

SPRi AI Brief

인공지능 산업의 최신 동향

2024년 10월호

CONTENTS

I. 인공지능 산업 동향 브리프

1. 정책/법제

- ▷ 미·영·EU, 법적 구속력 갖춘 유럽평의회 AI 국제조약에 서명 1
- ▷ 미국 캘리포니아 주지사, AI 규제법안 「SB1047」에 거부권 행사 2
- ▷ 호주 의회, 동의 없는 딥페이크 음란물 공유를 처벌하는 법안 통과 3
- ▷ UN, '인류를 위한 AI 거버넌스' 최종 보고서 발표 4

2. 기업/산업

- ▷ 앤스로픽과 오픈AI, 미국 AI 안전연구소와 모델 평가 합의 5
- ▷ 오픈AI, 추론에 특화된 AI 모델 'o1-프리뷰' 출시 6
- ▷ 메타의 AI 모델 '라마', 다운로드 수 3억 5천만 회 달성하며 활발한 생태계 형성 7
- ▷ 구글, AI 신기능 '젬스'와 이미지 생성 모델 '이마젠 3' 출시 8
- ▷ 구글, C2PA 표준 적용으로 AI 생성물의 투명성 향상 추진 9
- ▷ 마이크로소프트, 오픈소스 소형 언어모델 '파이 3.5' 공개 10
- ▷ 하이퍼라이트, 오류를 자체 수정하는 '리플렉션 70B' 오픈소스 모델 공개 11

3. 기술/연구

- ▷ 영국 옥스퍼드대 연구 결과, 글로벌 AI 칩 분포의 양극화 현상 심각 12
- ▷ 메타, LLM의 품질과 정확성을 평가하는 '자가학습 평가자' 개발 13
- ▷ 코히어 연구, LLM 사전학습에 코드 데이터 포함 학습시 LLM의 성능 향상 확인 14
- ▷ 중국 연구진, 재판 시뮬레이션으로 LLM의 법률 역량 향상하는 기법 개발 15
- ▷ AI 연구자들, 벤치마크 '챗봇 아레나'의 편향과 투명성 부족 지적 16

4. 인력/교육

- ▷ 영국 정부, AI 교육기업 대상 '콘텐츠 스토어' 프로젝트 발표 17
- ▷ 유고브 조사 결과, 미국 근로자들 AI의 일자리 영향에 엇갈린 의견 표시 18
- ▷ IBM 기업가치연구소, '생성 AI 시대 인적 잠재력 재해석' 보고서 발간 19
- ▷ 서비스나우, AI 도입으로 영국에서 61만 개 일자리 창출 전망 20

II. 주요 행사

- ▷ Cypher 2024 21
- ▷ AI World Congress 2024 21
- ▷ ML and AI Model Development and Governance 21

I . 인공지능 산업 동향 브리프

미·영·EU, 법적 구속력 갖춘 유럽평의회 AI 국제조약에 서명

KEY Contents

- 미국, 영국, EU가 법적 구속력을 가진 유럽평의회 AI 국제조약에 서명했으며, 서명국은 자국 내 조약 이행을 위한 후속 조치 채택 필요
- 그러나 조약에서 규정한 원칙과 의무가 지나치게 광범위하며, 민간 부문에 대한 규제 적용 여부는 당사국에 일임했다는 점에서 조약의 실효성에 대한 우려도 존재

● 유럽평의회 법적 구속력을 갖춘 AI 국제조약에 10개국 우선 서명 참여

- 미국, 영국, 유럽연합이 2024년 9월 5일 유럽평의회(Council of Europe)*의 AI 국제조약 ‘AI와 인권·민주주의·법치에 관한 기본 협약’**에 서명
 - * 46개 회원국으로 구성된 유럽 민주주의·인권·법치 수호 기구
 - ** Framework Convention on Artificial Intelligence and human rights, democracy and the rule of law
- 안도라, 조지아, 아이슬란드, 노르웨이, 몰도바 공화국, 산마리노, 이스라엘도 조약에 서명했으며, 전 세계 국가들도 조약에 서명하고 이행을 약속 가능
- 조약은 유럽평의회 회원국 46개국과 비회원국 11개국을 포함한 총 57개국의 협의를 거쳐 2024년 5월 17일 유럽평의회 각료위원회에서 채택
- AI 시스템으로 인해 영향을 받는 사람들의 인권 보호에 중점을 둔 이번 조약은 2024년 8월 발효된 EU AI 법과는 별개
- 인권·민주주의·법치와 일치하는 AI 시스템의 사용을 목표로 하는 이번 조약은 법적 구속력을 갖춘 첫 국제조약으로, 서명국은 조약 이행을 위한 자국 내 입법이나 행정 조치 채택 필요
 - 조약은 AI 시스템의 수명주기 전반에서 인간의 존엄성과 개인의 자율성 존중, 투명성, 책임성, 평등과 차별금지, 개인정보보호와 같은 기본원칙을 제시
 - AI 시스템 수명주기 내 활동으로 인한 피해에 대하여 효과적인 구제 수단을 보장하고 AI 시스템이 인권·민주주의·법치에 미치는 위험과 영향을 평가 및 완화하는 조치를 요구

