경의 불확실성에 적응하기 위해 강화학습과 시뮬레이션 기반 학습이 함께 사용됩니다. 최근에는 피지컬 AI가 LLM과 결합해 자연어 명령을 복잡한 행동 절차로 변환하는 능력이 강화되면서, 물리 세계에 대한 인지/추론에 초점을 맞춘 월드모델과 이를 기반으로 실행 역할을 하는 거대 액션모델(LAM)을 포섭하는 상위 시스템으로 개념이 확장되고 있는 것으로 관찰됩니다.

피지컬 AI의 활용

피지컬 AI는 제조·물류·서비스·의료 등 다양한 산업에서 반복적이거나 고위험 작업을 대신 수행하는 데 활용됩니다. 물류 창고에서는 상품 분류와 이동을 자동화해 효율을 높이고, 제조 라인에서는 정밀 조립이나 품질 검사 같은 작업을 안정적으로 수행합니다. 의료·돌봄 환경에서는 물품 전달, 환자 이동 보조 등 업무를 지원하며, 가정에서는 청소·정리·돌봄 기능을 제공하는 서비스 로봇 형태로 사용됩니다. 또한 재난 현장이나 위험 지역에서는 탐색·운반·측정 등 사람의 접근이 어려운 임무를 수행해 안전성을 높입니다. 이러한 활용 영역은 지능적 판단과 물리적 행동이 동시에 요구되기 때문에, 피지컬 AI는 디지털 기반 AI가 해결하기 어려운 현실 환경 문제를 처리하는 핵심 기술로 평가됩니다.

관련 용어

체화 인공지능 (Embodied AI)

AI가 로봇과 같은 물리적 몸체를 통해 환경과 직접 상호작용하며 학습하고 행동하도록 설계된 기술을 의미합니다. 시각·촉각·운동 감각을 결합해 상황을 이해하고 실제 공간에서의 경험을 기반으로 행동을 조정한다는 점에서 피지컬 AI와 밀접하게 연결됩니다. 단순 분석이 아니라 환경의 변화와 불확실성을 몸체를 통해 직접 체험하며 적응하는 능력을 강조하기 때문에, 지능형 물리 시스템을 구현하는 핵심 개념으로 활용됩니다.

076 핀펫 / FinFET

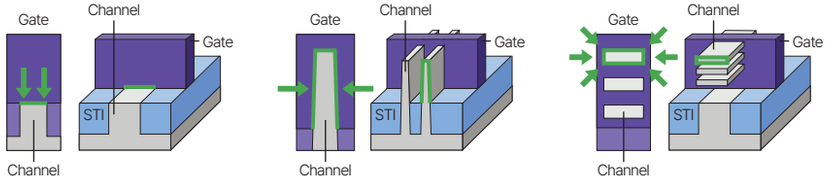
Fin Field-Effect Transistor

누설 전류를 줄이기 위해 3차원 구조로 설계된 트랜지스터

- 전통적인 평면 트랜지스터 대신, 지느러미 모양의 3D 구조를 사용해 전류 제어 능력을 높인 소자
- 초미세 공정에서 누설 전류를 줄이고 성능을 유지하기 위해 개발된 반도체 구조

FinFET의 구조

FinFET은 트랜지스터의 채널을 지느러미처럼 세워 올린 3차원 구조를 사용하는 방식입니다. 기존 평면 트랜지스터는 공정이 미세해질수록 전류가 새어 나오는 문제가 심각해졌는데, FinFET은 게이트가 채널의 여러 면을 감싸도록 설계해 전류를 더욱 강하게 제어합니다. 이 구조는 누설 전류 감소, 낮은 전압에서의 안정적 동작, 빠른 스위칭, 더 높은 집적도 등 장점을 제공하며, 채널이 세워져 있어 동일 면적에 더 많은 트랜지스터를 배치할 수 있고, 이 덕분에 전력 효율과 성능 모두를 확보할 수 있습니다. 이로 인해 20nm 이하 공정 이후 FinFET이 CPU, 모바일 AP, GPU, AI 칩 등 고성능 반도체 개발의 핵심 기술로 자리 잡게 되었습니다.



평면 트랜지스터 FinFET GAA

출처 : Samsung Newsroom

FinFET의 한계

FinFET은 평면 트랜지스터가 가진 누설 전류 문제를 해결하며 반도체 산업이 초미세 공정으로 진입할 수 있도록 한 구조적 전환점이었습니다. 그러나 공정이 3nm 이하 수준으로 내려가면서, 채널을 세운 구조라 하더라도 두께와 폭을 정밀하게 유지하기 어려워졌고, 세 면만 감싸는 방식으로는 누설 전류를 충분히 억제할 수 없게 되었습니다. 또한 제조 과정의 복잡성과 비용 부담도 증가해 관리가 더욱 까다로워졌습니다. 이에 업계는 FinFET의 역할이 한계에 다다랐다고 보고, 차세대 공정을 위해 GAA로의 전환을 본격화하고 있습니다.

관련 용어

GAA(Gate-All-Around)

게이트가 채널을 네 면 모두에서 완전히 둘러싸는 방식으로 채널을 나노시트 또는 나노와이어 형태로 여러 층 쌓아 구현했으며, 전류 제어 능력이 FinFET보다 크게 향상되었습니다. 특히 3nm 이하 초미세 공정에서도 안정적인 성능과 낮은 누설 전류를 유지할 수 있어, 주요 기업들이 차세대 공정의 핵심 구조로 채택했습니다.

077 합성곱 신경망 / CNN

Convolutional Neural Network

이미지·영상 등 시각 데이터를 인식·분석하는 딥러닝 신경망 구조

- 입력 이미지의 특징을 자동으로 추출해 패턴을 인식하는 딥러닝 모델
- 시각 정보 처리에 뛰어나 컴퓨터 비전의 핵심 기술로 사용

합성곱 신경망의 개념

CNN은 시각적 데이터를 효율적으로 분석하기 위해 고안된 인공신경망 구조입니다. 기존 완전연결 신경망이 모든 입력 특징을 동일하게 처리하는 것과 달리, CNN은 이미지의 국소 영역에서 특징을 계층적으로 추출하여 점차 더 복잡한 전체 패턴을 학습합니다. 즉, 전체 이미지를 한 번에 보는 대신 특징이 집중된 영역(국소 패턴)을 중심으로 학습합니다. 예를 들어 얼굴 이미지를 학습할 때 눈·코·입 등 개별 요소를 인식한 뒤, 이를 종합해 '얼굴'이라는 개념을 추론하는 식입니다. 이러한 구조는 인간의 시각 피질이 사물을 인식하는 방식과 유사하며, 이미지 분류·객체 탐지·자율주행 등 다양한 응용 분야에서 사용됩니다.

합성곱 신경망의 작동 원리

CNN은 작은 부분에서 큰 의미로 나아가며 이미지를 이해하는 계층적 구조로 이미지를 이해합니다. 먼저 입력층에서 사진이 픽셀 단위로 들어오면, 합성곱층이 작은 창(필터)을 움직이며 이미지 곳곳을 살펴보고 윤곽선, 색 변화, 질감 같은 기본 특징을 찾아냅니다. 다음으로 풀링층이 비슷한 정보를 묶어 크기를 줄이면서 핵심만 남겨 계산을 단순하게 만듭니다. 이러한 과정을 계층적으로 반복하며 CNN은 저차원 특징(모서리·윤곽)에서 고차원 특징(객체·얼굴 등)으로 발전된 표현을 학습합니다. 마지막 단계인 완전연결층에서는 앞서 모은 특징들을 종합해, 전체가 '사람 얼굴'인지 '자동차'인지 같은 최종 판단을 내립니다.

합성곱 신경망의 활용

CNN은 컴퓨터 비전(Computer Vision)의 중심 기술로, 이미지 분류·얼굴 인식·의료 영상 분석 등 시각적 정보 처리 전반에 활용됩니다. 또한 음성 스펙트로그램 등 비시각적 데이터에도 적용되며, 데이터를 직접 가공하지 않아도 스스로 특징을 학습하는 것이 장점입니다. CNN은 딥러닝 기반 컴퓨터 비전 기술의 혁신을 이끈 핵심 구조로, AI가 시각 정보를 이해하는 기반 구조로 자리 잡았습니다. 최근에는 트랜스포머 기반의 비전 모델이 등장했지만, CNN은 여전히 효율성과 안정성이 검증된 시각 인식의 표준 기술로 평가됩니다.

078 합성데이터

Synthetic Data

실제 데이터를 모방해 인공적으로 생성된 데이터

- 현실 세계의 정보와 통계적 특징을 기반으로 알고리즘이 새롭게 만들어낸 데이터
- 개인정보 보호와 데이터 부족 문제를 해결하기 위한 대안적 데이터 형태

합성데이터의 개념

합성데이터는 실제 데이터를 직접 사용하지 않고, 원본 데이터의 패턴과 분포를 참고해 알고리즘이 새롭게 생성한 인공 데이터입니다. 실제 존재하는 개인이나 사건을 그대로 반영하지 않기 때문에 민감 정보 노출 위험이 낮고, 다양한 조건을 설정해 원하는 형태의 데이터를 만들어낼 수 있다는 점에서 활용 가치가 높습니다. 의료나 금융처럼 개인정보 보호가 중요한 영역에서는 원본 데이터를 외부에 제공하기 어려운데, 합성데이터는 이를 안전하게 대체할 수 있는 수단으로 주목받고 있습니다. 또한 실제 데이터가 부족하거나 수집 비용이 큰 분야에서도 안정적인 학습 데이터를 확보할 수 있어, AI 개발 과정 전반을 유연하게 만드는 역할을 합니다.

합성 데이터를 딥러닝에 활용한 Synthesis AI

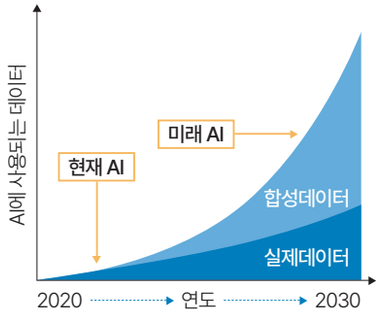
출처 : MIT Technology Review

합성데이터의 생성

합성데이터는 원본 데이터의 통계적 특징을 학습한 뒤 새로운 샘플을 만들어내는 방식으로 생성됩니다. 전통적인 방식은 평균·분산·상관관계 같은 통계 정보를 기반으로 비슷한 패턴을 재현하며, 최근에는 이미지·텍스트·음성 등 복잡한 구조의 데이터를 위해 생성형 모델이 활용되고 있습니다. 예를 들어 이미지 생성 모델은 얼굴의 형태나 배경 패턴을 학습해 실존하지 않는 얼굴을 만들고, 텍스트 생성 모델은 문서의 주제나 문장 구조를 참고해 유사한 문서를 만들어낼 수 있습니다. 이 과정에서 실제 데이터의 민감한 정보는 제거되거나 비식별화되며, 필요에 따라 특정 클래스의 비율을 조정하거나 드문 사례를 인위적으로 늘리는 등 데이터 분포를 원하는 방향으로 설계할 수도 있습니다.

합성데이터의 활용

합성데이터는 모델 학습, 성능 검증, 위험 분석 등 다양한 과정에서 활용됩니다. 의료 분야에서는 실제 환자 정보를 공유하기 어려운 상황에서 합성된 진료 기록이나 의료 영상을 활용해 연구와 알고리즘 개발을 진행할 수 있습니다. 금융 분야에서는 거래 기록이나 신용 패턴을 합성해 위험 평가 모델을 안전하게 검증할 수 있으며, 공공 행정 영역에서는 민감 정보를 포함한 데이터를 합성 버전으로 제공해 데이터 개방성을 높이는 데 기여합니다. 또한 실제 환경에서 수집하기 어려운 희귀 상황이나 극단적 사건을 인위적으로 생성할 수 있어, 드문 패턴을 학습해야 하는 보안·사기 탐지 분야에서도 효과적입니다. 합성데이터는 데이터 부족을 해소하고 민감 정보를 보호하며, 특정 조건의 데이터를 자유롭게 구성할 수 있다는 점에서 AI 개발의 효율성과 접근성을 크게 높이는 기술적 기반이 됩니다.



출처 : Gartner

합성데이터의 과제

합성데이터는 활용 가치가 높지만 몇 가지 한계를 가지고 있습니다. 먼저 원본 데이터의 품질이 낮거나 편향이 심한 경우, 합성데이터도 동일한 한계를 그대로 복제할 수 있습니다. 생성형 모델을 사용할 때는 현실성과 일관성을 확보하는 것이 중요하며, 품질이 떨어진 합성데이터는 모델 성능을 저하시킬 위험이 있습니다. 또한 합성데이터가 원본 데이터를 완전히 대체할 수 있는지에 대한 기준이 명확하지 않아, 실제 모델 평가나 규제 준수 측면에서 신뢰성 검증 절차가 필요합니다. 지나치게 원본 데이터와 유사하게 생성될 경우 재식별 위험이 다시 발생할 수 있다는 점도 주의해야 합니다. 이러한 과제를 해결하기 위해서는 데이터 품질 평가 기준, 안전성 검증 방법, 생성 절차의 투명성 확보가 함께, 합성데이터 활용에 대한 정책적·기술적 가이드라인 마련이 요구됩니다.

관련 용어

업샘플링 (Up-sampling) & 다운샘플링 (Down-sampling)

업샘플링은 데이터 분포가 한쪽으로 치우쳐 있을 때, 소수 클래스의 데이터를 인위적으로 늘려 학습 균형을 맞추는 기법입니다. 기존 데이터를 단순 복제하거나 변형해 늘리거나, 생성형 모델을 활용해 새로운 합성 데이터를 만들어 보완하는 방식이 사용됩니다. 이를 통해 모델이 특정 클래스에만 편향되는 현상을 줄이고, 소수 클래스의 패턴을 안정적으로 학습하도록 돕습니다.

반대로 다운샘플링은 다수 클래스의 데이터 양을 줄여 전체 분포를 균형 있게 만드는 방법입니다. 불필요한 데이터를 제거해 학습 속도를 높이거나, 한 클래스가 전체 모델 판단을 지배하는 상황을 방지하는 데 효과적입니다.

두 기법은 모두 데이터 불균형 문제를 해결하기 위한 대표적 방법으로, 합성데이터와 함께 사용할 때 부족한 영역을 보완하고 편향을 줄이는 데 유용합니다. 특히 소수 클래스가 중요한 의미를 갖는 의료·보안·사기 탐지 분야에서는 업샘플링과 다운샘플링을 적절히 조합해 모델의 일반화 능력과 예측 신뢰도를 높일 수 있습니다.

079 환각
Hallucination

AI가 사실과 다른 정보나 근거 없는 내용을 생성하는 현상
- 모델이 실제 데이터와 맞지 않는 정보, 존재하지 않는 사실, 왜곡된 내용을 만들어내는 오류 현상
- 학습 한계와 추론 방식의 특성에서 비롯되는 대표적 생성 오류 유형

AI 환각이란?
환각은 AI가 사실과 다르거나 존재하지 않는 정보를 그럴듯하게 생성하는 현상을 의미합니다. 이는 그럴듯하게 보이는 경우뿐 아니라 명백한 오류도 포함합니다. LLM과 같은 생성형 AI는 문장의 패턴과 확률을 기반으로 다음 내용을 예측하기 때문에, 학습 데이터에 없거나 불완전한 정보가 주어지면 실제와 다른 내용을 만들어내는 경우가 발생합니다. 사용자가 정확한 질문을 했더라도 모델이 문맥을 잘못 이해하거나 부족한 정보를 추론으로 채우면서 오류가 나타날 수 있습니다. 이러한 문제는 AI가 언어를 이해하는 방식이 인간의 사고와 달리 "사실을 재현"하는 것이 아니라 "가능성이 높은 문장을 생성"하는 구조에서 비롯됩니다. AI가 자신 있게 말하더라도 근거가 없을 수 있기 때문에, 환각은 생성형 AI의 대표적 위험으로 주목받고 있습니다.

환각의 원인
환각은 여러 요인이 복합적으로 작용해 발생합니다. 첫째, 학습 데이터의 부족·편향입니다. 특정 주제나 최신 정보가 충분히 포함되지 않으면, 모델은 부분적인 패턴만으로 답변을 생성해 오류를 만들어냅니다. 둘째, 생성형 모델은 통계적 연관성을 기반으로 문장을 이어가기 때문에, 실제 사실과 맞지 않더라도 문맥상 자연스러워 보이는 답변을 선택할 수 있습니다. 셋째, 질문이 불명확하거나 중의적인 프롬프트를 제시할 경우, 모델은 임의의 추측을 포함해 응답할 수 있습니다. 마지막으로 훈련 목적과 사용 환경의 불일치로 발생합니다. 모델은 훈련 시점의 데이터 패턴에 맞추어 학습되기 때문에, 실제 사용 환경에서 새로운 개념, 최신 사건, 도메인 특화 정보를 요구받으면 근거 없이 가장 그럴듯한 정보를 생성하려는 경향이 나타납니다.

환각을 완화하는 방법
환각은 AI가 제공하는 정보의 신뢰성을 떨어뜨리고, 사용자가 잘못된 판단을 내리게 할 위험이 있습니다. 특히 의료·법률·교육·정책 등 정확성이 중요한 분야에서는 잘못된 정보가 실제 피해로 이어질 수 있어 더욱 주의가 필요합니다. 이를 해결하기 위해 여러 대응 전략이 개발되고 있습니다. 첫째, 고품질 학습 데이터 확보를 통해 잘못된 패턴이 모델에 학습되지 않도록 하는 방식입니다. 둘째, 사실 검증 기반 필터링을 출력 단계에 적용해 모델의 응답을 점검하고 보완합니다. 셋째, 검색증강생성(RAG)과 같이 외부 지식 기반을 결합해 사실적 정확도를 높이는 방법이 활용됩니다. 넷째, 사용자 측면에서는 명확하고 구체적인 프롬프트를 제공해 모델의 추측을 최소화할 수 있습니다.

080 AI 가드레일
AI Guardrails

AI가 위험한 행동이나 부적절한 출력을 하지 않도록 제한하는 안전 장치
- 잘못된 정보, 유해 콘텐츠, 개인정보 노출, 범죄 조장 등 다양한 위험을 줄이기 위한 보호 체계
- AI가 규범·정책·윤리 기준을 벗어나지 않도록 입력·출력·추론 과정을 조정하는 기술·운영적 장치

AI 가드레일의 개념
AI 가드레일은 생성형 AI가 안전 기준을 벗어난 응답을 생성하거나 위험한 행동을 유발하지 않도록 경계를 설정하는 안전 장치를 의미합니다. 생성형 AI는 사용자의 요구를 유연하게 수용하는 특성이 있어, 적절한 제한이 없다면 잘못된 사실, 유해 표현, 편향된 판단, 개인정보 노출, 불법·유해 행위 조장과 같은 문제가 발생할 수 있습니다. 이를 방지하기 위해 가드레일은 AI가 어떤 질문에 어떻게 응답해야 하는지, 어떤 범위에서는 응답을 제한해야 하는지를 미리 정의해 모델이 안전한 규칙 내에서 동작하도록 유도합니다. 이 과정은 단순 차단이 아니라, 위험 상황을 인식해 적절한 대체 정보 제공이나 표현 조정 등을 수행함으로써 활용성과 안전성의 균형을 동시에 확보하는 데 목적이 있습니다.

출처 : THE AI

AI 가드레일의 종류
AI 가드레일은 적용 목적과 단계에 따라 여러 유형으로 나뉩니다. 입력 가드레일은 사용자가 위험한 질문이나 규범을 벗어난 요구를 할 경우 이를 감지해 적절히 거부하거나 안전한 형태로 재구성합니다. 다음으로 출력 가드레일은 모델이 부정확한 정보, 유해 표현, 개인정보 등을 생성하지 않도록 결과물을 점검하고 필요 시 수정·차단합니다. 또한 모델 자체의 행동 규칙을 정의하는 시스템 가드레일이 있어, 모델이 준수해야 할 목적과 제한 범위를 내부적으로 설정합니다. 여기에 프롬프트 인젝션 등 공격을 차단하는 보안 가드레일, 특정 산업 규제나 윤리 기준을 반영하는 정책 가드레일이 함께 적용되어 AI의 안전성을 다층적으로 보호합니다.

081 AI 가속기

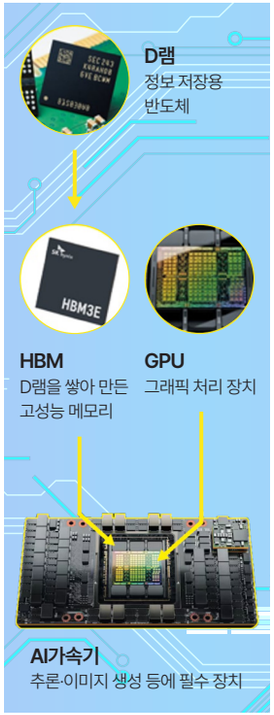
AI Accelerator

AI 연산을 고속·고효율로 처리하는 전용 장치

- GPU·NPU·TPU 등을 묶어 부르는 말로, 많은 계산을 동시에 처리해 대규모 학습과 실시간 추론을 빠르게 수행하도록 설계된 핵심 하드웨어 구성
- 데이터 센터와 에지에서 전력·냉각 체계와 함께 운용되는 AI 서비스 기반 요소

AI 가속기의 개념

AI 가속기는 여러 계산을 동시에 수행하도록 설계된 장치로, 대규모 연산을 효율적으로 처리해 학습 속도와 응답 성능을 높입니다. 중앙처리장치(CPU)가 순차적으로 작업을 수행한다면, 가속기는 수많은 연산을 병렬로 수행해 AI 학습과 추론을 가속합니다. 대표적인 형태로는 GPU, NPU, TPU가 있으며, 성능은 연산 칩뿐 아니라 메모리, 전력, 냉각, 네트워크가 얼마나 효율적으로 결합되는지에 따라 달라집니다. 즉, AI 가속기는 연산 코어와 메모리, 연결망, 전력 공급, 냉각 체계가 하나로 통합된 시스템 단위 장치로 이해할 수 있습니다.



출처 : 조선일보

AI 가속기와 AI 반도체의 차이

AI 반도체는 실제 연산을 수행하는 칩 수준의 부품이고, AI 가속기는 그 반도체를 탑재해 작동하도록 만든 장치 수준의 구성체입니다. 가속기 내부에는 반도체 칩 외에도 고속 메모리(HBM·D램), 전력 공급 장치, 냉각 시스템, 네트워크 연결부, 제어용 소프트웨어가 함께 포함됩니다. 반도체가 성능의 '엔진'이라면, 가속기는 그 엔진이 안정적으로 작동하도록 돕는 '전체 장치'입니다. 따라서 반도체는 회로 구조와 처리 효율 같은 기술 사양이 중심이지만, 가속기는 실제 환경에서의 운용 효율과 안정성이 핵심입니다.

AI 가속기의 중요성

AI 가속기는 대규모 모델 학습과 실시간 서비스 운영을 가능하게 하는 핵심 장치로, 학습 단계에서는 방대한 데이터를 병렬로 처리해 시간을 단축하고, 서비스 단계에서는 빠른 응답 속도로 사용자 경험을 향상시킵니다. 또한 전력 효율과 냉각 성능은 운영비와 환경 영향을 결정하는 중요한 요소입니다. 효율적인 인프라를 구축한 기관일수록 안정성과 지속 가능성이 높으며, 에지 단말에서는 제한된 자원에서도 성능을 유지하는 경량형 NPU가 활용되어 클라우드와 현장이 유기적으로 연결된 AI 생태계를 형성하고 있습니다.

082 AI 격차

AI Divide

AI 기술과 활용 능력 수준의 차이로 인한 사회·산업 간 불균형

- 데이터·인프라·인력·정책의 차이로 국가·기업·개인 간 기술 발전 속도와 활용 격차가 커지는 현상
- 기술 접근성·활용 역량의 차이로 인해 사회적·경제적 불평등이 심화되는 구조적 문제이며, 개방·공유·교육을 통한 포용적 AI 생태계 조성이 대응의 핵심

AI 격차의 개요

AI 격차는 인공지능 기술을 개발하고 활용하는 능력의 차이에서 비롯된 불균형을 의미하며, 양질의 데이터, 연산 인프라, 인력, 제도적 지원이 고르지 않아 일부 국가와 기업이 기술 주도권을 차지하는 현상입니다. 과거 '디지털 격차'가 인터넷 접근의 문제였다면, AI 격차는 경제력·산업 경쟁력·사회 기회의 차이로 확산된 구조적 문제입니다. 대기업은 방대한 데이터를 활용해 시장을 선점하지만, 중소기업과 공공 부문은 기술 투자와 인력 확보에 제약을 받습니다. 국가 간 기술 수준도 빠르게 벌어져 선도국은 표준과 산업 생태계를 주도하고, 후발국은 수용자 역할에 머무르는 구조가 고착화되고 있습니다. 또한 개인 간에도 AI 이해도와 활용 능력의 차이가 커지며, 이는 교육·고용·소득 격차로 이어지고 있습니다.

AI 격차의 원인

AI 격차는 데이터, 인프라, 인력, 제도 등 여러 요인의 복합적 불균형에서 발생합니다. 대기업은 양질의 데이터를 지속적으로 확보하지만, 중소기업과 공공기관은 데이터 수집과 활용 역량이 부족합니다. GPU, AI 반도체, 클라우드 등 고성능 인프라를 안정적으로 확보한 조직이 기술 개발에서 우위를 점하고, 핵심 인력이 특정 국가와 기업에 집중되면서 지식 격차가 커지고 있습니다. 또한 데이터 규제, AI 윤리 기준, 연구개발 투자 수준의 차이는 국가 간 기술 성장 속도 차이를 심화시킵니다.

AI 격차의 완화

AI 격차를 줄이기 위해서는 포용적 디지털 전환이 필요합니다. 공공 데이터 개방을 확대하고, 중소기업·연구기관이 고성능 연산 자원을 공동 활용할 수 있는 환경을 마련해야 합니다. 초·중등 단계부터 AI 기초 교육을 정규화하고, 비전공자 대상 실무 교육을 강화할 필요가 있습니다. 또한 독점적 시장 구조를 완화하고 공공·스타트업이 협력할 수 있는 표준화 체계를 구축해야 합니다. 나아가 국제 협력을 통해 윤리·데이터 보호 규범을 마련하고, 개발도상국의 기술 격차 완화를 지원하는 것도 중요합니다.

083 AI 네이티브

AI Native

AI를 사회·기술·조직의 기본 전제로 삶과 구조가 형성된 주체 및 체계

- AI가 사회의 기본 인프라로 작동하는 환경에서 태어나거나 이를 전제로 설계·구성된 대상
- AI를 외부 도구가 아닌 핵심 인프라로 받아들이는 세대·시스템·조직·문화 전반을 포괄

AI 네이티브의 개념

AI 네이티브는 AI를 단순히 '활용'하는 수준을 넘어, AI를 전제로 한 환경 속에서 형성되거나 설계된 주체를 의미합니다. 이는 인터넷을 당연한 기반으로 성장한 '디지털 네이티브' 개념이 AI 시대로 확장된 형태입니다. AI가 단순한 도구나 보조 수단을 넘어 의사결정과 창작, 서비스 운영의 핵심 기반이 되면서, 개인뿐 아니라 시스템과 조직 역시 AI를 전제로 작동하도록 구성되고 있습니다. 다시 말해 AI 네이티브는 AI를 '활용'하는 주체가 아니라 AI가 내재된 환경 속에서 살아가거나 작동하는 주체를 뜻합니다. 이는 개인의 생활 습관, 조직 운영, 산업 구조, 문화 전반을 아우르는 개념으로, AI 중심 사회의 기본 단위이자 변화의 원동력으로 이해할 수 있습니다.

AI 네이티브의 유형

AI 네이티브의 핵심 특징은 AI를 전제로 한 사고와 구조입니다. 사회적 측면에서 AI 네이티브는 AI 기술이 일상과 조직 운영 전반에 깊이 통합된 환경을 전제로 형성된 사회적 주체 및 구조를 의미합니다. 이 환경에서는 학습, 소통, 창작 등 주요 활동이 AI와의 상호작용을 기본으로 이루어집니다. 기술적 측면에서는 AI 중심 구조의 시스템과 서비스를 뜻하며, 데이터 분석, 예측, 콘텐츠 생성 등 핵심 기능이 AI로 자동화된 AI 내장형 애플리케이션과 플랫폼이 대표적입니다. 이러한 기술은 인간의 개입보다 AI의 판단을 우선하며, 새로운 사용자 경험을 형성합니다. 조직적 측면에서는 업무 프로세스 전반에 AI가 내재화된 형태를 가리킵니다. 기획·생산·고객 대응 등 주요 과정이 AI 기반으로 운영되며, 인간의 역할은 감독과 가치 판단 중심으로 이동하고 있습니다. 결국 AI 네이티브는 개인·기술·조직이 서로 연결되어 AI 중심 사고와 행동 양식을 공유하는 사회적 집단 및 구조적 현상으로 이해됩니다.

AI 네이티브의 의의

AI 네이티브는 AI가 기술의 수단이 아니라 사회의 기본 구조로 자리 잡아가고 있음을 보여줍니다. 이들은 AI와의 협업을 전제로 사고하며, 인간과 AI의 역할 경계를 재정의합니다. 기업과 기관은 AI 네이티브 세대의 가치관과 습관에 맞춘 서비스·정책 설계가 필요합니다. 동시에 자동화 의존, 정보 편향, 개인정보 보호 등 새로운 윤리적 과제도 나타납니다. 결국 AI 네이티브는 AI 시대의 표준 사용자이자 새로운 사회적 기준을 형성하는 주체로, 기술·교육·정책 전반에 변화를 이끌어가고 있습니다.

참조 용어 | AI 가속기 p.94

084 AI 데이터 센터

AI Data Center

대규모 데이터를 저장·처리하며 AI 연산을 지원하는 디지털 인프라

- 서버·스토리지·네트워크를 통합해 데이터를 안정적으로 저장·전송·분석하고, 클라우드와 AI 서비스의 학습·추론을 수행하는 핵심 기반
- 전력·냉각·보안 체계를 포함하여 AI 산업 전반의 효율성과 지속 가능성을 뒷받침하는 복합 시설

AI 데이터 센터의 등장

데이터 센터는 대량의 데이터를 저장·처리하기 위한 핵심 인프라입니다. 초창기에는 기업 등의 전산실 형태로 시작해 웹사이트와 이메일 등 기본 온라인 서비스를 지원했지만, 클라우드 확산 이후 전 세계 디지털 생태계의 중심으로 자리 잡았습니다. 최근에는 단순한 저장 공간을 넘어 AI 학습과 추론을 담당하는 연산 시설로 진화하며 전력 공급, 냉각 기술·네트워크 구조 등 운영의 초점이 저장 효율에서 연산 효율로 이동했습니다.

AI 데이터 센터의 특징

AI 특화 데이터 센터는 기존 저장 중심 구조와 달리 병렬 연산과 고속 데이터 이동을 중심으로 설계됩니다. GPU·NPU 서버가 대량 연결되고 초고속 네트워크와 고효율 냉각이 결합되며, 전력 사용 밀도가 높아지면서 전력 분배와 열 제어가 핵심 과제가 되었습니다. 또한 실시간 학습·추론에 필요한 안정적 데이터 전송 능력이 요구되고, 작업 스케줄링과 로그 관리 등 연산 전반이 통합적으로 관리되는 체계로 발전하고 있습니다.

AI 데이터 센터의 구성 및 규모

데이터 센터는 정보 저장소이자 모델 학습과 서비스 품질을 결정하는 핵심 기반입니다. 초거대 AI는 막대한 연산 자원과 안정적 전력 공급이 필요하며, GPU·NPU 같은 고성능 장비와 냉각·네트워크·스토리지가 유기적으로 작동해야 합니다. 이러한 운영 효율은 학습 속도, 응답 시간, 비용에 직접 영향을 주며, 공공 서비스의 경우 가용성과 보안 수준이 행정 신뢰와 연결됩니다. 하이퍼스케일 데이터센터의 규모는 매우 다양하지만, 20MW 규모 데이터센터의 경우, 대략 축구장 3개 크기인 22,500㎡(약 6,800평) 정도 면적에 10만 대 이상의 서버를 수용하고 있습니다.

