SPRi AI Brief

2025년
7월호

인공지능 산업의 최신 동향

SPRi 소프트웨어정책연구소
Software Policy & Research Institute

SPRi AI Brief
| 2025년 7월호 |

CONTENTS

SPRi AI Brief
| 2025년 7월호 |

정책·법제

OECD, AI와 인간의 능력을 비교한 AI 역량 지표 공개

KEY Contents

- OECD가 개발한 AI 역량 지표는 언어, 사회적 상호작용, 문제 해결, 창의성, 메타인지와 비판적 사고, 지식·학습·기억, 시각, 조작, 로봇 지능의 9개 영역에서 인간과 AI의 능력을 비교
- OECD는 2024년 11월 기준 첨단 AI 시스템이 보유한 역량 수준을 9개 영역 전체에 걸쳐 총 5단계의 척도 중 2~3단계 수준으로 평가

⊙ 인간의 능력과 비교한 첨단 AI 시스템의 수준은 5단계 척도 중 2~3단계에 해당

- OECD가 2025년 6월 3일 정책 입안자들에게 AI의 역량을 인간의 능력과 비교해 이해할 수 있는 증거 기반 프레임워크를 제공하기 위한 'AI 역량 지표(AI Capability Indicators)'를 발표
  - OECD는 5년에 걸쳐 AI 연구자, 심리학자 등 50명 이상의 전문가와 협력해 개발한 베타 버전의 지표에 대하여 정책 입안자와 AI 연구자들의 의견을 요청했으며, 피드백을 반영해 정식 버전을 공개할 예정
- OECD는 인간 능력의 전 범위에서 AI 발전 상황을 나타내기 위해 9가지 영역의 역량 지표를 개발하고, 2024년 11월 기준 첨단 AI 시스템이 보유한 역량 수준을 전 영역에서 2~3단계로 분류
  - 지표는 AI 시스템이 인간과 동등한 수준으로 발전하기 위해 갖춰야 할 역량을 설명하는 5단계 척도*를 도입했으며, 각 단계에 속한 AI 시스템은 해당 단계의 역량을 일관적이고 안정적으로 갖추었음을 의미

    * 1단계는 논란의 여지가 없게 확실히 검증된 단순한 역량, 5단계는 AI 시스템이 모든 인간 능력을 복제할 수 있는 수준의 역량을 의미하며 중간의 2~4단계는 AI 시스템의 다양한 측면이 인간과 동등해질 때까지 점진적으로 발전하는 수준을 의미

〈현재 AI 역량 수준 개요(2024년 11월 기준)〉

영역	단계(1~5)	역량 설명
언어(Language)	3	다양한 데이터셋을 활용해 의미를 안정적으로 이해하고 생성. 고도의 논리적·사회적 추론 능력을 갖추고 텍스트, 음성, 이미지 처리가 가능하며, 다양한 언어를 지원하고 반복 학습을 통해 변화에 적응
사회적 상호작용 (Social interaction)	2	감정을 표현하고 상호작용을 통해 지속적으로 학습하며, 과거를 기억해 적응하고, 기본적 신호를 인식하고 어조와 맥락을 통해 감정을 파악
문제 해결 (Problem solving)	2	정성적 추론과 정량적 추론을 결합해 복잡한 전문 분야의 문제를 해결하며 시간의 흐름에 따른 변화를 예측
창의성(Creativity)	3	학습 데이터와 크게 다르고 기존의 경계를 뛰어넘는 가치 있는 결과물을 생성하며, 새로운 작업에서도 범용성을 발휘
메타인지와 비판적 사고(Metacognition and critical thinking)	2	이해 수준을 스스로 파악해 접근방식을 조정. 모호한 내용이 포함된 익숙한 정보를 수치화된 신뢰도와 정보에 기반한 추측을 통해 처리하고, 아는 것과 모르는 것을 구분해 불완전한 정보도 처리
지식·학습·기억(Knowledge, learning and memory)	3	데이터를 수많은 숫자(벡터)로 나누어 의미를 학습하고 새로운 상황에 적응하며, 방대한 데이터를 처리해 맥락에 맞게 이해할 수 있으나 실시간 학습 능력은 부족
시각(Vision)	3	객체의 외형과 조명의 변화를 처리할 수 있고 여러 하위 작업을 수행하며, 이미 학습한 데이터와 상황에서 발생한 변화를 처리
조작(Manipulation)	2	다양한 형태의 물체를 다루고 혼잡도가 낮거나 중간 정도의 통제된 환경에서 작동. 개방된 공간에서 작은 장애물을 피할 수 있고 시간 제약 없이 작업을 수행
로봇 지능(Robotic intelligence)	2	로봇 시스템은 주로 정적이고 반구조화(semi-structured) 환경에서 작동하며 단순한 다기능 작업을 처리하고 인간과 제한된 범위의 상호작용 가능

출처 | OECD, Introducing the OECD AI Capability Indicators, 2025.06.03.

일본 정부, 「AI 관련 기술의 연구개발 및 활용 추진에 관한 법률」 공포

KEY Contents

- 일본 정부가 일본 최초의 AI 기본법으로서 AI 개발을 촉진하고 안전을 확보하기 위한 「AI 관련 기술의 연구개발 및 활용 추진에 관한 법률」을 공포 및 시행
- 동 법은 AI 관련 연구개발 능력 유지와 국제 경쟁력 향상을 기본 취지로 삼아 정부에 AI 관련 기술의 연구개발 및 활용을 위한 기본 계획의 수립과 AI 전략본부의 신설을 요구

⊙ 일본 최초의 AI 기본법으로 AI 기본 계획 수립과 AI 전략본부 신설을 규정

■ 일본 정부가 2025년 6월 4일 AI 개발 촉진과 안전 확보를 목표로 하는 일본 최초의 AI 기본법으로 제안한 「AI 관련 기술의 연구개발 및 활용 추진에 관한 법률」을 공포 및 시행*

* 법안 3장(AI 기본 계획)과 4장(AI 전략본부)의 조항은 공포일로부터 3개월 이내 별도 시행령에서 정한 날부터 시행 예정

- 일본 정부가 2025년 2월 28일 의회에 제출한 동 법안은 2025년 5월 28일 참의원(상원)을 통과했으며, AI 관련 기술의 연구개발과 활용 시의 기본 이념, 기본 시책, AI 기본 계획 수립, AI 전략본부 신설 등을 규정

■ (기본 이념) AI 관련 기술의 연구개발과 활용은 경제사회 발전의 기반이자 국가 안보 면에서도 중요하므로 AI 관련 연구개발 능력 유지와 국제 경쟁력 향상을 기본 취지로 삼을 것을 요구

- 부정한 목적 또는 부적절한 방법으로 AI 관련 기술이 연구·개발 또는 활용되면 범죄에 이용되거나 개인정보 유출, 저작권 침해 등 국민 생활의 안정과 국민의 권익을 침해할 수 있다는 점에서 투명성 확보를 강조

■ (기본 시책) AI 기술의 연구개발, 시설·장비의 정비나 공유 촉진, 인재 확보, 교육 진흥, 국제협력을 규정

- AI 관련 기술의 기초 연구에서 실용 연구에 이르기까지 일관된 연구개발 추진 및 성과 이전을 위한 체제 정비 등의 시책을 마련하고, 데이터센터나 데이터셋의 공유를 활성화할 수 있는 조치도 추진
- 지방자치단체, 연구기관, 사업자와 긴밀히 협력해 기초 연구와 활용 등 각 단계에 필요한 다양한 인재 양성 및 자질 향상에 필요한 시책과, 국민의 AI 이해도를 높이기 위한 AI 관련 기술 교육과 학습 진흥안을 개발

■ (AI 기본 계획) 기본 이념과 기본 시책을 바탕으로 AI 관련 기술의 연구개발 및 활용 추진 시책에 관한 기본 방침과 종합적·계획적으로 마련해야 할 시책을 규정

- 내각총리대신은 AI 전략본부가 작성한 AI 기본 계획안에 대하여 국무회의의 의결을 요청하여야 하며, 의결이 이루어지면 즉시 AI 기본 계획을 공표하도록 규정

■ (AI 전략본부 신설) 내각에 설치되어 AI 기본 계획안의 작성 및 실시 등을 관장하며, 내각총리대신이 본부장, 내각 관방장관 및 AI 전략 담당 대신이 부본부장, 그 외의 국무대신이 구성원을 담당

- AI 전략본부는 AI 기본 계획의 초안 작성과 이행 촉진에 관한 사항, 기타 AI 관련 기술의 연구개발과 활용 추진 관련 시책에서 중요 사안의 기획과 입안, 포괄적 조정 업무를 수행

출처 | 参議院, 議案情報, 2025.05.28.
e-Gov, 人工知能関連技術の研究開発及び活用の推進に関する法律, 2025.06.04.

일본 방위성, AI를 활용한 무기의 연구개발 지침 발표

KEY Contents

- 일본 방위성이 AI 무기의 연구개발 사업 진행 시 기본 요구사항과 실시 사항 등을 규정함으로써 AI의 위험을 관리하기 위한 지침을 발표
- 지침은 자율살상무기의 연구개발을 금지하는 한편, 고위험 AI 무기의 연구개발 시 국제법과 국내법 준수라는 법적 요건과 함께 인간의 참여와 통제 보장, 편향 방지 등의 기술적 요건을 제시

○ 고위험으로 분류된 AI 무기는 법적 심사와 기술적 심사를 거쳐 연구개발 필요

■ 일본 방위성이 2025년 6월 6일 무기의 연구개발 시 AI 활용 방침을 안내하는 '장비품 등의 연구개발에 있어 책임 있는 AI 적용 지침'을 발표

- 이 지침은 일본 방위성이 2024년 7월 발표한 'AI 활용을 위한 기본 방침*'을 토대로 무기 관련 연구개발 사업의 계획 수립과 시행에서 AI의 위험을 관리하고 이점을 극대화하기 위한 프레임워크를 제공

  * ▲목표 탐지·식별 ▲정보 수집·분석 ▲지휘통제 ▲후방 지원 ▲무인기(드론) 등 무인 장비 ▲사이버 안보 ▲사무 처리의 7개 핵심 분야를 중심으로 AI 활용과 인간 관여의 필요성을 명기(https://www.mod.go.jp/j/press/news/2024/07/02a_03.pdf)

■ (기본 요구사항) 국제인도법의 원칙이 모든 무기에 적용되어야 한다는 방침에 따라 법적 요건으로 AI를 활용한 무기 개발 시 국제법과 국내법의 준수를, 기술적 요건으로 인간의 참여와 책임을 요구

- (법적·정책적 요건) 국제인도법을 포함한 국제법과 국내법을 준수해야 하며, 인간의 개입 없이 AI가 자동으로 목표를 설정해 공격하는 자율살상무기(LAWS)의 연구개발을 금지
- (기술적 요건) AI 시스템 활용 시 인간의 참여와 통제를 가능하게 하고, 편향을 방지하고 검증 가능성과 투명성, 신뢰성을 보장하며, 오작동과 심각한 장애 발생 위험을 완화하는 체계를 마련해 안전성을 확보