● 민간 부문 규제를 당사국에 맡긴 조약의 실효성에 대한 우려도 존재

- 조약 초안 작성에 참여한 비영리기구 ECNL(European Center for Not-for-Profit Law Stichting)의 법률 전문가 프란체스카 파누치(Francesca Fanucci)는 조약의 법적 확실성과 실효성에 우려를 제기
 - 조약에서 규정한 원칙과 의무가 너무 광범위하고 단서 조항이 많으며, 국가 안보 목적으로 사용되는 AI 시스템은 조약 적용 대상에서 제외될 뿐 아니라, 공공 부문과 달리 민간 부문에 규제 적용 여부와 방법을 선택할 권한을 당사국에 일임해 이중 잣대를 적용했다는 지적

출처: Council of Europe, Council of Europe opens first ever global treaty on AI for signature, 2024.09.05.
Reuters, US, Britain, EU to sign first international AI treaty, 2024.09.06.

미국 캘리포니아 주지사, AI 규제법안 「SB1047」에 거부권 행사

KEY Contents

- 미국 캘리포니아주 하원에서 AI 기업의 기반 모델 개발 시 다양한 안전 조치를 요구하고 AI로 인한 심각한 피해 발생 시 AI 기업을 고소할 수 있는 AI 규제법 「SB1047」이 통과
- 그러나 개빈 뉴섬 캘리포니아 주지사는 법안이 크고 비싼 AI 모델 규제에만 중점을 두고 있다며 기술의 위협에 대응한 최선의 접근방식이 아니라는 이유로 거부권을 행사

● 캘리포니아주의 AI 규제법 「SB1047」, 주지사의 거부권 행사로 입법 무산

- 미국 캘리포니아주 하원에서 2024년 8월 28일 AI 규제법 「SB1047」*이 통과되었으나 개빈 뉴섬 (Gavin Newsom) 캘리포니아 주지사의 거부권 행사로 입법 무산

* 법안의 정식 명칭은 〈The Safe and Secure Innovation for Frontier Artificial Intelligence Models Act〉

- 법안은 캘리포니아에서 운영되는 AI 기업의 기반모델 개발 시 다양한 안전 조치를 요구하여, AI 기업은 모델 위험에 대한 안전 평가를 하고 예상치 못한 위험 발생 시 모델 전체를 신속히 중지할 수 있는 기능을 구현해야 하며, 훈련 후 모델 변조를 방지하는 조치를 취할 필요
- 법안은 또한 사망이나 재산 피해 등 AI로 인한 심각한 피해 발생 시 캘리포니아주 법무부 장관에게 AI 개발기업을 고소할 수 있는 권한을 부여