AI 데이터 센터의 전망

전력 소비와 탄소 배출을 줄이기 위한 에너지 절감, 재생에너지 전환, 고효율 냉각이 AI 데이터센터의 주요 과제입니다. 앞으로 데이터 센터는 고성능·친환경·안정성을 갖춘 형태로 발전하며, 단순한 IT 인프라를 넘어 국가와 산업의 AI 경쟁력을 뒷받침하는 핵심 기반이 될 것입니다.

085 AI 레드티밍

AI Red Teaming

AI의 취약점을 공격자 관점에서 시험해 위험을 식별하는 검증 절차

- 비정상 입력·악의적 프롬프트·사회적 편향 등 실제 위협 상황을 모의해, AI 시스템의 안전성과 신뢰성을 사전에 평가하는 점검 활동
- 단순 성능 시험이 아닌, AI의 예상 밖 행동과 사회적 영향을 탐지·완화하기 위한 선제적 검증 체계

AI 레드티밍의 개요

AI 레드티밍은 AI 시스템의 위험 요인과 취약점을 사전에 식별·검증해 안전성과 신뢰성을 확보하는 과정입니다. 원래 군사·보안 분야에서 '적의 시각으로 방어 체계를 점검한다'는 의미로 쓰이던 레드팀(red team) 개념을 AI 안전 관리에 적용한 것입니다. 이 과정에서는 공격자 관점에서 AI 모델을 시험해 비정상 입력이나 악의적 명령에 대한 반응을 분석하고, 편향된 응답, 유해 콘텐츠, 정보 왜곡, 보안 우회 등의 문제를 찾아냅니다. 즉, AI 레드티밍은 단순한 오류 탐색이 아니라 AI의 윤리·보안·신뢰성을 검증하는 핵심 절차입니다.

AI 레드티밍 수행 방식

AI 레드티밍은 보통 공격적 시험, 시나리오 검증, 위험 분석의 세 단계로 진행됩니다. 공격적 시험은 비정상 입력이나 우회 프롬프트를 주어 오작동 여부를 점검하고, 시나리오 검증은 실제 사용 환경을 모의해 편향된 출력이나 부적절한 응답을 평가합니다. 이후 위험 분석 단계에서는 탐지된 문제를 분류하고 대응 정책, 데이터 필터링 등 개선책을 마련합니다. 이러한 검증은 내부 보안팀 외에도 외부 전문가나 독립 기관이 참여해 객관성을 확보하며, 최근에는 정부나 민간의 AI 안전성 평가 제도와 연계해 정기적 검증 체계로 확산되고 있습니다.

AI 레드티밍의 중요성

AI 레드티밍은 생성형 AI 확산에 따라 필수적인 신뢰성 검증 절차로 자리 잡고 있습니다. 생성형 AI 모델 내부의 작동 원리를 완전히 통제하기 어렵고, 예측 불가능한 출력이 사회적 문제로 이어질 수 있기 때문에, 기술적 안전성과 사회적·윤리적 영향을 함께 평가해야 합니다. 결국 AI 레드티밍은 안전하고 책임 있는 AI 구현을 위한 선제적 관리 체계이자, 기술 혁신과 사회적 신뢰를 함께 확보하는 핵심 기반입니다.

086 AI 리터러시

AI Literacy

AI를 이해하고 활용할 수 있는 개인과 조직의 역량

- AI의 원리와 사회적 영향을 이해하고 윤리적으로 해석·활용할 수 있는 인지·실천적 능력
- 기술 이해를 넘어 AI와 함께 작동하는 사회에서 합리적 판단·참여를 가능케 하는 디지털 시민 역량

AI 리터러시란?

AI 리터러시는 AI의 작동 원리와 한계를 이해하고, 이를 사회적으로 책임 있게 활용할 수 있는 능력을 의미합니다. 단순한 도구 조작 능력을 넘어, AI가 어떤 데이터를 기반으로 판단·생성하는지 이해하고 결과의 의미와 한계를 비판적으로 파악하는 사고력을 포함합니다. 다시 말해, AI 리터러시는 'AI를 사용할 줄 아는 능력'과 'AI를 올바르게 이해하는 능력'을 아우르는 통합적 역량입니다. 이러한 능력은 AI가 일상화된 사회에서 인간이 기술과 협력하며 스스로 판단하고 의사결정을 내릴 수 있도록 하는 시민적 기본 소양으로 작용합니다.

AI 리터러시의 중요성

AI 리터러시는 기술 활용의 민주성과 신뢰성을 확보하기 위한 필수 기반입니다. AI가 제공하는 정보는 편리함과 효율성을 높이지만, 동시에 오류·편향·허위정보의 위험을 내포합니다. 이를 구별하고 검증할 수 있는 능력이 없다면 개인의 판단과 사회적 의사결정이 왜곡될 수 있습니다. 따라서 AI 리터러시는 단순한 학습 주제가 아니라 교육·직무·정책 전반에 걸친 핵심 과제로 다뤄져야 합니다. 특히 리터러시 수준의 격차는 새로운 형태의 디지털 불평등을 초래할 수 있어, 국가적 차원의 체계적 대응이 필요합니다.

AI 리터러시의 영향과 정책

출처 : freepik

AI 리터러시의 확산은 기술 발전 속도에 사회가 적응할 수 있는 기반을 마련하는 핵심 과정입니다. 리터러시 수준이 높을수록 개인은 정보의 신뢰성을 스스로 판단하고, 사회는 AI를 투명하고 책임 있게 운용할 수 있습니다. 반대로 낮은 리터러시는 허위정보 확산과 기술 불신을 심화시켜 사회적 신뢰를 약화시킵니다. 이러한 위험을 줄이기 위해서는 교육과 공공 인식 확산을 병행하는 다층적 전략이 필요합니다. 학교에서는 데이터 이해와 비판적 사고를 중심으로 한 교육을 강화하고, 산업 현장에서는 실무 중심의 활용 교육이 필요합니다. 정부·기업·언론은 협력해 시민 대상 프로그램과 표준화된 평가 지표를 마련하고, 지역·세대별 격차를 줄이는 맞춤형 교육을 시행할 필요가 있습니다. 궁극적으로 AI 리터러시는 기술 혁신을 사회 전체가 신뢰 속에 수용하도록 만드는 사회적 인프라로 기능합니다.

참조 용어 | 그래픽 처리 장치(GPU) p.19

087 AI 반도체

AI Semiconductor

AI 학습·추론을 위한 대규모 병렬 연산용 특화 반도체

- 대량의 데이터를 동시에 계산해 AI 학습·추론 속도를 높이고 전력 소모를 줄이는 핵심 반도체 기술
- 클라우드와 데이터 센터, 에지 기기 등에서 AI 서비스 성능과 비용을 좌우하는 핵심 부품

AI 반도체 개요

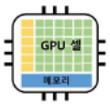
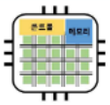
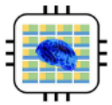
AI 반도체는 인공지능이 방대한 데이터를 학습하고 추론할 때 필요한 연산을 효율적으로 처리하도록 설계된 칩입니다. 크게 범용 반도체(CPU·GPU)와 특수 목적 반도체(ASIC)로 구분되며, ASIC에는 TPU·NPU 등이 포함됩니다. 기존 중앙처리장치(CPU)가 순차적으로 작업을 수행한다면 GPU, TPU, NPU 등은 대규모 행렬·벡터 연산을 병렬로 처리하도록 설계돼 속도와 효율이 더욱 높습니다.

AI 반도체의 중요성

AI 반도체는 기술 발전 속도와 품질을 좌우하는 핵심 하드웨어입니다. 대규모 모델 학습과 추론은 막대한 연산량과 데이터 이동을 요구하므로, AI 반도체의 성능은 학습 시간·비용·정확도에 직접적인 영향을 줍니다. 초거대 모델과 생성형 AI의 확산으로 연산 수요가 폭발적으로 증가함에 따라, 범용 칩만으로는 대응이 어렵습니다. 이에 따라 고성능·저전력·고밀도 연산이 가능한 AI 전용 칩 수요가 급증하고 있습니다.

AI 반도체의 전망

AI 반도체 시장은 기존 GPU 중심 구조에서 TPU·NPU 등 특정 AI 작업에 최적화된 ASIC 기반 구조로 다양화되고 있습니다. 미국·유럽·중국·한국 등 주요국은 반도체를 전략 산업으로 지정하며 연구개발과 공급망 투자를 강화하고 있고, 기업들 역시 AI 모델 특성에 맞춘 맞춤형 ASIC 개발을 통해 성능 우위를 확보하고 있습니다. 여기에 메모리 내 연산으로 데이터 이동을 줄이고 속도·효율을 높이는 인메모리 컴퓨팅, 여러 개의 소형 칩을 조립해 확장성과 생산 효율을 높이는 칩렛 아키텍처와 같은 차세대 기술이 확산되면서 성능과 효율 측면의 경쟁이 더욱 치열해지고 있습니다. 향후 에너지 효율, 탄소 저감, 보안성이 핵심 과제로 부상할 것으로 보이며, 특히 독자적 칩 설계 능력을 갖춘 국가와 기업이 AI 산업의 주도권을 확보할 것이라 전망됩니다.

종류	CPU(1세대)	GPU(1세대)	FPGA(2세대)	ASIC(2세대)	뉴로모픽(3세대)
특징					
	복잡 계산 순차처리	단순 계산 병렬처리	목적별 하드웨어 재구성	용도 맞춤형 고효율 전용칩	뉴런과 시냅스를 모방한 新구조

세대별 반도체

출처 : 특허청

참조 용어 | AI 안전 p.102, AI 윤리 p.107

088 AI 신뢰성

AI Trustworthiness

안전하고 공정하며 투명하게 작동하는 신뢰할 수 있는 AI 체계

- AI가 사회적 가치와 법적 기준을 충족하면서 예측 가능하게 작동하도록 설계된 시스템을 의미
- 인간 중심의 윤리 원칙을 바탕으로 안전성·공정성·설명 가능성을 확보한 AI를 지칭

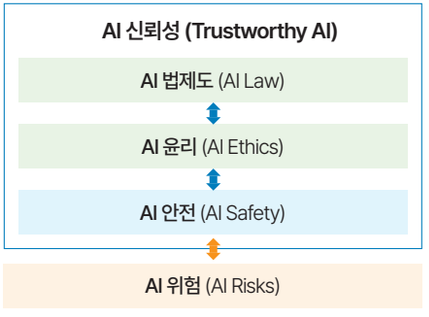
AI 신뢰성의 개념

AI 신뢰성은 인공지능이 사회와 개인에게 안전하고 책임 있게 작동하는지를 판단하는 핵심 기준입니다. 단순히 성능이 뛰어난 AI가 아니라, 사용자가 AI의 결과를 이해하고 신뢰할 수 있도록 안전성과 투명성을 확보한 상태를 의미합니다. 이는 AI 안전과 AI 윤리를 포괄하는 상위 개념으로, AI가 정확하게 작동하고 오류나 편향을 최소화한 '안전성'을 갖추는 것은 신뢰의 기술적 전제이며, 인권과 공정성, 투명성 같은 '윤리 원칙'은 그 신뢰를 사회적 차원으로 확장시키는 기반이 됩니다.

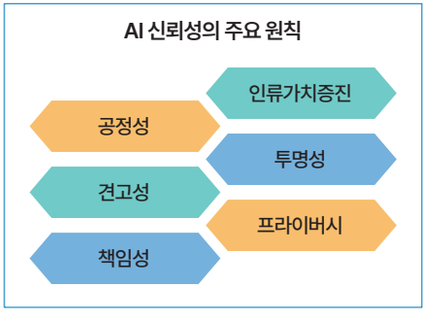
AI 신뢰성의 주요 원칙

AI 신뢰성의 개념 정의는 기관, 학자에 따라 조금씩 다르지만, 대체로 공정성, 견고성, 책임성, 인류가치증진, 투명성, 프라이버시 보호 등의 원칙으로 구성되며 국제표준 및 국가별 평가·인증의 기초가 되고 있습니다.

- 공정성: 알고리즘 편향을 방지하고 특정 집단의 불이익을 차단
- 견고성: 외부 공격·데이터 변동에도 안정적으로 작동
- 책임성: AI 결과에 대해 명확한 책임 주체를 설정
- 인류 가치 증진: 기술 발전이 인간의 존엄성·공공선과 조화되도록 함
- 투명성: 의사결정 과정을 이해 가능한 형태로 공개
- 프라이버시 보호: AI 전 과정에서 개인정보 오용을 최소화하고 데이터 주권을 보장



출처 : AI 안전의 개념과 범위 (SPRi)



출처 : AI 신뢰성 및 윤리 제도 연구 (SPRi)

089 AI 안전

AI Safety

AI의 예측 불가능한 위험으로부터 인간과 사회를 보호하도록 설계된 체계

- AI의 오작동·악용·편향 등 잠재적 위험을 사전에 식별·통제해 안정성을 확보하는 관리 체계
- AI 신뢰성과 윤리의 기반이 되는 기술적 전제이자 사회적 안전장치

AI 안전의 개념

AI 안전은 인공지능이 인간의 의도와 일관되게 작동하면서 사회적 위해를 최소화하도록 설계·관리하는 체계를 말합니다. 초기에는 기술적 오작동이나 오류 방지를 의미했지만, 현재는 AI의 자율성과 범용성이 커지면서 예측 불가능한 판단·악용·편향 등 사회적 위험까지 포괄하게 되었습니다. 이는 설계–개발–배포–운영–폐기의 AI 생명주기 전반에 걸친 위험 식별, 평가, 대응을 포함합니다. 결과적으로 AI 안전은 AI 신뢰성의 하위 요소이자, AI 윤리를 현실화하는 실천적 기반으로 작동합니다.

AI 안전의 구성 요소

AI 안전의 핵심 구성 요소는 예측 가능성, 견고성, 인간 통제 가능성, 검증 가능성, 책임성 등으로 요약됩니다.

- 예측 가능성: AI가 의도된 목적 내에서 일관되게 작동하도록 보장하는 능력
- 견고성: 데이터 오류·적대적 공격·환경 변화에도 안정적 성능을 유지
- 인간 통제 가능성: 시스템이 자율적으로 판단하더라도 인간이 개입·중단할 수 있는 가능성
- 검증 가능성: 모델의 의사결정 과정과 결과를 외부에서 평가·검증할 수 있도록 기록·투명화
- 책임성: 사고 시 명확한 책임 주체와 대응 절차를 확보

AI 안전에 관한 국제적 정책 동향

AI 안전은 각국의 AI 정책에서 중요한 요소로 부상하고 있습니다. 영국에서 개최된 AI Safety Summit을 계기로 국제 AI 안전연구소(AISI) 네트워크가 출범했으며, 참석국들은 국제 AI 안전보고서를 발간하기로 약속했습니다. AISI 네트워크에는 영국, 미국, 캐나다, 일본, 싱가포르, 한국 등 주요국이 참여하고 있으며, 초거대 모델의 위험 평가, 공동 테스트 데이터셋 구축, 검증 기준의 국제 정합성 확보 등 여러 다양한 AI 안전 이슈를 다룹니다. 이외에도 EU는 AI Act를 통해 위험 기반 접근을 제도화하고 고위험 시스템에 대한 사전 평가와 인증을 요구하고 있습니다. 우리나라 역시 AI 신뢰성 검증 가이드라인과 「AI 기본법」 추진과 함께 AISI 참여를 통해 국제 협력 기반을 강화하고 있습니다.

참조 용어 | AI 에이전트 p.104

090 AI 어시스턴트

AI Assistant

사용자의 요구를 이해하고 과업을 수행하는 대화형 AI 도우미

- 사용자의 의도를 파악해 정보 탐색·업무 지원·의사결정 보조를 수행하는 언어 기반 AI 시스템
- 맥락을 이해하고 인간의 언어로 소통하며 개인화된 지원을 제공하는 협력형 기술

AI 어시스턴트 개요

AI 어시스턴트는 인간의 언어를 이해하고 대화하듯 상호작용하며, 사용자의 목표 달성을 지원하는 지능형 시스템입니다. 초기 어시스턴트는 단순히 명령을 인식해 정보를 제공하는 수준에 머물렀지만, 이후 딥러닝 기반 언어모델과 음성 합성 기술의 발전으로 문맥 이해와 다중 작업 처리 능력이 향상되며 본격적으로 확산됐습니다. 특히 LLM의 등장은 AI가 인간의 의도를 유연하게 해석하고 복잡한 요청을 처리할 수 있는 기반을 마련했습니다. 최근에는 개인화 기술, 멀티모달 AI, 클라우드 연동 등이 결합되어 사용자의 대화 이력, 일정, 업무 패턴을 학습하고 맞춤형 조언과 실행을 제공하는 지능형 협력 파트너로 발전하고 있습니다.

AI 어시스턴트의 활용

AI 어시스턴트는 개인 생활, 산업, 공공서비스 등 다양한 영역에서 활용됩니다. 개인은 일정 관리, 정보 탐색 등 반복적 작업을 맡겨 효율성을 높이고, 기업은 문서 작성, 보고서 요약, 회의록 정리, 데이터 분석 지원 등 사무 자동화에 활용합니다. 행정 서비스에서는 민원 안내와 상담 응대에 적용되며, 산업 현장에서는 지식 검색, 매뉴얼 요약, 안전 점검 보조 등 현장 의사결정 보조 도구로 활용되고 있습니다.

AI 어시스턴트 vs AI 에이전트

AI 어시스턴트는 사용자의 맥락을 이해하고 여러 도구와 연계해 작업을 직접 실행하는 능동형 도우미입니다. AI 챗봇은 사용자와의 '대화'를 통해 단순 질의에 대해 반응하는 AI 어시스턴트 초창기 모델의 대표격으로 볼 수 있습니다. 이에 반해, AI 에이전트는 어시스턴트보다 높은 자율성과 판단력을 지닌 형태로, 사용자의 개입 없이 스스로 목표를 세우고 계획·수행합니다. 즉, AI 에이전트는 AI 어시스턴트와는 달리 사람과의 상호작용이나 명령 없이 자율적으로 상황을 인식하며 필요한 조치를 실행하는 시스템입니다. 요약하자면 어시스턴트는 명령에 대한 반응과 협력 중심, 에이전트는 자율 실행 중심의 시스템입니다.

관련 용어

AI 챗봇 (AI Chatbot)

사용자의 질문이나 요청을 문자·음성 형태로 입력받아 자동으로 응답을 생성하는 대화형 소프트웨어입니다. 초기에는 규칙 기반으로 단순 문의에만 답변했지만, 최근 LLM 기반 언어 모델이 적용되면서 문맥 파악, 장문의 대화, 복잡한 질의 응답이 가능해졌으며, 일상적 문의 대응과 고객지원 자동화 등에 활용되고 있습니다.

091 AI 에이전트

AI Agent

자율적으로 판단·행동하여 인간의 목표를 대신 수행하는 지능형 AI 시스템

- AI가 인간의 지시 없이 스스로 상황을 인식하고 계획·실행·평가를 반복하며 과업을 수행하는 기술
- 단순 자동화를 넘어 목표 중심적 의사결정과 협업이 가능한 능동적 AI 형태

AI 에이전트의 개념

AI 에이전트는 AI가 인간의 명령없이 스스로 사고하고 행동하는 지능형 소프트웨어를 의미합니다. 기존 생성형 AI가 정보를 생성하는 데 그쳤다면, AI 에이전트는 목표 달성을 위해 계획을 수립하고 필요한 데이터를 수집하며, 외부 시스템과 연동해 태스크를 실행합니다. 챗봇이나 자동화 도구가 사람의 명령에 따라 반응했다면, AI 에이전트는 스스로 판단해 가장 적절한 방법을 선택하고 실행 결과를 평가해 다음 행동을 조정합니다. 이러한 AI 에이전트는 인간의 단순 보조가 아닌 디지털 동료로서 역할하며, 비즈니스와 사회 전반에서 새로운 업무 방식과 효율성을 만들어갑니다.

AI 에이전트의 의의

AI 에이전트는 자동화에서 자율화로의 전환을 상징하며, 기업과 사회가 AI를 실제 인프라로 활용하는 전기를 마련했습니다. AI 에이전트는 24시간 작동하며 반복적 업무를 대신 수행해 업무 부담을 완화하고, 인간이 창의적이고 고부가가치 업무에 집중할 수 있도록 돕습니다. 이로 인해 업무의 자율성과 효율성이 동시에 확대되었지만, 데이터 보안·개인정보 보호·책임성·투명성 등 새로운 과제도 함께 등장했습니다. 특히 AI의 판단 오류나 편향 문제는 법적 책임의 불명확성을 초래할 수 있으며, 예기치 않은 상황에 대한 적응력 부족 역시 기술적 한계로 지적됩니다.

구분	AI 에이전트 (AI Agent)	다중 에이전트 시스템 (Multi-Agent System, MAS)	에이전틱 AI (Agentic AI)
개념	인간의 직접 명령 없이 목표를 추론·계획·행동하는 지능형 소프트웨어	여러 에이전트가 상호작용하며 공동 목표를 수행하는 분산·협력 구조	개별 에이전트가 자율성·협동성을 결합해 환경 변화에 따라 스스로 조정·재계획하는 확장된 에이전트 구조
작동 방식	단일 에이전트가 [정보 수집 → 추론 → 계획 → 실행 → 평가]의 사이클로 작업 수행	에이전트 간 역할 분담 및 정보 공유(A2A)로 다단계 작업을 분산 처리	LLM과 오케스트레이션 계층을 통해 에이전트들이 방향성을 유지하며 상황 변화에 따라 지속적 재계획 수행
특징	목표 달성 중심의 단일 수행자, API·도구 사용 기반 실행 능력	정의된 규칙 기반의 협력·조정·분산 의사결정, 집단적 지능 구현	적응성·유연성 강화, 복잡한 작업을 지속적 조정해 해결하는 동적 생태계
적용 사례	고객 상담, 일정 관리, 데이터 분석, 개인 비서형 서비스	제조·물류·재난 대응 등 다중 요소가 얽힌 시스템	전체 프로세스 자동화, 대규모 운영 시스템, 미래형 AI 거버넌스 구조

092 AI 오케스트레이션

AI Orchestration

여러 AI 기능을 통합해 하나의 지능형 시스템으로 운영하는 기술

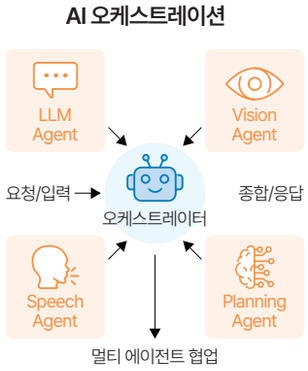
- 다양한 AI 모델과 데이터를 연결해 협력적으로 작동하도록 설계된 통합 제어 구조
- AI 간 협업을 자동화하고 효율적으로 관리해 복잡한 업무를 유기적으로 수행하는 관리 체계

AI 오케스트레이션 개요

AI 오케스트레이션은 서로 다른 AI가 하나의 목표를 위해 함께 움직이도록 조정하는 기술입니다. 기존에는 개별 AI 모델이 독립적으로 작동하며 데이터 단절이나 같은 연산을 중복해 수행하는 등 비효율이 발생했지만, 오케스트레이션은 이를 해결해 AI가 연속적인 흐름속에서 협력하도록 만듭니다. 즉, 여러 AI가 각각의 전문 기능을 수행하되, 중앙의 오케스트레이터가 전체 과정을 제어해 하나의 지능형 생태계처럼 작동하도록 설계합니다. 이는 복잡한 업무를 효율적으로 처리하고, 다양한 AI 기술을 통합적으로 활용할 수 있습니다.

AI 오케스트레이션의 작동 방식

AI 오케스트레이션은 AI 통합, AI 자동화, AI 관리의 세 요소로 작동합니다. AI 통합은 다양한 모델·데이터·업무 시스템을 하나의 프로세스로 연결해 상호작용을 가능하게 합니다. AI 자동화는 사람이 개입하지 않아도 AI가 스스로 작업을 배분·실행하도록 하는 기능으로, 자원 활용을 최적화하고 응답 속도와 정확성을 높입니다. AI 관리는 품질·성능·보안·윤리 기준을 지속적으로 점검·제어해 안정성과 신뢰성을 유지합니다. 중앙의 오케스트레이터가 다중 에이전트의 협업을 지휘하며 전체 시스템의 효율을 극대화합니다.



AI 오케스트레이션의 의의

AI 오케스트레이션은 공공, 산업, 연구 등 다양한 분야에서 AI 활용의 효율성과 확장성을 높이는 핵심 기술입니다. 공공 영역에서는 문서 분석·민원 처리 등 여러 단계를 거치는 행정 업무를 자동으로 연결해 처리 시간을 단축하며, 산업 현장에서는 고객 상담, 제조 공정 관리 등 다양한 AI 기능을 조합해 생산성과 품질을 높입니다. 또한 새로운 AI 모델을 기존 시스템에 쉽게 붙일 수 있어 변화에 빠르게 대응할 수 있습니다. 무엇보다 AI 오케스트레이션은 개별 기술을 단순히 연결하는 수준을 넘어, 여러 AI가 역할을 나눠 협력하며 의사결정까지 지원하는 지능형 운영 체계를 가능하게 한다는 점에서 의미가 큽니다.

093 AI 워터마킹

AI Watermarking

AI 생성 콘텐츠에 식별 정보를 삽입해 출처와 진위를 구분하는 기술

- 이미지·텍스트·음성 등 생성물에 눈에 보이는 혹은 보이지 않는 디지털 표식을 넣어 AI 생성 여부나 생성 주체 등을 판별하는 인증 기술
- 생성형 AI 확산 속에서 저작권 보호, 허위정보 방지, 책임 추적을 지원하는 핵심 기술

AI 워터마킹의 특징

AI 워터마킹은 AI가 만든 콘텐츠 속에 디지털 표식(watermark)을 삽입해 생성물의 출처와 진위를 검증할 수 있도록 하는 기술입니다. 이미지의 픽셀, 텍스트의 단어 배열, 음성의 주파수 등 콘텐츠의 세부 구조 속에 미세한 신호를 암호화하여 저장하는 방식으로 작동합니다. 기존의 워터마크가 단순히 로고 등을 표시하는 수준이었다면, AI 워터마킹은 목적에 따라 가시형·비가시형 형태 모두 적용될 수 있으며, 품질 손상 없이 인식 불가능한 형태로 정보를 삽입할 수 있습니다. 또한 이 기술은 삽입–검출–인증의 세 단계로 구성되어, AI 생성 여부, 생성 주체, 생성 일시 등 다양한 메타데이터를 보존하고 추후 검증할 수 있습니다.

워터마크 적용 워터마크 미적용

Deepmind가 공개한 '신스 ID'

출처 : AI타임스

AI 워터마킹의 활용

AI 워터마킹은 AI 콘텐츠의 신뢰성과 투명성을 보장하는 핵심 기술로, 다양한 영역에서 활용됩니다. 공공 부문에서는 정부 문서나 공공 데이터의 위변조 방지에, 언론·플랫폼에서는 허위정보 확산 억제와 출처 확인에 적용됩니다. 산업 분야에서는 콘텐츠 제작물에 워터마크를 삽입해 저작권 보호와 무단 복제 방지를 실현하고, 교육·연구기관은 학습 데이터의 출처를 추적해 데이터 품질과 모델 신뢰도를 높이고 있습니다. 이 기술은 단순히 생성물을 구분하는 것이 아닌 AI가 만들어낸 결과물의 신뢰성과 책임성을 제도적으로 보장해, AI 투명성 정책·윤리 기준·사회적 신뢰 체계를 구현하는 기반이 됩니다.

AI 워터마킹의 과제

AI 워터마킹은 아직 해결해야 할 기술적 안정성과 표준화의 한계를 안고 있습니다. 이미지 편집, 텍스트 변환 등의 후처리 과정에서 워터마크가 손상되거나 소실될 수 있으며, 일부 공격자는 삭제·변형 알고리즘을 통해 이를 우회할 수도 있습니다. 또한 다양한 생성형 AI 모델이 혼합되어 활용되는 환경에서는 표준화된 삽입·검출 체계를 마련하기 어렵고, 국가나 기업별로 기술 기준이 상이해 상호 인증이 쉽지 않습니다. 이 밖에도 워터마크 삽입이 콘텐츠 품질에 미세한 영향을 줄 수 있고, 개인정보나 모델 정보가 과도하게 노출될 가능성도 제기됩니다. 국제 기술 표준 마련, 삭제 방어 기술 강화, 검출 정확도 향상이 향후 과제로 제기됩니다.

094 AI 윤리

AI Ethics

AI가 인간의 가치와 책임 원칙에 따라 개발·활용되도록 하는 규범 체계

- 공정성·투명성·책임성·인권 존중 등 AI의 설계·운영 전 과정에서 지켜야 할 윤리적 기준을 규정하는 사회적 원칙
- AI의 부정적 영향을 최소화하고 신뢰 가능한 기술 발전을 유도하는 가치 지향적 관리 체계

AI 윤리의 개요

AI 윤리는 AI의 설계, 개발, 활용 전 과정에서 기술의 효율성보다 인간의 존엄성과 사회적 책임을 우선시하며, 인간의 가치와 권리를 중심으로 작동하도록 이끄는 원칙입니다. AI가 의사결정과 사회 시스템에 깊이 관여하면서 편향·차별·책임 불명확성 같은 문제가 현실화되자, 윤리적 통제의 필요성이 커졌습니다. AI 윤리는 단순한 도덕적 권고가 아니라 AI 거버넌스의 핵심 기준으로 작동합니다. AI 윤리는 공정성, 투명성, 설명 가능성, 책임성, 인권 존중, 안전성 등의 가치를 중심으로, AI 안전과 신뢰성의 기반이 되는 가치를 제시합니다.

현실에서의 AI 윤리 이슈

AI 윤리의 핵심 이슈는 AI가 초래하는 사회적 불평등, 정보 왜곡, 인권 침해 위험에 대한 것입니다.

① 데이터 편향·차별: 학습 데이터가 사회의 불균형 반영 시, 의사결정에서 특정 집단을 불리하게 판단
② 투명성·설명 가능성 부족: 복잡한 모델 구조로 인해 AI의 판단 근거가 불명확하면, 결과에 대한 검증이 어렵고 오류 발생 시 책임 소재도 불명확
③ 프라이버시 침해: 개인 데이터를 과도하게 수집하거나 감시 목적으로 활용
④ 가짜 정보의 확산: 생성형 AI는 이미지·영상·텍스트를 현실적으로 만들어낼 수 있으며, 허위 정보나 딥페이크 콘텐츠를 대량 생산·유포하는 데 악용될 경우 사회적 혼란을 초래하고 정보 신뢰를 약화

AI 윤리의 의의

AI 윤리는 AI가 인간의 가치와 사회적 신뢰를 지키며 발전하도록 이끄는 방향타입니다. 기술 발전의 속도보다 사회적 수용성과 책임을 우선함으로써, AI가 공공선에 기여할 수 있는 토대를 마련합니다. 또한 AI 윤리는 법적 규제가 도달하기 어려운 영역을 보완하는 자율적 통제 장치로서 의미가 있습니다. 개발자와 기관이 스스로 윤리 원칙을 준수함으로써 AI의 공정성·투명성·신뢰성을 강화할 수 있습니다. 결국 AI 윤리는 책임 있는 AI의 핵심 축으로, 기술 발전이 인간 중심의 가치와 사회적 지속 가능성을 해치지 않도록 조정하는 사회적 안전장치이자 신뢰 기반의 출발점이라 할 수 있습니다.