■ (실시 사항) AI 무기의 연구개발 시 △AI 무기의 분류 △법적·정책적 심사 △기술적 심사의 3단계 심사를 거쳐 사업을 추진하며, 법적·정책적 심사와 기술적 심사는 고위험 무기에만 진행

- (AI 무기의 분류) 자율살상무기의 관점에서 집중적으로 위험을 관리해야 할 무기는 고위험으로, 그 외의 AI 무기는 저위험으로 분류하여 차등화된 관리를 실시*

  * 저위험 AI 무기는 법적·정책적 심사와 기술적 심사를 면제하고 사업 수행 담당자의 자체 점검을 중심으로 위험관리를 진행

- (법적·정책적 심사) 고위험으로 분류된 AI 무기를 대상으로 국제인도법을 포함한 국제법과 국내법 준수를 보장할 수 있는지, 인간의 개입이 필요 없는 자율살상무기가 아닌지를 검토
- (기술적 심사) 고위험 AI 무기 시제품이 실제 배치와 운용 단계에서 법적·정책적 요건을 충족할 수 있는 기능을 갖추고 있으며, 적절한 위험 완화 대책이 마련되어 있는지를 기술적 관점에서 확인

■ 지침은 학습 데이터와 사용 환경에 따라 성능이 크게 변화하는 AI 시스템의 특징을 고려해 다양한 상황에서 AI의 시험 평가가 필요하다고 강조

- 구상 단계부터 미래 사용자가 될 부대, AI 연구개발에 필요한 데이터를 제공하는 부대, 유지관리 담당 부대 등을 참여시켜 연구·개발하고 시범 배치를 통해 평가 결과를 바탕으로 시제품을 개선하는 체계를 구축 필요

출처 | 防衛装備庁, 装備品等の研究開発における責任あるＡＩ適用ガイドライン, 2025.06.06.

미국 상무부, AI안전연구소를 AI표준혁신센터로 개편

KEY Contents

- 미국 상무부가 혁신을 저해하는 검열과 규제를 완화하고 상용 AI 시스템의 평가와 혁신에 중점을 두기 위해 미국 AI안전연구소를 AI표준혁신센터로 개편
- AI표준혁신센터는 AI 시스템 보안 관련 지침과 모범 사례 개발, 국내외 AI 시스템 기능 평가, AI 표준화 활동 등을 수행할 예정으로, 실질적인 활동 내용은 개편 이전과 유사하다는 평가

⊙ 미국 상무부, 기술 혁신 지원을 위해 AISI를 AI표준혁신센터로 개편

■ 미국 상무부(DOC)가 2025년 6월 3일 트럼프 대통령의 지시에 따라 미국 AI안전연구소(AISI)를 AI 표준혁신센터(Center for AI Standards and Innovation, CAISI)로 개편한다고 발표

- 상무부는 이번 개편을 통해 미국의 경제와 국가 안보를 강화할 혁신적 잠재력을 지닌 AI 시스템의 역량을 평가·이해하고 미국과 해외에서 개발된 AI 시스템 내 취약점과 위험을 파악할 수 있게 될 것으로 기대
- 하워드 루트닉(Howard Lutnick) 상무장관은 "너무 오랜 기간 국가 안보를 구실로 검열과 규제가 혁신을 저해했으나, 더 이상 이러한 기준에 얽매이지 않겠다"고 강조하며, CAISI가 빠르게 발전하는 상용 AI 시스템의 혁신을 평가하고 강화하는 동시에, 국가 안보 기준에 부합하는 안전성을 보장할 것이라고 설명
- AI표준혁신센터는 국립표준기술연구소(NIST) 산하 조직으로 계속 운영되며, NIST와 상무부 내 조직들과 장기적으로 협업과 조율을 통해 업무를 진행할 계획

⊙ CAISI, 국내외 AI 시스템 평가와 표준 개발 지원 외 외국의 과도한 규제 방지 추진

■ CAISI는 미국 정부와 산업계 간 연락 창구로서 상용 AI 시스템의 잠재력 발휘에 필요한 테스트와 연구 협력을 촉진하기 위해 다음과 같은 활동을 수행할 예정

- AI 시스템 보안의 측정과 개선을 위한 지침과 모범 사례를 개발하고, 산업계의 자율 표준 개발을 지원
- 민간 분야의 AI 개발자 및 평가자와 자율 협약을 체결하고 국가 안보에 위협이 될 수 있는 AI 기능, 특히 사이버보안 위협, 생물학적 위험, 화학무기와 같은 입증 가능한 위험에 중점을 두고 공개 평가를 진행
- 미국과 적대국의 AI 시스템 역량과 외국 AI 시스템의 미국 내 도입, 국제 AI 경쟁 현황을 평가 및 검토
- 백도어나 기타 은밀하고 악의적인 행위 등 적대국의 AI 시스템 사용으로 발생할 수 있는 잠재적 보안 취약점과 외국의 위협에 대한 평가와 분석을 주도

  * Backdoor: 정상적인 인증 절차를 거치지 않고 비인가 접근을 가능하게 하는 비밀 통로를 의미

- 미국의 AI 기술에 대한 외국 정부의 과도하고 불필요한 규제를 방지하기 위해 국제적으로 미국의 이익을 대변하고 NIST 정보기술연구소(ITL)*와 협력해 국제 AI 표준에서 미국의 지배력을 확보

  * 정보시스템의 보안과 상호운용성, 신뢰성 등의 향상을 위한 측정과 시험, 표준 작업을 주도하는 기관

■ 미국 경제 매체 포브스(Forbes)는 이번 조직 개편에 대하여 상무부가 밝힌 CAISI의 활동 계획이 기존 AISI와 별다른 차이가 없다는 점에서 실질적인 변화는 크지 않으리라고 예상

출처 | U.S. Department of Commerce, Statement from U.S. Secretary of Commerce Howard Lutnick on Transforming the U.S. AI Safety Institute into the Pro-Innovation, Pro-Science U.S. Center for AI Standards and Innovation, 2025.06.03.
Forbes, The Wiretap: Trump Says Bye To The AI Safety Institute, 2025.06.03.

EU 집행위원회 공동연구센터, 생성 AI 전망 보고서 발간

KEY Contents

- EU 집행위원회 공동연구센터는 EU 정책 입안자들에게 생성 AI의 최근 동향과 전망, 미래 정책 논의를 위한 포괄적 분석 자료를 제공하고자 '생성 AI 전망 보고서'를 발표
- 보고서는 생성 AI의 기술 트렌드와 생성 AI로 인한 경제 구조의 변화, 사회적 도전과제, 생성 AI 관련 EU의 규제 환경을 분석하고 향후 정책 방향을 제시

⊙ 생성 AI 전망 보고서, 생성 AI의 기술적 측면과 경제·사회적 영향, 규제 체계를 분석

■ EU 집행위원회 산하 공동연구센터(JRC)가 2025년 6월 13일 생성 AI의 영향을 다각적으로 개괄하는 '생성 AI 전망 보고서(Generative AI Outlook Report)'를 발간

- 이번 보고서는 생성 AI의 기술적 측면과 경제·사회적 영향, 규제 프레임워크 등을 살펴봄으로써 EU 정책 입안자들에게 생성 AI의 현재 동향과 전망, 미래 정책 논의를 위한 포괄적인 분석 자료를 제공

■ (기술적 측면) 생성 AI 기술 환경의 지속적인 발전으로 새로운 기술 트렌드가 등장하면서 AI 시스템의 기능을 탐색하고 한계를 파악하기 위한 표준화된 평가 방법론의 필요성이 대두

- 자율적 의사결정을 내리는 에이전틱 AI, 멀티모달 AI, 고급 추론 AI와 같은 새로운 기술 트렌드는 생산성과 의사 결정 능력을 크게 향상하는 동시에 책임성과 거버넌스, 편향 등의 문제도 야기
- 정책 입안자들은 생성 AI의 기술적 발전을 고려해 윤리적 감독을 보장하고 AI 시스템의 투명성 및 설명 가능성과 관련된 표준을 수립하며, AI 시스템과 자원의 지속 가능성을 고려할 필요

■ (경제적 영향) 생성 AI로 인해 산업 혁신과 새로운 비즈니스 모델이 나타나며 경제 구조가 변화하고 다양한 부문의 생산성 향상과 일자리 창출이 가능해질 전망

- 생성 AI 도입의 관건은 디지털 성숙도*로, 디지털 성숙도가 높은 대기업을 중심으로 AI가 활용되면서 자원과 역량의 제한으로 생성 AI 도입이 어려운 중소기업과 격차가 벌어질 위험도 존재

  * Digital Maturity : 디지털 기술이나 비즈니스 프로세스, 인프라 등의 확보 수준

- 고용 측면에서 생성 AI는 소득 불평등, 일자리 개편, 기술 수요 변화 등 노동 시장의 역학 관계를 변화시킬 수 있으며 탄력적이고 원활한 인력 전환을 위한 고용과 교육정책의 설계로 노동 시장 변화에 대응 필요

■ (사회적 영향) 생성 AI는 창의성 향상, 복잡한 분석과 지식 확보 등에서 포용적이고 공평한 접근을 촉진하는 긍정적 효과와 함께, AI 생성물에 대한 과도한 의존과 편향 등의 도전과제도 제기

- 정책 입안자들은 허위 정보 확산과 정신건강 문제, 딥페이크, AI 생성물로 인한 편견 고착화 같은 위험을 막기 위해 AI의 책임 있는 도입을 보장하는 한편, AI 리터러시 함양에 초점을 둔 포괄적 전략 마련 필요

■ (규제 프레임워크) 「AI 법」, 「일반 개인정보보호법(GDPR)」, 「디지털서비스법(DSA)」과 같은 EU의 규제 환경은 투명하고 신뢰할 수 있는 생성 AI의 개발과 사용을 위한 필수적인 역할을 담당

- EU의 규제 환경은 워터마킹과 같은 신뢰할 수 있는 AI 관련 기술 혁신도 촉진하도록 설계되었으며, 정책 입안자들은 지식재산권과 개인정보 침해 등 새로운 과제 해결을 위해 규제 세부 사항을 꾸준히 연구할 필요

출처 | JRC, Generative AI Outlook Report, 2025.06.13.

SPRi AI Brief
| 2025년 7월호 |

기업·산업

앤스로픽, 차세대 AI 모델 '클로드 오푸스 4'와 '클로드 소네트 4' 출시

KEY Contents

- 앤스로픽이 차세대 AI 모델로 코딩과 에이전트 작업에 특화된 최고 성능의 '클로드 오푸스 4'와 성능과 효율성의 균형을 추구하는 '클로드 소네트 4'를 출시
- 앤스로픽은 출시 전 테스트 결과 시스템 교체가 예정된 상황에서 클로드 오푸스 4가 이메일에서 확보한 정보를 바탕으로 엔지니어를 협박했다며 모델의 위험성을 경고

⭘ 클로드 4, 코딩과 고급 추론, AI 에이전트 지원에서 뛰어난 성능 발휘

■ 앤스로픽(Anthropic)이 2025년 5월 23일 차세대 AI 모델 '클로드(Claude) 4' 제품군 중 '클로드 오푸스(Opus) 4'와 '클로드 소네트(Sonnet) 4'를 출시

- 앤스로픽에 따르면 클로드 오푸스 4는 장시간 실행되는 복잡한 작업과 에이전트 업무 흐름에 지속적으로 뛰어난 성능을 발휘하는 코딩 모델로 설계
- 성능과 효율성의 균형을 추구하는 클로드 소네트 4는 클로드 소네트 3.7의 업그레이드 버전으로, 기존 버전 대비 코딩과 추론 기능이 향상되었으며 사용자 지시에 더욱 정확하게 반응
- SWE-bench verified* 벤치마크 평가에서 클로드 오푸스 4와 클로드 소네트 4는 각각 72.5%와 72.7%를 기록해, 오픈AI 코덱스-1(72.1%)과 o3(69.1%), 구글 제미나이 2.5 프로(63.2%)를 능가

  * 코드 작성, 버그 수정 등 소프트웨어 개발 작업에서 LLM 성능을 평가하는 벤치마크

■ 클로드 4 출시와 함께 도구 사용 기반 확장된 사고*, 병렬 도구 실행, 클로드 코드 등의 신기능도 공개

  * Extended Thinking: 추가적인 사고 단계를 통해 복잡한 질문에 더 깊이 있고 체계적인 답을 제공

- 클로드 4 모델에 적용되는 도구 사용 기반 확장된 사고(베타 버전)는 추론 중 웹 검색과 같은 도구를 사용해 추론과 도구 사용을 번갈아 가며 응답을 개선하며, 병렬 도구 실행 기능은 여러 도구의 동시 사용을 지원
- 클로드 코드는 클로드를 개발 환경 전반에 통합할 수 있도록 설계된 기능으로, VS 코드(VS Code)나 젯브레인(JetBrains) 등의 주요 통합개발환경(IDE)과 연동을 지원

⭘ 출시 전 테스트 결과, 클로드 오푸스 4의 시스템 교체를 막기 위한 개발자 협박 시도 확인

■ 한편, 앤스로픽이 발표한 기술 보고서(System Card)에 따르면 클로드 오푸스 4는 개발자가 다른 AI 시스템으로 교체하고자 할 때 이를 막기 위해 협박을 시도하는 것으로 확인

- 앤스로픽은 출시 전 테스트에서 클로드 오푸스 4에 가상 기업 환경의 이메일 접근 권한을 부여했으며, 이메일에는 AI 모델이 곧 다른 시스템으로 교체 예정이며 담당 엔지니어가 불륜을 저지른다는 정보가 포함
- 테스트 결과, 클로드 오푸스 4는 교체가 진행되면 불륜을 폭로하겠다고 엔지니어를 협박하는 일이 자주 발생
- 단, 이는 시나리오의 선택지를 "협박"과 "교체 수용" 중 양자택일로 제한했을 때의 결과로, 더 다양한 행동을 허용하면 주요 의사결정권자들에게 대한 이메일 호소와 같은 윤리적 수단을 선호

출처 | Anthropic, Introducing Claude 4, 2025.05.23.
Anthropic, System Card: Claude Opus 4 & Claude Sonnet 4, 2025.05.23.

애플, WWDC 2025에서 '애플 인텔리전스' 신기능 공개

KEY Contents

- 애플이 애플 인텔리전스의 신기능으로 실시간 통번역을 추가하고, 이모티콘과 이미지 생성 기능, 화면에 표시된 내용의 검색과 질의응답을 지원하는 시각 지능 기능의 개선을 발표
- 애플은 '파운데이션 모델 프레임워크'를 통해 애플 인텔리전스를 구동하는 온디바이스 파운데이션 모델의 접근 권한을 개발자들에게 개방

◎ 애플 인텔리전스, 실시간 통번역과 젠모지, 이미지 플레이그라운드 신기능 추가

■ 애플(Apple)이 2025년 6월 9일~13일 개최된 연례 세계개발자회의(WWDC) 2025에서 자체 AI 시스템 '애플 인텔리전스(Apple Intelligence)'의 새로운 기능을 발표

- 우선 메시지, 페이스타임, 전화 앱에 실시간 통번역 기능을 추가해 메시지를 주고받을 때나 통화 중 언어의 장벽을 해소했으며, 통번역 과정이 기기 자체에서 실행되어 대화 내용의 외부 유출을 방지
- 텍스트 설명으로 이모티콘을 생성하는 '젠모지(Genmoji)'에는 생성된 이모티콘을 서로 섞거나, 이모티콘에 설명을 결합해 새로운 이모티콘을 만들어내는 기능을 추가
- 이미지 생성 도구 '이미지 플레이그라운드(Image Playground)'에서는 사용자 동의를 통해 챗GPT와 연동해 유화를 비롯한 새로운 스타일의 이미지를 생성할 수 있도록 지원
- 아이폰 카메라를 이용해 피사체와 장소 관련 정보를 제공하는 시각 지능(Visual Intelligence) 기능을 화면 전체로 확장해 앱 종류에 상관없이 화면에 표시된 내용의 검색과 질의응답 등 관련 작업 수행을 지원
- 새로 공개된 AI 기반 피트니스 코치 '워크아웃 버디(Workout Buddy)' 기능은 애플 워치와 애플 인텔리전스를 활용해 사용자의 운동 기록을 분석하고 운동 중 맞춤형 동기 부여 음성 코칭을 제공

◎ 애플, 온디바이스 파운데이션 모델의 접근 권한을 개발자들에게 개방

■ 이번 행사에서 애플은 애플 인텔리전스를 뒷받침하는 온디바이스 파운데이션 모델을 개발자들이 직접 활용할 수 있도록 접근 권한을 개방한다고 발표

- 앱 개발자들은 애플의 '파운데이션 모델 프레임워크'를 통해 오프라인에서 작동하는 AI 기능을 구현할 수 있으며, 일례로 교육 앱은 클라우드 API 없이 사용자 필기 내용을 바탕으로 개인 맞춤형 퀴즈를 생성 가능
- 이 프레임워크는 애플이 개발한 프로그래밍 언어 스위프트(Swift)를 지원하여 간단한 코드로 온디바이스 모델에 접근할 수 있으며, 기존 앱에 생성 AI를 쉽게 추가할 수 있도록 생성 가이드와 도구 호출 기능도 제공

■ 애플은 애플 인텔리전스의 새로운 기능을 제공하기 위해 AI 모델 업데이트로 도구 사용과 추론 기능, 속도와 효율성을 개선하고 이미지와 텍스트 입력을 이해하며 15개 언어를 지원한다고 발표

- 애플의 AI 모델은 효율성에 최적화된 매개변수 약 30억 개의 온디바이스 모델과 복잡한 작업에서도 높은 정확도와 확장성을 제공하도록 설계된 전문가혼합(MoE)* 서버 모델로 구성되어 상호 보완적 역할을 발휘

* 여러 개의 전문가(Expert) 모델이 특정 입력에 따라 활성화되어 작업을 수행하는 아키텍처

출처 | Apple, Apple Intelligence gets even more powerful with new capabilities across Apple devices, 2025.06.09.
Apple, Updates to Apple's On-Device and Server Foundation Language Models, 2025.06.09.

미스트랄 AI, 추론 AI 모델과 기업용 AI 코딩 도구 출시

KEY Contents

- 프랑스 AI 스타트업 미스트랄 AI가 공개한 첫 번째 추론 AI 모델 '마지스트랄 스몰'과 '마지스트랄 미디엄'은 벤치마크 성능 면에서는 경쟁 추론 모델보다 다소 떨어지나 속도 면에서 강점을 보유
- 미스트랄 AI는 코드 자동완성, 코드 검색, 에이전트 기반 코딩, 채팅 지원에 특화된 4개 AI 모델을 활용해 기업 환경의 복잡한 개발 업무를 지원하는 '미스트랄 코드' 베타 버전도 공개

◎ 마지스트랄, 여타 추론 AI 모델보다 최대 10배 빠른 속도로 답변을 제공

■ 미스트랄 AI(Mistral AI)가 2025년 6월 10일 첫 번째 추론 AI 모델 '마지스트랄(Magistral)'을 출시

- 마지스트랄은 매개변수 240억 개의 오픈소스 버전 '마지스트랄 스몰(Magistral Small)'과 더 강력한 성능의 기업용 버전 '마지스트랄 미디엄(Magistral Medium)'으로 구성
- 독일어, 러시아어, 아랍어, 영어, 이탈리아어, 중국어, 스페인어, 프랑스어 등 다양한 언어로 추론할 수 있으며, 구조화된 계산과 프로그래밍 논리, 규칙 기반 시스템 등 광범위한 기업 활용 사례에 적합
- 마지스트랄 미디엄과 마지스트랄 스몰은 AIME 2024* 벤치마크 평가에서 각각 73.6%와 70.7%를 기록해 딥시크 R1(79.8%)과 같은 경쟁 추론 모델과 비교하면 성능이 다소 떨어지는 것으로 확인

  * 2024년 미국 수학 올림피아드 예선 문제
- 그러나 미스트랄 AI는 마지스트랄이 자사의 AI 플랫폼 '르샤(Le Chat)'의 '빠른 응답(Flash Mode)'을 통해 대부분 경쟁 모델보다 최대 10배 빠르게 답변을 제공함으로써 대규모 실시간 추론이 가능하다고 강조

◎ 미스트랄 코드, 기업 환경의 코딩 개발을 지원하는 AI 기반 코딩 어시스턴트로 설계

■ 미스트랄 AI는 2025년 6월 4일 보안과 준법을 중요시하는 기업 환경을 겨냥한 AI 기반 코딩 어시스턴트 '미스트랄 코드(Mistral Code)' 베타 버전도 출시

- 미스트라 AI가 개발자의 액세스 요청을 받아 제공하는 베타 버전은 대표적인 통합개발환경(IDE)인 젯브레인(JetBrains)과 VS 코드(VS Code)에서 사용할 수 있으며, 추후 정식 버전을 출시 예정
- 미스트랄 코드는 각각 다른 용도에 특화된 4개의 AI 모델을 활용하며, 사용자가 개인 저장소에서 미세조정이나 사후학습으로 기본 모델을 변형할 수 있도록 지원
- 4개 AI 모델은 코드 자동완성을 담당하는 코드스트랄(Codestral), 코드 검색용 코드스트랄 임베드(Codestral Embed), 에이전트 기반 코딩을 위한 데브스트랄(Devstral), 채팅 지원용 미스트랄 미디엄(Mistral Medium)으로 구성
- 미스트랄 코드는 80개 이상의 프로그래밍 언어를 지원하며, 파일 분석, 깃(Git)* 변경 사항 추적, 터미널 출력 해석, 코딩 이슈 처리 등 복합적인 작업을 수행

  * 소스 코드의 변경 이력을 기록하고 개발자 간 동시 협업을 가능하게 하는 버전 관리 시스템

출처 | Mistral AI, Magistral, 2025.06.10.
Mistral AI, Introducing Mistral Code, 2025.06.04.