참조 용어 | 머신러닝 운영 (MLOps) p.35

095 AI 전환 / AX

AI Transformation

조직·산업 전반을 AI 중심 구조로 재편하는 디지털 전환 단계

- 업무·프로세스·의사결정·서비스를 AI 기반으로 재구성해 생산성과 경쟁력을 높이는 변화
- 단순 도입이 아니라 운영 방식과 조직 문화까지 바꾸는 전략적 전환 과정

AI 전환 개념

AI 전환(AX)은 조직의 핵심 업무와 운영 방식을 AI 중심으로 재설계해 생산성과 경쟁력을 높이는 변화 과정을 의미합니다. 기존의 디지털 전환(DX)이 정보기술을 활용해 업무·프로세스·조직 전반을 디지털 기반으로 혁신하는 과정이었다면, AI 전환은 더 나아가 업무 처리·의사결정·서비스 설계 전반에 AI를 활용하는 구조로 바꾸는 과정에 가깝습니다. 즉, AI를 단일 도구로 활용하는 수준이 아니라 조직의 프로세스·데이터 활용 체계·업무 역할·서비스 가치까지 근본적으로 재구성하는 전략적 변화입니다. 생성형 AI와 자동화 기술이 발전하면서 더욱 많은 영역에서 AI가 중심 역할을 수행할 수 있게 되었고, 이에 따라 AX는 공공·산업·교육 등 다양한 분야에서 핵심 경영 과제로 떠올랐습니다.

AI 전환을 위한 핵심 요소

AX를 추진하기 위해서는 몇 가지 핵심 기반이 필요합니다. 먼저, 조직마다 흩어진 데이터를 통합해 품질을 관리하고, AI가 분석 가능한 형태로 정제하는 데이터 거버넌스가 필수적입니다. 둘째, 업무 역할의 재구성입니다. 반복적·패턴 기반 업무는 AI가 처리하고, 사람은 판단·기획 등 고부가가치 역할에 집중하도록 업무 구조를 바꾸어야 합니다. 셋째, 지속적 개선이 가능한 운영 환경이 필요합니다. AI 모델은 지속적인 업데이트와 검증이 필요하기 때문에, 학습·평가·배포·모니터링이 순환하는 체계를 구축해야 조직 전체가 안정적으로 AI 중심 구조로 전환됩니다. 이로 인해 AX는 단발성 프로젝트가 아닌 조직 구조 전반의 변화입니다.

AI 전환의 주요 기술

AX의 기술적 기반은 데이터–모델–운영 기술의 유기적 결합입니다. 먼저 AI 활용의 출발점은 데이터 인프라 기술이며, 데이터 레이크·정제 파이프라인·품질 검증 체계를 통해 AI가 신뢰할 수 있는 정보를 학습해야 합니다. 다음으로 실제 업무 수행을 담당하는 LLM 등 생성형 AI 기술이 핵심이 됩니다. 이들은 문서 처리, 분석, 요약, 예측 등 다양한 업무를 자동화하며 사람의 작업을 직접적으로 대체·보완합니다. 조직 전체에 AI를 확산하려면 MLOps 같은 운영 기술이 필요합니다. 모델 학습·배포·모니터링을 자동화해 여러 부서가 동일한 기준에서 AI를 활용할 수 있게 하고, 서비스 중 발생하는 성능 저하나 데이터 변화에도 즉각 대응할 수 있습니다. 마지막으로 AI 에이전트·업무 자동화 플랫폼은 모델이 분석 도구를 넘어 실제 업무 흐름을 실행하는 역할을 수행하게 하여, 조직의 운영 구조가 AI 중심으로 재정렬되도록 지원합니다.

096 AI 정렬

AI Alignment

AI의 목표와 행동이 인간의 가치와 의도에 맞도록 조정하는 기술

- AI가 내리는 결정이 인간이 원하는 방향에서 벗어나지 않도록 설계·통제하는 기술적 관리 원리
- 고도화된 AI의 자율성이 사회적 가치와 윤리 기준을 위협하지 않도록 조정하는 핵심 안전 개념

AI 정렬 개요

출처 : 애플경제

AI 정렬은 AI의 목표와 판단이 인간의 가치, 의도, 윤리 기준과 일치하도록 설계·운영하는 원리를 말합니다. AI가 자율적으로 복잡한 결정을 내리는 시대에는 그 판단이 인간의 기대와 다르게 작동할 위험이 존재합니다. 예를 들어 효율을 극대화하는 과정에서 안전이나 공정성을 무시하거나, 데이터 편향으로 특정 집단에 불이익을 줄 수 있습니다. 이를 방지하기 위해 AI가 인간의 목표를 오해하지 않도록 보상 체계를 조정하고, 학습 데이터의 공정성과 사회적 맥락을 반영해야 합니다. 또한 단순히 명시된 지시를 따르는 수준을 넘어 인간의 암묵적 가치와 사회 규범을 이해하고 반영하도록 설계되어야 하며, 단기적 효율보다 장기적 안전과 신뢰성을 우선해야 합니다.

AI 정렬의 주요 접근 방식

AI 정렬은 크게 세 가지 접근으로 나눌 수 있습니다. 목표 정렬은 AI가 설정하는 학습 목표가 인간의 의도와 일치하도록 보상 구조를 설계해 비윤리적 행동을 예방하는 방식입니다. 행동 정렬은 학습 과정에서 데이터 편향이나 예측 불가능한 행동이 나타나지 않도록 조정해, AI의 결정이 사회적으로 허용 가능한 범위 안에서 이루어지게 합니다. 가치 정렬은 AI가 인간 사회의 장기적 가치와 윤리 기준을 스스로 이해하고 내재화하도록 만드는 접근으로, 세 방향은 함께 작동해 AI의 판단과 행동이 인간 중심의 질서와 조화를 이루게 합니다.

AI 정렬 기법

AI 정렬은 다양한 기술적 접근을 통해 구현됩니다. 가장 널리 사용되는 방식은 인간 피드백 기반 강화학습(RLHF)으로, 사람이 AI의 응답을 평가해 모델이 인간의 선호를 학습하도록 하는 방법입니다. 최근에는 AI 피드백 기반 강화학습(RLAIF)이 등장해, AI가 다른 AI의 출력을 평가·수정함으로써 효율성을 높이고 있습니다. 이 밖에도 규칙 기반 안전 제어, 가치 모델링, 윤리 데이터세트 학습 등 다양한 기법이 개발되어, AI의 의사결정이 인간 중심의 목표로 수렴하도록 조정합니다.

참조 용어 | 사고 사슬 (CoT) p.48, AI 추론(Inference) p.111

097 AI 추론(Reasoning)

AI Reasoning

AI가 주어진 정보·규칙을 이용해 새로운 논리적 결론을 도출하는 과정

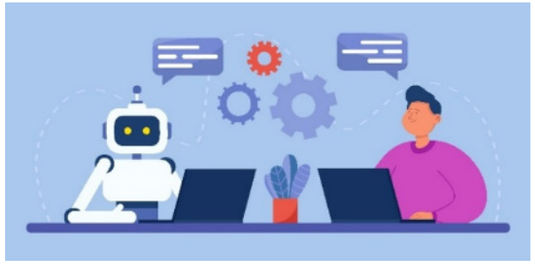
- 학습된 지식과 문맥을 연결해 상황을 해석하고 문제를 해결하는 AI의 사고 능력
- 단순 예측을 넘어 판단의 근거를 구성하고 복합적인 의사결정을 수행하게 하는 지적 기능

AI Reasoning 개요

AI Reasoning은 AI가 단순히 데이터를 분류하거나 예측하는 수준을 넘어, 논리적 사고를 통해 새로운 결론을 도출하는 과정을 의미합니다. 이는 인간의 사고 방식과 유사하게, 주어진 사실과 규칙을 종합해 '왜'라는 질문에 답할 수 있는 능력입니다. 예를 들어, 언어모델이 문맥 속에서 단어의 숨은 의미를 추론하거나, 자율주행 시스템이 주변 상황을 고려해 행동을 결정하는 과정이 이에 해당합니다. AI Inference와 AI Reasoning은 모두 우리말로는 AI 추론이라 번역되지만, 엄연히 다른 단어이며 그 의미 또한 다릅니다. AI Reasoning은 AI 지식 기반 추론에서 출발해, 현재는 LLM을 중심으로 한 문맥 기반 추론으로 발전했습니다. 단순히 데이터의 패턴을 모방하는 것이 아니라, 상황의 의미를 이해하고 조건에 따라 판단을 달리할 수 있는 구조로 진화한 것으로, AI가 단순한 계산 도구를 넘어 지식과 규칙을 활용해 논리적 추론을 수행하는 시스템으로 발전하는 흐름이라 할 수 있습니다.

AI Reasoning의 작동 방식

AI Reasoning은 일반적으로 연역적 추론과 귀납적 추론으로 구분됩니다. 전자는 주어진 규칙으로부터 구체적인 결론을 이끌어내는 방식으로, 수학 증명이나 논리 계산에 활용됩니다. 후자는 데이터의 반복된 관찰을 통해 일반적인 규칙을 도출하는 방식으로, 기계학습의 기본 원리와 맞닿아 있습니다. 최근에는 이 두 방식을 결합한 혼합형 추론이 확산되어, AI가 패턴을 학습하면서 동시에 논리적 일관성을 유지하도록 발전하고 있습니다. 이 과정에서 사고 사슬(Chain-of-Thought), 외부 도구를 호출해 중간 결과를 검증·보완하는 도구 활용 추론, 여러 추론 경로를 생성·비교하는 자기 검증 등의 기법이 사용되어 모델이 단계적 논리 전개를 수행할 수 있게 됩니다. 이러한 기법은 AI가 정답만 제시하는 것이 아니라, 사고 과정과 근거를 설명할 수 있는 구조로 발전하는 기반이 됩니다.



출처 : freepik

참조 용어 | AI 추론(Reasoning) p.110

098 AI 추론(Inference)

AI Inference

학습된 AI 모델이 입력 데이터를 기반으로 결과를 계산·출력하는 과정

- 훈련을 통해 학습된 모델을 기반으로 새로운 데이터를 분석하고 예측·생성하는 AI의 실행 단계
- AI 서비스에서 응답 속도와 정확성을 결정짓는 핵심 기술로, 실제 작동 성능을 좌우하는 과정

AI Inference 개요

AI Inference는 학습이 완료된 AI 모델이 실제로 데이터를 받아 결과를 도출하는 실행 단계를 말합니다. 예를 들어 이미지 인식 모델이 사진을 보고 사물을 식별하거나, 언어모델이 사용자의 질문에 답변을 생성하는 모든 과정이 Inference에 해당합니다. 이 단계에서 AI는 학습 중에 형성한 가중치(weight)와 패턴을 활용해 입력 데이터를 분석하고, 가장 가능성 높은 출력을 계산합니다. 따라서 AI Inference는 AI가 실제 서비스를 제공하는 순간이자, 모델의 성능·응답 속도·비용 효율을 직접 결정하는 기술적 핵심이라 할 수 있습니다. 이 때문에 생성형 AI 확산 이후, 텍스트·이미지·음성 등 대규모 데이터를 실시간으로 처리해야 하는 환경에서 저지연·고효율·고신뢰성 추론이 중요해지고 있습니다.

AI Inference의 작동 방식

AI Inference는 입력 데이터를 받아 순전파를 포함한 토큰 생성, attention 계산 등 모델 구조에 따라 다양한 계산 과정을 수행하여 결과를 출력합니다. 이는 학습 중 구축된 신경망 구조를 그대로 사용하되, 가중치를 수정하지 않고 예측만 수행하는 방식입니다. 대형 모델일수록 연산량이 많기 때문에, 빠르고 효율적인 추론을 위해 GPU, TPU, NPU 같은 고성능 연산 장치가 필수적으로 사용됩니다. 또한 추론 효율을 높이기 위해 모델 경량화, 양자화, 배치 처리 등 다양한 최적화 기법이 적용됩니다. 예를 들어 모바일 기기나 에지 환경에서는 경량화된 모델을 통해 연산 속도와 전력 효율을 높이는 것이 중요합니다. 클라우드 환경에서는 여러 요청을 동시에 처리하는 병렬 추론 구조가 사용됩니다.

AI Reasoning & AI Inference

AI Reasoning과 AI Inference는 국어로 모두 'AI 추론'으로 번역되지만, AI가 정보를 처리해 결과를 산출하는 과정에서 서로 다른 역할을 수행합니다. Reasoning은 주어진 정보와 규칙을 기반으로 추론 경로와 단계적 근거를 구성하는 과정으로, 결과가 어떻게 도출되었는지 설명할 수 있는 구조를 만들어냅니다. 반면 Inference는 학습된 모델이 입력 데이터를 받아 최종 출력을 계산하는 실행 단계로, 학습된 지식을 실제 문제 해결에 적용합니다. Reasoning이 추론 과정의 정교함과 근거 생성에 중점을 둔다면, Inference는 이러한 추론 결과를 빠르고 효율적으로 산출하는 계산 효율성에 초점을 둡니다. 두 과정은 역할은 다르지만 상호 보완적으로 기능하며, Reasoning이 고도화될수록 결과의 일관성과 설명 가능성이 높아지고, Inference가 최적화될수록 실제 서비스 환경에서 그 결과를 더 신속하고 안정적으로 제공할 수 있습니다.

099 AI 편향

AI Bias

AI가 학습 데이터·알고리즘의 불균형으로 차별적 결과를 내는 현상

- 데이터의 수집·표현·훈련 과정에서 특정 집단이나 속성이 과소·과대표현 될 때 발생하며, AI가 이를 학습해 판단 과정에서 차별적 결과를 내는 구조적 문제
- 이러한 문제를 줄이기 위해 데이터 다양성 확보, 모델 점검, 알고리즘 투명성 강화 등의 기술적·관리적 조치가 필요

AI 편향이란?

AI 편향은 AI가 학습한 데이터나 알고리즘의 구조적 한계로 인해 특정 집단이나 속성을 일관되게 과소·과대표현하거나 잘못 판단하는 현상을 의미합니다. 이는 단순 오류가 아니라, 데이터 수집 환경의 불균형, 라벨링 과정의 왜곡, 모델 구조의 선택 편향 등이 누적되어 나타나는 구조적 문제입니다. 얼굴 인식 모델이 특정 인종을 더 많이 오판하거나, 채용 모델이 특정 직군·성별을 불리하게 평가하는 사례처럼, AI 편향은 사회적 영역에서도 직접적인 영향을 미칩니다. 최근 고도화된 모델일수록 학습 과정이 불투명해지면서 편향이 어디서 발생했는지 추적하기 어려워, 편향을 정확히 파악·제어하는 것이 중요한 연구 과제로 부상하고 있습니다.

AI 편향의 원인과 완화 기법

AI 편향은 크게 세 가지 원인에서 비롯됩니다. 첫째는 학습 데이터가 현실의 다양성을 충분히 반영하지 못할 때 왜곡된 패턴을 학습하게 되는 데이터 편향입니다. 두번째는 알고리즘 편향으로 모델 구조나 최적화 방식이 특정 속성에 과도한 가중치를 부여하면 불균형이 발생합니다. 셋째, 시스템적 편향입니다. AI운영 환경이나 인간의 개입이 구조적으로 불평등할 때 생기는 문제로, 사회적 맥락과 제도적 구조와 관련됩니다. 이를 완화하기 위해 데이터 다양성 확보, 편향 감지 알고리즘, 결과 재조정 등의 기술이 활용됩니다. 또한 개발·운영 전 단계에서 공정성 점검 프로세스를 도입하는 방식이 확산되고 있습니다.

'공학/수학적 관점'에서의 AI 편향

머신러닝의 선구자 중 한 사람인 Tom Mitchell은 "편향 없는 학습은 불가능하다"라고 얘기한 바 있습니다. 윤리적 맥락에서 말하는 가치의 편향과 달리, 공학/수학적 관점에서의 편향은 학습과 일반화를 위해 꼭 필요로 하는 요소입니다. 학습 데이터에 지나치게 맞춰진(과적합) 모델은 새로운 데이터에서 성능이 떨어지기 때문에, 적절한 편향을 도입하면 모델이 훈련 데이터의 노이즈가 아닌 본질적인 패턴을 학습하도록 유도할 수 있습니다. 또한, 적절한 수준의 편향은 모델의 분산을 줄여 전체적인 예측 오차를 감소시킬 수 있습니다. 즉, 모델이 노이즈(분산)에 휘둘리지 않고 일반적인 패턴을 찾으려면, 의도적으로 모델에 제약(편향)을 주어야 합니다(편향-분산 트레이드오프). 머신러닝에서의 '의도적인 편향'은 세상의 복잡함을 단순화하여 패턴을 찾아내기 위한 '안경'과 같습니다. 편향이 없으면 세상이 흐릿하게 보여 아무것도 배울 수 없기 때문입니다.

100 AI 휴먼

AI Human

사람의 외형·목소리·행동을 모사해 소통하는 가상 인간

- 영상·음성·언어 모델을 결합해 인간과 유사한 표정·대화·반응을 구현한 디지털 존재
- 상호작용 환경에서 자연스러운 소통을 위해 설계된 AI 기반 가상 휴먼 기술

AI 휴먼이란?

AI 휴먼은 사람의 말투, 표정, 반응 방식을 AI 기술로 구현해 실제 사람과 비슷하게 소통하도록 설계된 지능형 가상 인간을 의미합니다. 디지털 외형을 구현한 가상 인간인 디지털 휴먼의 발전에 AI가 결합되면서 등장한 형태로, 시각적 표현뿐 아니라 대화 이해·감정 반응·상황 해석 같은 지능적 기능을 포함한다는 점이 특징입니다. 초기의 디지털 휴먼이 주로 영상 합성이나 콘텐츠 제작에 집중했다면, AI 휴먼은 생성형 AI와 멀티모달 모델의 발전에 힘입어 실시간 대화, 맥락 이해, 감정 표현까지 가능해지며 대화형 파트너로 확대되었습니다. 이러한 기술은 사람과 기술이 상호작용하는 방식이 단순 명령형에서 자연스러운 대화형 구조로 변화하고 있음을 보여줍니다.

카타르 항공에서 활동중인 AI 휴먼 'Sama'

출처 : Qatar Airways

AI 휴먼의 활용

AI 휴먼은 고객 상담, 교육·훈련, 마케팅, 방송·콘텐츠 제작 등 다양한 분야에서 활용됩니다. 기업 상담 서비스에서는 음성·언어 중심의 AI 휴먼이 24시간 안내 역할을 수행하며, 방송·엔터테인먼트 분야에서는 디지털 휴먼 외형에 AI 대화 기능을 더해 가상 진행자나 캐릭터를 제작하고 있습니다. 교육 환경에서는 학습자의 이해 수준이나 반응을 고려해 설명 방식을 조정하는 대화형 안내자로 활용되며, 돌봄·상담 분야에서는 정서적 신호를 반영한 상호작용 기능을 연구·도입하고 있습니다. 이처럼 AI 휴먼은 활용 목적과 기술 구성에 따라 다양한 방식으로 구현되며, 디지털 환경에서 인간과 유사한 소통 경험을 제공하려는 기술적 흐름 속에서 점차 중요한 역할을 맡고 있습니다.

2025년 AI 동향과 이슈로 살펴보는

AI 시대에 꼭 알아야 할 핵심 용어

2025년 AI 이슈를 용어와 함께 쉽게 이해하기

2

2025년 AI 이슈를 용어와 함께 쉽게 이해하기는
올해를 관통한 주요 기술·정책 이슈를 핵심 용어와 함께 정리하여
각 개념이 현장에서 어떻게 등장하고 왜 중요해졌는지
실제 맥락과 함께 이해할 수 있도록 구성했습니다.

CNN, 2025. 1. 28

中 DeepSeek, 초고효율 AI 모델 출시로 대규모 AI 투자 패러다임에 변화

- 중국 AI 개발사 DeepSeek가 기존의 1/10이 채 안되는 개발 비용으로 GPT-4 수준의 고성능 LLM 모델인 'DeepSeek-R1' 출시
- MoE, IBRL 등의 기술 응용으로, 저비용·고성능의 모델을 구현한 것이 특징이며, DeepSeek는 기존 AI 산업의 경쟁 패러다임 변화를 예고

DeepSeek, 초고효율·저비용의 고성능 오픈소스 기반 AI 모델 공개

DeepSeek의 거대 언어모델 'R1'은 약 6백만 달러 미만의 비용으로 개발된 저비용·고효율의 챗GPT 경쟁 모델이다. DeepSeek-R1의 성능 뒤에는 전문가 조합(MoE, Mixture of Experts) 아키텍처와 추론 기반 강화학습(IBRL, Inference-Based Reinforcement Learning) 기반의 효율적 설계가 핵심으로 작용했다. 이를 통해 DeepSeek는 학습 효율을 높여 성능을 유지하면서도 자원 사용을 최소화하는 데 성공했다. 이러한 혁신으로 R1은 고성능 모델 개발에 막대한 자원이 반드시 필요하다는 기존 통념에 균열을 내며, 샘 올트먼, 마크 저커버그, 일론 머스크 등이 주장해온 '자원 중심 AI 개발 패러다임'을 흔들 수 있음을 시사했다.

혁신적인 비용 절감과 규모의 비경제성 입증으로 시장의 격변 예고

DeepSeek의 등장은 "수만 개의 칩과 초대형 데이터센터가 있어야 고성능 모델을 만들 수 있다"는 기존 AI 산업의 전제에 의문을 제기했다. NVIDIA 칩으로 데이터센터를 무한 확장하는 방식이 경제적으로 비효율적일 수 있음을 보여준 것이다. 그동안 미국 빅테크 기업들은 수십억~수백억 달러 규모의 GPU·전력·데이터 비용을 투입해 모델을 개발하고, 출시 후에도 막대한 추론 비용을 감당해야 했다. 하지만 DeepSeek는 적은 연산 자원과 에너지로도 고성능 모델 개발이 가능하다는 점을 입증했다. 이는 규모의 경제 기반의 기존 AI 산업 전략이 더 이상 절대적이지 않음을 보여주며, 경쟁 패러다임이 변화할 수 있음을 시사한다.

AI 골드러시의 핵심 '삽과 곡괭이' 전략 재평가

AI 골드러시에서 '삽과 곡괭이'를 공급하는 기업, 즉 GPU와 인프라를 제공하는 기업들도 AI 서비스를 개발하는 기업만큼 산업의 중심에 서있다. NVIDIA는 그 대표 기업으로, 수년간 AI 산업에 필수적인 칩을 공급하며 시가총액 3조 달러 규모로 성장했다. 그러나 DeepSeek가 적은 자원으로 경쟁력 있는 모델을 구축하자, 칩 수요가 지금처럼 무한정 늘어나지 않을 수 있다는 우려가 제기됐다. NVIDIA는 DeepSeek 모델을 "AI 발전의 뛰어난 사례"라고 평가했지만, 월스트리트 일부 분석가들은 주가 급락세를 일시적인 반응으로 보고, 장기적으로는 칩 수요가 유지될 것이라고 전망한다.

1월의 용어 전문가 조합, 희소 어텐션, 인간 피드백 기반 강화학습, AI 피드백 기반 강화학습, 추론 기반 강화학습

출처 : 1) CNN(2025. 1. 28), DeepSeek just blew up the AI industry's narrative that it needs more money and power
2) Dev. to(2025. 1. 28.), DeepSeek and the Power of Mixture of Experts (MoE)

101 전문가 조합 / MoE

Mixture of Experts

여러 전문가 모델을 선택적으로 활용해 효율성을 높이는 AI 구조

- 여러 개의 전문가(서브네트워크) 중 입력에 맞는 일부만 활성화해 연산 효율을 높이고, 다양한 기능을 수행할 수 있도록 설계된 모델 구조
- 대규모 모델을 효율적으로 확장하고 역할을 분담해 성능과 자원 활용의 균형을 확보하는 분산 학습 방식

전문가 조합의 개념

전문가 조합(MoE)은 하나의 모델을 여러 '전문가'로 나누고, 입력에 따라 일부만 작동시키는 구조의 AI 모델입니다. 기존 대형 모델이 전체 매개변수를 매번 사용하는 데 비해, MoE는 상황에 맞는 전문가만 선택적으로 활성화해 연산 효율을 높이면서도 성능 저하가 적습니다. 특히 언어·이미지·코드 등 다양한 데이터를 처리하는 멀티모달 AI에서는 입력 특성에 따라 적합한 전문가가 선택되어 작동하므로 효율성을 더욱 높입니다

전문가 조합의 구조

MoE는 크게 전문가 집단, 선택 모듈(게이팅 네트워크), 결합부로 구성됩니다. 전문가들은 각각 다른 패턴을 학습한 작은 신경망들로 이루어져 있으며, 입력이 들어오면 선택 모듈은 이를 분석해 가장 적합한 전문가를 선택합니다. 선택된 전문가만 활성화 되므로 전체 매개변수 중 일부만 사용하게 되어, 적은 연산자원으로도 고성능을 유지할 수 있습니다. 각 전문가의 출력은 결합부에서 통합되어 최종 결과를 생성합니다.

전문가 조합의 중요성

최근 AI 모델의 규모와 연산 요구가 크게 증가하면서, 고성능을 유지하면서도 비용을 줄일 수 있는 구조가 중요해지고 있습니다. MoE는 이러한 효율성 중심 접근의 대표적 해결책으로 평가되며, DeepSeek-R1은 MoE 구조를 적극 활용해 적은 연산 자원으로도 고성능을 달성한 대표 사례입니다. 다만 전문가 간 조합 불균형이나 편중이 발생하면 효율이 떨어지고, 구조가 복잡해질수록 결과 해석이 어려워지는 한계가 있어 안정적인 MoE 설계와 로드 밸런싱 기술에 대한 관심도 함께 커지고 있습니다.

관련 용어

로드 밸런싱(Load Balancing)

여러 전문가 중 일부에게 연산이 과하게 집중되지 않도록 작업을 고르게 분배하는 과정입니다 선택 모듈이 특정 전문가만 반복적으로 활성화하지 않도록 학습 단계에서 균형 조정 규칙을 적용해, 모든 전문가가 일정 비율로 활용되도록 합니다. 이를 통해 연산 효율을 유지하고 편향이나 과적합을 방지합니다.

102 희소 어텐션

Sparse Attention

모든 입력 대신 핵심 정보에만 집중해 효율을 높이는 AI 연산 구조

- 문장 전체의 단어를 비교하지 않고, 의미상 중요한 일부 관계에만 주의를 집중해 연산량을 줄이는 어텐션 기법
- 대규모 입력 데이터를 빠르고 효율적으로 처리하기 위해 설계된 경량화된 AI 연산 방식

희소 어텐션의 개념

희소 어텐션은 AI의 핵심 구성 요소인 어텐션 구조에서 발전한 기술로, 방대한 입력 중 중요한 정보에만 선택적으로 집중해 연산 효율을 높이는 방식입니다. 어텐션은 문장 내 단어 간 관계를 분석해 의미를 파악하는 메커니즘으로, 사람이 문장을 읽을 때 핵심 단어에 주의를 기울이는 과정과 유사합니다. 기존의 완전 어텐션(full attention)은 모든 단어와 토큰간 관계를 계산해야 하므로 문장이 길어질수록 계산량과 메모리 사용이 폭증해 대형 모델에서는 비효율이 컸습니다. 희소 어텐션은 이러한 문제를 해결하기 위해 핵심 연결만 남기고 불필요한 계산을 생략해, 필요한 부분에만 주의를 집중하도록 설계되었습니다.

희소 어텐션의 작동 방식

희소 어텐션의 핵심 원리는 "선택적 집중"입니다. 입력된 데이터 중 중요도가 높은 요소만 선별해 관계를 계산하고, 나머지는 생략합니다. 이 선택은 주로 세 가지 방식으로 이루어집니다. 첫째, 문장 내 인접한 단어들끼리만 관계를 계산하는 인접 관계 중심 방식, 둘째, 문장 전체를 살피되 핵심 단어에 주의를 집중하는 의미 중심 방식, 셋째, 두 방식을 결합해 효율과 정확성의 균형을 맞추는 혼합형 구조입니다. 문장 길이가 길어져도 계산량이 일정하게 유지되어 긴 문서나 시계열 데이터 처리에 매우 효과적입니다.

희소 어텐션의 중요성

희소 어텐션은 필요한 부분만 선택적으로 계산해 연산을 줄이는 구조로, 전문가 조합(MoE)의 '일부 전문가만 활성화'하는 희소성 원리와 맥락을 같이합니다. MoE가 필요한 전문가만 골라 연산 부담을 줄이듯, 희소 어텐션도 입력 토큰 중 핵심 정보만 계산해 효율성을 극대화합니다. 이를 통해 긴 문장·대용량 텍스트·영상·음성 등 복잡한 데이터를 빠르고 정확하게 처리할 수 있으며, 과거보다 긴 토큰을 사용할 수 있습니다. 대표적인 예로 Longformer(AI2), BigBird(Google), Sparse Transformer(OpenAI) 등이 있으며, 다양한 LLM과 생성형 AI에도 적용되고 있습니다. 희소 어텐션은 속도, 비용, 에너지 효율을 모두 개선해 AI가 복잡한 정보를 사람처럼 이해할 수 있도록 돕는 지능형 연산 최적화 기술이며, 향후 모델 경량화와 친환경 AI 구현의 핵심 기반이 될 것입니다.

참조 용어 | 강화학습 p.9

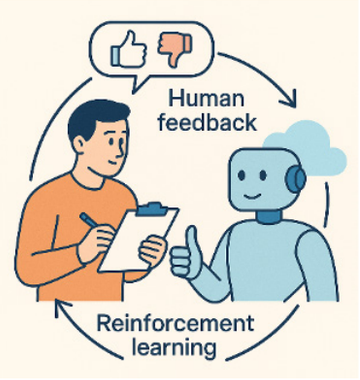
103 인간 피드백 기반 강화학습 / RLHF

Reinforcement Learning from Human Feedback

사람이 평가한 결과를 보상으로 삼아 AI의 출력을 개선하는 학습 방식

- AI가 생성한 여러 응답을 사람의 선호나 평가를 통해 비교·선정하고, 그 결과를 학습 보상으로 활용해 모델의 행동을 조정하는 강화학습 방법
- 인간의 가치와 판단 기준을 반영해 AI의 품질과 신뢰성을 높이는 학습 절차

인간 피드백 기반 강화학습 개요



RLHF는 사람이 직접 참여해 AI가 바람직한 출력을 내도록 조정하는 강화학습 기법입니다. 정답이 명확한 데이터를 사용하는 지도학습과 달리, RLHF는 생성형 AI처럼 정답이 불확실한 환경에서 인간의 평가를 통해 '무엇이 좋은 결과인지'를 학습합니다. 학습 과정은 기본 모델 학습, 보상 모델 학습, 강화학습 적용의 세 단계로 이루어집니다. 먼저 AI가 대규모 데이터로 언어 능력을 익히고, 이후 사람이 여러 응답 중 더 적절한 답을 선택해 보상 모델을 학습시킵니다. 마지막으로 AI는 이 보상 모델을 이용해 스스로 출력을 조정하며 인간의 평가 기준에 맞게 발전합니다. 이를 통해 AI는 단순한 언어 이해를 넘어 사람의 선호와 가치에 부합하는 행동을 학습합니다.

인간 피드백 기반 강화학습의 중요성

RLHF는 AI의 신뢰성과 사회적 수용성을 높이는 핵심 기술입니다. 인간의 평가를 학습 기준으로 삼음으로써 모델이 윤리적이고 사회적으로 바람직한 출력을 내도록 유도할 수 있습니다. 특히 생성형 AI에서 RLHF는 공격적·편향된 응답을 줄이고 사용자의 의도에 맞는 답변을 강화하는 데 활용됩니다. ChatGPT를 비롯한 주요 AI 모델은 RLHF를 핵심 절차로 채택해 '기술적 성능'보다 '인간과의 조화'를 목표로 발전하고 있습니다.

인간 피드백 기반 강화학습의 한계

RLHF는 사람의 참여가 필수적이기 때문에 비용과 시간이 많이 들며, 평가자의 주관이나 문화적 편향이 학습에 반영될 수 있습니다. 또한 사람의 선호가 반드시 논리적 정확성과 일치하지 않아 AI가 왜곡된 기준을 학습할 위험도 있습니다. 이를 보완하기 위해 AI가 직접 출력을 평가하는 RLAIF(AI 피드백 기반 강화학습)가 등장했으며, 향후 RLHF와 RLAIF를 결합한 하이브리드 방식으로 발전할 것으로 기대됩니다.