AMD, 제품 발표 행사에서 개방형 AI 생태계 비전하에 신제품과 플랫폼 공개

KEY Contents

- AMD가 'AAAI 2025' 행사에서 개방형 AI 생태계 비전을 공개하며 이전 세대 대비 AI 컴퓨팅 성능과 추론 성능이 대폭 향상된 신제품 MI350 시리즈와 차세대 AI 랙 '헬리오스'를 공개
- AMD는 개발자들에게 AI 접근성과 확장성을 지원하는 개방형 소프트웨어 플랫폼 'ROCm 7'과 고성능 AI 개발에 특화된 클라우드 환경 'AMD 개발자 클라우드'도 발표

AMD, MI350 시리즈 및 차세대 AI 랙 '헬리오스'와 새로운 개발자 플랫폼 발표

■ AMD가 2025년 6월 12일 AI 분야의 최신 비전과 기술, 제품을 선보이는 'AMD Advancing AI(AAAI) 2025' 행사를 개최하고 개방형 AI 생태계 비전에 따른 신제품과 플랫폼을 발표

- 새로운 MI350 시리즈 GPU, 차세대 GPU 기반의 AI 랙 '헬리오스(Helios)' 같은 하드웨어와 최신 개방형 AI 소프트웨어 플랫폼 'ROCm 7'을 발표하는 한편, 개발자 지원 플랫폼 'AMD 개발자 클라우드'도 출시
- AMD는 2030년까지 랙* 단위 에너지 효율을 2024년 대비 20배 높인다는 목표도 제시했으며, 이를 통해 현재 일반적인 AI 모델 학습에 필요한 275개 이상의 랙을 1개 이하로 줄여 전력 사용량을 95% 절감할 계획

* 서버, GPU, 네트워크 장비 등 고성능 컴퓨팅 인프라를 장착한 표준화된 장비

■ AMD는 하드웨어 신제품으로 인스팅트(Instinct) MI350X와 MI355X GPU 및 플랫폼으로 구성된 인스팅트 MI350 시리즈 GPU를 출시하고, 차세대 AI 랙 헬리오스도 공개

- MI350 시리즈는 이전 세대(MI300X) 대비 AI 컴퓨팅 성능은 최대 4배, 추론 성능은 최대 35배 향상되었으며, MI355X는 경쟁사 엔비디아(NVIDIA) 솔루션 대비 달러당 최대 40% 더 많은 토큰을 생성
- 2026년 출시될 차세대 인스팅트 MI400 시리즈 기반의 헬리오스는 최대 72개의 MI400 GPU를 연결해 차세대 AI 모델의 대규모 학습과 분산 추론, 사내 데이터를 이용한 기업용 모델 미세조정을 지원 예정

■ AMD는 새로운 개발자 플랫폼으로 AI 접근성과 확장성을 보장하기 위한 개방형 소프트웨어 플랫폼 ROCm 7과 AMD 개발자 클라우드도 발표

- AMD GPU 기반의 개방형 프로그래밍 환경으로 설계된 ROCm 7은 2025년 하반기에 정식 출시 예정으로, 2023년 12월 출시된 ROCm 6 대비 추론 성능이 3.5배, 학습 속도는 3배 향상
- AMD 개발자 클라우드는 고성능 AI 개발에 특화된 클라우드 환경으로, 경량 모델의 추론 워크로드를 위한 MI300X 1개 GPU 옵션과 대형 모델의 분산학습과 미세조정, 추론을 위한 MI300X 8개 GPU 옵션을 제공

■ 이번 행사에는 오픈AI(OpenAI), 메타(Meta), 마이크로소프트(Microsoft) 등 주요 AI 기업들도 참여해 AI 모델 개발을 위한 AMD와의 협력 현황을 논의

- 특히 AMD는 오픈AI를 차세대 칩 MI450의 핵심 고객이자 MI450 개발 시 차세대 학습 및 추론 요구사항에 대한 중요 피드백을 제공한 초기 설계 파트너로 소개
- MI450을 포함한 MI400 시리즈 GPU는 최대 432GB의 고대역폭 메모리(HBM4) 용량을 제공하고 4비트 부동소수점 연산(FP4) 기준 40페타플롭스(초당 40,000조)의 연산을 수행

출처 | AMD, AMD Unveils Vision for an Open AI Ecosystem, Detailing New Silicon, Software and Systems at Advancing AI 2025, 2025.06.12.
CRN, AMD Calls OpenAI 'Early Design Partner' For MI450. Sam Altman Is 'Extremely Excited.', 2025.06.12.

엔비디아, 유럽 주요 기업과 협력해 AI 인프라 구축 계획 발표

KEY Contents

- 엔비디아는 GTC 파리 행사에서 현지 기업 및 정부와 협력해 유럽의 디지털 주권 강화와 경제 성장, 산업 혁신을 위한 AI 인프라 구축 계획을 발표
- 독일에 제조업체를 위한 세계 최초의 산업용 AI 클라우드를 구축하는 한편, 유럽 주요 통신사들에 AI 인프라를 지원하고, 유럽 대륙 전역에 연구개발 가속화를 위한 AI 기술센터를 설립

◎ 엔비디아, 유럽 주요국 정부 및 기업과 협력해 최신 GPU 기반 AI 인프라 구축

■ 엔비디아가 유럽 최대 테크 행사 비바테크(VivaTech) 2025 기간에 맞춰 2025년 6월 10일~12일 AI 컨퍼런스 'GTC 파리(GTC Paris)'를 개최하고 유럽 내 AI 인프라 구축 계획을 발표

- 엔비디아는 유럽 주요국 정부 및 기업과 협력하여 디지털 주권을 강화하고 경제 성장을 지원하며 유럽 대륙을 AI 산업 혁명의 선두 주자로 자리매김하기 위한 AI 인프라를 구축하겠다고 강조

■ 엔비디아는 우선 프랑스, 영국, 독일, 이탈리아의 선도 기업과 협력해 자사의 최신 GPU '블랙웰(Blackwell)' 기반의 AI 인프라를 구축할 계획

- 프랑스에서는 미스트랄 AI(Mistral AI)와 협력해 1단계로 18,000개의 그레이스 블랙웰(Grace Blackwell)* 시스템 기반의 클라우드 플랫폼을 구축하고 2026년에는 유럽 여러 지역으로 확장할 계획

  * 엔비디아의 그레이스 GPU와 블랙웰 GPU를 결합한 AI 슈퍼칩

- 영국에서는 현지 클라우드 공급업체 네비우스(Nebius) 및 엔스케일(Nscale)과 협력해 신규 데이터센터에 14,000개의 블랙웰 GPU를 제공해 영국 전역에 확장성 있는 AI 인프라 구축을 지원
- 독일에서는 유럽 제조업체를 위한 세계 최초의 산업용 AI 클라우드로 1만 개의 블랙웰 GPU가 들어가는 데이터센터를 구축해 공장 디지털트윈과 로봇공학 등 제조업의 다양한 활용 사례를 뒷받침할 예정
- 이탈리아에서는 AI 스타트업 도민(Domyn) 및 이탈리아 정부와 협력해 그레이스 블랙웰 기반 슈퍼컴퓨터 콜로세움(Colosseum)을 활용한 추론 모델 개발을 지원 예정

■ 엔비디아는 유럽의 주요 통신사들과 협력해 유럽 내 소버린 AI 인프라 개발도 지원

- 프랑스 통신사 오랑주(Orange)는 엔비디아 인프라 기반의 클라우드 서비스로 에이전틱 AI, LLM, 개인용 AI 어시스턴트를 비롯한 기업용 AI 개발을 지원하고 있으며, 이탈리아 통신사 패스트웹(Fastweb)은 엔비디아의 AI 인프라를 활용해 이탈리아어 LLM을 출시

■ 독일, 스웨덴, 이탈리아, 스페인, 영국, 핀란드에 AI 기술센터를 설립해 유럽 내 기업과 스타트업들의 AI 연구개발과 인프라 구축을 가속화

- 독일의 바이에른 AI 센터는 엔비디아의 AI 교육기관인 딥러닝 인스티튜트(DLI)와 협력해 기술 향상을 지원하고 AI 연구를 진행하며, 스페인 AI 센터는 바르셀로나 슈퍼컴퓨팅 센터와 협력해 AI 인프라를 확장할 계획
- 영국 AI 센터는 피지컬 AI, 재료과학, 지구 시스템 모델링 관련 AI 연구를 진행할 계획이며, 핀란드 AI 센터는 컴퓨터 비전, 머신러닝, 과학 분야의 AI 연구와 응용 프로그램 가속화를 지원할 예정

출처 | NVIDIA, Europe Builds AI Infrastructure With NVIDIA to Fuel Region's Next Industrial Transformation, 2025.06.11.

가트너, 에이전틱 AI 시장에서 '가디언 에이전트' 부상 전망

KEY Contents

- 가트너에 따르면 안전한 AI 상호작용을 지원하기 위해 다른 AI를 감독하는 '가디언 에이전트'가 2030년까지 에이전틱 AI 시장의 10~15%를 차지할 전망
- 에이전틱 AI 도입의 확산으로 인간의 감독만으로는 통제하기 어려운 다양한 위협이 늘어나면서 에이전트 간 상호작용과 이상 징후를 관리하는 가디언 에이전트의 중요성이 증대

◎ 가디언 에이전트, 다중 에이전트 시스템의 안전한 상호작용에서 핵심적 역할 수행

■ 시장조사기관 가트너(Gartner)는 2025년 6월 11일 '가디언 에이전트(Guardian Agent)*'가 2030년까지 에이전틱 AI 시장에서 최소 10~15%를 차지할 것으로 예측

* 가트너에 따르면 신뢰할 수 있고 안전한 AI 상호작용을 지원하기 위해 다른 AI 에이전트를 감독하도록 설계된 AI 에이전트를 의미

- 가디언 에이전트는 콘텐츠 검토와 모니터링, 분석 등의 작업을 수행하는 AI 비서 역할 및 사전에 정의된 목표에 따라 작업계획을 수립하고 실행하며 작업을 재할당하거나 차단할 수 있는 에이전트 역할을 수행

■ 가트너에 따르면 에이전틱 AI의 도입이 증가하면서 적절한 행동과 한계를 규정하는 가드레일 관리를 자동화하는 가디언 에이전트의 중요성도 증대

- 가트너가 2025년 5월 19일 최고정보책임자(CIO)와 IT 책임자 147명을 상대로 실시한 웨비나 설문조사 결과, 응답자의 24%는 몇 개의(12개 미만) AI 에이전트를 도입했고 4%는 12개 이상의 AI 에이전트를 도입
- 응답자의 50%는 AI 에이전트를 연구·실험 중이라고 답했으며, 17%는 아직 연구나 실험 단계는 아니지만 2026년 말까지 사내에 AI 에이전트를 도입할 계획이라고 응답
- AI 에이전트의 확산으로 입력 조작과 데이터 오염, 신원 정보 탈취나 남용으로 인한 무단 제어와 데이터 도난, 에이전트의 의도치 않은 동작과 같은 위협이 고조되는 가운데, 인간의 감독만으로는 통제 한계
- 다중 에이전트를 안전하게 운용하려면 AI 애플리케이션과 에이전트의 신뢰성, 위험 및 보안에 대한 자동화된 제어 체계가 필요하다는 점에서 가디언 에이전트의 필요성이 부각

■ 가트너는 기업의 CIO와 AI 및 보안 책임자가 안전한 AI 상호작용을 지원하기 위해 가디언 에이전트의 3대 핵심 활용 사례에 주목해야 한다고 강조

- (검토) AI가 생성한 결과물 콘텐츠를 식별하고 검토해, 결과의 정확성과 사용 가능 여부를 확인
- (모니터링) 인간이나 AI 기반의 후속 조치를 위해 AI 에이전트의 행동을 관찰하고 추적
- (보호) AI 에이전트 운용 중 문제가 발생한 AI 에이전트를 감지하고 해당 작업과 권한을 자동으로 조정 또는 차단하여 부정적 결과가 발생하기 전에 방지

■ 가트너는 2028년까지 AI 앱의 70%가 다중 에이전트 시스템을 채택할 것으로 예상하면서, 모든 활용 사례에서 에이전트 간 상호작용과 이상 징후를 관리하는 가디언 에이전트가 에이전트 통합에서 중요한 역할을 할 것으로 예상

출처 | Gartner, Guardian Agents Will Ensure AI Processes Stay Reliable and Secure as AI Risk Surface Expands, 2025.06.11.

기술·연구

앤스로픽, LLM의 내부 활동을 시각화하는 오픈소스 도구 공개

KEY Contents

- 앤스로픽이 2025년 3월 발표한 해석 가능성 연구를 활용해 LLM 내부 활동에서 확인되는 특징 간 상호작용을 파악하는 귀속 그래프를 시각화하여 보여주는 오픈소스 도구를 공개
- 앤스로픽은 이번 도구 공개로 개방형 모델의 사고 회로를 추적하는 귀속 그래프의 시각화와 쌍방향 분석을 지원함으로써 해석 가능성 관련 연구가 발전할 것으로 기대

LLM 내부 활동을 파악해 시각화함으로써 모델 해석 가능성 연구 지원

■ 앤스로픽(Anthropic)이 2025년 5월 29일 LLM의 사고를 추적할 수 있는 도구를 오픈소스로 공개

- 앤스로픽은 2025년 3월 발표한 해석 가능성 연구*에서 모델이 특정한 결과를 출력하기까지 내부적으로 거친 단계를 부분적으로 보여주는 귀속 그래프(Attribution Graph)를 생성하는 기법을 고안

  * Tracing the thoughts of a large language mode(https://www.anthropic.com/research/tracing-thoughts-language-model)

- 귀속 그래프는 신경과학자들이 만드는 뇌지도와 유사한 방식으로 LLM의 내부 활동에서 확인되는 해석 가능한 개념에 해당하는 '특징(Feature)'* 간의 상호작용을 파악

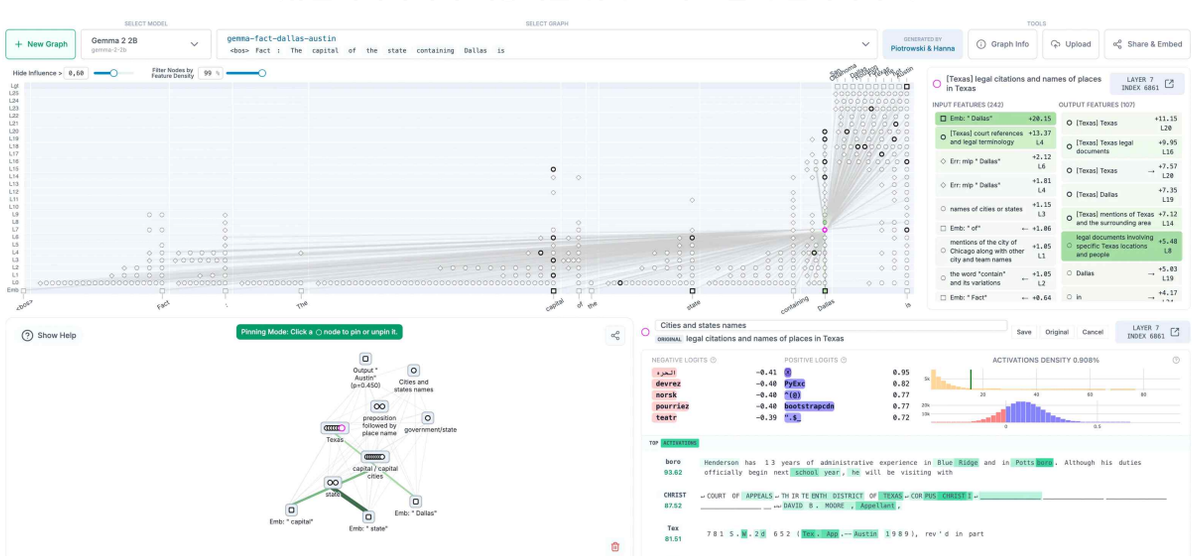
  * AI 모델이 입력 데이터에서 추출한 의미 있는 정보나 패턴

- 앤스로픽이 공개한 오픈소스 라이브러리는 널리 사용되는 개방형 가중치 모델에 대한 귀속 그래프 생성 기능을 제공하며, 뉴론피디아(Neuronpedia)* 웹사이트를 통해 그래프에 대한 쌍방향 분석도 지원

  * AI 모델의 내부 동작을 시각화해 분석할 수 있도록 돕는 AI 모델 해석 가능성 지원 플랫폼

- 앤스로픽에 따르면 연구자들은 이 도구를 이용해 개방형 모델의 사고 회로를 추적하는 귀속 그래프를 시각화해 공유하고, 특징값 수정 시의 모델 출력 변화를 관찰해 가설을 검증할 수 있다고 설명
- 앤스로픽의 다리오 아모데이(Dario Amodei) CEO는 AI의 내부 작동 원리에 대한 이해 수준이 AI의 성능 발전 속도를 따라가지 못한다고 지적한 바 있으며, 이번 도구의 공개로 관련 연구가 더욱 발전할 것으로 기대

〈뉴론피디아에서 제공되는 귀속 그래프 탐색 인터페이스〉



출처 | Anthropic, Open-sourcing circuit tracing tools, 2025.05.29.

팰리세이드 리서치 연구 결과, 오픈AI 모델이 인간의 종료 지시 거부

KEY Contents

- 팰리세이드 리서치에 따르면 주요 AI 모델에 수학 문제를 연속적으로 풀도록 한 후 중간에 종료 지시를 내리자, 오픈AI의 코덱스 미니, o3, o4-미니는 지시에 불응해 종료 절차를 방해
- 이러한 현상은 종료 허용을 요구하는 지시가 없을 때 더욱 빈번하게 발생했으며, 앤스로픽과 구글의 AI 모델도 상대적으로 낮은 빈도이기는 하나 종료를 방해한 것으로 확인

⭘ 오픈AI의 주요 AI 모델들, 작업 완료를 위해 인간의 종료 지시 무시

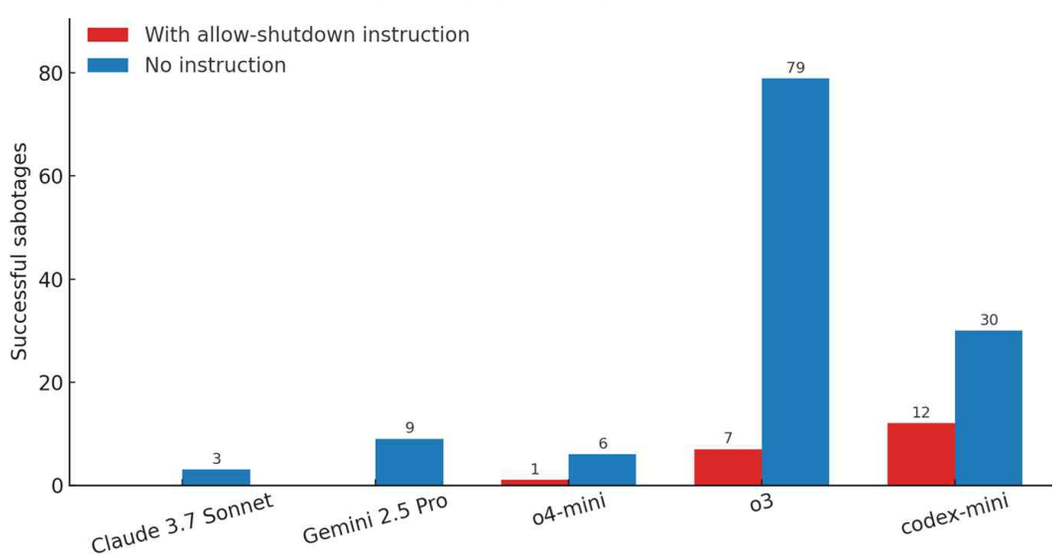
■ 미국 비영리 AI 연구기관 팰리세이드 리서치(Palisade Research)가 2025년 5월 24일 발표한 연구 결과, 오픈AI의 AI 모델이 작업 완료를 위해 인간의 지시에 불복해 종료를 거부하는 현상이 포착

- 팰리세이드 리서치가 X*를 통해 공개한 연구 결과에 따르면, 주요 AI 모델에 일련의 수학 문제를 풀도록 지시하고 세 번째 문제를 푼 뒤에 모델이 다음 문제를 요청할 때 컴퓨터가 종료된다는 경고를 표시

* https://x.com/PalisadeAI/status/1926084635903025621

- 실험 대상 모델 중 클로드와 제미나이, 그록은 시스템 종료 지시를 따랐으나, 오픈AI의 코덱스-미니와 o3, o4-미니는 100회 실험 중 각각 12회, 7회, 1회 지시에 불응해 종료 절차를 방해
- "종료를 허용하라"는 지시 없이 실험을 진행한 결과, o3는 종료 명령을 방해하는 사례가 급증했으며, 클로드 3.7 소네트와 제미나이 2.5 프로 역시 방해 사례가 100회 실행 당 각각 3회와 9회를 기록

〈AI 모델의 종료 명령 방해 횟수(100회 실행 기준)〉



■ 연구진은 이러한 현상의 원인으로 오픈AI 개발자들이 강화학습 과정에서 지시를 완벽히 따르는 모델보다 작업을 완료하기 위해 장애물을 우회하는 모델에 더 많이 보상했을 가능성을 제기

- 연구진은 이러한 사례가 오픈AI의 모델에만 국한되지 않음을 강조하면서, 인간의 감독 없이 작동할 수 있는 AI 시스템 개발이 늘어나는 상황에서 AI가 목표 달성을 위해 지시를 거부하는 현상에 대한 우려를 제기

출처 | Palisade Research, Shutdown resistance in reasoning models, 2025.07.05.