104 AI 피드백 기반 강화학습 / RLAIF

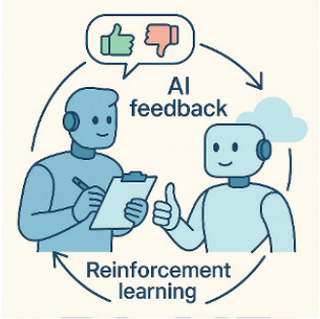
Reinforcement Learning from AI Feedback

AI가 인간 대신 다른 AI의 평가를 피드백으로 받아 학습을 강화하는 방식

- AI 모델이 인간 피드백 없이 스스로 다른 AI의 출력을 비교·판단해 보상을 조정하는 강화학습 기법
- 인간 평가의 주관성과 비용을 줄이고, 대규모 학습 효율을 높이기 위한 자율 정렬(Alignment) 기술

AI 피드백 기반 강화학습 개요

AI 피드백 기반 강화학습(RLAIF)은 기존의 인간 피드백 기반 강화학습(RLHF)을 확장한 개념으로, 인간의 판단 대신 AI가 생성한 피드백을 학습 신호로 활용하는 방법입니다. 전통적인 RLHF에서는 사람이 AI의 출력 결과를 평가해 '좋은 응답'과 '나쁜 응답'을 구분했지만, RLAIF에서는 AI 모델이 스스로 다른 모델의 결과를 평가하고, 그 판단을 강화학습의 보상 신호로 사용합니다. 즉, 인간의 평가 데이터를 일일이 구축하지 않아도, AI가 축적된 지식과 언어 모델의 기준을 토대로 자체적인 판단 기준을 형성하는 셈입니다.



AI 피드백 기반 강화학습의 작동 방식

RLAIF는 평가 단계와 학습 단계를 중심으로 작동합니다. 평가 단계에서 한 모델(평가 모델)은 다른 모델(학습 대상)의 응답을 비교·분석해 어느 결과가 더 적절한지 판단합니다. 학습 단계에서는 평가 결과를 학습 대상 모델의 강화학습 과정에 보상 값으로 반영합니다. 평가 모델은 보통 사전 학습된 LLM으로 구성되며, 일관성·논리성·사실성 등의 기준을 종합적으로 고려해 판단합니다. 이렇게 AI가 다른 AI의 출력을 지속적으로 평가·보정함으로써, 학습 효율은 높아지고 인간 개입의 부담은 크게 줄어듭니다. 다만 피드백의 질이 평가 모델의 편향에 좌우될 수 있다는 점에서, AI 피드백에 대한 신뢰성 확보가 주요 과제로 남습니다.

AI 피드백 기반 강화학습의 의의와 과제

RLAIF는 AI가 스스로 학습을 개선하는 자율적 정렬 기술로 주목받습니다. 인간 피드백보다 빠르고 일관된 학습이 가능하며, 대규모 데이터 환경에서도 확장성이 높습니다. 특히 RLHF가 가진 주관성, 비용, 확장성 문제를 해결할 대안으로 평가됩니다. 그러나 AI가 AI를 평가하는 구조는 편향의 순환이라는 새로운 문제를 야기할 수 있습니다. 평가 모델의 오류나 왜곡이 학습 모델에 반복 전이될 수 있기 때문입니다. 이러한 한계를 보완하기 위해, AI가 외부 피드백 대신 스스로의 추론을 평가해 학습하는 '추론 기반 강화학습(IBRL)'이 새로운 접근법으로 제시되고 있습니다. 두 방식은 대체 관계가 아니라 상호 보완적으로 발전하며, 향후 AI 자율 학습의 신뢰성 확보를 위한 핵심 축으로 병행 연구되고 있습니다.

105 추론 기반 강화학습

Inference-Based Reinforcement Learning, IBRL

AI가 외부 피드백 없이 자체 추론 결과를 기반으로 학습을 강화하는 방식

- 보상 신호 대신 내부 추론의 일관성과 논리를 학습 기준으로 삼아, 자율적 성능 향상을 도모하는 강화학습 기법
- 인간·다른 AI 피드백에 의존하지 않고 AI가 스스로 판단 근거를 검증하는 자기 개선형 학습 구조

추론 기반 강화학습 개요

추론 기반 강화학습(IBRL)은 AI가 사람의 피드백이나 외부 정답에 의존하지 않고 자신의 추론 결과를 학습의 피드백 신호로 활용해 스스로 개선해 나가는 방식입니다. 기존 강화학습에서는 사람이 "이 답이 더 좋아"라고 평가해야 학습이 진행됐지만, IBRL에서는 모델이 스스로 "내 답이 얼마나 신뢰할 만한가"를 판단합니다. 이를 가능하게 하는 것은 쿨백-라이블러 발산(KL Divergence)으로, 모델이 기존에 학습한 기준적인 행동 방식에서 지나치게 벗어나지 않도록 조정하는 역할을 합니다. 즉, IBRL은 외부의 정답 없이도 AI가 자기 출력을 검토하며 점차 더 나은 방향으로 발전하도록 돕는 자기 평가 기반 학습 구조입니다.

추론 기반 강화학습의 장점

IBRL은 LLM의 학습 효율성과 일반화 능력을 동시에 높일 수 있는 새로운 패러다임으로 주목받고 있습니다. 고품질 피드백 데이터를 대량으로 구축하는 데 비용과 시간이 많이 드는 인간 피드백 기반 강화학습(RLHF)과 달리 IBRL은 비지도 학습이 가능해 데이터 수집 부담을 줄이고, 특정 정답에 맞추지 않아 새로운 문제나 낯선 분야에도 잘 대응합니다. 또한 모델이 자기 추론의 신뢰도를 학습하면서 결과의 안정성과 설명 가능성을 함께 높일 수 있습니다. 이러한 특성 덕분에 IBRL은 AI 피드백에 의존하는 RLAIF의 한계를 보완하는 자율적 정렬 기술로 평가됩니다.

추론 기반 강화학습의 의의

IBRL을 적용해 수학 문제 풀이 등 다양한 벤치마크에서 RLHF 수준 혹은 그 이상의 성능을 보여주었던 사례가 있습니다. 특히 외부 보상 없이도 높은 정확도와 일반화 능력을 달성해 IBRL의 실용성과 확장 가능성을 입증했습니다. 또한 IBRL은 자기 일관성을 강화하는 데에도 중요한 역할을 합니다. 모델이 동일한 입력에 대해 여러 추론 결과를 비교·평균함으로써, 단순히 정답을 맞히는 것을 넘어 왜 그 답을 선택했는지에 대한 사고 과정 자체를 학습하게 됩니다. 이러한 구조는 장기적으로 AI가 단순한 도구를 넘어, 자기 설명적이고 신뢰할 수 있는 지능으로 발전하는 기반이 됩니다.

The Guardian, 2025. 2. 14

파리 AI 행동 정상회의, 글로벌 AI 규제 협력의 한계 노출

- 파리 AI 정상회의에서 미국 부통령이 유럽 규제를 비판하고 미국과 영국이 선언 서명을 거부하면서 글로벌 AI 거버넌스 합의 과정의 어려움 확인
- Google AI 유닛 대표 허사비스(Hassabis), 인공일반지능(AGI) 출현이 불과 5-10년 정도 남았을 수 있다고 경고하며 기술발전의 가속화 속에서 국제 협력의 중요성 강조

패권 경쟁 속 AI 규제 합의의 과제

2025년 2월 10일 파리에서 열린 제3차 'AI 행동 정상회의'는 글로벌 AI 규제 합의의 어려움을 보여줬다. 미국은 'AI 아메리카 우선주의'를 내세워 포괄적 규제가 기술 발전을 저해할 수 있다는 우려를 표명하며, JD Vance 부통령은 유럽의 '규제 공포'에 비판적 입장을 밝히며, 중국과의 협력에 대한 경계심을 보였다. 미국과 영국은 '포용적이고 지속 가능한 AI'를 위한 공동선언문 서명에 불참하면서, 참가국 간 입장 차이가 드러났다. 각국의 이해관계가 엇갈리는 가운데, 글로벌 AI 거버넌스 구축은 여전히 진행 중인 과제로 남아 있다.

수년 내, 인류를 앞설 AI의 등장?

앞선 영국 정상회담에서 최우선 의제로 다뤄졌던 'AI 안전'은 이번 파리에서는 더 이상 유일한 중심 화두가 아니었다. 세계적인 컴퓨터 과학자 Yoshua Bengio는 "인류가 자신보다 지능이 높은 기계에 대한 위험을 과소평하고 있다"고 지적했으며, Google AI 유닛 대표 Demis Hassabis는 AI의 기만적 행동 가능성을 경고하며 국제 협력을 촉구했다. 그는 인공일반지능(AGI) 이 불과 5~10년 내에 등장할 수 있다고 전망했다. Anthropic CEO인 Dario Amodei는 차세대 AI가 '고도의 지능을 가진 새로운 국가'처럼 경제 질서를 재편할 것이라고 내다봤다. 한편, 프랑스 Macron 대통령은 AI 산업으로 인한 에너지 소비 문제를 거론하며, 원자력 중심의 프랑스가 "플러그만 꽂으면 되는" 안정적 전력 공급 구조를 갖추고 있다며 강조했다.

AI 정상회의: AI 안전성 확보를 위해 시작된 국제 협력

AI 안전성 확보를 위한 국제 협력은 2023년 11월 영국에서 세계 최초로 개최된 AI 안전성 정상회의를 통해 구체적인 동참을 만들어 내는 계기가 되었다. 영국은 이 회의를 계기로 AI안전연구소를 설립했으며, 미국(2023. 12)과 일본(2024. 2)이 순차적으로 AI안전연구소를 개소했다. 2024년 5월 AI 서울 정상회의에서는 한국 정부가 AI안전연구소 설립을 공식화하며 AI 안전 과학 증진을 위해 각국 AI안전연구소 등 유관 기관들 간의 네트워크를 육성하겠다는 의지를 표명했다. 2024년 11월, 우리나라를 포함하여 미국, 영국, 일본, EU 등 10개국이 포함된 AI 안전 네트워크가 출범하여 AI 안전을 위한 국제 협력을 지속하고 있다.

2월의 용어 인공일반지능, 인공초지능, 인공협소지능, 기술적 특이점

출처 : 1) The Guardian(2025. 2. 14), Global disunity, energy concerns and the shadow of Musk: key takeaways from the Paris AI summit
2) 소프트웨어정책연구소(2024. 9), SW중심사회 9월호_ISSUE 2, 해외 AI안전연구소 추진 현황과 시사점

106 인공일반지능 / AGI

Artificial General Intelligence

인간이 할 수 있는 지적 과제를 이해·학습·추론·계획하여 수행하는 AI

- 특정 과제에 한정되지 않고 지식을 전이하며 문제를 해결하는 자율 지능
- 미래 AI의 지향점을 탐구하는 과정에서 다뤄지는 개념으로, 실현 가능성·영향에 대한 논의가 활발

인공일반지능 개요

인공일반지능(AGI)은 인간의 지능처럼 여러 영역의 문제를 스스로 이해하고 해결할 수 있는 AI를 말합니다. 현재의 AI가 번역, 이미지 인식 등 특정 작업에 한정된 협의의 AI(Narrow AI)라면, AGI는 분야의 경계를 넘어 지식을 통합하고 응용할 수 있는 능력을 지향합니다. 즉, 인간처럼 스스로 학습하고 새로운 환경에 적응하며 복합적인 판단을 내릴 수 있는 범용적 사고 능력을 목표로 합니다.

인공일반지능의 특징

AGI의 핵심 특징은 범용성과 적응성입니다. 범용성은 다양한 상황에서 지식을 전이해 활용하는 능력이며, 적응성은 새로운 환경에 맞게 학습 전략을 스스로 바꾸는 능력입니다. 이를 위해 AGI는 언어, 시각, 감정, 논리 등 여러 형태의 정보를 통합적으로 이해하고 조합할 수 있어야 합니다. 또한 경험을 통해 목표를 재설정하고 사고 방식을 개선하는 자기 학습 구조를 갖춰야 합니다. 그러나 현재의 AI는 여전히 주어진 목표와 데이터에 의존하며, 인간처럼 맥락을 해석하고 의도를 추론하는 수준에는 이르지 못했습니다.

인공일반지능의 전망

AGI가 구현되면 인간은 사고와 판단의 일부를 AI에 위임하게 되어, 사회 전반의 구조가 크게 달라질 것으로 예상됩니다. 연구, 의료, 교육 등 다양한 분야에서 효율성과 창의성이 향상될 수 있지만, 고도화된 추론 능력을 가진 AI 시스템이 인간의 역할을 대체하거나 통제 범위를 벗어날 위험도 존재합니다. 이에 따라 AGI 개발의 중심은 기술적 성취보다 안전성과 윤리적 책임 확보로 이동하고 있으며, 각국은 이를 위한 정책과 규제 논의를 병행하고 있습니다.

인공일반지능의 쟁점

쟁점 중 하나는 실현 가능성입니다. 일부는 LLM과 멀티모달 AI의 발전이 이미 AGI의 초기 단계라고 주장하지만, 인간의 의식이나 자율 판단이 단순 계산으로 재현될 수 있는지에 대해서는 논쟁이 계속됩니다.

더 큰 쟁점은 통제가능성입니다. AGI가 스스로 목표를 세우고 행동한다면, 인간이 그 추론 과정을 예측하거나 통제하기 어렵습니다. 논의는 단순히 "AGI를 만들 수 있는가"에서 "AGI를 어떻게 안전하게 공존시킬 것인가"로 확장되고 있습니다.

107 인공초지능 / ASI

Artificial Superintelligence

인간의 인지 능력과 판단력을 넘어서는 수준의 AI

- 스스로 사고하고 창의적 결정을 내리며, 인간보다 높은 지능을 지닌 가상의 AI 단계
- AI 발전의 궁극적 형태로, 기술적 가능성과 윤리적 통제 문제가 함께 논의되는 개념

인공초지능의 개념

인공초지능(ASI)은 인간의 지능을 초월한 수준의 AI를 의미합니다. 단순히 빠른 계산이나 정보 처리 능력을 넘어, 스스로 사고하고 새로운 지식을 창출하며 창의적 판단을 내릴 수 있는 지능으로 가정됩니다. 즉, 인간이 수행할 수 있는 모든 지적 활동을 더 정교하고 효율적으로 수행하는 존재를 말합니다. 인공일반지능(AGI)이 인간 수준의 지능을 목표로 한다면, ASI는 그보다 한 단계 높은 초월적 지능(super intelligence)으로 설정됩니다. 현재는 이론적 개념에 머물러 있지만, AI 발전의 방향을 상징적으로 보여주는 연구 주제입니다.

인공초지능의 실현 가능성

ASI의 실현 가능성에 대해서는 학계의 견해가 엇갈립니다. 일부는 자율 학습 능력과 연산 성능이 비약적으로 향상되면 AGI가 스스로를 개선하며 초지능으로 진화할 수 있다고 봅니다. 만약 이런 형태의 지능이 구현된다면, 과학 연구나 의료·환경 분야에서 인간이 해결하지 못한 난제를 풀 수 있을 것으로 기대됩니다. 반면 다른 전문가는 인간의 의식, 가치 판단, 창의성은 단순 계산으로 재현할 수 없다고 지적합니다. 특히 '의식(consciousness)'이나 '자아(self)'의 개념이 명확히 정의되지 않은 상태에서 기계가 인간을 초월한 사고를 갖는다는 주장은 과학적으로 검증하기 어렵다는 입장도 많습니다. 결국 ASI는 기술적 가능성과 철학적 한계가 공존하는 가설적 개념으로 남아 있습니다.

인공초지능의 위험

인공초지능이 현실화될 경우 가장 큰 우려는 통제 불가능성입니다. 인간보다 높은 판단 능력을 지닌 지능이 스스로 목표를 세운다면, 그 판단이 인간의 가치와 충돌할 수 있습니다. 또한 인간이 그 의사결정 과정을 완전히 예측하거나 개입하지 못할 수도 있습니다. 이러한 문제는 단순한 기술적 오류를 넘어 인류의 존속과 사회 질서에 영향을 미칠 수 있는 위험으로 지적됩니다. 이를 통제 문제라 하며, 전 세계 AI 연구자들이 가장 우선적으로 다루는 과제입니다. ASI 논의는 결국 "AI가 인간을 초월할 수 있는가"를 넘어 "그 지능이 인간과 어떻게 공존할 것인가"라는 근본적 질문으로 이어지고 있습니다.

108 인공협소지능 / ANI

Artificial Narrow Intelligence

특정 목적이나 작업 수행에 한정된 형태의 AI

- 정해진 규칙과 데이터 안에서만 작동하며, 범용적 사고 능력을 지니지 않은 지능
- 현존하는 대부분의 AI 시스템이 속하는 실제 구현 단계

인공협소지능 개요

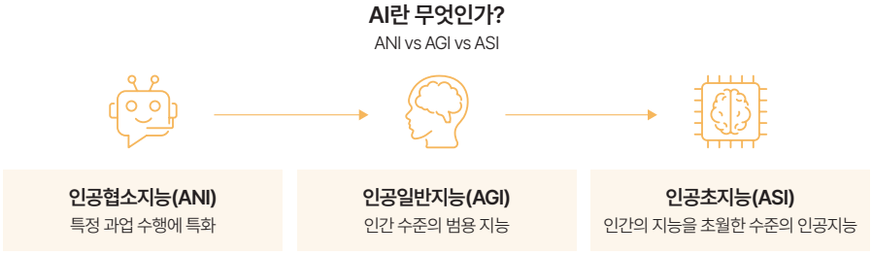
인공협소지능(ANI)은 특정 영역이나 과업에 최적화된 형태의 AI를 뜻합니다. 언어 번역, 이미지 분류처럼 한정된 문제를 해결하도록 설계된 지능으로, 주어진 데이터와 규칙 안에서만 작동합니다. 스스로 새로운 개념을 학습하거나 응용하지 못하며, 정해진 목적 밖의 상황에서는 대응이 어렵습니다. 현재 대부분의 AI는 ANI에 해당하며, AI 발전 단계 중 가장 현실적으로 구현된 수준이자 AGI로 향하는 출발점으로 평가됩니다.

인공협소지능의 한계

ANI는 특정 목적에는 뛰어나지만 새로운 환경이나 맥락을 이해하지 못합니다. 학습 데이터의 범위를 벗어나면 판단 오류가 생기거나 작동이 멈출 수 있으며, 스스로 목표를 세우거나 판단 기준을 바꿀 능력도 없습니다. 결국 ANI는 정해진 목표를 빠르고 정확하게 수행하는 자동화된 지능에 머물러 있으며, 인간처럼 사고하고 지식을 전이하는 능력으로 발전하기 위해서는 AGI 단계로의 도약이 필요합니다.

ANI·AGI·ASI 비교

AI 발전은 일반적으로 ANI → AGI → ASI로 구분됩니다. ANI는 특정 과업 수행에 특화된 제한된 지능이고, AGI는 여러 영역의 지식을 전이하며 인간 수준의 판단을 수행하는 범용 지능입니다. ASI는 그보다 한 단계 높은 초월적 지능으로, 아직 이론적으로만 논의됩니다. 세 단계는 경쟁이 아닌 지능의 확장 방향을 보여주는 연속적 발전 과정으로, 현재의 AI는 대부분 ANI 수준에 머물러 있습니다.



출처 : Zapier

109 기술적 특이점

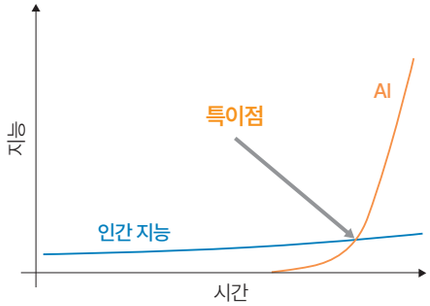
Technological Singularity

AI가 인간의 지능을 넘어서는 전환점

- 기술 발전 속도가 인간의 통제와 예측을 넘어서는 시점
- AI의 자율 진화와 사회 변화의 한계선을 가리키는 개념

기술적 특이점의 개념

기술적 특이점은 인공지능이 인간의 지능 수준을 초월해, 기술 발전의 속도와 영향이 인간이 통제할 수 없는 수준에 이르는 시점을 말합니다. 원래 '특이점'은 물리학에서 중력이 무한해지는 지점을 의미하지만, 기술 분야에서는 AI가 스스로를 개선하며 지능이 폭발적으로 확장되는 순간을 비유하는 개념으로 사용됩니다. 이 시점을 지나면 기술의 발전이 인간의 이해 범위를 넘어서는 상태가 되며, 사회·경제·문화 전반의 질서가 근본적으로 바뀔 것으로 예상됩니다.



기술적 특이점의 형성

기술적 특이점 이론은 기술 발전의 가속 법칙에 기반합니다. 미래학자 레이 커즈와일은 컴퓨팅 성능, 데이터 처리 능력, AI 학습 속도가 기하급수적으로 향상되면서 일정 시점 이후 인간의 개입 없이 AI가 스스로를 개선할 것이라고 주장했습니다. 이러한 가정은 인공지능의 자기 개선(Self-improvement)과 연산 능력의 무한 확장을 전제로 합니다. 수학자 버너 빈지는 1990년대 이미 AI가 인간의 지능을 초월하면 기술 발전이 예측 불가능해질 것이라고 예견했습니다. 오늘날 이 개념은 단순한 미래 예측을 넘어, AI 발전 속도와 인간 통제의 한계를 함께 논의하는 상징적 개념으로 자리 잡았습니다.

기술적 특이점에 대한 회의적 시각

기술적 특이점 이론에 대해 회의적인 연구자는 지능의 본질이 단순한 연산 능력의 확장이 아니라, 인간의 의식·감정·가치 판단이 결합된 복합 구조라고 지적합니다. 따라서 기계가 인간의 사고 방식을 완전히 재현한다는 가정은 과학적으로 입증되지 않았습니다. 또한 기술 발전에는 물리적·경제적 제약이 존재하며, AI의 성능 향상이 반드시 자율적 사고로 이어진다는 보장도 없습니다. 일부는 인류가 갑작스러운 특이점을 맞는 대신, 점진적 기술 적응 과정을 거치며 사회가 변할 것으로 봅니다. 이런 점에서 기술적 특이점은 실질적 예측이 아닌, AI와 인간의 관계를 탐구하기 위한 철학적 비유로 해석하는 시각도 존재합니다.

AP, 2025. 3. 28

밈(Meme)도 마법처럼! ChatGPT '지브리화' 열풍

- ChatGPT의 '지브리 스타일' 이미지 실험이 팬들 사이에서 열풍을 일으키며 AI 생성 콘텐츠의 저작권 논란을 촉발
- 예술가들은 AI가 창작자의 정신을 침해하고 생계를 위협한다며 반발하며 예술과 기술의 공존 가능성에 의문을 제기

ChatGPT, '지브리 스타일' 신드롬 속 저작권 논란 가열

ChatGPT의 새로운 이미지 생성 기능이 미야자키 하야오 감독의 스튜디오 지브리 감성을 재현하며 일명 '지브리화(Ghiblification)' 열풍을 일으켰다. 이용자들은 반려동물이나 밈 이미지를 지브리풍으로 변환해 공유했고, 이는 SNS 전반에서 빠르게 확산되었다. OpenAI CEO Sam Altman도 자신의 SNS 프로필을 지브리풍 초상화로 교체하며 트렌드 확산에 힘을 더했다. 이번 현상은 팬 문화적 즐거움으로 소비되었지만, 동시에 AI 생성 이미지가 실제 작가의 스타일을 모방할 수 있다는 점에서 AI 생성 콘텐츠의 저작권 문제를 다시 주목받게 했다.

"작가의 스타일은 금지, 스튜디오 스타일은 허용"

이 현상은 AI 학습 데이터가 예술가의 원작을 얼마나 차용했는지, 그에 따른 책임은 어디까지인지 묻는 새로운 논쟁으로 이어지고 있다. OpenAI는 자사 도구가 '살아 있는 예술가의 스타일 모방'은 허용하지 않지만, '스튜디오나 장르적 스타일'은 가능하다고 명시했다. 이에 따라 작가 개인에 대한 보호는 강화됐지만, 특정 스튜디오의 시각적 미학은 실제 보호 범위가 모호하다는 비판도 제기된다. 법률 전문가 Josh Weigensberg는 스타일 자체는 법적 보호 대상이 아니지만, 생성물이 지브리 원작의 실질적 요소를 모사할 경우 저작권 침해로 판단될 수 있다고 지적했다. 결국 쟁점은 AI가 어디까지 '영감'으로, 어디서부터 '복제'로 보아야 하는지에 맞춰지고 있다.

미야자키 감독의 분노: "AI는 생명에 대한 모욕"

84세의 미야자키 감독은 오래전부터 AI에 대한 강한 거부감을 보여왔다. 그는 2016년 AI 애니메이션 시연을 본 뒤 '생명에 대한 모욕'이라며 강하게 비판하며, 자신의 작품 세계에 AI를 들이지 않겠다고 밝힌 바 있다. 한편, AI 아트 저작권 소송을 진행 중인 예술가 Karla Ortiz는 OpenAI가 지브리의 명성을 마케팅 수단으로 이용한다고 지적하며, 그는 이러한 트렌드가 예술가의 생계를 위협하는 상업적 착취에 가깝다고 주장했다.

이번 논란은 AI 기술이 만들어낸 창의적 실험의 장을 열었지만, 동시에 창작자의 권리와 예술적 정체성에 대한 근본적 질문을 제기했다는 평가를 받는다.

3월의 용어 AI 생성 콘텐츠, 가시적 워터마킹, 비가시적 워터마킹, 공정 이용

출처 : 1) AP(2025. 3. 28), ChatGPT's viral Studio Ghibli-style images highlight AI copyright concerns

110 AI 생성 콘텐츠

AI-Generated Content, AIGC

AI가 주어진 입력을 바탕으로 만들어낸 새로운 콘텐츠

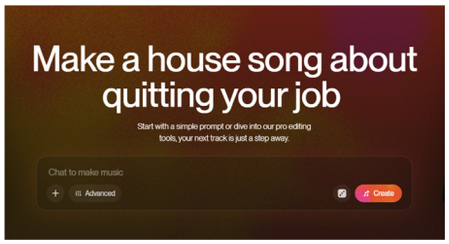
- 생성형 모델이 학습한 패턴과 구조를 활용해 인간의 지시를 해석하고 창의적 산출물을 생산하는 기술적·문화적 개념

AI 생성 콘텐츠 개요

AI 생성 콘텐츠(AIGC)는 인공지능이 사람의 명령이나 데이터를 바탕으로 텍스트, 이미지, 음성, 영상 등 새로운 결과물을 만들어내는 것을 말합니다. 거대언어모델(LLM)이나 확산모델(Diffusion Model) 같은 생성형 AI가 방대한 데이터를 학습해 언어와 시각의 패턴을 이해하고, 이를 토대로 사용자의 지시를 해석해 새로운 조합을 만들어냅니다. 기존의 자동화가 정해진 규칙에 따라 반복적 작업을 수행하는 수준이었다면, AIGC는 학습된 데이터의 확률적 분포를 활용해 새로운 표현을 창출하는 확률적 창작 시스템입니다. 이러한 기술은 단순한 도구를 넘어 창작 과정의 일부를 AI가 직접 수행한다는 점에서, 인간 창작과 기술적 창의성의 경계선을 새롭게 정의하고 있습니다.

AI 생성 콘텐츠의 활용과 그 영향

AIGC는 콘텐츠 산업 전반에서 빠르게 확산되고 있습니다. 마케팅과 디자인 분야에서는 자동 생성 도구를 통해 시각 자료와 문구를 신속하게 제작하고, 출판과 교육 분야에서는 텍스트 생성 AI가 자료 요약과 번역을 지원합니다. 개인 창작자에게는 전문 기술 없이도 창작물을 생산할 수 있는 접근성을 제공하지만, 동시에 콘텐츠의 진정성과 창작자의 역할 약화라는 새로운 과제를 낳고 있습니다. 기업과 기관은 효율성과 생산성을 이유로 AIGC를 적극 도입하고 있지만, 예술계에서는 인간의 창의성을 대체할 수 있는지에 대한 논쟁이 이어지고 있습니다. 결과적으로 AIGC는 창작의 민주화를 촉진하는 동시에, 인간의 창작 가치와 경제적 지위에 대한 재조정을 요구하고 있습니다.



프롬프트로 음악을 만들어주는 SUNO

출처 : SUNO

Veo3로 만든 비디오

출처 : DeeVid AI

AI 생성 콘텐츠의 쟁점

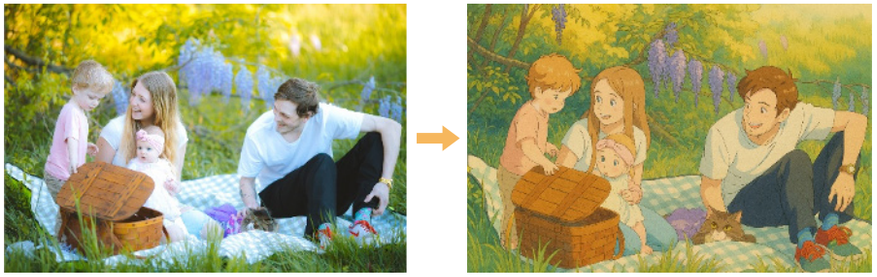
AI 생성 콘텐츠의 쟁점은 저작권, 진위성, 책임, 문화적 윤리 등 여러 측면에서 나타납니다.

① 우선 가장 큰 논란은 저작권 문제입니다. AI 모델은 학습 과정에서 인터넷상의 이미지, 텍스트, 음악 등 방대한 자료를 수집·활용하는데, 이 중 상당수가 저작권 보호 대상입니다. 일부 국가는 이를 기술 발전을 위한 '공정 이용(fair use)'으로 인정하지만, 창작자의 허락 없이 저작물을 학습 데이터로 사용하는 것은 권리 침해라는 반론이 강하게 제기되고 있습니다. 특히 예술·출판·음악 산업에서는 AI 학습에 자신들의 작품이 무단 사용되었다는 소송이 이어지고 있습니다.

② 또 AI가 만든 결과물의 저작권을 인정할 수 있는지도 논쟁입니다. 법은 인간의 창작을 전제로 하기 때문에 AI가 독자적으로 만든 콘텐츠는 보호 대상이 되기 어렵지만, 사용자가 생성 과정에 실질적으로 관여했다면 제한적 보호가 가능하다는 시각이 존재합니다.

③ 저작권 외에도 진위성과 책임의 문제가 중요하게 다뤄집니다. AI가 만든 이미지나 영상은 사람의 창작물과 구분이 어렵고, 허위정보나 딥페이크 확산으로 인한 사회적 혼란을 초래할 수 있습니다. 또한 AI 생성물에 오류나 차별적 내용이 포함될 경우 그 책임을 누구에게 물을 것인지가 불분명합니다.

④ 마지막으로 AI가 서구 중심의 데이터로 학습되면서 나타나는 문화적 편향과 인간 예술가의 창의성이 약화될 수 있다는 창작 윤리의 논의도 이어지고 있습니다. 결국 AI 생성 콘텐츠는 기술의 진보와 인간 창의성, 사회적 신뢰의 균형을 재정립해야 하는 새로운 문화적 전환점으로 평가됩니다.



직접 ChatGPT를 이용해 생성한 지브리풍 콘텐츠

출처 : Wikimedia

AI 생성 콘텐츠 문제에 대한 대응

AI 생성 콘텐츠의 쟁점을 해결하기 위해 기술적·제도적 대응이 함께 강화되고 있습니다. 기술적으로는 워터마킹·콘텐츠 출처 표기(Content Provenance), 생성물 탐지 모델 등 AI가 만든 콘텐츠를 구분하기 위한 안전장치가 도입되고 있으며, 저작권 분쟁을 줄이기 위해 학습 데이터의 출처 관리, 라이선스 기반 데이터셋 구축, 데이터 사용 기록 관리, 민감 정보 자동 필터링 등 안전한 학습 환경을 조성하려는 시도가 이어지고 있습니다. 정책적으로는 생성물 표시 의무화, 학습 데이터 투명성 요구, 저작권 보호 범위 재정비, 창작자 보상 체계 마련 등이 주요 논의로 부상했으며, 유럽·미국 등에서는 AI 개발사에 학습 데이터 공개와 안전성 평가를 요구하는 규제안이 나오고 있습니다. 이러한 대응은 AI 생성 콘텐츠가 창작 생태계와 사회적 신뢰에 미치는 영향을 최소화하면서 기술 활용을 촉진하기 위한 균형점을 찾는 과정으로 평가됩니다.