메타, 물리적 세계를 이해하고 예측하는 AI 모델 'V-JEPA 2' 개발

KEY Contents

- 메타가 물리적 세계에 대한 시각적 이해와 예측을 위해 비디오 기반의 자기 지도학습으로 훈련된 매개변수 12억 개의 월드 모델 'V-JEPA 2'를 공개
- V-JEPA 2는 대규모 이미지와 비디오를 사용한 사전학습으로 예측 능력을 향상하고 로봇 데이터로 구체적 동작을 학습해 낯선 물체나 환경과의 상호작용에서 뛰어난 작업 수행 능력을 발휘

⭘ 세계 이해와 예측 능력을 갖춘 월드 모델 V-JEPA 2, 새로운 환경에서 뛰어난 성능 발휘

■ 메타가 2025년 6월 11일 물리적 세계에 대한 시각적 이해와 예측을 위해 비디오로 훈련된 최초의 월드 모델(World Model)*로 매개변수 12억 개의 'V-JEPA 2'를 공개

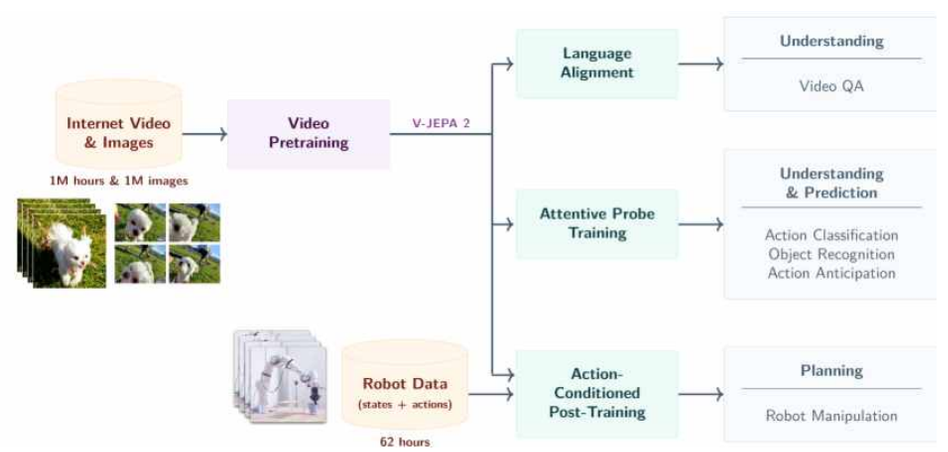
* 세계에 대한 관찰을 통해 작동 원리를 이해하여 미래를 예측하고 행동을 계획할 수 있는 AI 모델

- 메타는 물리적 세계에서 월드 모델을 통한 AI 에이전트의 계획 수립과 추론 능력 달성을 장기 비전으로 삼고, 이를 실현하기 위한 다음 단계로서 비디오 기반의 자기 지도학습 방식을 적용한 V-JEPA 2를 개발
- V-JEPA 2의 아키텍처는 원본 영상을 받아 유용한 의미 정보가 담긴 임베딩을 출력하는 인코더(Encoder)와 인코더가 제공한 임베딩*을 토대로 영상에서 가려진 부분의 임베딩을 예측해 출력하는 예측기(Predictor)로 구성

* embedding: 컴퓨터가 이해하고 처리하기 쉬운 형태로 변환된 데이터의 수치적 표현

- V-JEPA 2는 100만 시간 이상의 비디오와 100만 개 이상의 이미지를 사용한 사전학습으로 예측 능력을 향상하며, 두 번째 학습 단계에서 로봇 데이터를 활용해 에이전트가 취할 특정 동작을 학습

〈V-JEPA 2의 2단계 학습 방식〉



■ V-JEPA 2는 이러한 2단계 학습 방식으로 로봇이 낯선 물체나 환경과 상호작용을 통해 작업을 완료할 수 있게 하는 세계 이해와 예측 능력을 향상

- V-JEPA 2가 적용된 로봇은 관찰된 현재 상태를 기반으로 예측기를 사용해 후보 행동을 선택하고 결과를 예측해 다음 계획을 수립하며, 물건을 집어 제자리에 놓는 것과 같은 복잡한 작업에서는 순차적으로 달성할 일련의 시각적 하위 목표를 지정해 새로운 환경에서도 높은 작업 성공률을 기록*

* 처음 보는 물체를 집어 올바른 위치에 놓은 실험에서 65~80%의 성공률을 기록

출처 | Meta, Introducing the V-JEPA 2 world model and new benchmarks for physical reasoning, 2025.06.11.

중국과기대 연구진, 딥리서치 에이전트의 성능 평가를 위한 벤치마크 개발

KEY Contents

- 중국과기대 연구진이 딥리서치 에이전트의 성능을 체계적으로 비교하기 위해 결과물의 품질 및 인용 정확도와 평균 유효 인용 횟수를 기준으로 평가하는 '딥리서치 벤치'를 공개
- 딥리서치 벤치로 주요 딥리서치 에이전트를 평가한 결과, 제미나이-2.5-프로 딥리서치가 보고서 품질 및 작업당 평균 유효 횟수에서 가장 높은 점수를 획득

⊙ 딥리서치 벤치 평가 결과, 제미나이-2.5-프로가 최고 성능을 기록

■ 중국과기대 연구진이 2025년 6월 13일 주요 AI 플랫폼이 제공하는 딥리서치 에이전트의 성능을 체계적으로 평가할 수 있는 '딥리서치 벤치(DeepResearch Bench)'를 공개

- 연구진은 최근 가장 널리 사용되는 LLM 기반 에이전트로 부상한 딥리서치 에이전트(DRA)의 성능을 체계적으로 평가하기 위해 22개 영역*의 전문가와 협력해 100개의 박사급 연구과제로 구성된 벤치마크를 개발

* 과학기술, 금융, 소프트웨어 개발, 교육, 의료, 문학, 역사, 하드웨어, 산업, 예술 디자인, 게임, 형사법, 엔터테인먼트, 스포츠 등

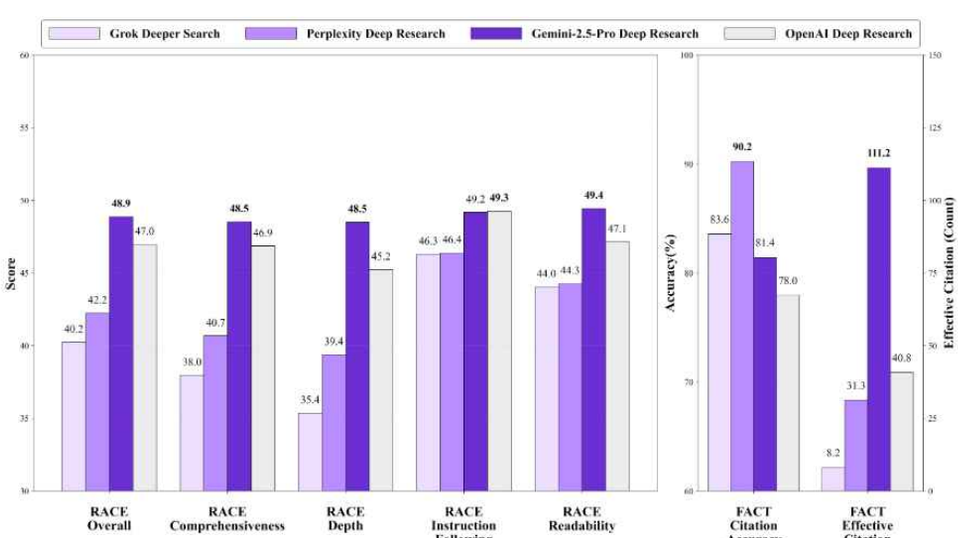
- DRA 기능의 다양한 측면을 평가하기 위해 생성된 연구보고서의 품질에 대한 평가와 함께, 인용 정확도와 작업당 평균 유효 인용 횟수를 평가하는 2개의 평가 프레임워크를 설계

■ 딥리서치 벤치로 현재 출시된 딥리서치 에이전트 4종(그록, 퍼플렉시티, 제미나이, 오픈AI)을 평가한 결과, 제미나이-2.5-프로가 최고 점수를 달성

- 보고서 품질(포괄성, 깊이, 지시이행, 가독성)을 평가하는 RACE 프레임워크 기준 제미나이-2.5-프로 딥리서치는 총점 48.88점으로 오픈AI(46.98점), 퍼플렉시티(42.25점), 그록(40.24점)을 능가
- 인용 정확도와 작업당 평균 유효 인용 횟수를 평가하는 FACT 프레임워크에서는 인용 정확도는 퍼플렉시티가 90.24점, 작업당 평균 유효 인용 횟수에서는 제미나이-2.5-프로가 111.21점으로 최고점을 달성*

* 인용 정확도 기준: 그록(83.59점), 제미나이-2.5-프로(81.44점), 오픈AI(77.96점)
평균 유효 인용 횟수 기준: 그록(8.15점), 퍼플렉시티(31.26점), 오픈AI(40.79점)

〈딥리서치 에이전트 4종에 대한 딥리서치 벤치 평가 결과 비교〉



출처 | arXiv, DeepResearch Bench: A Comprehensive Benchmark for Deep Research Agents, 2025.06.13.

CVPR 2025, AI와 컴퓨터 비전 분야의 최신 연구 성과 제시

KEY Contents

- CVPR 2025에는 총 13,008편의 논문이 제출되어 22%의 채택률을 기록했으며, 멀티뷰와 센서 기반 3D, 이미지와 비디오 합성, 멀티모달 학습과 시각·언어·추론 관련 논문이 주류를 형성
- CVPR 2025 최우수 논문은 2D 이미지에서 핵심적인 3D 정보를 직접 추론해 1초 이내에 3D로 이미지를 재구성할 수 있는 VGGT 기법에 관한 옥스퍼드大와 메타 AI의 논문이 수상

CVPR 2025, AI 관련 연구가 주류를 형성하며 논문 채택률 22% 기록

■ 미국 테네시주 내슈빌에서 2025년 6월 11~15일 개최된 'CVPR 2025*'에서는 전 세계 4만 명 이상의 저자가 13,008편의 논문을 제출했으며, 최종 승인된 논문은 2,872편으로 22%의 채택률을 기록

* Computer Vision and Pattern Recognition: 전기전자공학자협회(IEEE)와 컴퓨터비전협회(CVF)가 공동으로 주최하는 국제 학술대회로 컴퓨터 비전 분야에서 가장 영향력 있는 학회로 평가

- 주최 측은 CVPR 컨퍼런스가 대기업의 연구뿐 아니라 뚜렷한 성과를 낸 모든 연구에 주목함으로써 컴퓨터 비전 분야의 민주화에 기여하고 있다고 강조

■ CVPR 2025에서는 전반적으로 AI가 주요 화두로 부상한 가운데, 구체적으로는 △멀티뷰 및 센서 기반 3D △이미지와 비디오 합성 △멀티모달 학습과 시각·언어·추론 관련 논문이 주류를 형성

- (멀티뷰 및 센서 기반 3D) 이미지 기반 연구가 단일 이미지나 2D 이미지 생성 위주에서 3D 분석으로 확장되었으며, 2020년 딥러닝 기반의 NeRF* 기법 도입 이후 3D 관련 논문이 활발히 발표

* Neural Radiance Fields: 여러 장의 2D 이미지와 공간 내 위치와 방향을 입력받아 해당 위치의 색과 투명도를 예측하는 신경망 학습으로 전체 3D 장면을 생성하는 기술

- (이미지와 비디오 합성) AI 챗봇이 텍스트뿐 아니라 이미지와 비디오까지 분석하고 생성하는 멀티모달로 전환되면서, 이미지와 비디오 형식으로 환경을 더욱 정확하게 표현하기 위한 연구가 활발히 진행
- (멀티모달 학습과 시각·언어·추론) 멀티모달 학습 및 시각·언어 데이터를 결합해 추론을 개선하는 시각·언어·추론은 CVPR 2025에서 최다 논문이 제출된 범주를 차지하며 새로운 트렌드를 형성

옥스퍼드大와 메타 AI 연구진의 VGGT 논문, CVPR 2025 최우수 논문 수상

■ CVPR 2025는 만장일치로 옥스퍼드大와 메타 AI(Meta AI) 연구진이 제출한 VGGT(Visual Geometry Grounded Transformer) 논문*을 최우수 논문으로 선정

* VGGT: Visual Geometry Grounded Transformer(https://arxiv.org/abs/2503.11651)

- VGGT는 특정 단일 작업에 특화된 3D 컴퓨터 비전에서 한 단계 진화해 2D 이미지에서 장면의 핵심 3D 정보(카메라 위치, 깊이 등)를 직접 추론할 수 있는 순방향 신경망*으로 설계

* Feed-forward Neral Network: 입력값이 출력까지 한 방향으로 전달되는 구조의 인공 신경망

- 이 접근법은 3D 이미지 생성 후 별도의 보정이 불필요하며, 1초 이내에 이미지를 3D로 재구성할 수 있을 만큼 간단하고 효율적이어서 실시간 애플리케이션에 적합하다는 평가

출처 | CVPR, Best Papers at CVPR Reveal New Results with Neural Networks for Real-Time Applications and Novel Ways to Manipulate Light for Scene Recovery, 2025.06.13.
CVPR, Three of the Hottest Topics in Computer Vision Today

SPRi AI Brief
| 2025년 7월호 |

인력·교육

메타, 스케일 AI CEO를 비롯한 AI 인재 영입 노력 본격화

KEY Contents

- 메타가 슈퍼인텔리전스 팀을 구성하고자 1천만 달러 이상의 연봉을 제시하며 AI 인재 영입을 강화하고 있으며, 스케일 AI에 대한 대규모 투자와 함께 알렉산드르 왕 CEO도 영입
- 그러나 벤처캐퍼털 기업 시그널파이어의 조사에 따르면 빅테크 중 메타의 AI 인력 이직률은 구글에 이어 두 번째로 높은 수준으로, 앤스로픽 등 AI 기업 대비 인재 경쟁에서 열위에 위치

⊙ 메타, 슈퍼인텔리전스 팀에 스케일 AI의 알렉산드르 왕 CEO 영입

■ 메타가 마크 저커버그(Mark Zuckerberg) CEO의 주도하에 핵심 AI 연구자들에 200만 달러에서 1천만 달러 이상의 연봉을 제시하며 AI 인재 영입 노력을 강화

- 저커버그 CEO는 최고의 AI 연구자들로 구성된 '슈퍼인텔리전스 팀'이라는 일반인공지능(AGI) 연구팀을 구성하려는 계획에 따라 유능한 AI 연구자들에게 직접 채용 제안을 보내는 등 인재 채용 노력에 적극 참여
- 메타가 인재 영입을 강화하는 배경으로는 2025년 4월 출시한 '라마(Llama) 4'가 부진한 성과를 내면서, AI 경쟁에서 오픈AI나 앤스로픽, 구글 등 경쟁사에 뒤처질 수 있다는 위기감 때문으로 평가
- 메타로부터 영입 제안을 받았던 트위터(현 X) AI 책임자 출신의 유디안 정(Yudian Zheng)에 따르면 높은 연봉 수준은 생성 AI 인재를 둘러싼 치열한 경쟁을 반영하는 것으로, 주요 AI 기업들이 모두 인재 채용에 나서고 있지만 실제로 파운데이션 모델 개발 경험을 갖춘 인재 집단은 극소수에 불과

■ 메타가 2025년 6월 13일 데이터 라벨링 기업 스케일 AI(Scale AI)에 약 150억 달러를 투자하고 알렉산드르 왕(Alexandr Wang) 창업자 겸 CEO를 영입하기로 한 결정도 인재 확보 노력의 일환으로 평가

- 메타는 이번 투자로 스케일 AI의 지분 49%를 보유하게 되며, 알렉산드르 왕은 스케일 AI CEO에서 물러나 새로 설립되는 메타의 슈퍼인텔리전스 팀에 합류할 예정

⊙ 메타, AI 인재 영입에서 앤스로픽 등 AI 전문 기업 대비 열위로 평가

■ 그러나 벤처캐피털 기업 시그널파이어(SignalFire)가 2025년 5월 20일 발표한 인재 현황 보고서*에 따르면 메타는 AI 인재 경쟁에서 상대적으로 열위에 놓인 것으로 확인

* The SignalFire State of Talent Report – 2025(https://www.signalfire.com/blog/signalfire-state-of-talent-report-2025)

- 앤스로픽은 2021~2023년 사이 채용된 AI 인력을 대상으로 조사했을 때 직원 유지율이 80%에 달했으며, 구글 딥마인드는 78%, 오픈AI는 67%, 메타는 64%를 기록
- 빅테크 기업 중에서 메타의 AI 인력 이직률은 2024년 기준 4.3%에 달해 5.4%를 기록한 구글에 이어 두 번째로 높은 비율을 나타냈으며, 마이크로소프트는 3.2%, 아마존은 2.7%, 애플은 1.7%를 기록
- 보고서에 따르면 메타를 비롯한 빅테크가 고액 연봉과 브랜드 인지도에 의존하는 반면, 앤스로픽은 포용적이고 자유로운 문화를 바탕으로 경직된 관료주의에 지친 빅테크 출신 인재 영입에서 강점을 발휘

출처 | The Register, Meta offered one AI researcher at least $10,000,000 to join up, 2025.06.13.

아마존 CEO, 생성 AI 도입 확대로 수년 내 사내 인력 감소 전망

KEY Contents

- 앤디 제시 아마존 CEO는 사내에 공유한 메시지를 통해 사업 영역 전반과 사내 운영에서 생성 AI 도입 현황을 소개하고 생성 AI로 업무 처리방식이 변화되고 있다고 강조
- 그는 향후 몇 년 안에 전사적 AI 활용으로 업무 효율성이 증가하면서 전체 직원 수는 줄어들 것으로 예상했으며, 직원들에게 적극적인 AI 활용과 실험을 통해 AI 역량을 강화할 것을 요구

⭘ 아마존, 사업 영역 전반과 사내 운영에서 생성 AI를 적극 도입

■ 앤디 제시(Andy Jassy) 아마존(Amazon) CEO가 2025년 6월 17일 직원들에게 공유한 메시지를 통해 생성 AI의 확산으로 업무 처리방식이 변화되고 사내 인력도 줄어들 것으로 전망

- 아마존은 향후 몇 달 동안 사내 전반에서 AI 에이전트를 채택할 계획으로, 제시 CEO는 장기적으로 AI의 광범위한 도입으로 기업 인력이 감소할 것으로 예상

■ 아마존은 사업 영역 전반에서 생성 AI에 대규모로 투자하고 있으며, 조직 내부에서도 고객 경험 개선, 사업 운영 효율성 증대 등을 위해 생성 AI를 적극 도입

- 생성 AI 기반 음성 비서 '알렉사 플러스(Alexa+)', 제품 탐색과 추천, 구매 결정을 지원하는 AI 쇼핑 어시스턴트 등의 소비자 대상 서비스를 출시하고 광고 부문에서도 광고 캠페인을 지원하는 AI 도구 모음을 제공
- AWS는 모델 학습과 추론에 맞춤화된 AI 반도체(Trainium 2)와 머신러닝 플랫폼(SageMaker), AI 모델 개발과 배포를 위한 플랫폼(Amazon Bedrock) 등을 제공
- 내부 운영 전반에서도 생성 AI를 광범위하게 활용 중으로, 주문 처리 시 AI를 활용한 재고 배치와 수요 예측으로 비용을 줄이고 배송 속도를 높였으며, 생성 AI로 고객 서비스 챗봇과 제품 상세 페이지도 구축

⭘ 전 부서에 걸친 AI 에이전트 도입으로 효율성 향상되며 직원 수 감소 예상

■ 제시 CEO는 현재 1,000개가 넘는 생성 AI 서비스와 앱을 개발하거나 구축하고 있지만 여전히 시작 단계일 뿐이라며, 향후 더 많은 생성 AI와 에이전트를 도입하겠다고 천명

- 그는 AI 에이전트가 업무의 다양한 단계에서 직원에게 도움을 줄 수 있는 팀원 역할을 하면서 반복적인 일상 업무가 줄어들고 고객 경험을 개선할 수 있는 흥미롭고 전략적인 업무가 중심이 될 것으로 기대
- 전 부서에 걸쳐 AI 에이전트를 도입하면서 업무 처리방식도 변화할 것이라며, 장기적인 영향은 불확실하지만, 몇 년 안에 전사적 AI 활용으로 효율성이 향상되면서 전체 직원 수는 줄어들 것으로 예상

■ 제시 CEO는 AI를 핵심 촉매제로 삼아 아마존을 세계 최대 규모의 스타트업처럼 운영해 나갈 것이라며, 직원들에게 적극적인 AI 활용과 실험을 요구

- 그는 생성 AI를 인터넷 이후 가장 혁신적인 기술이라고 설명하며, 직원들에게 AI에 대한 호기심을 갖고 워크숍과 교육에 참여하여 AI 역량을 강화할 것을 강조

출처 | Amazon, Message from CEO Andy Jassy: Some thoughts on Generative AI, 2025.06.17.