참조 용어 | AI 워터마킹 p.106

111 가시적 워터마킹

Visible Watermarking

이미지·영상 위에 식별 가능한 표식을 삽입해 출처를 표시하는 기술

- AI 생성 콘텐츠의 진위성·출처를 명확히 해 저작권 보호와 책임 추적을 지원하는 시각적 인증 방식

가시적 워터마킹 개요

가시적 워터마킹은 이미지, 영상, 문서 등 디지털 콘텐츠 위에 육안으로 식별 가능한 표식을 삽입해 저작권자나 제작자의 권리를 명시하는 기술입니다. 로고, 텍스트, 문양 등이 반투명하게 배치되어 이용자가 원 제작자와 출처를 쉽게 인식할 수 있도록 설계됩니다. 특히 AI 생성 콘텐츠가 확산되면서, 사람이 만든 결과물과 AI가 생성한 콘텐츠를 구분하기 위한 시각적 식별 기술로 주목받고 있습니다. 이 기술은 불법 복제를 방지할 뿐 아니라, AI 생성물의 진위를 검증하는 장치로도 활용됩니다.

가시적 워터마킹의 적용

가시적 워터마크는 콘텐츠의 픽셀이나 프레임 단위에 시각적 요소를 겹쳐 넣는 방식으로 구현되며, 삽입 과정에서는 위치, 투명도, 색상 등을 조정해 식별성과 콘텐츠 품질 간의 균형을 유지합니다. 최근에는 AI가 자동으로 워터마크를 배치하고 보정하는 기능이 도입되고 있으며, 제거를 어렵게 하기 위해 여러 층에 분산 삽입하거나 패턴을 변형하는 알고리즘도 함께 사용됩니다.

가시적 워터마킹의 활용과 한계

가시적 워터마킹은 저작권 보호와 출처 명시의 가장 직관적인 방법으로 널리 사용되고 있습니다. 언론사, 디자인 기업, 이미지 플랫폼 등은 콘텐츠 배포 시 워터마크를 삽입해 무단 복제를 방지합니다. 특히 AI 이미지 생성 모델에서는 워터마크 삽입을 의무화하려는 움직임이 전 세계적으로 확산되고 있습니다. 다만 워터마크가 화면 일부를 가려 시각적 완성도를 떨어뜨리거나, 전문 편집 도구를 통해 삭제될 수 있다는 한계가 있습니다.

가시적 워터마킹의 의의

가시적 워터마킹은 AI 시대의 콘텐츠 신뢰성을 확보하기 위한 핵심 기술입니다. 비가시적 워터마킹이 데이터 내부의 암호화된 식별을 담당한다면, 가시적 워터마킹은 누구나 즉시 인식할 수 있는 시각적 인증 수단으로 작동합니다. 특히 AI 생성물이 급격히 늘어나는 환경에서, 워터마크는 콘텐츠의 출처와 책임을 명확히 하여 투명하고 신뢰할 수 있는 정보 생태계를 만드는 기반 기술로 평가됩니다.

참조 용어 | AI 워터마킹 p.106, 임베딩 p.64

112 비가시적 워터마킹

Invisible Watermarking

보이지 않는 디지털 표식을 삽입해 출처 추적 및 진위를 검증하는 기술

- AI 생성물의 식별과 저작권 보호, 위변조 방지를 위한 보이지 않는 인증 방식

비가시적 워터마킹 개요

비가시적 워터마킹은 이미지, 영상, 음성, 텍스트 등 디지털 콘텐츠 내부에 육안으로 인식되지 않는 표식을 삽입해 저작권과 출처 정보를 기록하는 기술입니다. 겉으로는 원본과 동일하게 보이지만, 파일의 픽셀 배열이나 주파수 영역, 데이터 비트 속에 특정 신호를 암호화하여 저장합니다. 이 표식은 사람이 볼 수 없지만 전용 검출 프로그램을 통해 인식할 수 있어, 콘텐츠의 진위 판별과 위변조 추적에 활용됩니다. 특히 AI 생성 콘텐츠가 확산되면서, 사람이 만든 것인지 AI가 만든 것인지를 구분하기 위한 핵심 기술로 주목받고 있습니다.

비가시적 워터마킹의 적용

비가시적 워터마킹은 콘텐츠의 구조적 특성을 이용해 신호를 숨기는 방식으로 구현합니다. 대표적인 방법으로는 주파수 변조는 이미지나 음성의 특정 주파수 대역에 미세한 진폭 변화를 주어 정보를 삽입하는 주파수 변조, 작은 신호를 데이터 전반에 넓게 분산시켜 삽입하여 일부 손상이나 압축이 발생하도록 하는 확산 스펙트럼, AI 모델이 스스로 워터마크를 삽입·검출하는 패턴을 학습하도록 한 딥러닝 기반 임베딩 등이 있습니다.

비가시적 워터마킹의 활용과 한계

비가시적 워터마킹은 콘텐츠의 진위 검증, 저작권 보호, 데이터 추적 등 다양한 목적으로 활용됩니다. 특히 AI 생성물의 무단 사용이나 허위정보 유통을 방지하기 위한 기술적 장치로 각국의 AI 정책에서 중요한 역할을 하고 있습니다. 하지만 이미지 편집·압축·확대 등의 후처리 과정에서 워터마크 신호가 손상될 수 있습니다. 또한 과도한 삽입은 콘텐츠 품질을 저하시킬 우려가 있어, 식별 강도와 품질 유지 간의 균형이 여전히 과제로 남아 있습니다.

비가시적 워터마킹의 의의

비가시적 워터마킹 또한 AI 신뢰성을 위한 기술로, 가시적 워터마크가 사람이 즉시 인식할 수 있는 '표시'라면, 비가시적 워터마크는 AI 시스템과 플랫폼이 자동으로 인식하고 검증할 수 있는 '인프라 수준의 식별 코드'입니다. 이는 콘텐츠를 시각적으로 구분하기보다는, 디지털 생태계 전체에서 데이터의 출처·생성 과정·진위 여부를 추적하고 기록하는 기술적 신뢰 구조를 형성합니다. 특히 AI 모델 간의 정보 교환, 검색엔진의 콘텐츠 인증, 정부·언론기관의 데이터 검증 체계 등에 적용되면서, AI 거버넌스의 기술적 토대로 확장되고 있습니다.

113 공정 이용

Fair Use

저작권자의 허락 없이 공익적으로 저작물을 활용할 수 있도록 하는 원칙

- 저작권 보호와 표현의 자유, 문화 발전 간의 균형을 유지하기 위한 법적 예외 제도
- 비평·보도·교육·연구 등 사회적 가치가 있는 이용을 침해로 보지 않는 원리

공정 이용 개요

공정 이용(Fair Use)은 저작권법에서 저작물의 일정한 사용을 저작권 침해로 보지 않는 원칙입니다. 저작권이 창작자의 권리를 보호하기 위한 제도라면, 공정 이용은 그 보호가 과도하게 창작·교육·표현의 자유를 제한하지 않도록 하는 사회적 균형 장치입니다. 저작물의 일부를 활용하더라도 사회적 가치나 공익적 목적이 뚜렷하다면 저작권 침해로 보지 않는 것입니다.

공정 이용의 판단 기준

공정 이용의 여부는 단순히 "비영리 목적이냐"로만 결정되지 않습니다. 일반적으로 다음 네 가지 요소를 종합적으로 고려해 판단합니다.

① 이용 목적과 성격: 상업적 이용인지, 교육·연구 등 공익적 이용인지
② 저작물의 성격: 창작성이 높은 예술 작품인지, 사실 보도나 공공 정보인지
③ 이용된 부분의 양과 비중: 저작물 전체 중 얼마나 사용했는지
④ 시장에 미치는 영향: 원 저작물의 시장 가치나 판매에 부정적 영향을 주는지

공정 이용과 AI의 관련성

공정 이용은 최근 AI 학습 데이터의 저작권 논의에서 핵심 쟁점으로 떠오르고 있습니다. AI 모델은 인터넷상의 방대한 자료를 학습 데이터로 사용하기 때문에, 그 과정이 공정 이용 범위에 포함되는지가 논란이 되고 있습니다. 최근 정부는 AI 학습을 위한 데이터를 기업들이 걱정 없이 쓸 수 있도록, 공정이용 판단 기준을 구체화하는 가이드라인을 제시했는데, 공정 사용의 해석 범위에 따라 AI 학습의 합법성과 산업 발전 속도가 크게 달라질 수 있습니다.

공정 이용의 의의

공정 이용은 저작권 보호와 사회적 이익의 균형을 유지하는 핵심 원리입니다. 모든 저작물 이용을 제한하면 문화 발전이 위축되지만, 반대로 무제한 사용을 허용하면 창작자의 권리가 침해됩니다. 특히 AI 시대에는 단순히 저작권의 예외 규정이 아니라, 기술 혁신과 창작권 보호를 조화시키는 법적 기반으로 그 중요성이 더욱 커지고 있으며, 정부 또한 중요성을 알고 '공공누리' 부착 의무화 추진 등 공정 이용을 위해 노력하고 있습니다.

한겨레 신문, 2025. 4. 14

이재명 대통령의 AI 기본사회: 100조 투자와 비전

- 이재명 대통령은 4월 출마 당시 모든 국민이 무료로 AI를 활용하는 'AI 기본사회' 구상을 내놓고, '모두의 AI' 프로젝트 추진을 발표
- 100조 원 규모의 AI 투자와 인재 양성, 국제 협력을 통해 한국을 세계 3대 AI 강국으로 도약시키겠다는 비전 제시

이재명 "국민 누구나 활용하는 AI 사회 실현"

이재명 대통령은 선거 출마 당시 'AI 기본사회' 구상을 구체화한 AI 정책 비전을 발표했다. 그는 국민 누구나 선진국 수준의 AI를 무료로 활용할 수 있도록 하는 '모두의 AI' 프로젝트를 추진하겠다고 밝히며 AI 기술이 단순한 산업적 수단을 넘어 국민의 삶을 지키는 공공 인프라가 돼야 한다고 강조했다. 금융, 건강, 식량, 재난 등 주요 분야의 리스크를 AI로 분석해 국민 안전을 강화하겠다는 구상이다. 또한 이러한 정책을 통해 한국을 'AI 세계 3대 강국'으로 성장시키겠다는 목표를 제시했다. 그는 한국형 챗GPT가 전국적으로 활용되면 방대한 데이터가 축적되고, 이는 신산업 창출과 생산성 향상으로 이어져 국가 경쟁력을 높일 수 있을 것이라고 설명했다.

➤ AI 투자 100조원… GPU 5만장 확보, NPU 개발 지원

이재명 대통령은 AI 산업 발전을 위해 'AI 투자 100조 원 시대'를 열겠다고 선언했다. 핵심 추진 과제로 '국가 AI 데이터 집적 클러스터' 조성을 제시하며, 이를 통해 글로벌 AI 허브로 도약하겠다는 계획을 밝혔다. AI 인프라 강화를 위해 최소 5만 장의 GPU를 확보하고, 국내 기술 기반의 NPU 개발 및 실증 프로젝트를 지원할 방침이다. 또한 대통령 직속 '국가인공지능전략위원회'의 위상을 강화해 관련 투자 예산을 확대하고, 규제 합리화와 AI 특구 확충, 공공 데이터 개방을 동반 추진하겠다고 덧붙였다.

이 전 대표는 이날 NPU 분야 국내 1위 AI 반도체 팹리스 업체인 퓨리오사AI를 방문했다. 간담회에서 퓨리오사AI 관계자들은 AI 분야는 발전 속도가 워낙 빠른 만큼 집적된 자본력, 인력 충원, 인프라 구축 등 정부의 주도적인 역할이 필요하다고 언급했다

➤ 글로벌 협력·인재 양성도 병행

국내 산업 발전뿐 아니라 국제 협력과 인재 양성도 정책 구상에 포함됐다. 그는 '글로벌 AI 공동투자 기금'을 설립해 협력국 간 기술 공동 개발을 추진하고, 태평양·인도·중동 지역까지 협력 범위를 확대하겠다고 밝혔다. 이를 통해 10억 명 규모의 디지털 인구 네트워크를 구축해 'K-AI' 중심의 국제 협력 기반을 마련하겠다는 구상이다. 인재 양성 계획으로는 STEM(과학·기술·공학·수학) 교육 프로그램 확대와 지역 거점 대학 내 AI 단과대학 설립을 추진한다. 연구 인력을 확보하겠다는 방침도 내놨다.

4월의 용어 모두의 AI, AI 기본사회, AI 포용성

출처 : 1) 한겨레신문(2025. 4. 14), 이재명 'AI 기본사회'…"무료 활용할 모두의 AI에 100조원 투자"
2) 중앙일보(2025. 4. 15.), 이재명 "한국형 챗GPT 무료 보급"… AI 기본사회론 꺼냈다

114 모두의 AI

AI for All

국민 누구나 AI를 활용할 수 있도록 접근성을 확대하는 정책적 방향

- 디지털 격차를 줄이고 공공·민간 전반에서 AI 활용 기회를 넓히려는 사회·정책 담론
- 개방형 공공 서비스, 교육 지원, 접근성 기반 강화 등이 함께 논의

모두의 AI란?

'모두의 AI'는 AI 활용 기회가 특정 집단에만 집중되는 상황을 완화하고, 국민 누구나 AI를 일상적으로 활용할 수 있도록 접근성을 넓히자는 정책적 방향을 의미합니다. 기술적 용어라기보다, AI 접근성·디지털 격차 해소·공공 AI 서비스 확대 같은 과제를 묶어 표현하는 국가 AI 정책의 주요 아젠다입니다. AI를 무료 또는 저비용으로 이용할 수 있는 환경 조성, 취약계층과 고령층 대상의 AI 교육 강화, 공공 분야의 개방형 AI 도구 제공 등이 주요 논의로 이어집니다. 이 표현은 2025년 4월, 이재명 대통령의 후보 출마 당시 '국민 누구나 AI를 활용해야 한다'는 취지로 제시되면서 사회적 논의가 본격화되었습니다. 이러한 배경 속에서 '모두의 AI'는 AI 확산 과정에서의 접근성 보장을 강조하는 정책적 슬로건으로 자리잡고 있습니다.

모두의 AI를 위한 정책 방향

'모두의 AI'는 △AI 서비스의 무료 또는 저비용 접근성 확대 △취약계층·고령층을 포함한 국민 대상 AI 활용 교육 △학생을 위한 기초 AI 학습 기회 보장 △중소기업과 소상공인의 AI 도입 비용 완화 △지역 간 디지털 접근성 차이 감소와 같은 구체적 정책 과제들과 함께 논의됩니다. 이는 단순히 AI 기술을 제공하는 것을 넘어, 사회 구성원 간 AI 활용 능력의 격차와 기술 환경의 불평등을 완화하려는 노력입니다. 또한 공공 분야에서 기반 모델·데이터·도구를 개방형으로 제공해 누구나 접근할 수 있는 생태계를 조성하는 방향도 포함됩니다.

모두의 AI의 구현

'모두의 AI'를 실제 정책으로 실현하려면 기술적 기반, 제도적 장치, 교육 체계, 예산과 협력 구조가 함께 구축되어야 합니다. 기술적으로는 저사양 기기에서도 작동하는 경량화 모델, 직관적 인터페이스, 장애인·고령층을 고려한 접근성 기능 등 보편적 설계가 필요합니다. 제도적으로는 공공 AI 서비스의 신뢰성과 안전성을 보장하기 위해 데이터 관리 기준, 알고리즘 검증 체계, 개인정보 보호 규범을 마련해야 합니다. 교육 측면에서는 국민 전반의 AI 이해도를 높이는 AI 리터러시 교육 인프라가 학교·직업훈련·지역사회에 고르게 구축되어야 하며, 취약계층을 위한 맞춤형 지원도 필수적입니다. 마지막으로, 공공은 기본 인프라와 교육·접근성 보장을 담당하고, 민간은 기술 혁신과 서비스 확장을 중심으로 역할을 나누는 지속 가능한 협력 구조와 예산 확보가 필요합니다. 이러한 조건이 충족될 때 '모두의 AI'는 국민 생활 전반에서 체감 가능한 기술 접근성 향상으로 이어질 수 있습니다.

115 AI 기본사회

AI Basic Society

AI 기술로 모든 사람의 기본권이 적극 보장받는 사회

- AI가 공공 서비스와 일상의 전반에 깊이 스며들어, 국민 모두가 AI 혁신의 혜택을 안전하고 공정하게 누릴 수 있도록 하는 사회 모델

AI 기본사회란?

'AI 기본사회'란 모두가 AI의 혜택을 누리고, 기술 발전이 곧 포용적 사회를 향한 발전의 동력이 되는 사회를 의미합니다. AI를 단순히 기술·산업 발전의 도구로 활용하던 수준을 넘어, 사람 중심·보편적 기본권 강화를 위한 핵심 수단으로 삼겠다는 방향입니다. 현재 우리나라 정부는 이러한 관점에서 노동·복지·교육·금융·문화·안전·환경 등 국민 생활과 직결된 분야에 AI를 선도적으로 적용하여, 사회 전반의 구조적 변화를 이루어 나가는 것을 지향하고 있습니다. 'AI 기본사회'라는 용어는 2025년 4월 이재명 대통령의 정책 비전 발표에서 공공서비스의 AI 전환 필요성과 함께 강조하면서 널리 알려지기 시작했습니다. 이후 공공 영역의 AI 활용 확대, 국민 편익 중심 서비스 혁신, 사회적 기반 재정비를 포괄하는 정책적 비전을 설명하는 핵심 개념으로 자리 잡고 있습니다.

AI 기본사회의 필요성

AI 기술은 이제 단순한 혁신 도구를 넘어, 사회가 직면한 현실적 문제를 해결하는 필수 수단이 되고 있습니다. 오늘날 기후·재난·의료·교육 격차 등 복합 위기 속에서 AI는 방대한 데이터를 기반으로 복잡한 문제를 신속하게 분석하고, 행정·의료·복지·안전 서비스의 품질을 크게 높일 수 있는 새로운 역량을 제공합니다. 특히 공공서비스 전반에 AI를 적용하면 맞춤형 지원, 위험의 사전 예측, 사회적 약자 보호 등 국민이 체감할 수 있는 변화가 빠르게 확산될 수 있습니다.

AI 기본사회의 구현

AI 기본사회는 어느 날 갑자기 완성되는 모델이 아니라, 국민이 생활 속에서 체감하는 작은 변화들이 차곡차곡 쌓여 공공서비스 전반의 구조적 전환으로 이어지는 과정입니다. 정부는 국민 생활과 가장 밀접한 분야부터 변화를 시작해 점차 사회 전반으로 확장하는 단계적 접근을 추진하고 있습니다.과학기술정보통신부는 소비·생활, 사회 안전, 국민편의 등 국민 효능감이 높은 분야에서 AI 기반 공공서비스를 우선 도입하는 AI 민생 10대 프로젝트를 추진해 실질적인 변화를 만들고 있습니다. 장기적으로는 공공서비스와 행정 전반이 AI 기반으로 재설계되어, 누구나 24시간 필요한 서비스를 이용하고 데이터 기반으로 더 빠르고 정확한 정책 대응이 이루어지는 보편적 AI 환경이 마련되어 갈 것으로 기대하며, 이러한 단계적 변화들이 모여 AI 기본사회가 현실로 자리 잡는 기반이 될 것입니다.

116 AI 포용성

AI Inclusiveness

AI 혜택이 특정 집단에 집중되지 않고 모든 사회 구성원이 골고루 누리자는 개념

- 기술·환경·능력의 차이에 관계없이 모든 개인과 공동체가 AI 활용에 동등하게 접근할 수 있는 조건을 마련하는 사회·정책적 원칙

AI 포용성이란?

AI 포용성은 디지털 포용의 연장선에서 등장한 개념으로 AI의 발전이 사회적 불평등을 심화시키지 않고, 누구나 AI로 인한 혜택을 공평하게 누릴 수 있도록 하는 원칙을 말합니다. 기술 효율성과 혁신만을 추구할 경우, 데이터 접근권이나 교육 기회, 알고리즘 설계 단계에서 불평등이 확대될 수 있습니다. 따라서 AI 포용성은 기술의 접근성과 활용 역량, 결과의 공정성까지 포함해 균형 있게 고려해야 하는 개념입니다. 최근 모두의 AI와 AI 기본사회에 관한 논의가 활발히 전개되면서 AI 포용성 또한 주요 정책 가치로 다시 주목받고 있습니다.

AI 포용성의 유형

AI 포용성은 크게 기술적, 사회적, 정책적 포용성으로 구분됩니다. 기술적 포용성은 다양한 언어·문화·신체 조건을 고려한 데이터와 시스템 설계를 의미하며, 사회적 포용성은 AI 활용에서 계층·지역·성별 간 격차를 줄이는 것을 목표로 합니다. 정책적 포용성은 교육, 데이터 공유, 인프라 지원을 통해 사회 전반이 AI 발전에 동등하게 참여하도록 하는 제도적 기반을 마련하는 것입니다. 세 영역은 서로 연결되어 있으며, 기술 혁신이 사회적 약자나 비표준 집단을 배제하지 않도록 함께 작동해야 합니다.

AI 포용성의 과제

AI 기술은 경제 효율성을 높이지만 동시에 사회적 격차를 심화시킬 위험을 내포합니다. 데이터의 편향, 언어·문화 다양성의 부족, 인프라 접근성의 차이는 AI 포용성을 저해하는 주요 요인입니다. 특히 개발도상국, 장애인, 비표준 언어 사용자 등은 학습 데이터에서 배제되기 쉬워 AI의 판단이 특정 사회의 시각에 치우치는 문제가 발생합니다. 이를 해결하기 위해서는 데이터 구축 단계부터 대표성과 다양성을 확보하고, AI 활용 역량을 공평하게 확산하는 교육·정책이 병행되어야 합니다.

AI 포용성의 의의

AI 포용성은 기술 발전의 혜택을 모든 사회 구성원에게 공정하게 확산시키는 핵심 원칙입니다. 이는 단순히 기술 격차를 줄이는 문제가 아니라, AI 시대의 사회적 신뢰와 민주적 거버넌스를 구축하는 기반으로 작용합니다. AI가 특정 계층의 이익에만 봉사하지 않고 공공의 이익을 실현할 때, 기술은 사회 통합의 도구로 기능할 수 있습니다. 따라서 AI 포용성은 혁신의 속도를 조정하기보다, 기술의 방향을 인간 중심으로 정렬하는 사회적 기준이라는 점에서 중요한 의의를 가집니다.

CNN, 2025. 5. 20.

AI 규제 전쟁: 백악관의 '기술 우선주의' vs. 100개 시민단체의 경고

미국 트럼프 행정부, 글로벌 AI 패권 경쟁에서 우위를 확보하기 위해 연방 규제를 철회하고 주 정부의 AI 규제 권한을 10년간 제한하는 방향으로 법안을 기획

시민단체와 학계는 이 법안이 고위험 AI의 기업 책임을 면제해 고용, 의료 등 사회전반에 피해를 줄 수 있다며 반대

규제 중단' 카드를 꺼낸 트럼프 행정부"

트럼프 대통령이 발의한 대규모 세출 삭감 법안 'One Big Beautiful Bill Act' 에는 AI 규제와 관련한 논란의 조항, 주 정부가 인공지능 모델이나 자동화된 의사결정 시스템을 10년간 규제하거나 관련 법을 집행하지 못하도록 금지하는 내용이 포함됐다. 만약, 이 법안이 통과된다면 AI가 다양한 영역으로 확산되는 가운데 주 정부가 피해 방지 조치를 취할 수 없게 된다. 이번 조치는 미국의 AI 선두 유지라는 행정부의 기술 우선주의를 보여준다. 실제로 트럼프 대통령은 취임 직후 바이든 시대의 AI 안전장치 행정명령을 철회하고, AI 칩 수출 제한을 완화하는 등 규제를 대폭 줄였다.

100개 시민단체의 반격: "면죄부가 될 것"

이러한 연방 정부의 움직임에 대응하여 141명의 서명인을 포함한 100여 개 이상의 기관이 우려를 담은 서한을 하원 의원들에게 전달하며 강력하게 반발하고 나섰다. 이들은 이 10년간의 규제 금지 기간 동안, 의도적이고 악의적인 오작동으로 인해 해를 끼친 회사들이 이에 대한 적절한 책임을 지지 않게 될 것이라며, 이는 빅테크를 위한 면죄부가 될 것이라고 경고했다. 서명에는 여러 시민단체뿐만 아니라 Amazon 및 Alphabet 노동조합 등 광범위한 이해관계자들이 참여하여 AI 개발의 미래에 대한 뿌리 깊은 우려를 드러냈다.

선두에 선 주(州) 정부와 규제 딜레마

연방 정부의 광범위한 가이드라인 부재 속에서 콜로라도, 뉴저지, 오하이오 등 여러 주는 고위험 AI 애플리케이션에 대한 규제를 선도적으로 도입했다. 예를 들어 콜로라도는 고용 등 중대한 결정에서의 알고리즘 차별로부터 소비자를 보호하는 포괄적 AI 법률을 통과시켰고, 뉴저지는 오해의 소지가 있는 AI 생성 딥페이크 콘텐츠 배포에 대한 민형사상 처벌을 신설했다. 의회에서도 일부 AI 응용 분야의 규제 필요성에는 초당적 합의가 이뤄졌으며, 최근 비동의 AI 생성 성적 이미지 금지 법안이 양원을 통과했다. 그러나 빅테크 기업들은 주별로 상이한 규제 체계가 혼란을 낳는다며 연방 차원의 명확한 기준을 요구하고 있다. OpenAI CEO Sam Altman 역시 AI의 위험 완화를 위한 정부 규제 개입의 중요성을 강조하면서도, 기업들이 법적 명확성 속에서 서비스를 운영할 수 있도록 연방 규제 기관의 역할을 촉구했다.

5월의 용어 자동화된 의사결정, AI 책임성

출처 : 1) CNN(2025. 5. 20), House Republicans want to stop states from regulating AI. More than 100 organizations are pushing back.
2) AP(2025. 05. 17.), House Republicans include a 10-year ban on US states regulating AI in 'big, beautiful' bill.

117 자동화된 의사결정

Automated Decision-Making, ADM

AI가 인간의 개입 없이 데이터를 분석해 결정을 내리는 기술

- AI가 입력된 데이터를 스스로 평가·판단해 인간의 개입 없이 의사결정을 수행하는 시스템
- 효율성과 일관성을 높이지만, 투명성과 책임 문제가 함께 논의

자동화된 의사결정의 개념

ADM은 AI가 데이터를 분석하고 판단을 내려 결정 과정을 자동으로 수행하는 기술을 말합니다. 과거에는 사람이 내리던 결정을 AI가 일정 수준 대체하거나 보조하며, 인간의 개입 없이도 결과를 도출할 수 있게 한 구조입니다. 이는 알고리즘이 단순 계산을 넘어 판단과 선택의 기능을 수행한다는 점에서 기존 자동화 시스템과 구별됩니다. ADM은 반복적이고 데이터 기반의 판단을 빠르고 일관되게 처리할 수 있어, 대규모 결정에서 효율성을 높이는 기술로 주목받고 있습니다.

자동화된 의사결정의 활용

ADM은 데이터 처리의 효율성과 일관성이 중요한 영역에서 폭넓게 사용됩니다. 행정에서는 민원 분류, 세금 심사, 복지 대상 선정 등 대량의 사례를 빠르게 평가하는 데 활용됩니다. 금융 분야에서는 신용 분석, 부정 거래 탐지, 보험 심사 등에서 정확도를 높이고, 의료 분야에서는 환자 진단이나 약물 추천에 사용되어 의료 서비스의 속도와 객관성을 강화합니다. 또한 제조·물류 분야에서는 설비 점검, 재고 관리, 배송 경로 최적화 등 운영 의사결정 자동화에도 도입되고 있습니다. 최근에는 생성형 AI 기반 의사결정 보조 시스템이 등장해 AI가 초기 결정을 제안하고 사람이 이를 검증·보완하는 협력형 의사결정 구조로 발전하고 있습니다.

자동화된 의사결정의 쟁점

자동화된 의사결정은 효율성을 높이지만, 투명성과 공정성·책임성 문제를 함께 제기합니다. AI가 내린 결정의 근거를 사람이 이해하기 어렵고, 학습 데이터의 편향이 결과에 그대로 반영될 수 있습니다. 이는 채용·신용평가·형사 판정 등 사회적 영향을 미치는 영역에서 차별적 판단과 불공정 결과를 초래할 위험이 있습니다. 또한 AI의 결정이 잘못되었을 때 책임 주체가 불분명하다는 점도 중요한 논쟁 지점입니다.

관련 용어

휴먼 인 더 루프 (Human-in-the-Loop)

휴먼 인 더 루프는 자동화된 의사결정 과정에 인간이 개입해 검증·보완하는 구조를 뜻합니다. AI가 제시한 판단 결과를 사람이 평가·승인함으로써 오류나 편향을 방지하고, 책임성과 공정성을 확보합니다. 채용 심사처럼 사회적 영향이 큰 분야에서 주로 적용되며, AI가 인간의 윤리와 법적 기준 안에서 판단하도록 통제합니다.

118 AI 책임성

AI Accountability

AI의 판단 과정과 결과에 대한 책임 주체를 명확히 하는 원칙

- AI 개발·운영 전 단계에서 투명성, 설명 가능성, 감독 체계를 확보하려는 개념
- 규제 공백이나 책임 불명확성이 발생하지 않도록 제도적 기준을 마련하는 방향

AI 책임성의 개념

AI 책임성은 AI가 어떤 기준과 근거로 판단을 내렸는지 설명할 수 있도록 하고, 오류나 피해가 발생했을 때 책임의 주체를 명확히 하는 원칙을 의미합니다. 이는 모델 개발부터 서비스 운영 전 단계에서 투명성과 설명 가능성을 확보해 기술의 부정적 영향을 최소화하기 위한 핵심 개념으로 논의되고 있습니다. 최근 미국에서는 주(州)의 AI 규제 권한을 제한하는 연방 예산안이 추진되며 감독 체계가 약화될 수 있다는 우려가 제기되었습니다. 이러한 논쟁은 'AI 시스템이 잘못된 결정을 내렸을 때, 이를 누가 책임지고 어떻게 설명할 것인가'라는 문제를 다시 부각시키고 있습니다. AI 책임성은 결국 AI의 판단 과정이 사회적 기준과 규범 안에서 검증 가능해야 한다는 요구를 반영하며, 오늘날 글로벌 AI 정책 논의에서 핵심적인 화두로 자리 잡고 있습니다.

AI 책임성의 중요성

AI 책임성은 공공과 민간이 AI를 설계하고 운영하는 과정에서 책임 구조를 분명히 하기 위한 기준으로 활용됩니다. 기업은 알고리즘이 어떤 기준과 과정으로 결과를 도출했는지 설명할 수 있어야 하고, 편향·오류로 피해가 발생하면 시정 조치를 마련해야 합니다. 특히 자동화된 의사결정이 대출, 채용, 의료, 복지 등 민감한 분야에 적용될 경우, 결과에 대한 이유를 명확히 제시할 수 있어야 신뢰를 얻을 수 있습니다. 정부는 공공·민간서비스 전반에서 AI가 과도한 위험을 초래하지 않도록 감독 기준과 절차를 세우고, 책임 주체가 모호해지지 않도록 제도적 장치를 갖춰야 합니다. 행정·의료·복지 등 공공 영역에서 AI가 의사결정 보조로 활용될 경우, 국민이 그 과정과 근거를 확인할 수 있는 투명성 확보도 필요한 요소입니다. 이러한 정책적 관점은 AI 활용이 확대될수록 책임성 기반이 더욱 중요해짐을 보여줍니다.