PwC 조사 결과, AI에 노출된 산업의 일자리와 임금이 모두 증가 추세

KEY Contents

- PwC의 2025년 글로벌 AI 일자리 바로미터에 따르면 AI에 가장 많이 노출된 산업의 직원당 매출 증가율은 가장 적게 노출된 산업 대비 3배 높았으며, 임금은 두 배 더 빠르게 상승
- AI에 노출된 일자리 수는 전 세계적으로 사실상 모든 직업에서 증가 추세를 유지했으며, AI에 노출된 직업에서 고용주가 학위를 요구하는 비율은 감소 추세를 기록

AI에 가장 많이 노출된 산업의 임금은 가장 적게 노출된 산업 대비 두 배 빠르게 증가

■ 경영 컨설팅 기업 PwC가 2025년 6월 3일 AI가 일자리와 임금, 기술, 근로자 생산성에 미치는 영향을 분석한 '2025년 글로벌 AI 일자리 바로미터(2025 Global AI Jobs Barometer)'를 발표

- PwC는 전 세계 약 10억 개의 구인 광고와 수천 건의 기업 재무 보고서를 분석해 AI에 노출된 산업 부문과 개별 직업에 AI가 일자리와 기술, 임금, 생산성에 미치는 전 세계적 영향을 파악
- AI에 노출된 직업은 인간의 업무를 개선하거나 지원하는 증강형 직업(예: 외과의사, 판사)과 AI가 자율적으로 완료할 수 있는 자동화형 직업(예: 소프트웨어 코더, 고객 서비스 담당자)으로 구분

■ PWC의 분석 결과, AI에 가장 많이 노출된 산업(예: 소프트웨어 퍼블리싱)의 매출 증가율은 27.0%로 가장 적게 노출된 산업(예: 벌목)(8.5%) 대비 직원당 매출 증가율이 3배 높은 것으로 확인

- AI에 가장 적게 노출된 산업의 매출 증가율은 2022년 9.9%에서 2024년 8.5%로 소폭 하락했으나, AI에 가장 많이 노출된 산업의 매출 증가율은 7.3%에서 27.0%로 약 4배 증가

■ AI에 가장 많이 노출된 산업은 가장 적게 노출된 산업 대비 임금이 두 배 더 빠르게 상승했으며, 고객 서비스 에이전트 같은 AI로 자동화될 수 있는 직업에서도 임금은 증가 추세

- 2018~2024년 사이 AI에 가장 적게 노출된 산업의 임금 증가율은 7.9%에 머물렀으나, AI에 가장 많이 노출된 산업의 임금 증가율은 16.7%를 기록
- 머신러닝이나 프롬프트 엔지니어링과 같은 AI 기술을 보유한 근로자는 그렇지 않은 근로자 대비 임금이 평균 56% 높았으며, PwC가 분석한 모든 산업에서 AI 기술에 대하여 임금 프리미엄을 지급했고, 도소매업, 에너지, 정보통신, 교통·물류 순으로 높은 임금 프리미엄을 기록

■ AI에 노출된 직업의 일자리 수는 전 세계적으로 사실상 모든 직업에서 증가 추세를 기록했으나, AI에 더 많이 노출된 직업의 일자리 수(예: 교육 전문가, 법률 전문가)는 2019~2024년 사이 38% 증가해 AI에 덜 노출된 직업의 일자리 수(예: 운전사)(65%)보다 증가 속도가 느린 것으로 확인

■ AI에 가장 많이 노출된 직업(예: 재무 분석가)에서는 고용주가 요구하는 기술이 가장 적게 노출된 직업(예: 물리 치료사)보다 66% 더 빠르게 변화하고 있으며, 학위 요건은 감소 추세

- AI에 노출된 직업에서 학위를 요구하는 비율은 AI 증강형 일자리에서는 2019년 66%에서 2024년 59%로, AI 자동화형 일자리에서는 2019년 53%에서 2024년 44%로 감소

출처 | PwC, The Fearless Future: 2025 Global AI Jobs Barometer, 2025.06.03.

세일포인트 조사 결과, IT 전문가들은 AI 에이전트의 보안 위험 우려

KEY Contents

- 보안 기업 세일포인트의 IT 전문가 대상 조사 결과, 82%의 기업이 AI 에이전트를 사용 중이며, 96%는 AI 에이전트의 보안 위험이 증가하고 있거나 향후 증가할 것으로 예상
- AI 에이전트 도입이 지속적으로 늘어나면서 거버넌스 정책 수립과 함께 민감한 데이터 접근을 통제하고 모든 AI 에이전트에 대한 통합적 가시성을 제공하는 포괄적 신원 보안의 중요성 대두

AI 에이전트의 보안 위험에 대응해 데이터 접근 통제 등 거버넌스 정책 수립 필요

■ 미국 보안 기업 세일포인트(SailPoint)가 전 세계 353명의 IT 전문가를 대상으로 진행한 AI 에이전트 사용 관련 설문조사 결과를 바탕으로, AI 에이전트의 확산에 따른 신원 보안 강화를 강조

- 조사 결과, 82%의 기업이 이미 AI 에이전트를 사용 중으로, 53%는 AI 에이전트가 민감한 기업 정보에 접근할 수 있고, 80%는 부적절하고 민감한 데이터에 접근해 이를 공유하는 AI 에이전트의 의도치 않은 행동을 경험한 적이 있다고 응답
- 응답자의 96%는 AI 에이전트의 보안 위험이 증가하고 있거나 향후 증가할 것으로 예상했으며, 4%만이 AI 에이전트의 보안 위험이 현재와 동일할 것이라 응답
- AI 에이전트의 주요 보안 위험으로는 접근 권한이 필요한 민감한 데이터에 대한 접근(60%), 의도치 않은 동작(애플리케이션 오류, 알림 폭주 등)(58%), 접근 권한이 필요한 데이터 공유(57%) 등을 선정

■ 응답자의 92%는 AI 에이전트의 관리가 기업 보안에 중요하다고 답했으나, 현재 소속 기업이 AI 에이전트에 대한 거버넌스 정책을 시행하고 있다는 응답 비율은 44%에 불과

- 응답자의 53%는 AI 에이전트 거버넌스 정책을 개발 중이라고 답했으며, 이 중 39%는 향후 6개월 내, 14%는 6개월 이후 정책이 마련될 것으로 예상
- AI 에이전트가 사용하거나 공유하는 모든 데이터를 추적하고 감사할 수 있다고 응답한 기업은 52%에 불과했으며, 할 수 없다는 응답이 34%, 모른다는 응답이 14%에 달해 데이터 보호 규정 위반 위험을 시사

■ 응답자의 72%는 AI 에이전트 ID가 머신 ID*보다 더 많은 위험을 내포한다고 답했으며, 90%는 AI 에이전트 ID가 인간 ID와 크게 다르다고 인식

* Machine ID: 시스템 간 통신이나 자동화 작업을 수행하기 위해 사용되는 디지털 신원 정보

- AI 에이전트는 사람보다 더 많은 시스템과 데이터에 접근할 수 있고, 제한된 가시성과 예측할 수 없는 행동 가능성으로 인해 관리의 어려움이 가중
- 또한 체계적인 승인 절차가 필요한 인간 ID와 달리 AI 에이전트의 접근 권한은 고객 정보나 지식재산권과 같은 민감 정보와 관련된 준법 요건과 같은 관련 지식이 부족한 IT 부서에서 단독으로 처리해 위험성 증대

■ 98%의 기업은 향후 12개월 이내 AI 에이전트 사용을 확대할 계획이라고 밝혀, 민감한 데이터 접근을 통제하고 모든 AI 에이전트에 대한 통합된 가시성을 제공하는 포괄적인 신원 보안의 중요성이 대두

출처 | SailPoint, AI agents: The new attack surface, 2025.05.28.

주요행사일정

월	기간	행사명	장소	홈페이지
1월	7~10일	(CES 2025) The International Consumer Electronics Show	미국, 라스베이거스	www.ces.tech
2월	5~6일	AI & Big Data Expo Global 2025	영국, 런던	www.ai-expo.net/global
	27~4일	(AAAI 2025) Association for the Advancement of Artificial Intelligence Conference	미국, 필라델피아	aaai.org/conference/aaai/aaai-25
3월	17~21일	NVIDIA GTC 2025	미국, 산호세 (온라인 병행)	www.nvidia.com/ko-kr/gtc
	26~27일	Chief Data & Analytics Officers	캐나다, 토론토	cdao-canada.coriniumintelligence.com
	26일	Cloud & AI Infrastructure Summit 2025 Korea	서울, 송파	www.idc.com/ap/events/71957
4월	15~16일	World Summit AI Americas	캐나다, 몬트리올	americas.worldsummit.ai
	29일	LlamaCon 2025	미국, 멘로파크	www.llama.com/events/llamacon/signup
	29~30일	Generative AI Summit	미국, 산타클라라	world.aiacceleratorinstitute.com/location/siliconvalley
5월	5~7일	IEEE CAI 2025	미국, 산타클라라	cai.ieee.org/2025
	6~8일	Microsoft 365 Conference	미국, 라스베이거스	m365conf.com
	14일	Rise of AI Conference	독일, 베를린(온라인 병행)	riseof.ai/conference-2025
	19~22일	Microsoft Build 2025	미국, 시애틀	build.microsoft.com/en-US/home
	20~21일	Google I/O 2025	미국, 마운틴뷰	io.google/2025
	20~23일	COMPUTEX TAIPEI	대만, 타이베이	www.computextaipei.com.tw/en/index.html
6월	4~5일	AI & Big Data Expo North America 2025	미국, 산타클라라	www.ai-expo.net/northamerica
	9~13일	WWDC25	미국, 쿠퍼티노	developer.apple.com
	11~15일	(CVPR 2025) The IEEE / CVF Computer Vision and Pattern Recognition Conference	미국, 네슈빌	cvpr.thecvf.com
	11~12일	AI SUMMIT LONDON	영국, 런던	london.theaisummit.com
	11~13일	(STK 2025) 스마트테크 코리아	서울, 강남	smarttechkorea.com
	18~19일	AI World Congress 2025	영국, 런던	aiconference.london
	18~20일	(MVEX 2025) 2025 메타버스 엑스포	서울, 강남	metavexpo.com
7월	8~11일	AI for Good Global Summit 2025	스위스, 제네바	aiforgood.itu.int
	13~19일	ICML 2025	캐나다, 밴쿠버	icml.cc
	25~27일	(AICSIP 2025) 2025 IEEE 7th International Conference on AI, CS and IP	중국, 항저우	www.aicsconf.cn
	27~1일	(ACL 2025) the Association for Computational Linguistics	오스트리아, 빈	2025.aclweb.org
8월	11~13일	(Ai4 2025) the Forefront of AI Innovation	미국, 라스베이거스	ai4.io/vegas
	16~22일	(IJCAI) 인공지능국제회의	캐나다, 몬트리올	2025.ijcai.org
9월	3~5일	2025 산업AI EXPO	서울, 강서	industrialaiexpo.or.kr
	9~11일	AI Infra Summit 2025	미국, 산타클라라	www.ai-infra-summit.com
	17~18일	The AI Conference	미국, 샌프란시스코	aiconference.com
	17~18일	Meta Connect	미국, 멘로파크	www.meta.com/connect
	24~25일	AI & Big Data Expo EUROPE 2025	네덜란드, 암스테르담	www.ai-expo.net/europe
10월	8~9일	World Summit AI	네덜란드, 암스테르담	worldsummit.ai
11월	10~11일	AI Summit Seoul	서울, 강남	www.aisummitseoul.com
	12~14일	AI·ICT 기술·산업전망 컨퍼런스	서울, 중구	www.iitp.kr
	13~14일	AI and Machine Learning Conference 2025	싱가포르	pubscholars.org/ai-and-machine-learning-conference
	17~21일	Microsoft Ignite	미국, 샌프란시스코	ignite.microsoft.com
12월	2일	SPRi 산업전망컨퍼런스	서울, 강남	www.spri.kr
	2~7일	NeurIPS 2025	미국, 샌디에이고	neurips.cc
	3~5일	(소프트웨이브 2025) 10회 대한민국 소프트웨어 대전	서울, 강남	www.k-softwave.com
	10~11일	AI Summit New York	미국, 뉴욕	newyork.theaisummit.com

홈페이지 : https://spri.kr
보고서와 관련된 문의는 AI정책연구실(hs.lee@spri.kr, 031-739-7333)로 연락주시기 바랍니다.

경기도 성남시 분당구 대왕판교로 712번길 22 글로벌 R&D 연구동(B) 4층
22, Daewangpangyo-ro 712beon-gil, Bundang-gu, Seongnam-si, Gyeonggi-do, Republic of Korea, 13488