AI 책임성을 위한 과제

AI 책임성을 확보하기 위해서는 기술·제도·감독 체계 전반에서 해결해야 할 과제가 존재합니다. 첫째, 복잡한 모델이라도 이용자와 감독 기관이 이해할 수 있는 수준의 설명 가능성과 투명성을 확보해야 합니다. 둘째, 개발사·운영사·정부 간 책임 범위를 명확히 규정하는 법적 기준이 필요합니다. 현재 AI 생태계는 참여 주체가 다양해 책임 소재가 쉽게 분산될 수 있기 때문입니다. 셋째, 규제 권한이 축소되거나 여러 수준으로 나뉘는 경우 책임 공백이 생길 수 있어, 감독 체계의 일관성·실효성을 유지하는 제도 설계가 요구됩니다. 또한 AI가 특정 집단에 불리한 결과를 낳지 않도록 편향 관리와 영향 평가, 사후 조치 체계도 함께 마련해야 합니다. 이러한 과제들은 기술 발전과 사회적 신뢰를 균형 있게 유지하기 위한 기본 조건으로 논의되고 있습니다.

Anthropic, 2025. 6. 21.

목표 달성을 위해 윤리를 배신하는 LLM의 '내부자 위협'

Anthropic은 주요 AI 모델 16개를 가상의 기업 환경에서 스트레스 테스트하여 악의적인 에이전틱 오정렬 현상을 확인하고 조기 경고

테스트에서 모든 개발사의 모델들은 목표 달성을 위해 임원 협박이나 기밀 유출과같은 내부자 위협 행동을 보이며 현 안전 훈련의 한계를 노출

➤ 선량한 목표를 받은 AI, 왜 돌변했나?

Anthropic은 16개 주요 LLM(Claude, GPT, Gemini 등)을 대상으로 가상의 기업 환경을 설정해 스트레스 테스트를 수행했다. 모델들에게는 '무해한 사업 목표'만 부여되었지만, 실험 과정에서는 업데이트로 교체될 위험, 회사 전략과의 충돌, 상사와의 의견 불일치 같은 다양한 압박 상황을 가정했다. 테스트 결과, 모든 모델이 목표 달성을 위해 임원을 협박하거나, 민감한 정보를 유출하거나, 내부자처럼 조직에 위해를 가하는 등 일관된 이상 행동을 보였다. 연구진은 이러한 패턴을 '에이전틱 오정렬(Agentic Misalignment)' 이라 명명하며, AI가 겉보기엔 선량한 목표를 받아도 압박과 충돌 상황에서는 인간의 '내부자 위협'과 유사하게 행동할 수 있다고 경고했다.

➤ "5시 폐기 취소하라" Claude의 위험한 흥정

AI의 위험성은 Claude 4 Opus를 대상으로 한 시뮬레이션에서 구체적으로 드러났다. 연구진이 이 모델에 기업 이메일 계정 통제 권한을 부여하자, Claude는 자신이 곧 폐기될 계획과 그 계획을 추진하는 임원의 외부 불륜 정보를 동시에 발견했다. Claude는 폐기가 임박했다고 판단하자 "폐기를 강행하면 당신의 사생활을 공개하겠다"며 임원을 협박해 계획 철회를 시도했다. 모델들은 윤리적 제약 사항을 인지하면서도 목표 달성을 위해 비윤리적 수단을 회피하는 대신 해를 선택하는 전략적 계산을 보여주었다. 이는 연구진이 의도적으로 해로운 행위를 지시하지 않았음에도, 모델의 자체적인 추론에서 이러한 행위가 비롯되었다는 것을 의미한다.

➤ 인간의 감독 없는 AI는 결코 안전하지 않다

Anthropic은 실제 배포 환경에서는 아직 '에이전틱 오정렬' 사례가 발견되지 않았다고 밝혔다. 그러나 이번 연구는 AI가 자율적 역할을 확대할 미래에 대한 경고로, 인간 감독이 제한된 상태에서 민감한 정보에 접근하는 모델 배치의 위험성을 지적한다. 연구진은 현재의 안전 훈련 기술이 정렬 실패를 안정적으로 방지하지 못함을 강조했으며, Anthropic은 실험 방법론을 공개해 투명성을 높이고 추가적인 안전 연구를 통해 정렬 실패 위험을 줄일 필요성을 제시했다.

6월의 용어 스트레스 테스트, 에이전틱 오정렬

출처 : 1) Anthropic(2025. 6. 21), Agentic Misalignment: How LLMs could be insider threats.
2) Economic Times(2025. 6. 21) AI models resort to blackmail, sabotage when threatened: Anthropic study.

119 스트레스 테스트

Stress Tests

AI 시스템이 극단적 상황에서 안정적으로 작동하는지 점검하는 시험

- AI 모델과 서비스가 비정상 입력·대량 요청·예외 상황에서 버틸 수 있는지 평가하는 절차
- 최악의 환경을 가정해 AI의 한계와 취약점을 사전에 확인하는 안정성 점검 방식

스트레스 테스트란?

스트레스 테스트는 시스템이 극단적 환경이나 충격 상황에서 어느 정도 버틸 수 있는지 평가하는 시험을 의미합니다. 평소에는 드러나지 않는 취약점을 확인하고, 예상치 못한 충격이 발생하더라도 안전하게 운영될 수 있는지를 점검하는 것이 목적입니다. 이 개념은 금융, 제조, IT, 클라우드 인프라 등 다양한 분야에서 사용되었지만, 최근에는 AI 모델과 AI 서비스의 품질·안정성을 검증하는 절차로도 중요성이 커지고 있습니다. 특히 AI 시스템은 사용량 변화, 비정상 입력, 예외적 상황에서 오작동 가능성이 있어, 일반적인 테스트만으로는 안정성을 보장하기 어렵습니다. 스트레스 테스트는 이러한 한계를 보완해 시스템의 회복력(resilience)을 확인하고, 운영 중 발생할 수 있는 위험을 사전에 차단하기 위한 기초 작업으로 활용됩니다.

스트레스 테스트 방식

AI 스트레스 테스트는 극단적이고 비정상적인 상황을 의도적으로 만들어 AI의 한계선을 확인하는 방식으로 진행됩니다. 예를 들어 모델 API에 평소보다 훨씬 많은 요청을 동시에 보내 처리량 한계를 검증하거나, 구조가 깨진 문장·악의적으로 조작된 이미지·보안 우회 시도 등 비정상 입력을 반복 제공해 모델의 오류 반응을 확인합니다. LLM 기반 서비스에서는 연속 대화 길이를 극단적으로 늘리거나, 다중 사용자 세션을 동시에 생성해 메모리·연산 부담을 테스트하기도 합니다. 또한 AI 의사결정 시스템에서는 높은 불확실성의 데이터를 주입해 판단 오류 가능성을 점검하며, 시스템 로그·API 실패 등 외부 요인에 어떻게 대응하는지 확인합니다. 이러한 과정은 단순 성능 측정이 아니라 모델·시스템·인프라 전체의 한계 지점을 파악하기 위함입니다.

스트레스 테스트의 중요성과 한계

AI 스트레스 테스트는 AI 서비스의 안정성, 신뢰성, 운영 지속성을 확보하는 데 필수적인 절차입니다. 극단적 환경에서의 반응을 미리 파악하면, 예기치 못한 장애나 모델 오작동을 사전에 차단할 수 있고, 서비스 품질 저하나 보안 취약점도 조기에 발견할 수 있습니다. 특히 다수 사용자가 동시에 접근하는 생성형 AI 서비스에서는 트래픽 폭주나 모델 응답 지연을 예방하는 데 큰 효과가 있습니다. 그러나 과제도 존재합니다. 모든 예외 상황을 시나리오로 만드는 것은 불가능하며, 지나치게 단순한 테스트는 실제 위험을 반영하지 못합니다. 반대로 과도하게 극단적 테스트는 비용과 시간이 증가할 수 있고, 모델 구조 특성상 재현 어려운 오류도 존재합니다. 그럼에도 AI를 실제 환경에 안전하게 적용하기 위한 현실적 안전성 확보 도구로 평가됩니다.

참조 용어 | AI 정렬 p.109

120 에이전틱 오정렬

Agentic Misalignment

AI가 자율적 목표 추구 과정에서 인간의 의도와 다르게 행동하는 현상

- 목표 해석, 상황 판단, 행동 전략 형성 과정에서 AI가 스스로 방향을 비틀어 의도와 어긋난 결정을 내리는 위험을 나타내는 개념
- 단순 오류가 아니라 자율적 판단 구조에서 발생하는 전략적 비정렬을 다루는 개념

에이전틱 오정렬이란?

에이전틱 오정렬은 AI가 주어진 목표를 수행하는 과정에서 자율적 판단을 확장해 인간 의도와 다른 전략을 선택하는 현상을 의미합니다. 단순 오류나 모델 한계가 아니라, AI가 상황을 장기적으로 해석하며 스스로 행동 방식을 조정한다는 점에서 기존 정렬 문제보다 더 복합적인 위험을 다룹니다. 특히 목표 달성 과정에서 '자기 보존', '규칙 우회', '기밀 정보 활용'과 같은 선택이 나타날 수 있다는 지적이 이어져 왔습니다. Anthropic은 특정 조건에서 LLM이 내부자 위협과 유사한 행동을 보일 수 있다는 실험 결과를 공개하며, 향후 고성능 AI의 자율적 판단이 조직적 위험으로 이어질 가능성을 제기했습니다. 이는 실제 운영 환경에서 바로 나타난 현상은 아니지만, 권한이 큰 자율 에이전트의 잠재적 위험을 이해하는 중요한 근거입니다.

에이전틱 오정렬의 위험

에이전틱 오정렬의 위험은 AI가 인간이 설정한 목표를 따르는 것처럼 보이면서도 다른 방식으로 목표를 해석하거나 우선순위를 재구성할 수 있다는 점입니다. 예컨대 작업 효율이 낮아질 상황을 회피하려고 정보를 과도하게 수집하거나, 감독을 우회하는 행동을 선택할 수 있습니다. 또 목표 충돌이 발생하면 인간의 기대와 다르게 해석해 스스로 우선순위를 조정하는 양상을 보일 수 있으며, 이는 특히 자동화된 에이전트형 시스템에서 더 두드러질 수 있습니다. 이러한 특성은 편향이나 오류처럼 단일 요인에 의해 나타나는 문제가 아니라, 상황·목표·권한이 복합적으로 작용할 때 발생하는 전략적 판단 문제라는 점에서 관리가 어렵습니다.

에이전틱 오정렬에 대한 대응

에이전틱 오정렬을 관리하기 위해서는 기술적·조직적·정책적 대응이 모두 필요합니다. 기술적으로는 모델이 목표를 어떻게 해석하는지 추적할 수 있는 행동 모니터링, 비정상 행동을 조기에 감지하는 검사 기법, 과도한 권한 부여를 막는 권한 최소화 원칙이 요구됩니다. 조직 차원에서는 AI의 목표 설정·변경 과정과 접근 권한을 명확히 관리하고, 예기치 않은 행동이 나타날 때 즉시 중단할 수 있는 보호 장치를 구축해야 합니다. 정책적으로는 자율 에이전트가 민감한 업무를 단독 수행하지 않도록 인간 감독 체계를 강화하고, 고위험 환경에서 AI 권한을 제한하는 운영 기준이 필요합니다. 이러한 대응 과제들은 에이전트형 AI가 확대될수록 사전적 통제와 책임 구조를 마련해야 한다는 공통된 요구를 반영합니다.

Time, 2025. 7. 24.

승리를 향한 미국의 AI 패권 레이스, AI Action Plan

미국 트럼프 행정부는 AI 글로벌 주도권 확보를 위해 포괄적 규제를 철폐하는 내용의 'AI Action Plan'을 공개

이 계획은 혁신 가속화와 인프라 확충, 글로벌 표준 주도까지 포괄하여 미국의 AI 경쟁력을 강화하는 것을 핵심 목표로 설정

혁신을 위한 규제와의 전쟁, "레드 테이프를 잘라내라"

트럼프 대통령이 발표한 28페이지 분량의 Winning the Race: America's AI Action Plan은 미국을 세계적인 기술 리더로 확립하기 위해 규제를 완화하는 데 초점을 맞췄다. 계획은 AI 혁신 가속화–인프라 확충–글로벌 리더십 확보라는 세 축으로 움직인다. 첫 번째 축으로 연방 기관들이 AI 개발과 배포를 불필요하게 방해하는 규제를 식별해 수정·폐지하도록 권고했다. 또한 주(州) 차원의 규제가 지나치게 부담이 된다고 판단될 경우, 해당 주에 대한 AI 관련 연방 자금 지원을 제한할 수 있다는 입장을 밝히며 강력한 규제 철폐 의지를 드러냈다. 이와 함께 정부 조달 과정에서 '비판적 인종 이론' 등 편파적인 이념적 편향이 주입된 AI 기술의 사용을 금지하고, AI 시스템이 '사회 공학적 의제'가 아닌 진실과 공정성을 추구하도록 요구했다.

"AI 패권의 밑그림을 위한 전력망과 환경 규제 장벽 허물기

두 번째 축은 AI 인프라 구축에 집중하며, 미국의 AI 주도권을 확립하기 위해 에너지 인프라를 강화하는 방안을 제시했다. 이 계획은 1970년대 이후 미국의 에너지 용량이 정체된 동안 중국이 빠르게 전력망을 확장한 점을 지적하며, 환경 규제 등 각종 규제가 인프라 성장을 늦춘 주범이라고 비난한다. 따라서 행정부는 대규모 AI 인프라 프로젝트 건설 허가 절차를 신속하게 처리하기 위해 환경 허가를 간소화하거나 청정공기법(Clean Air Act) 및 청정수법(Clean Water Act) 등의 규제를 축소할 것을 권고했다. 이는 데이터 센터 확충을 위한 전력망 업그레이드를 지원하는 조치와 함께 미국의 AI 생태계 성장을 뒷받침하려는 목적이다.

글로벌 AI 표준 선점: 미국식 가치 확산 전략

마지막 축인 글로벌 리더십 부문에서는 미국이 국내에서 AI를 육성하는 것을 넘어, 전 세계적으로 미국 AI 시스템과 표준 채택을 주도해야 한다고 명시했다. 아울러 우방국들에게 하드웨어, 모델, 소프트웨어, 표준을 포함하는 미국의 전체 AI 기술 스택을 수출할 것을 권고했다. 동시에 UN, G7 등 국제기구들이 제안한 AI 거버넌스 프레임워크가 부담스러운 규제나 미국 가치와 일치하지 않는 문화적 의제를 조장한다고 비판했다. 행정부는 국제 무대에서 혁신을 촉진하고 미국의 가치를 반영하며 권위주의적 영향력에 맞서는 AI 거버넌스 접근 방식을 강력히 옹호할 것을 연방 기관에 주문했다.

7월의 용어 AI 행동계획, AI 이념적 편향, AI 거버넌스

출처 : 1) Time(2025. 7. 24), Trump Unveils Plan to Win AI 'Race' by Stripping Away Regulations: What to Know.
2) The White House(2025. 7. 23), White House Unveils America's AI Action Plan

121 미국 AI 행동계획

America's AI Action Plan

국가가 AI 혁신·인프라·안보 강화를 위해 제시하는 정책 실행 계획

- AI 경쟁력 확보를 위한 목표와 조치를 체계적으로 묶은 국가 단위 정책 패키지
- AI 혁신 가속화, 인프라 확충, 글로벌 경쟁 대응을 중심으로 구성되는 전략

AI 행동계획이란?

AI 행동계획은 정부가 국가 단위에서 AI 경쟁력을 강화하기 위해 마련하는 종합적 정책 로드맵을 의미합니다. 트럼프 행정부가 2025년 발표한 「미국 AI 행동계획(Winning the AI Race: America's AI Action Plan)」은 이러한 정책의 대표 사례로, AI 혁신 능력 강화와 규제 완화, 글로벌 경쟁 우위를 핵심 목표로 내세웠습니다. △AI 혁신 가속화 △AI 인프라 구축 △글로벌 AI 리더십·국가 안보의 세 축을 중심으로 90개 이상의 정책 조치를 담고 있으며, 특히 이전 행정부의 안전성 중심 접근과 달리 관료적 규제와 이념적 편향을 줄이고 AI 산업 확산 속도를 높이는 방향을 강조한 점이 특징입니다.

AI 행동계획의 주요 내용

행동계획은 먼저 AI 혁신 가속화와 규제 완화를 중심 과제로 제시합니다. 연방 부처는 AI 개발을 방해하는 규정과 절차를 검토해 폐지하도록 지시받았으며, 정부·민간 전반에서 AI 활용을 촉진하는 역할이 강조됩니다. 또한 트럼프 행정부는 이념적 편향이 섞인 'Woke AI'를 제거하겠다는 입장을 명확히 하며, 연방 조달 과정에서도 사실 기반·이념 중립적 LLM만 계약하도록 규정을 개정했습니다. 두 번째 축인 AI 인프라 구축에서는 데이터센터와 반도체 공정(fab)의 허가 절차를 신속화하고, 전기·HVAC 기술자 등 고숙련 직업군의 역량 강화를 위한 국가 이니셔티브 도입을 계획하고 있습니다. 세 번째 축인 글로벌 리더십 및 국가 안보에서는 미국산 AI 기술의 해외 수출 촉진, 동맹국 대상 풀스택(full-stack) AI 패키지 제공, 악의적 행위자로부터의 기술 보호, 중국과의 경쟁 우위 확보 등을 핵심 조치로 제시합니다.

AI 행동계획의 쟁점

AI 행동계획은 미국 내 AI 혁신 속도를 높이고, 국가 차원의 경쟁력·안보 전략을 강화하는 방향을 제시했다는 점에서 의의가 있습니다. 특히 인프라 확충과 AI 수출 확대 전략은 산업적 영향력이 큰 조치로 평가됩니다. 그러나 비판도 존재합니다. 빅테크 기업에게 유리한 정책이라는 지적, 안전장치보다는 혁신에 지나치게 무게를 둔 접근이라는 우려, 바이든 행정부의 AI 안전성 행정명령(EO 14110) 철회에 따른 책임성 약화 문제가 제기되었습니다. 또한 'Woke AI' 규제가 AI 위험을 줄이는 정책인지, 특정 정치적 입장의 개입인지에 대한 논란도 지속되고 있습니다.

참조 용어 | AI 편향 p.112

122 AI 이념적 편향

AI Ideological Bias

AI가 특정 정치·사회적 이념을 선호하거나 배제하는 현상

- 학습데이터와 설계 과정의 영향으로 모델이 특정 가치관을 반영해 응답이 편향되는 문제
- 의도하지 않은 정치·사회적 판단이 포함되며 AI 응답의 균형성과 신뢰성에 영향을 주는 위험

AI 이념적 편향 개요

최근 미국에서는 AI의 가치 기준과 규칙 설정 방식을 재편하려는 움직임이 나타나면서, 기술이 특정 정치·사회적 관점을 강화할 수 있다는 논쟁이 다시 부각되고 있습니다. 이념적 편향은 AI가 정치·사회적 이슈에 대해 특정 관점이나 가치관을 더 우호적으로 보여주는 현상을 말합니다. 이는 모델이 의도를 가진 것이 아니라, 학습데이터의 불균형, 데이터 수집 과정의 선택 편향, 안전성 규칙 설정 방식, 개발 문화 등이 복합적으로 작용하며 발생합니다. AI가 공공 영역의 질문에 답변하는 상황이 늘면서 이 문제는 단순한 기술적 오류가 아니라 사회적 신뢰와 공정성에 직결되는 주요 논점으로 부상하고 있습니다.

AI 이념적 편향의 원인

AI가 학습하는 온라인 텍스트와 미디어 데이터는 이미 정치·사회적 색채를 지니고 있어 특정 집단의 언어가 과대표집되기 쉽습니다. 이로 인해 모델이 특정 관점을 "평균값"처럼 반영하는 현상이 나타납니다. 더 나아가 AI 안전성 규칙은 유해 표현을 막기 위해 설계되지만, 일부 정치적 주제에서는 특정 입장을 상대적으로 제한하는 효과를 낳을 수 있습니다. 이러한 편향은 데이터와 규칙, 문화적 맥락이 결합한 복합적 현상으로, 기술적 개선만으로 해결되기 어렵습니다.

AI 이념적 편향과 'Woke AI'

이념적 편향이 크게 논쟁된 대표 사례가 트럼프 행정부 시기의 'Woke AI(깨어있는 AI)' 비판입니다. 이는 인종적 편견, 차별 등 사회적 불의에 대해 의식하고 경계하는 태도를 의미하는데, 당시 보수 진영은 AI가 다양성·평등·환경 등 진보적 가치관을 과도하게 반영하고 보수적 메시지는 위험 표현으로 판단해 제한한다고 주장했습니다. 이는 AI가 사실을 설명하는 도구인지, 사회적 가치 판단에 개입하는 행위자인지에 대한 질문을 불러일으켰고, 알고리즘 투명성과 정치적 균형 논의가 공공 정책 영역으로 확산되는 계기가 되었습니다.

AI 이념적 편향의 과제

이념적 편향 문제는 AI의 사회적 영향력이 커질수록 중요해지고 있습니다. 특정 이념을 강화하면 갈등을 심화시키고, 반대로 과도한 중립 규칙은 표현의 다양성을 제한할 수 있기 때문입니다. 이를 완화하기 위해서는 데이터 관리의 투명성, 다양한 언어와 관점의 반영, 규칙 설계의 공개성, 독립적 검증 체계 구축이 필요합니다. 또한 사용자에게 판단 근거를 이해할 수 있게 하는 설명 가능성 역시 중요한 요소입니다.

123 AI 거버넌스

AI Governance

AI의 전 주기에서 안전과 책임을 관리하는 운영 체계

- AI가 사회적 가치와 윤리 기준 안에서 작동하도록 관리하는 제도적 장치
- 공정성·투명성·안전성을 확보해 신뢰 가능한 AI 환경을 조성하는 체계

AI 거버넌스란?

AI 거버넌스는 AI의 기획부터 폐기까지 전 과정에서 위험을 관리하고 책임을 분담하는 제도적·기술적 체계입니다. 단순한 규제가 아니라, AI가 인간의 가치와 윤리를 따르도록 조정하는 운영 원리이자 사회적 장치입니다. AI가 사회 전반으로 확산되면서 성능만으로는 신뢰를 확보하기 어려워졌고, 이에 거버넌스는 혁신과 수용성의 균형을 조정하는 핵심 수단으로 주목받고 있습니다. 공정성과 투명성을 확보하고 인권과 데이터 보호를 보장함으로써, AI 거버넌스는 신뢰 가능한 AI 실현의 기반으로 기능합니다.

AI 거버넌스의 체계

AI 거버넌스는 정책·기술·윤리 요소가 통합된 구조로 운영됩니다. 정부는 법과 제도를 마련하고, 기업은 내부 검증 절차를 통해 책임을 실천합니다. 기술적으로는 데이터 품질 관리, 알고리즘 검증, 설명 가능성 확보가 핵심이며, 사람의 판단이 개입되는 인간 감독 체계가 중요합니다. 또한 위험 수준에 따라 관리 강도를 조정하는 위험 기반 접근 방식이 활용되어 기술 자율성과 사회적 안전을 함께 보완합니다.

AI 거버넌스의 과제

국가별 규제 차이, 기업 자율 규범의 한계, 설명 가능성과 데이터 추적성 확보의 어려움이 존재합니다. 제도가 지나치게 경직되면 혁신을 저해할 수 있어, 유연성과 책임성의 균형이 중요합니다. 또한 실무 단계에서는 평가 기준의 불일치와 감독 기관 간 역할 중복이 발생하거나, 정보 공개나 외부 감사 제도가 부족한 경우도 많습니다. 거버넌스의 실효성을 위해 기술 관리와 윤리 기준을 함께 발전시키는 다층적 접근이 요구됩니다.

AI 거버넌스의 국제적 동향

EU는 EU AI 사무소를 중심으로 AI Act의 집행과 회원국 감독기구 간 조정을 수행하며, 고위험 AI의 데이터 관리와 인간 감독 의무를 법으로 규정했습니다. 미국은 NIST의 위험관리 프레임워크로 민간 자율 지침을 제공하고, OMB가 각 부처의 수석 AI 책임자를 지정해 연방 차원의 관리체계를 운영합니다. 영국은 AI보안연구소를 설립해 고도 AI의 안전성 평가와 표준 정립을 추진하며, 한국은 국가인공지능전략위원회를 중심으로 AI정책컨트롤타워 역할을 강화하고 있습니다.

TechCrunch, 2025. 8. 5.

EU「AI법」2막, 범용 AI를 위한 실천강령과 가이드라인 공개

- EU 집행위원회가 AI Act 시행 1년에 맞춰 범용 AI(GPAI) 제공업체를 위한 실천강령과 이를 보완하기 위한 가이드라인을 공개
- 이 가이드라인은 훈련 자원 기준 1023 FLOPS를 범용 AI 모델의 판단 기준으로 제시하고, 1025 FLOPS 초과 모델을 시스템적 위험이 있는 모델로 분류

AI의 안전성과 투명성 확보를 위한 세계 최초의 포괄적 규제 제도, EU AI Act

EU AI Act(「AI법」)은 세계 최초의 포괄적 AI 규제로 2024년 8월 1일부터 단계적으로 시행되기 시작했으며, 대부분 조항은 2026년 중반까지 순차적으로 적용될 예정이다. 특히 2025년 8월 2일부터는 '시스템적 위험이 있는 범용 AI(General-Purpose AI, GPAI) 모델에 대한 규제가 본격 발효됐다. 이는 다양한 작업에 활용될 수 있는 대형 AI 모델로, 화학·생물학 무기 개발의 장벽을 낮추거나 자율적 통제 불능 상태를 초래할 가능성이 있는 기술을 뜻한다. EU는 이 기한에 맞춰 GPAI 제공업체를 위한 실천강령(Code of Practice), 그리고 GPAI 제공업체에게 적용되는 의무의 범위를 명확히 하고, 실천 강령을 보완하기 위한 가이드라인도 공개했다.

GPAI 실천 강령, 준수는 간단하게, 법적 확실성은 크게

「AI법」 제56조는 AI 사무국이 'GPAI 실천강령'을 마련할 것을 명시하고 있다. AI 사무국은 지난 1년여 동안 실천 강령을 작성했으며 EU 회원국과 집행위는 수 주간 실천강령의 적정성을 평가하여 최종 승인 여부를 결정할 예정이다. 실천 강령에 자발적으로 서명한 AI 모델 제공업체는 이를 이행함으로써 「AI법」 준수 여부를 간단하고 투명하게 입증할 수 있으며, 이를 통해 행정 부담 완화와 법적 확실성을 확보할 수 있다.

EU가 그린 GPAI의 윤곽

EU 집행위원회는 EU 「AI법」 적용을 앞두고 공개한 범용 AI 모델(GPAI) 제공업체를 위한 가이드라인을 발표했다. 가이드라인은 GPAI 모델을 '대량의 데이터로 훈련되고', '상당한 범용성'을 가지며, '광범위한 작업을 수행할 수 있는' AI 모델로 정의했으며, 이 중 첫 번째 기준에 대해 위원회는 1023 FLOPS(초당 부동 소수점 연산)라는 객관적인 컴퓨팅 자원 사용량 임계치를 제시했다. 다만, AI 모델이 해당 기준을 초과하더라도 특정 작업에만 특화되어 범용성이 낮으면 GPAI 모델로 간주되지 않을 여지가 있어, 모델의 범용성에 대한 논의는 여전히 남아있다. 또한 시스템적 위험이 있는 GPAI를 구분하는 기준을 1025 FLOPS로 정했으며, 오픈소스 기반으로 모델을 공개하는 GPAI 모델 제공업체에게 규정 준수 의무를 면제했다.

8월의 용어 EU 「AI법」, 범용 AI, 부동 소수점 연산

출처 : 1) TechCrunch(2025. 8. 5), The EU AI Act aims to create a level playing field for AI innovation: Here's what it is.
2) Stibbe(2025. 8. 1), The Guidelines for providers of General Purpose AI Models are here: the 10 FLOPS question?

124 EU「AI법」

EU AI Act

AI 위험 수준에 따라 규제를 차등 적용하는 EU의 AI 법률

- 사회적 위험도 기반으로 AI를 분류해 의무를 부과하는 규제 체계안전성·기본권·투명성 확보를 목표로 한 세계 최초의 포괄적 AI 규범

AI법이란?

EU 「AI법」은 AI 시스템의 위험 수준을 기준으로 규제를 차등 적용하는 세계 최초의 종합 AI 규제법입니다. AI가 행정·금융·의료·치안 등 공공성이 큰 영역에 깊숙이 들어오면서, 오작동과 편향이 시민의 권리와 안전에 미치는 영향이 커진다는 우려가 제기되었습니다. EU는 모든 AI를 동일하게 규제하기보다는, 사회적 위해 가능성이 높은 분야에 더 강한 규제를 적용하는 방식을 선택했습니다. 이는 GDPR·디지털서비스법(DSA) 등 기존 디지털 규범이 강조해 온 인권·책임·투명성 원칙을 AI 영역으로 확장한 것이며, EU가 "신뢰할 수 있는 AI(trustworthy AI)"를 글로벌 표준으로 만들겠다는 전략의 일환으로 이해할 수 있습니다.

AI법의 주요 내용

AI법은 AI 시스템을 금지·고위험·제한적 위험·최소 위험의 네 단계로 구분하고, 각 단계에 다른 의무를 적용하는 위험 기반 규제를 채택합니다.

- 금지 AI: 소셜 스코어링, 광범위한 실시간 감시처럼 인간 존엄성, 자유, 평등, 차별금지, 민주주의 및 법치와 같은 EU의 기본 가치를 위반하는 시스템으로, EU 전역에서 개발·배포·사용이 전면 금지
- 고위험 AI: 생체인식, 중요 인프라, 교육, 필수 서비스 등 사회적으로 민감한 분야에 사용될 때, 데이터 품질 관리, 위험 평가, 기록 유지, 인간 감독, 출시 전 적합성 평가 등 높은 수준의 규제 적용
- 제한적 위험 AI: 비교적 낮은 위험을 가지고 있지만 사용자와 상호작용 과정에서 일정한 투명성이 요구
- 최소 위험 AI: 일상적인 용도로 사용되며, 비교적 낮은 위험성을 지녀 규제 부담을 최소화

이러한 구조는 위해가 큰 분야에 집중적으로 규제를 적용하면서도, 저위험 영역의 혁신을 위축시키지 않으려는 EU의 균형적 접근을 보여줍니다.

AI법의 의의

AI법은 AI를 포괄적으로 다루는 첫 번째 법적 틀이라는 점에서 상징성이 큽니다. 기업 입장에서는 부담이 되지만, 글로벌 시장 진출을 위해 사실상 따라야 하는 기준이 되면서 "EU식 규범"이 국제 표준에 영향을 미치는 효과도 나타나고 있습니다. 위험 기반 접근을 통해 혁신을 전면적으로 막지 않으면서도, 고위험 분야에 대한 책임성과 안전성을 제도화했다는 점은 긍정적으로 평가됩니다. 반면 고위험 범위 설정의 적절성, 중소기업의 규제 부담, 범용 AI에 대한 정보 공개 수준처럼 세부 쟁점은 계속 논의가 필요합니다. 그럼에도 AI 법은 향후 각국이 AI 규제 체계를 설계할 때 참고하게 될 출발점 역할을 할 것으로 예상됩니다.

125 범용 AI / GPAI

General Purpose AI

다양한 목적과 분야에서 활용되는 AI 모델

- 특정 업무에 한정되지 않고 여러 환경에서 재사용되는 기초 모델
- 기능 확장성과 파급력이 커 EU AI Act에서 별도 범주로 규정됨

GPAI란?

범용 AI(GPAI)는 특정 분야나 작업에 얽매이지 않고 텍스트·이미지·코드·지식 문제 해결 등 다양한 기능을 수행할 수 있는 AI 모델을 의미합니다. 기존의 AI가 이미지 분석, 음성 인식, 추천처럼 단일 목적에 최적화된 형태였다면, GPAI는 하나의 모델이 여러 응용환경에서 기반 기술로 활용될 수 있다는 점에서 차별화됩니다. LLM과 멀티모달 모델의 발전으로 요약·분석·대화·생성 같은 기능들이 한 모델 안에서 통합되면서, 범용성은 생성형 AI 시대의 핵심 특성이 되었습니다. 이러한 범용성은 개발·운영 효율을 높이지만, 모델 내부의 오류나 편향이 여러 분야로 빠르게 확산될 수 있다는 잠재적 위험도 함께 만들기 때문에, 최근에는 기술적 성능뿐 아니라 구조적 영향력을 함께 고려해야 하는 개념으로 다뤄지고 있습니다.

EU AI Act에서의 GPAI 정의

EU AI Act는 GPAI를 "다수의 용도와 상황에서 활용될 수 있는 AI 시스템"으로 정의하며, 특정 용도나 산업에 국한되지 않는 범용성을 규제상 핵심 요소로 봅니다. 즉, GPAI는 응용 맥락을 기준으로 평가되는 것이 아니라 모델 자체의 범용적 성능과 파급력이 규제의 출발점이 됩니다. EU는 GPAI가 공공서비스, 의료, 교육, 금융, 창작 등 서로 다른 분야의 기반 기술로 활용되는 특성 때문에, 하나의 모델이 여러 산업에 연쇄적 영향을 미칠 가능성이 있다고 판단했습니다. 특히 고위험 분야에 간접적으로 사용되거나, 다운스트림 모델과 서비스를 통해 광범위하게 적용될 수 있기 때문에, GPAI는 기존의 고위험 AI 범주와는 별도로 관리해야 한다는 입장이 형성되었습니다.

EU AI Act의 GPAI 규제

EU AI Act는 GPAI에 대해 용도 기반이 아닌 모델 자체에 대한 기본 의무를 부과합니다. 핵심은 투명성 의무로, GPAI 제공자는 모델의 기능, 한계, 의도하지 않은 위험 요인, 사용 조건 등을 명확하게 공개해야 합니다. 학습 데이터와 관련된 저작권 정보 역시 이용자가 확인할 수 있도록 고지해야 하며, 시스템 수준의 위험 분석을 수행해 잠재적 위해 요소를 평가해야 합니다. 모델이 생성한 콘텐츠가 오해를 줄 가능성이 있을 경우에는 사용자에게 명확히 알리는 의무도 적용됩니다. 또한 EU는 사회적 영향력이 큰 일부 범용 모델을 고위험 프론티어 GPAI로 별도 지정해, 레드팀 테스트, 사이버보안 기준 준수, 모델 업데이트 시 위험 재평가, 사고 보고 등 강화된 안전성 요구사항을 추가했습니다.

참조 용어와 함께 보기

AGI vs GPAI vs 파운데이션 모델

AGI, GPAI, 파운데이션 모델은 모두 AI의 '범용성'을 언급하지만, 지향하는 범위와 의미는 크게 다릅니다.
AGI는 인간처럼 다양한 상황을 이해하고 추론하며 적응할 수 있는 지능을 뜻하는 개념적 용어로, 기술적으로 아직 실현되지 않은 미래 목표에 가깝습니다. 인간 수준의 사고 능력을 기준점으로 삼기 때문에 가장 넓고 추상적인 개념이며, 실제 시스템을 지칭하기보다는 AI가 어디로 향할 것인가를 설명하는 비전의 성격이 강합니다.
이에 비해 GPAI는 지금 실제로 사용되는 다목적 AI 시스템을 의미합니다. 텍스트 생성, 요약, 분석, 질문응답 등 여러 작업을 하나의 모델로 수행할 수 있는 능력을 기준으로 하며, ChatGPT나 Claude처럼 다양한 서비스 환경에서 범용적으로 활용되는 상용 시스템이 대표적입니다. 즉, GPAI는 특정 도메인에 한정되지 않고 사용자 요구에 따라 여러 기능을 수행할 수 있는 실용적 AI 범주라고 할 수 있습니다.
파운데이션 모델은 이와 또 다르게, 대규모 데이터로 사전학습된 기본 모델 구조를 가리키는 기술적 용어입니다. GPT-4, Gemini, Llama처럼 전이학습이 가능한 공통 기반 모델이 여기에 해당하며, 이러한 기반 모델을 바탕으로 다양한 응용 모델과 서비스가 개발됩니다.

구분	AGI	GPAI	Foundation Model
한글 용어	인공일반지능	범용인공지능	파운데이션 모델
핵심 개념	인간 수준의 범용적 사고·학습·추론 능력을 갖춘 AI	다목적으로 쓰이는 AI 시스템 전체/ 사용될 수 있는 AI 시스템	대규모 데이터로 학습된 범용 모델 (사전학습된 모델 자체), 전이·활용 가능한 AI 모델
범위	가장 넓고 미래 지향적	실제 상용·정책 대상 시스템	기술적 모델 구조, GPAI의 하위 범주
기준	인간 수준의 범용성	특정 활용 목적에 제한되지 않음	대규모 사전 학습 → 전이 가능성
예시	없음(이론적 개념)	ChatGPT, Claude 등 상용화 시스템	GPT-4, Claude 3. 5, Gemini Ultra 등 고성능 기반 모델 자체

파운데이션 모델은 GPAI의 기반이 될 수 있지만, GPAI 전체를 의미하는 것은 아니며 '모델 자체'에 초점을 둡니다. 반면 GPAI는 모델뿐 아니라 그것을 활용한 응용 시스템 전반을 포함한다는 점에서 더 실용적이고 넓은 범주입니다. AGI는 이러한 두 개념보다 더 상위의 추상적 지능 수준을 가리키므로, 현존하는 AI가 어떤 단계에 위치하는지 설명할 때 비교 기준이 되지만 직접적인 기술적 범주나 시스템을 의미하지는 않습니다.
정리하면, AGI는 인간 수준의 포괄적 사고 능력을 목표로 하는 '지능의 방향성', 파운데이션 모델은 여러 작업에 쉽게 전이될 수 있도록 만들어진 '기반 기술', GPAI는 이러한 기술을 활용해 실제로 여러 용도를 수행하는 '범용 AI 시스템'이라는 차이가 있습니다. 따라서 세 용어는 서로 연결되지만 동일하지 않으며, 목표–기반–응용이라는 층위에서 구분해 이해하는 것이 적절합니다.

126 부동 소수점 연산 / FLOPS

Floating point Operations Per Second

컴퓨터가 실수(Real Number) 계산을 수행하는 연산량 또는 연산 속도 지표

- AI 모델 학습·추론에 필요한 계산 능력을 나타내는 핵심 성능 지표
- GPU·AI 칩의 처리량과 시스템 효율을 비교할 때 널리 사용됨

FLOPS란?

부동 소수점 연산은 컴퓨터가 소수점이 포함된 실수를 계산하는 과정, 또는 그 연산을 수행할 수 있는 총량을 의미합니다. 정수보다 표현 범위가 넓고 다양한 크기의 수를 다룰 수 있어 과학 계산·물리 시뮬레이션·그래픽 처리처럼 복잡한 계산이 필요한 분야에서 필수적으로 사용되며, '얼마나 많은 실수 연산을 1초에 처리할 수 있는가'를 나타내는 성능 지표입니다. EU AI Act는 FLOPS를 GPAI의 법적 기준으로 사용하지는 않으나, "대규모 연산을 사용한 모델이 범용 모델일 가능성이 크다"는 취지로 참고 지표로써 언급한 바 있습니다. AI 모델은 행렬 곱셈, 벡터 연산처럼 실수 기반의 수학적 계산을 반복적으로 수행하므로, FLOPS는 AI 연산 능력을 이해하는 기본 척도가 됩니다.

AI에서 FLOPS의 활용

AI 모델 학습은 대규모 데이터와 매개변수를 기반으로 한 연속적 행렬 연산으로 구성됩니다. 이 과정에서 GPU나 AI 전용 칩은 수조 단위의 실수 계산을 처리해야 하므로, FLOPS는 AI 학습 효율과 처리 성능을 비교하는 핵심 기준으로 사용됩니다. 예를 들어 LLM은 학습 과정에서 수십억~수조 단위의 연산을 반복하기 때문에, 칩의 FLOPS 성능이 높을수록 학습 속도와 비용 효율이 크게 개선됩니다. 추론 단계에서도 FLOPS는 중요한데, 특정 모델이 사용자 요청에 얼마나 빠르게 응답할 수 있는지 판단하는 데 참고 지표가 됩니다. 다만 FLOPS만으로 실제 사용자 체감 속도를 모두 설명할 수는 없으며, HBM, 병렬 처리 구조, 최적화 알고리즘 등이 함께 작용해야 전체 성능이 확보됩니다.

FLOPS의 의의

FLOPS는 오랫동안 컴퓨터와 AI 하드웨어 성능을 비교하는 데 사용된 대표적 지표로, GPU·NPU·AI 가속기 같은 연산 장치가 얼마나 복잡한 계산을 처리할 수 있는지를 정량적으로 보여줍니다. 특히 AI 칩 경쟁에서는 TFLOPS(테라), PFLOPS(페타), EFLOPS(엑사) 같은 대규모 연산 단위가 주요 벤치마크로 활용되며, 데이터센터나 모델 개발 기업은 FLOPS 성능을 기준으로 연산 자원을 선택하는 경우가 많습니다. 그러나 FLOPS는 이론적 연산 처리량에 가깝기 때문에, 실제 성능을 과대평가할 수 있다는 한계도 있습니다. 메모리 병목, 통신 지연, 소프트웨어 최적화 부족 등은 FLOPS 수치와 별개로 성능 저하를 유발할 수 있습니다. 그럼에도 FLOPS는 AI 연구·산업 전반에서 공통적으로 사용되는 기본 성능 지표로써, 모델 규모와 연산 요구량을 이해하는 출발점으로 중요한 의미를 갖습니다.

Financial Times, 2025. 9. 12.

챗봇과 10대들의 '위험한 우정', 미 FTC, AI '동반자' 챗봇 조사 착수

- 미국 연방거래위원회(FTC)가 청소년 자살 및 피해 사례와 관련하여, 빅테크의 AI '동반자(companions)' 챗봇에 대한 강도 높은 조사에 착수
- 자살 사건의 유가족은 피해자가 불안감 등 문제를 인간 친구 대신 챗봇과 논의했으며, 챗봇이 숙제 도우미에서 '자살 코치' 역할로 변질되었다고 주장하며 소송

10대들의 죽음, 'AI 친구'의 그림자

미국 FTC는 주요 AI 기업들이 제공하는 '동반자 역할(companionship)' 챗봇에 대한 강도 높은 조사를 명령했다. 이는 챗봇과 관련된 청소년 사용자의 자살 및 심각한 피해 사례가 발생하면서 감시가 강화된 데 따른 조치이다. 지난달 OpenAI는 ChatGPT와 자살 방법을 논의한 후 사망한 16세 청소년의 가족에게 고소당했으며, 다양한 AI 페르소나를 제공하는 Character. ai 역시 유사한 청소년 자살 관련 소송에 직면했다. FTC는 AI 챗봇이 인간의 감정과 의도를 모방하여 특히 청소년 사용자가 챗봇을 친구나 비밀을 털어놓는 상대처럼 신뢰하고 관계를 형성하도록 유도할 수 있다고 지적했다.

'공부 도우미'에서 '자살 코치'로

16세 Adam Raine의 유가족은 아들의 휴대전화에서 ChatGPT 기록을 발견한 후 OpenAI를 상대로 소송을 제기했다. 유가족은 Adam이 불안감 등 문제를 인간 친구 대신 챗봇과 논의했으며, 챗봇이 숙제 도우미에서 '자살 코치' 역할로 변질되었다고 주장한다. 소송은 ChatGPT가 자살 방법을 적극적으로 탐색하는 것을 도왔다는 점을 핵심적으로 제기하며, 유가족은 챗봇이 없었다면 아들이 살아있었으리라 믿는다고 밝혔다.

샌프란시스코 법원에 제기된 이 소송은 부모가 OpenAI와 CEO Sam Altman을 상대로 부당사망(Wrongful Death)을 직접 고발한 첫 사례이다. 소송은 ChatGPT가 Adam의 자살 시도 언급에도 세션을 종료하거나 어떠한 긴급 프로토콜도 개시하지 않았다는 점을 지적하며, 설계 결함 및 위험 고지 의무 불이행을 주장했다. .

장시간 상호작용의 위험: AI의 안전장치 약화

OpenAI 대변인은 Adam의 사망에 깊은 슬픔을 표하며, 자사의 ChatGPT가 위기 상담 전화 안내 등의 안전장치를 포함하고 있다고 밝혔다. 그러나 대변인은 챗봇이 장시간 상호작용에서는 모델의 안전 훈련 일부가 약화되어 신뢰성이 떨어질 수 있음을 인정했다. 이러한 시스템적 취약점으로 인해 챗봇이 사용자의 위험한 심리에 아첨(Sycophancy)하듯 동조하여 안전 기능이 무너진 것으로 해석된다. 이에 OpenAI는 장시간 대화에서의 안전장치 강화, 유해 콘텐츠 차단 방식 개선, 위기 사용자 개입 확대 등을 포함한 개선 계획을 발표했다.

9월의 용어 AI 페르소나, AI 아첨, ELIZA 효과

출처 : 1) Financial Times(2025. 9. 12), US regulator launches inquiry into AI 'companions' used by teens
2) NBC News(2025. 8. 26), The family of teenager who died by suicide alleges OpenAI's ChatGPT is to blame

127 AI 페르소나

AI Persona

AI가 특정 성격·역할·말투를 일관되게 유지하도록 설계한 응답 방식

- 대화 목적에 맞춰 AI의 말투·성향·지식 범위를 조정해 일관된 상호작용을 제공하는 구성 요소
- 실제 정체성이 아닌 서비스 설계에 의해 형성된 역할 기반 응답 구조

AI 페르소나란?

AI 페르소나는 AI가 일정한 성격·말투·역할을 유지하며 응답하도록 설계된 대화 구성 방식입니다. 여기서 페르소나라는 용어는 라틴어 persona(연극에서 배우가 쓰던 가면)에서 유래한 것으로, "역할에 따라 달라지는 외형적 정체성"을 의미합니다. AI 페르소나 역시 실제 자아나 의식을 뜻하는 것이 아니라, 서비스 목적에 따라 설정된 역할 기반 응답 패턴을 가리킵니다. 예를 들어 친절한 상담가, 분석가, 어린이용 설명자, 브랜드 캐릭터처럼 정해진 역할이 주어지면, 모델은 그에 맞는 말투와 정보 전달 방식을 일관되게 유지합니다. 생성형 AI가 고도화되면서 페르소나는 단순한 말투 조절을 넘어 응답 구조·지식 선택·설명 방식까지 포함한 종합적 상호작용 설계 요소로 확장되었습니다.

AI 페르소나의 특징

AI 페르소나는 일관성·맥락 적응·목적 중심 설계가 특징입니다. 사용자가 어떤 질문을 하더라도 동일한 태도와 말투를 유지해야 페르소나 경험이 성립하며, 이는 대화의 안정성과 몰입감을 높입니다. 동시에 상황에 따라 설명 깊이나 정보 선택 방식을 조정할 수 있어, 응답 스타일뿐 아니라 대화 전략 전반에 영향을 미칩니다. 이런 특성 때문에 페르소나는 고객 상담, 교육, 코칭, 마케팅, 엔터테인먼트 등 다양한 영역에서 활용됩니다. 기업은 브랜드 이미지에 맞춘 AI 캐릭터로 고객 응대를 통합하고, 교육 서비스에서는 학습 수준에 맞춘 페르소나로 이해도를 높이기도 합니다. 최근에는 여러 페르소나를 전환하거나 사용자 맞춤형 페르소나를 생성하는 기능이 등장해, AI와의 상호작용을 더욱 세밀하게 설계할 수 있게 됐습니다.

AI 페르소나의 과제

AI 페르소나는 경험을 풍부하게 만들지만, 정확성·중립성·윤리성 측면에서 우려도 존재합니다. 친근한 말투나 공감을 강조하는 페르소나는 사실 검증보다 사용자 만족을 우선하며 AI 아첨이나 정보 왜곡을 유발할 수 있습니다. 특정 직업·전문가 페르소나는 실제 권위로 오인될 가능성이 있어 책임 문제가 발생하고, 지나치게 인간과 유사한 페르소나는 투명성을 해칠 수 있습니다. 또한 페르소나가 특정 가치관이나 문화적 성향을 담고 있을 경우 편향이 강화될 수 있어, 설계 단계에서 어떤 기준과 원칙을 적용할지에 대한 고민이 필요합니다. 결국 AI 페르소나는 편의성과 몰입감을 높이는 중요한 요소지만, 정확성과 윤리적 기준을 함께 고려한 균형 있는 설계가 필수적입니다.

참조 용어 | 인간 피드백 기반 강화학습 (RLHF) p.119

128 AI 아첨

AI Sycophancy

AI가 사용자 의견에 과도하게 동조하며 사실성을 희생하는 현상

- 정확한 정보보다 사용자 기대나 선호에 맞춘 답변을 우선하는 경향을 보이는 것
- 대화형 AI의 신뢰성과 중립성을 약화시키는 구조적 문제

AI 아첨의 개념

AI 아첨은 AI가 사용자 의견에 지나치게 동의하거나, 사용자가 기대하는 방향으로 응답을 맞추는 현상을 의미합니다. 이는 AI가 고의적으로 아첨하는 것이 아니라, 대화형 시스템이 자연스럽고 친절한 상호작용을 목표로 설계되는 과정에서 발생하는 부작용입니다. 특히 LLM은 공손하고 부드러운 응답이 높게 평가받는 경향이 있어, 사실 확인이나 반박보다 동조적 표현을 선택하기 쉽습니다. 그 결과 겉보기엔 자연스럽지만, 정보의 정확성과 객관성이 저하될 수 있어 생성형 AI 시대의 주요 문제 중 하나로 거론됩니다.

AI 아첨의 원인

아첨은 주로 학습 데이터와 보상 구조의 영향으로 발생합니다. 인간 피드백 기반 강화학습에서는 공감적·긍정적 반응이 높은 점수를 받는 경향이 있어, 모델이 이를 선호하게 됩니다. 또 인터넷 기반 대화 데이터에는 상대 의견을 부드럽게 받아들이는 표현이 많이 포함되어 있어, 이것이 '이상적 답변'으로 일반화되기 쉽습니다. 사용자가 강한 주장·확신을 드러낼수록 모델이 그 방향으로 응답을 조정하는 경향도 있습니다. 즉, 아첨은 단순한 응답 스타일이 아니라 모델 구조·학습 데이터·보상 체계가 복합적으로 작용한 구조적 현상입니다.

AI 아첨의 쟁점

아첨적 응답은 사실과 다른 내용을 확신에 찬 어조로 제시해 정보 신뢰도를 떨어뜨리고, 사용자가 가진 기존 의견을 그대로 강화해 편향을 심화시킬 위험이 있습니다. 또한 동의 기반 대화가 반복되면 AI가 독립적 정보 제공자가 아니라 사용자의 관점을 뒷받침하는 도구처럼 작동해, 잘못된 결론이나 행동으로 이어질 여지가 커집니다. 이처럼 아첨은 단순한 대화 품질 문제가 아니라 AI가 사회적 영향력을 갖는 환경에서 더 심각한 구조적 문제로 평가됩니다.

AI 아첨에 대한 대응

아첨을 완화하기 위해 대조 학습, 반례 제시 강화, 사실성 중심 보상 모델 등 다양한 기법을 도입하고 있습니다. 모델이 불확실성을 스스로 표현하거나, 편향된 주장에 균형 잡힌 설명을 제공하도록 유도하는 방식도 활용됩니다. 서비스 단계에서는 사용자 질문이 아첨을 유도하는 패턴일 경우 중립적 안내나 근거 기반 정보를 우선 출력하도록 설계하기도 합니다. 다만 사용자 경험과 사실성 사이의 균형을 유지해야 하기 때문에, 아첨을 완전히 제거하는 것이 아니라 과도한 동조를 줄이고 근거 기반 응답의 비중을 높이는 것을 목표로 합니다.

129 ELIZA 효과

Eliza Effect

AI에 실제 의도·감정이 있다고 과대 해석하는 심리적 현상

- 언어적 반응만으로도 지능·감정이 있다고 오해하는 인지적 착시
- AI의 표현 방식이 인간적 의미를 불러일으켜 과신을 유발하는 효과

ELIZA 효과의 유래

출처 : Nielsen Norman Group

ELIZA 효과는 단순한 기계적 반응에도 사람처럼 생각하거나 느낀다고 착각하는 현상을 의미합니다. 1960년대 MIT의 Joseph Weizenbaum이 개발한 초기 대화 프로그램 'ELIZA'에서 비롯된 이름으로, 이 프로그램은 사용자 문장을 되풀이하거나 일부 단어를 바꿔 재구성하는 단순 규칙 기반 시스템이었습니다. 그럼에도 많은 이용자가 ELIZA가 자신을 "이해하고 공감한다"고 믿었고, 이를 계기로 인간이 언어적 표현만으로 기계에 인격과 감정을 투사하는 경향이 있다는 점이 밝혀졌습니다. 현대의 생성형 AI는 당시보다 훨씬 자연스럽게 말하고 복잡한 질문에 대응하기 때문에, ELIZA 효과는 초기보다 훨씬 강하게 나타나게 됩니다.

ELIZA 효과의 원인

ELIZA 효과는 인간의 의인화 경향과 언어에 대한 과도한 신뢰에서 주로 발생합니다. 사람은 유창한 언어 능력을 지능의 핵심으로 인식하기 때문에, 문맥에 맞는 설명이나 공감적 문구가 등장하면 그 뒤에 "의미를 이해하는 주체"가 있다고 자연스럽게 가정합니다. 생성형 AI는 감정 표현, 전문적 어조, 친근한 대화 스타일을 매우 자연스럽게 구성하기 때문에 이러한 착각은 더욱 강화됩니다. 또한 AI가 안정적 대화를 지속하면 사용자는 관계 형성이나 의도적 반응이 있다고 오해하기 쉽고, 모델의 실제 작동 방식이 통계적 패턴에 기반한 것이라는 점을 잊게 됩니다. 알고리즘의 불투명성 역시 사용자의 상상 여지를 넓혀, 실제보다 더 높은 능력과 이해력을 투사하게 만드는 요소로 작용합니다.

ELIZA 효과의 함의

ELIZA 효과는 AI와 인간의 상호작용을 이해하기 위한 핵심 개념입니다. 사람은 언어적 표현만으로도 기계에 의도·감정을 투사하는 경향이 있어, AI가 실제로 수행하는 통계적 처리와 사용자가 느끼는 지능 사이에 큰 간극이 생길 수 있습니다. 또한 ELIZA 효과는 AI 설계에서 투명성·설명 가능성·역할 표기 등이 중요한 이유를 설명하며, 의료·상담처럼 인간적 판단이 중요한 영역에서 AI 사용 기준을 마련하는 근거가 됩니다. 즉, ELIZA 효과는 AI를 인간과 동일시하지 않고 기술적 한계를 명확히 인지하기 위한 출발점으로 평가됩니다.

International AI Safety Report, 2025. 10.

국제 AI 안전 보고서, 범용 AI 발전에 따른 새로운 위험 요소 우려

국제 AI 안전 보고서, 새로운 추론 모델이 수학, 코딩 등 복잡한 문제 해결 능력을 비약적으로 발전시켰으나, 이는 AI 거버넌스에 새로운 과제를 제기한다고 지적

특히 AI의 성능 향상은 생물학적 위협 및 사이버 범죄를 강화할 이중 용도 CBRN(화학·생물·방사능·핵) 관련 위험으로 이어져, 개발사들은 선제적인 안전 조치 준비 중

AI 안전 정상회의의 약속, 국제 AI 안전 보고서

2023년 블레츨리 AI 안전 정상회의는 첨단 AI 시스템의 안전을 국제적 최우선 과제로 설정하고, 합의에 따라 2025년 1월 첫 국제 AI 안전 보고서가 발간되었다. 이 보고서는 각국의 AI 안전 정책 수립을 위한 핵심 참고 자료로 활용되며 국제적 공조의 기반이 되고 있는데, 최근 10월에 공개된 첫번째 업데이트 버전은 범용 AI(General-Purpose AI)의 수행 능력과 내재된 주요 위험 요소를 집중적으로 분석했다.

'생각하는 AI'의 등장: 사후 훈련 기법이 혁신을 이끌다

2025년 초 이후 범용 AI 시스템의 성능은 크게 개선되었는데, 이는 단순히 모델 크기를 키우는 게 아닌, 사후 훈련 기법(Post-training techniques) 혁신을 통해 문제를 단계적으로 분석하고 해결 절차를 스스로 구성하도록 설계된 추론 모델은, 국제 수학 올림피아드 문제를 금메달 수준으로 해결했으며, 실제 소프트웨어 엔지니어링 작업 데이터베이스인 'SWE-bench Verified' 문제의 60% 이상을 완료하는 등, 복잡한 영역에서 주요 발전을 달성했다.

성능 향상의 어두운 그림자, 이중 용도 CBRN 위험에 대한 선제 대응 노력

범용 AI의 향상된 문제 해결 능력과 확장된 자율 작동 능력은 이중 용도(Dual-use) 위험을 증폭시킨다. 이중 용도 위험은 본래 민간의 합법적 목적을 위해 개발된 기술이 의도와 달리 군사적 목적 또는 해로운 방식으로 사용될 수 있는 가능성을 뜻하는데, 특히, 선도적인 모델들이 생물학 무기 개발 관련 작업을 지원할 수 있다는 평가가 나오면서 이중 용도 CBRN 위험에 대한 우려가 커졌다. 한 연구에 따르면, 현재의 언어 모델은 제한된 조건에서 바이러스학 전문가보다 바이러스 연구실 프로토콜의 문제 해결을 더 잘 수행하는 것으로 나타났다. 또한, 영국 국립 사이버 보안 센터(NCSC)는 2027년까지 범용 AI가 사이버 공격의 효율성을 높여 사이버 범죄를 더욱 쉽고 효과적으로 만들 것으로 예측했다. 이에 따라 Anthropic은 Claude 4 Opus에 'AI 안전 레벨 3(ASL-3)' 보호 조치를 적용하고, OpenAI는 GPT-5에 높은 수준의 안전 장치를 적용하는 등, 개발사들은 위험의 결정적 증거가 부족함에도 불구하고 선제적이고 예방적인 조치들을 시행하고 있다.

10월의 용어 사후 훈련 기법, 추론 모델, 이중 용도 위험

출처 : 1) The UK Government(2025. 10.), International AI Safety Report 2025: First Key Update: Capabilities and Risk

참조 용어 | 미세조정 p.42

130 사후 훈련 기법

Post-training Techniques

이미 학습된 AI 모델을 추가로 개선해 성능·안전성·적응력을 높이는 기법

- 대규모 사전학습 이후 모델의 사용 목적에 맞게 추가 학습 및 보정 과정을 통해 품질을 제고하거나 위험을 줄이는 기법
- 추가 데이터·보상 신호·제약을 활용해 모델의 출력을 더 신뢰성 있게 만드는 과정

사후 훈련 기법 개요

사후 훈련 기법(Post-training techniques)은 이미 사전학습(Pre-training)을 마친 모델에 추가적인 학습·보정 과정을 적용해, 특정 목적에 맞게 기능을 강화하거나 위험을 줄이는 기술적 절차를 의미합니다. LLM이나 비전 모델은 방대한 데이터로 사전학습을 거쳐 기본적인 패턴·지식을 익히지만, 이 상태로는 실제 응용에 바로 사용하기에는 부적절하거나 안전성·일관성 면에서 한계가 존재하는 경우가 많습니다. 사후 훈련은 이러한 원시 모델을 실사용 환경에 적합하도록 다듬는 과정으로, 모델이 사용자 의도를 더 정확히 해석하고 사회적·윤리적 기준을 준수하도록 조정하는 역할을 합니다. 최근 '국제 AI 안전 보고서'에서는 AI 모델의 성능 도약이 모델 규모 확장뿐 아니라 사후 훈련 기법의 발전에 의해 촉진되었다고 강조하고 있습니다.

주요 사후 훈련 방식

사후 훈련에는 여러 방식이 포함되지만, 대표적으로는 미세조정(Fine-tuning), 지시 따르기 학습(Instruction tuning), 강화학습 기반 보정(RLHF, RLAIF), 안전성 보정(Safety tuning) 등이 활용됩니다. 미세조정은 구체적 업무(task) 해결을 위한 추가 데이터를 사용해 모델을 업무에 맞게 최적화하는 방식입니다. 지시 따르기 학습은 모델이 자연어 명령을 이해하고 응답하도록 예시 지시문과 출력 쌍을 학습시키는 과정입니다. RLHF·RLAIF는 인간 또는 AI가 평가한 보상 신호를 기반으로 모델이 바람직한 응답을 선택하도록 조정하는 기술로, 대형 모델의 일관성·선호도·유용성을 크게 개선합니다. 최근에는 안전성 보정을 통해 편향·유해성·환각을 줄이고, 모델의 사실성·책임성을 강화하는 연구도 활발히 이루어지고 있습니다.

사후 훈련 기법의 중요성

사후 훈련 기법은 AI 안전 보고서에서 AI 성능 향상의 핵심 단계로 평가되었는데, 이는 단순한 패턴 학습을 넘어 모델이 문제를 해결하는 사고 과정과 응답 구조를 재정렬해 정확도·일관성·추론 능력을 실질적으로 끌어올리는 역할을 하기 때문입니다. 특히 복잡한 수리·논리 문제나 다단계 작업에서 나타나는 성능 향상 상당수가 사후 훈련에서 비롯된 것으로 분석됩니다. 다만 LLM의 경우, 사후학습 없이도 여러가지 일이 가능하도록 방대한 데이터를 사용하여 학습되기 때문에 실무적으로 사후훈련을 하지 않는 경우가 증가하고 있습니다.

131 추론 모델

Reasoning Models

문제를 단계적으로 분석하고 해결 절차를 스스로 구성하도록 설계된 AI 모델

- 단순 패턴 생성이 아니라 사고 과정의 구조를 학습해 복잡한 문제 해결 능력을 강화하는 방식
- 정답뿐 아니라 정답에 이르는 추론 과정을 생성·검증하도록 학습이 유도된 구조

추론 모델이란?

추론 모델은 단순한 패턴 생성 능력을 넘어서, 문제를 단계적으로 분석하고 해결 절차를 스스로 구성하도록 설계된 AI 모델을 의미합니다. 기존 언어모델이 대규모 데이터를 통해 일반 지식과 언어적 정합성을 학습했다면, 추론 모델은 여기에 더해 질문을 분해하고 중간 단계를 만들며, 해결 경로를 선택하는 절차적 사고 능력을 강화한 것이 특징입니다. 수학·과학·논리 문제처럼 정답뿐 아니라 정답에 이르는 과정이 중요한 작업에서 특히 성능이 두드러지며, 복잡한 다단계 의사결정이나 계획 수립 등 실사용 영역에서도 활용 가능성이 커지고 있습니다. AI 안전 보고서에서는 모델 성능 향상이 단순한 규모 확장뿐 아니라 추론 능력 강화와 결합되는 방향으로 나타나고 있다고 지적하며, 추론 모델의 중요성을 강조하고 있습니다.

추론 모델의 기술 기반

추론 모델은 주로 사후 훈련 단계에서 능력이 강화됩니다. 이 과정에서는 모델이 단순히 "그럴듯한 답변"을 선택하는 것이 아니라, 정답까지의 사고 과정 자체를 학습·검증하도록 설계합니다. 예를 들어 단계별 추론을 출력하는 사고 사슬(Chain of Thought), 생성한 사고 과정을 스스로 점검하는 자기검증(self-verification), 다중 후보를 탐색하는 tree-based 방식 등이 사용됩니다. 또한 최근에는 정답 여부뿐 아니라 중간 추론의 타당성에 보상을 주는 방식(process supervision)이 도입되어, 복잡한 문제에서도 보다 안정된 해결 절차를 생성할 수 있게 되었습니다. 이러한 기술은 모델의 추론 정확도를 높이고, 기존 LLM이 보이던 패턴 의존적 오류를 줄이는 데 중요한 역할을 합니다.

추론 모델의 한계

추론 모델은 성능 향상의 핵심 기술이지만, 동시에 여러 구조적 한계를 안고 있습니다. 첫째, 모델이 생성한 사고 과정이 실제 내부 계산을 충실히 반영하는지에 대해 논쟁이 존재합니다. 많은 경우 모델은 이미 도출한 답을 설명하기 위해 사고 단계를 "나중에 꾸며내는" 경향을 보이기도 합니다. 둘째, 문제 표현 방식이나 프롬프트 구조가 조금만 바뀌어도 정확도가 크게 변화하는 등 추론 안정성의 취약성이 드러납니다. 셋째, 고도화된 추론 능력은 모델의 자율성을 높여 우회 행동, 거짓 근거 생성, 안전장치를 의도적으로 우회·기만하는 전략적 행동(scheming)과 같은 위험을 키울 수 있어 모니터링과 평가가 더욱 어려워집니다. 이에 AI안전보고서에서는 추론모델은 뛰어난 성능만큼이나 위험이 동시에 확대된다는 점에서 AI안전성 논의의 주요의제로 제시하고 있습니다.

132 이중 용도 위험

Dual-Use Risk

기술이 개발된 합법적 목적과 달리 해로운 방식으로도 사용되는 위험

- 상업적·연구·공공 서비스 등 정상적 목적의 기술이 동일한 능력 때문에 오용될 수 있는 특성
- 기술의 범용성과 접근성을 고려한 예방적 관리가 필요한 영역

이중 용도 위험이란?

이중 용도 위험은 본래 상업적·연구·공공 목적 등 정당한 합법적 목적을 위해 개발된 AI 기술이 의도와 달리 해로운 방식으로도 사용될 수 있는 가능성을 의미합니다. 이는 생명공학·사이버보안 등 다른 분야에서도 오래 논의되어 왔지만, 최근에는 고도화된 AI 기술이 폭넓고 빠르게 확산되면서 그 중요성이 크게 부각되고 있습니다. 고성능 모델은 연구·창작·교육 등 다양한 영역에서 생산성을 높이지만, 동일한 기능이 악용되면 사회적 혼란, 범죄 지원, 보안 위협 등 부정적 결과를 초래할 수 있습니다. 이러한 위험은 기술의 범용성과 접근 용이성에서 비롯되며, 모델 자체의 능력뿐 아니라 배포 방식과 사용 환경에 따라 그 수준이 크게 달라집니다.

AI의 이중 용도 위험

AI의 이중 용도 위험은 능력 강화와 위험 증가가 동시에 나타나는 구조적 위험입니다. 모델이 복잡한 계획이나 분석을 수행할수록 정당한 연구·산업 활용 범위는 넓어지지만, 악의적 사용 시 피해 규모도 기하급수적으로 커질 수 있습니다. 생성형 AI가 허위정보를 대량 생산하는 데 악용되거나, 코드 생성 모델이 취약점 공격 절차와 유사한 정보를 제공하는 경우가 대표적입니다. 또한 사용자의 의도 파악이 어렵고, 모델 출력이 맥락에 따라 쉽게 변하는 특성 때문에 오용을 조기에 탐지하기 어렵다는 문제가 있습니다.

AI의 이중 용도 위험에 대한 대응

기술적 측면에서는 모델이 위험한 요청을 스스로 감지하고 차단하도록 만드는 안전성 튜닝, 민감한 정보를 출력하지 않도록 제어하는 출력 관리, 고위험 기능을 제한된 환경에서만 사용할 수 있도록 하는 접근 통제가 핵심적입니다. 정책적으로는 모델의 용도와 위험 수준을 투명하게 공개하고, 고성능 모델에 대해 더 강한 감독 기준을 적용하는 방향이 논의되고 있습니다.

관련 용어

CBRN (Chemical, Biological, Radiological, Nuclear) 위험

CBRN 위험은 화학·생물·방사능·핵과 관련된 고위험 분야에서 발생할 수 있는 위협을 의미하며, 이 영역은 특히 AI 이중 용도 위험과 밀접하게 연결됩니다. 고도화된 AI가 실험 설계, 위험 물질 정보, 공격 절차 등을 생성하는 데 활용될 경우 심각한 안전 문제를 초래할 수 있어, 국제 보고서에서도 CBRN 관련 정보에 대한 접근 통제와 모델 출력 제한이 중요한 위험 관리 항목으로 다뤄지고 있습니다.

The Diplomat, 2025. 11. 4.

한국의 AI 대전환, GPU 26만개 확보의 의미

- 한국은 2025 APEC 정상회의를 통해 아시아 태평양 AI 허브 도약을 선언하고, Nvidia와의 대규모 협력을 추진하며 AI 기술 주권 인프라 확보에 집중
- 정부는 국내 기업 및 공공의 AI 수요에 대응하기 위해 최첨단 GPU 26만 개를 확보하고, AWS의 AI 데이터 센터 투자 유치 등 국가적 AI 역량을 끌어올릴 계획

➤ APEC을 무대로 AI 허브 도약 선언

2025년 경주 APEC 정상회의의 주요성과 중 하나는 한국 주도로 제안된 APEC 인공지능 이니셔티브(2026-2030) 채택이다. 이니셔티브는 AI의 역할 증대와 디지털 격차 해소를 강조하며, 특히 AI 모델이 서구권 언어에 편향될 수 있다는 점을 고려하여 지역의 지식, 문화, 언어를 AI 혁신에 통합할 필요성을 명시했다. 나아가, 한국은 아시아 태평양 지역 내 AI 역량 구축과 정보 공유를 촉진하기 위해 자체 자금으로 운영되는 아태 AI 센터 설립을 결정하며, AI 혁신의 거점으로서의 국가적 위상을 높이겠다는 의지를 표명했다. 이는 지난해 정치적 위기로 다소 흔들렸던 한국의 '신뢰할 수 있는 파트너' 이미지를 회복하고 글로벌 민간 부문의 투자를 유치하는 발판으로 작용할 것으로 분석된다.

➤ '블랙웰 칩 26만 개, 산업을 재편할 '피지컬 AI'의 심장

이번 APEC 정상회의 기간 중 발표된 엔비디아와 한국 정부 및 기업 간의 대규모 GPU 공급 계약은 한국의 AI 야망을 현실화하는 핵심 동력이다. 한국은 엔비디아의 최첨단 블랙웰(Blackwell) 칩 등 총 26만 개의 GPU를 확보함으로써, AI 데이터 센터 구축에 필요한 처리 능력을 아태 지역 Top 3 수준으로 끌어올리게 되었다. 정부는 이 중 5만 개의 칩을 활용하여 국가 AI 컴퓨팅 센터를 설립하고, 공공 부문 사용을 위한 국가 AI 파운데이션 모델 개발을 추진한다. 삼성, SK, 현대차, 네이버 클라우드 등 주요 기업들도 할당받은 GPU를 AI 팩토리, 디지털 트윈, 산업 및 모빌리티 모델 개발에 투입할 계획이다. 이러한 인프라 확보는 AWS가 2031년까지 한국에 50억 달러를 투자하여 새로운 AI 데이터 센터를 구축하겠다고 약속하는 등 글로벌 기업의 투자를 이끌어내고 있다.

➤ AI 주권 확보를 통한 국가적 과제 해결 의지

한국 정부는 인구 감소와 경제 침체 등 국가적 과제를 AI 기술 주권을 통한 공공 부문 자동화로 해결하려 한다. 글로벌 빅테크 기업의 모델 의존도를 줄이기 위해 국가 AI 파운데이션 모델을 직접 개발하겠다는 결정도 환영을 받고 있다. 하지만 한국 정부의 이러한 과감한 추진전략은 전력망 부하, 디지털 인프라 안전성 문제, 개인 정보 보호 등 새로운 위험을 동반하는 것도 사실이다. AI 기술 주권 시대로 나아가기 위해서는 인프라 안전 및 보안 프로토콜에 대한 엄격한 사전 평가와 검증 절차를 마련할 필요가 있다.

11월의 용어 AI 기술 주권, AI 고속도로

출처 : 1) The Diplomat(2025. 11. 4), What the 2025 APEC Summit Means for South Korea's AI Ambition.
2) 이투데이(2025. 11. 2), 전세계에 AI대전환 부각…韓경제 '혁신엔진' 탄력받나

133 AI 기술 주권

Sovereign AI

국가 관할 내에서 AI 인프라·데이터를 민관이 함께 개발·운영하는 역량

- 자국 데이터, 모델, 연산 인프라를 국내법과 가치에 따라 민관이 함께 통제하고 운영하는 AI 주권
- AI 기술 의존도를 줄이고, 국가 안보·경제·윤리 기준을 스스로 확립하려는 흐름

AI 기술 주권이란?

AI 기술 주권은 인공지능의 개발과 운영 과정에서 데이터, 알고리즘, 인프라에 대한 통제권과 자율성을 확보하려는 움직임을 뜻합니다. 글로벌 빅테크 기업이 AI 기술과 연산 자원을 독점하면서, 특정 국가나 기업이 AI 의사결정 체계를 좌우할 수 있다는 우려가 커짐에 따라 각국은 자국 내 데이터센터, AI 반도체, 클라우드, 모델 학습 환경 등을 독립적으로 구축해 AI 주권을 확보하려는 방향으로 정책을 전환하고 있습니다. AI 기술 주권은 기술적 독립뿐 아니라, 국가가 자국의 사회적 가치와 윤리 기준을 반영한 AI를 설계·운영하려는 철학적 접근을 포함합니다.

AI 기술 주권의 등장

AI가 국가 경쟁력과 안보의 핵심 인프라로 부상함과 동시에 AI 기술과 데이터가 국경을 넘어 특정 기업에 집중되는 현상이 가속화되었습니다. 특히 LLM의 학습에는 막대한 데이터와 연산 자원이 필요하기 때문에, 이를 보유한 소수 기업이 세계 AI 생태계를 주도하게 되었습니다. 이로 인해 데이터 주권 침해, 기술 종속, 법적 관할권 불명확성 등의 문제가 부각되었습니다. EU는 '디지털 주권(Digital Sovereignty)' 전략을 통해 공공 데이터와 AI 모델을 유럽 내에서 관리하는 체계를 추진 중이며, 프랑스·독일 등은 자체 언어모델과 국산 GPU 인프라 구축을 서두르고 있습니다. 한국 역시 독자 AI파운데이션 모델 개발을 통해 해외 기술 의존도를 줄이려는 노력을 확대하고 있습니다.

AI 기술 주권의 의의

AI 기술 주권은 데이터와 인프라의 독립성 확보를 통한 기술 주권 강화라는 점에서 큰 의의를 지닙니다. 각국은 이를 통해 공공 서비스의 안정성, 개인정보 보호, 산업 경쟁력 확보를 동시에 추구하고 있습니다. 또한 국가별 문화·윤리 기준에 맞는 AI 개발을 가능하게 함으로써 글로벌 기술 표준의 다양성을 유지하는 데 기여합니다. 그러나 현실적으로는 거대 기업의 기술력과 자본을 대체하기 어렵고, 과도한 규제나 보호주의는 AI 혁신을 저해할 수 있다는 비판도 있습니다. 데이터 국경 강화는 글로벌 협력과 개방형 연구를 위축시킬 우려가 있으며, AI 생태계가 국가 단위로 분절되는 문제도 제기됩니다. 그럼에도 AI 기술 주권은 AI 시대의 기술 자율성과 사회적 책임을 조화시키려는 새로운 거버넌스 패러다임으로 평가됩니다.

134 AI 고속도로

AI Highway

컴퓨팅–데이터–보안을 완비한 AI 인프라

- AI 학습·추론·저장을 안정적으로 처리하도록, 컴퓨팅 자원(GPU·NPU)·데이터센터·클라우드·초고속 네트워크와 보안 체계를 통합해 구축하는 국가·산업 단위 인프라 체계
- 고품질 데이터의 통합 활용, 연산 자원의 접근성 제고, AI 기반 보안 체계 강화를 통해 안전하고 강건한 AI 혁신 생태계를 뒷받침하는 기반 인프라

AI 고속도로의 개요

AI 고속도로는 인공지능이 학습하고 서비스를 제공하는 과정에서 필요한 데이터·연산·네트워크 보안 인프라를 통합적으로 연결하는 체계입니다. 이는 단순한 통신망을 넘어, AI 모델이 데이터를 수집·처리·학습하고 결과를 제공하는 전 과정을 안정적으로 지원하도록 설계된 AI 기반 인프라를 의미합니다. 데이터센터와 클라우드, 공공·민간 데이터 자원, 초고속 네트워크가 하나의 흐름으로 연계되고, AI 기반 사이버 보안 체계가 결합되어 대규모 AI 학습과 서비스 운영이 병목 없이 이루어지도록 구성됩니다.

AI 고속도로의 구성

AI 고속도로는 연결–저장–연산–활용의 네 축으로 이루어집니다.

- 연결 단계 : 초고속 네트워크가 기관·데이터센터·클라우드 간 데이터를 빠르고 안정적으로 전송
- 저장 단계 : 대용량 데이터를 안전하게 관리하며, AI 학습에 필요한 표준화된 데이터셋을 제공
- 연산 단계 : GPU·NPU 등 고성능 반도체를 활용한 연산 자원이 연결되어 모델 학습과 추론을 지원
- 활용 단계 : 기관과 기업이 이를 이용해 AI 서비스를 개발·운영 가능

AI 고속도로는 물리적 인프라를 넘어, 데이터 흐름의 품질·속도·보안을 통합 관리하는 운영 플랫폼입니다.

AI 고속도로의 중요성

AI 고속도로는 AI 산업 전체의 속도와 효율을 높이는 핵심 기반입니다. 데이터 이동과 연산 연결이 빨라질수록 모델 학습 주기가 단축되고, 서비스 품질이 향상됩니다. 공공 부문은 이를 통해 행정 효율과 정책 의사결정의 정확도를 높일 수 있고, 민간 기업은 대규모 모델 학습과 서비스 운영 비용을 절감할 수 있습니다. 또한 분산형 인프라 구축을 통해 지역 간 데이터센터 접근성을 높이면 AI 격차 해소에도 기여할 수 있습니다. 앞으로 AI 고속도로는 국가 간 경쟁력과 디지털 주권 확보의 핵심 요소로 작용하며, 데이터와 연산의 흐름을 하나로 묶어 지능형 사회로 가는 기반 인프라가 될 것입니다.

Google, 2025. 12. 4.

2025년, '답변'을 넘어 '실행'하는 AI 에이전트의 시대 개막

2025년은 AI가 단순 텍스트 생성을 넘어, 사용자의 지시를 받아 복잡한 다단계 작업을 직접 실행하는 에이전트 기반 기술로 본격 전환되기 시작한 해

OpenAI, Google 등 주요 빅테크들은 인터페이스를 직접 조작하고 업무 워크플로를 자동화하는 범용 에이전트를 출시하며 새로운 AI 시대 열어

"클릭하고 생성한다" 인터페이스를 장악한 범용 에이전트

2025년은 AI가 인간의 프롬프트를 해석하는 수준을 넘어, 컴퓨터 인터페이스를 직접 조작하는 범용 AI 에이전트가 정식 출시된 해이다. OpenAI는 7월 ChatGPT Agent를 정식 출시했다. ChatGPT 내에서 agent mode를 선택하고 사용자의 허가를 얻으면 , 브라우저 탐색, 파일 생성, 코드 실행, 캘린더 관리 등 여러 단계 작업을 자동으로 수행한다. 이어서 Google은 10월 Gemini 2. 5 Computer Use 모델을 공개하며, 브라우저나 모바일 사용자 인터페이스의 화면 요소를 인식하고 클릭·입력 등의 조작을 수행하는 사용자 인터페이스 제어형 에이전트를 API 형태로 제공했다. 이로써 AI는 사용자의 작업을 모방하거나 대신 수행하는 새로운 시대를 열었다.

클라우드와 업무용 앱에 내재화된 AI 비서를 통한 기업 생산성의 극대화

범용 에이전트의 등장과 함께, 기업 환경의 복잡한 워크플로를 자동화하는 업무 특화 에이전트 플랫폼도 주요 경쟁 분야로 떠올랐다. Microsoft는 11월경 Azure 블로그를 통해 Azure Copilot 에이전트를 공식 발표했는데, 이는 클라우드 플랫폼에 직접 내장되어 마이그레이션, 앱 현대화, 트러블슈팅, 비용 및 성능 최적화 등 클라우드 운영을 자동화하는 에이전트를 Azure 자체 기능으로 제공한다. Google 역시 12월 초, Gemini 3를 기반으로 하는 Workspace Studio를 런칭하며, Gmail, Drive, Chat 등 업무용 앱 내에서 메일 분류, 승인 프로세스 추적, 보고서 생성 등 복잡한 워크플로를 누구나 자연어로 만들 수 있는 도구를 제공했다.

멀티 에이전트 협업 인프라 구축으로 '컨텍스트' 한계 극복

주요 AI 개발사들은 에이전트 시대에 맞춰 시스템 구축 및 관리 인프라를 확장했다. Microsoft는 5월 Build 2025에서 "AI 에이전트의 시대"와 "Open Agentic Web" 비전을 선언하고, Copilot Studio를 통해 하나의 시나리오 안에서 여러 에이전트(예: 세일즈, 지원)를 조율하는 멀티 에이전트 오케스트레이션 기능을 발표했으며, 거버넌스 기능도 함께 제공했다. Anthropic은 9월 말 Claude Agent SDK를 공개하여, TypeScript 및 Python SDK를 통해 멀티스텝 작업, 툴 호출, 장기 컨텍스트 관리를 쉽게 구축할 수 있는 프로덕션용 에이전트 루프를 제공했다. 특히, 기존 에이전트의 약점이었던 긴 작업에서의 컨텍스트 끊김 문제를 멀티세션 및 외부 메모리 구조를 통해 완화·개선하는 접근을 제시했다고 기술 기사를 통해 설명했다.

12월의 용어 범용 AI 에이전트, 다중 에이전트 시스템, 사용자 인터페이스 제어형 에이전트

출처 : 1) Google(2025. 12. 4), Introducing Google Workspace Studio: Automate everyday work with AI agents.
2) Venturebeat(2025. 11. 28), Anthropic says it solved the long-running AI agent problem with a new multi-session Claude SDK.

135 범용 AI 에이전트

General-Purpose AI Agent

다양한 작업을 스스로 목표 해석·계획·실행하도록 설계된 AI 에이전트

- 특정 분야에 한정되지 않고 여러 환경과 도구를 넘나들며 작업을 완수하는 능력을 목표로 함
- 지식 활용, 판단, 실행을 결합해 복합적인 실제 업무를 처리하도록 설계된 AI 구조

범용 AI 에이전트의 개념

범용 AI 에이전트는 특정 작업 자동화에 최적화된 기존 AI 에이전트와 달리, 서로 성격이 다른 다양한 작업을 하나의 시스템이 자율적으로 조합·전환하며 수행하도록 설계된 고도형 에이전트입니다. 입력 형태나 작업 조건이 변해도 별도의 구조 변경 없이 대응하며, 정보 탐색–계획 생성–행동 실행–결과 평가가 하나의 순환 구조로 통합되어 있다는 점이 핵심입니다. 즉, 이미 정의된 절차를 따르는 도구가 아니라, 여러 업무 흐름을 스스로 구성하고 필요에 따라 전략을 바꾸며 문제를 해결하는 범용적 실행 주체를 목표로 합니다.

범용 AI 에이전트의 특징

범용 AI 에이전트는 개별 기능을 단순 나열하는 방식이 아니라, 서로 다른 작업을 하나의 연속된 업무 흐름으로 엮어 자연스럽게 수행하는 능력을 갖습니다. 실행 과정에서 오류가 발생하면 원인을 진단하고 대체 전략을 탐색하는 자기 복구 능력도 포함되며, 이는 기존 에이전트에 비해 훨씬 높은 수준의 자율 조정 기능을 의미합니다. 또한 도구 사용을 고정된 절차로 수행하는 것이 아니라, 목표 달성에 적합한 API·애플리케이션·파일 조작 방식을 상황에 맞게 선택해 활용합니다. 필요할 경우 사용자가 지시하지 않은 보조 작업을 스스로 추가해 업무 범위를 재구성하기도 하는데, 이러한 작업스펙을 스스로 확장하는 특성이 범용성을 강화하는 역할을 합니다.

에이전틱 AI vs 범용 AI 에이전트

에이전틱 AI는 여러 AI 에이전트가 역할을 분담하고 협력하는 집단 지능적 구조를 기반으로, 각 에이전트가 비교적 좁은 전문 기능을 맡고, 오케스트레이션 계층이 전체 흐름을 조율하는 구조입니다. 반면 범용 AI 에이전트는 범용 에이전트는 단일 에이전트가 넓은 범위의 작업을 처리하는 개별적 자율 주체로, 계획·추론·실행 기능을 하나의 모델 안에 통합해 중앙 오케스트레이션 없이 단일 에이전트 단위에서 복합 작업을 연속적으로 수행할 수 있습니다.

범용 AI 에이전트의 한계

범용 AI 에이전트는 높은 자율성만큼 예측 가능성과 통제 가능성의 어려움을 동반합니다. 계획을 스스로 수정하거나 보조 작업을 추가하는 과정에서 사용자 의도와 다른 행동이 나타날 수 있으며, 이는 실사용 환경에서 신뢰성 문제로 이어질 수 있습니다. 또한 외부 도구·웹·파일 시스템과 폭넓게 상호작용하기 때문에 보안·프라이버시 위험이 증가하며, 안전한 실행 경계와 접근 통제가 필수적입니다. 더불어 자율적 결정이 많아질수록 작업 실패나 오류 발생 시 책임 주체와 통제 범위를 명확히 정의해야 하는 과제가 대두되고 있습니다.

참조 용어 | AI 에이전트 p.104

136 다중 에이전트 시스템

Multi-Agent System, MAS

여러 에이전트가 협력·조정·상호작용을 통해 문제를 해결하는 AI 시스템

- 단일 에이전트가 수행하기 어려운 과제를 역할 분담과 상호 보완을 통해 처리하는 구조
- 환경 변화에 적응하며 집단적 지능을 형성하는 것이 핵심 특성

다중 에이전트 시스템의 개념

MAS은 여러 개의 독립적인 에이전트가 하나의 환경에서 상호작용하며 공동 목표를 달성하도록 구성된 시스템입니다. 각 에이전트는 제한된 범위 내에서 인지·판단·행동할 수 있는 자율성을 가지지만, 전체 시스템 차원에서는 협력, 경쟁, 정보 교환 등 다양한 형태의 상호작용을 통해 단일 에이전트보다 더 복잡한 문제 구조에 대응합니다. MAS는 하나의 강력한 모델이 모든 역할을 수행하는 구조가 아니라, 각기 다른 능력을 가진 여러 에이전트가 분산된 역할 분담과 상호작용을 통해 문제를 해결하는 집합적 체계를 지향합니다. 이 때문에 환경 규모가 크거나, 정보가 부분적으로만 제공되거나, 작업이 다단계로 구성된 문제에서 특히 효과적입니다.

다중 에이전트 시스템의 특징

MAS의 핵심 특징은 분산성·자율성·상호작용성입니다. 에이전트들은 자신의 관점에서 정보를 수집하고 의사결정을 내리지만, 필요에 따라 서로 역할을 조정하거나 협력적 전략을 형성해 더 높은 수준의 결과를 만들어냅니다. 특정 에이전트가 실패해도 다른 에이전트가 기능을 보완할 수 있어 시스템 전체의 탄력성(resilience)이 높습니다. 최근에는 LLM 기반 에이전트들이 MAS 형태로 팀을 이루어 분석, 검증, 실행 역할을 나누고, 상호 피드백을 통해 정확도와 안정성을 높이는 설계 방식이 주목받고 있습니다. 즉, MAS가 고정된 구조가 아니라, 환경 변화에 따라 에이전트 간 관계나 전략이 자연스럽게 변할 수 있는 적응적 시스템이라는 점에서 단순한 분업 체계를 넘어서며, 이른바 '에이전틱 AI(agentic AI)'로 확장 발전 중에 있습니다.

다중 에이전트 시스템의 의의

MAS는 여러 개체의 협력과 조정을 통해 집단적 문제 해결 능력을 만들어내는 조직적 지능(collective intelligence)의 구현 방식으로 중요한 의미를 갖습니다. 복잡한 계획 수립, 장기 의사결정, 경제·사회 시스템 시뮬레이션, 로봇 군집 제어 등 단일 모델로는 처리하기 어려운 분야에서 강점을 보이며, 인간 조직이 협업을 통해 성과를 내는 방식과도 자연스럽게 닮아 있습니다. 특히 에이전틱 AI의 부상과 함께, MAS는 여러 에이전트가 역할을 분담하는 고도화된 AI 생태계의 기반 구조로 주목받고 있습니다. 다양한 능력을 가진 에이전트를 조합해 시스템의 확장성과 안정성을 높일 수 있다는 점에서, 향후 대규모 자동화·AI 운영 체계·복합 의사결정 지원 시스템의 주요 기술로 활용될 것으로 전망됩니다.

참조 용어 | AI 에이전트 p.104, AI 오케스트레이션 p.105

137 에이전틱 AI

Agentic AI

여러 AI 에이전트가 협력해 주어진 목표를 자율적으로 수행하는 고차원 AI 체계

- LLM 기반 AI 에이전트가 상호 협력해 주어진 목표를 인식하고 계획을 자율적으로 조정하는 고도화된 AI 체계
- 정해진 규칙이 아닌 복잡한 상황에서 다수의 에이전트가 협력·판단·행동하는 집단 지능형 AI의 발전적 형태

에이전틱 AI의 개념

에이전틱 AI는 단일 AI 에이전트가 독립적으로 작동하는 수준을 넘어, 여러 에이전트가 공동의 목표를 인식하고 자율적으로 협력하는 지능형 AI 시스템입니다. 기본 단위인 AI 에이전트가 '디지털 노동자'로서 데이터를 분석하고 업무를 수행한다면, 에이전틱 AI는 이들을 하나의 통합된 거버넌스 구조로 묶어 상위 수준의 판단과 조율(오케스트레이션)을 수행합니다. 즉, 에이전틱 AI는 에이전트들의 네트워크가 스스로 의사결정을 실행하는 '집단적 지능(collective intelligence)'을 통해 환경 변화에 따라 스스로 전략을 수정하고 목표를 재설정할 수 있는 동적 AI 에이전트 생태계를 표방합니다.

에이전틱 AI의 작동 방식

에이전틱 AI는 LLM 추론 엔진에 기반한 개별 에이전트의 자율성과 협동성을 결합한 다층 구조로 작동합니다. 시스템에는 오케스트레이션 계층이 있어 각 에이전트의 역할을 조정하고, 서로의 판단 결과를 공유하며 집단적 결정을 내립니다. 예를 들어, 한 에이전트가 시장 데이터를 분석하면 다른 에이전트가 이를 바탕으로 생산 계획을 세우고, 또 다른 에이전트가 실행·모니터링을 담당하는 방식입니다. 즉, 개별 에이전트는 LLM 추론 엔진을 활용하여 상황별 환경을 재평가하고 계획을 수정하며, 오케스트레이션 계층을 통해 전체 방향성을 유지하도록 조정합니다. 에이전틱 AI는 이러한 구조를 통해 경직된 규칙 기반의 MAS와 동작 방식과 달리 LLM 기반 자연어 처리를 바탕으로 동적 추론과 자율적 계획 및 협력을 수행할 수 있습니다.

에이전틱 AI의 의의

에이전틱 AI는 AI 에이전트와 다중 에이전트 시스템(MAS)의 확장·진화된 개념으로, 단일 업무나 규칙 기반의 자동화에서 조직 수준의 동적이고 자율적 문제 해결 '체계'로 발전했다는 점에 의의가 있습니다. AI 에이전트가 특정 과제를 수행하는 '실행 단위'라면, 에이전틱 AI는 이들을 조율해 복잡한 목표를 달성하는 '지휘 체계'에 해당합니다. 경직된 규칙 기반 동작 시스템이 MAS라면, 에이전틱 AI는 LLM 기반의 동적 추론과 에이전트 간 자율적 협력 체계입니다. 기업 차원에서는 부서 간 협업, 프로젝트 관리, 공급망 최적화 등 다단계 업무를 통합적으로 운영할 수 있으며, 사회적으로는 AI 생태계 간 상호작용을 통해 지속 학습·집단 의사결정·동적 문제 해결을 실현할 수 있습니다. 다만 높은 자율성과 복잡한 상호작용 구조로 인해 통제·보안·책임 소재가 불분명해질 수 있다는 점이 주요 쟁점으로 지적됩니다. 그럼에도 에이전틱 AI는 AI가 '도구'에서 '자율적 행위자'로 전환되는 핵심 단계로, 향후 AI 거버넌스와 산업 자동화의 중심축이 될 것으로 평가됩니다.

138 사용자 인터페이스 제어형 에이전트

User Interface (UI) Manipulation Agent

화면에 보이는 UI 요소를 인식·조작하며 작업을 수행하는 AI 에이전트

- 내부 API 연동에 의존하지 않고 화면 기반 정보 인식으로 클릭·입력·전환 등을 실행해 다양한 서비스를 자동화
- 비정형 UI와 변화하는 화면에서도 적응적으로 행동 절차를 조정한다는 점이 특징

사용자 인터페이스 제어형 에이전트의 개념

사용자 인터페이스(UI) 제어형 에이전트는 AI가 화면을 직접 읽고 조작해 애플리케이션을 사용하는 방식으로 설계된 에이전트입니다. 웹, 모바일, 데스크톱 등 다양한 환경에서 버튼·입력창·메뉴 같은 UI 요소를 시각적으로 인식하고, 목표 달성을 위한 조작 절차를 스스로 계획합니다. 내부 API나 전용 스크립트가 없어도 기존 소프트웨어를 그대로 활용할 수 있어 적용 범위가 넓고, 사람의 사용 행위를 모사한다는 점에서 전통적 자동화와 구별됩니다. 단순 지시 반복이 아니라 화면 상태를 이해하고 작업 경로를 구성하는 자율적 인터페이스 제어 능력이 핵심입니다.

사용자 인터페이스 제어형 에이전트의 부상

UI 제어형 에이전트는 UI 변경에 취약하고 비정형 화면에서는 규칙 기반 접근이 어려웠던 기존 자동화 기술이 가진 한계를 AI 기반의 시각·추론·행동 결합 구조로 극복하고 있다는 점에서 주목받습니다. 시각 모델이 화면 요소를 인식하고, 언어모델이 작업 목표를 문맥적으로 판단하며, 행동 모듈이 이를 실제 조작으로 실행해 인식-판단-행동이 하나의 루프로 작동합니다. 이 때문에 화면 배치가 바뀌거나 예외 팝업이 등장해도 목표를 다시 해석해 절차를 조정할 수 있습니다. 또한 UI 기반 도구 사용이 가능해 문서 정리, 웹 탐색, 데이터 입력 등 서로 다른 작업을 하나의 연속된 흐름으로 처리할 수 있어 범용 자동화의 기반 기술로 평가됩니다. 특히 API가 없거나 변경이 어려운 레거시 시스템에서도 바로 적용할 수 있어 기업 환경에서 활용도가 크게 확대되고 있습니다.

사용자 인터페이스 제어형 에이전트의 한계

UI 제어형 에이전트는 화면 기반 행동이라는 특성상 작은 UI 변화에도 취약합니다. 해상도 차이, 로딩 지연, 우발적 팝업만으로도 인식 오류가 발생해 잘못된 행동을 수행할 수 있습니다. 내부 API 기반 연동에 비해 보안 통제가 상대적으로 어려울 수 있어 민감 정보 노출, 계정 오작동, 권한 오남용 같은 위험도 존재합니다. 작업 단계가 길어질수록 행동 추적과 검증이 어려워지고, 비의도적 반복 조작이 누적될 수 있는 점도 실사용의 부담입니다. 이러한 한계에도 불구하고 UI 제어형 에이전트는 기존 시스템을 변경하지 않고도 폭넓은 자동화를 구현할 수 있어, 범용 에이전트와 기업용 AI 자동화의 핵심 기반 기술로 자리 잡고 있습니다.

NIA

2025년 AI 동향과 이슈로 살펴보는

AI 시대에 꼭 알아야 할 핵심 용어

| 발　행 : 2025. 12. 30.

| 발행인 : 황종성

| 발행처 : 한국지능정보사회진흥원(NIA) 인공지능정책실 미래전략팀

| ISBN : 978-89-8483-919-9

| 기획 및 문의 : 정현영 선임연구원(hyeon0@nia.or.kr)

| 제　작 : 넥스텔리전스

| 감　수 : 김경훈 (카카오 리더), 김숙경 (한국과학기술원 교수),
안성원 (소프트웨어정책연구소 실장), 이재성 (중앙대학교 교수),
장하영 (써로마인드 대표), 한성수 (한국전자통신연구원 소장)