● 캘리포니아주 주지사, 최선의 접근방식이 아니라는 이유로 법안에 거부권 행사

- 하원을 통과한 이번 법안은 오픈소스 발전과 소규모 AI 기업의 발전을 저해할 수 있다는 기술업계의 의견을 반영해 개정된 버전
- 개정안은 AI 안전을 관할하는 별도의 기관을 설립하는 대신 규제 업무를 기존 당국에 맡기기로 했으며, 실질적 피해가 없어도 안전 규정을 준수하지 않으면 AI 기업을 처벌하도록 한 원안과 달리 규정 위반 시에도 실질적 피해를 끼친 경우에만 기업을 처벌하기로 변경
- 그러나 기술 기업과 벤처 투자사, 연방 의원들은 반대를 고수하며 법안 통과를 막기 위한 치열한 로비활동을 전개해 왔으며, 뉴섬 주지사는 결국 법안에 거부권을 행사
- 뉴섬 주지사는 이 법안이 가장 크고 비싼 모델에 대한 규제에 집중하고 있으며, AI 시스템이 위험한 상황에 사용되는지, 중요한 의사 결정에 관여하거나 민감한 데이터를 사용하는지는 고려하지 않는다고 지적
- 뉴섬 주지사는 소형 AI 모델이 민감한 데이터를 다루는 위험한 작업에 사용되거나 대형 AI 모델이 고객 서비스와 같은 위험이 낮은 업무에 활용될 수도 있다며, 「SB1047」 법안이 대중을 AI 기술의 위협에서 보호하기 위한 최선의 접근방식으로 볼 수 없다고 설명

☞ 출처: The New York Times, California Legislature Approves Bill Proposing Sweeping A.I. Restrictions, 2024.08.28.
SiliconAngle, California Gov. Gavin Newsom shoots down divisive AI safety bill SB 1047, 2024.09.29.

호주 의회, 동의 없는 딥페이크 음란물 공유를 처벌하는 법안 통과

KEY Contents

- 호주 의회가 딥페이크 음란물을 동의 없이 공유한 경우 최대 징역 6년, 직접 제작하여 공유한 경우 최대 7년의 징역형을 부과하는 법안을 가결
- 일부 의원은 딥페이크 음란물의 제작이 처벌 범위에 빠진 점을 지적하는 한편, 선거 관련 딥페이크에 대한 처벌 요구도 제기

○ 형법 개정안, 동의 없는 딥페이크 음란물 공유에 최대 7년의 징역형 부과

- 호주 의회에서 2024년 8월 21일 딥페이크(Deepfake) 음란물을 동의 없이 공유한 경우, 최대 7년의 징역형을 부과하는 법안이 통과
 - 「형법 개정안(딥페이크 음란물) 2024」*는 AI나 기타 기술을 이용한 디지털 생성물 등 노골적인 콘텐츠를 동의 없이 공유한 경우 최대 6년, 이를 직접 제작하여 공유한 경우에는 최대 7년의 징역형을 부과

* The Criminal Code Amendment(Deepfake Sexual Material) Bill 2024
 - 마크 드레퓔스(Mark Dreyfus) 연방 법무부 장관은 “동의 없이 공유되는 딥페이크 음란물은 매우 심각한 형태의 학대이며, 여성·소녀 대상이 압도적으로 많고 유해한 성별 고정 관념을 지속시키며 젠더 기반 폭력을 강화”한다고 지적
- 호주 연방 정부는 젠더 기반 폭력을 해결하려는 정책적 노력의 일환으로 2024년 6월 동 법안을 의회에 상정했으며, 이외에도 호주 국민 대상 개인정보 유출 행위를 불법화하는 별도의 법률 제정 계획을 수립

○ 딥페이크 음란물 제작 및 선거 영역까지 처벌 범위를 확대해야 한다는 지적도 제기

- 동의 없는 딥페이크 음란물의 공유를 금지하는 이번 법안은 의회 전반의 폭넓은 지지를 받았으나, 일각에서는 법안이 불충분하며 동의 없는 딥페이크 음란물의 제작 및 이를 제작하겠다는 위협도 범죄로 간주해야 한다고 주장
 - 래리사 워터스(Larissa Waters) 녹색당 상원 의원은 정부가 딥페이크 콘텐츠의 제작에 대한 처벌 규정을 법안에 포함하지 않은 점을 비판했으나, 정부는 이는 주와 준주*의 책임이라고 해명

* 준주(Territory)는 주(State)와 비슷하지만 완전한 주의 지위나 권한을 갖지 않는 지역, 호주에는 6개의 주와 2개의 준주가 있음
- 한편, 데이비드 포콕(David Pocock) 무소속 상원 의원은 딥페이크 처벌 범위가 선거와 관련된 영역까지 확대되어야 한다고 지적
 - 포콕 의원은 유권자를 오도하거나 속이기 위한 딥페이크가 전 세계적으로 폭발적으로 증가하고 있다며, 동의 없이 사용되는 딥페이크는 민주주의를 위협하므로 선거 보호를 위해 금지되어야 한다고 강조

출처: Attorney-General’s Department, New criminal laws to combat sexually explicit deepfakes, 2024.08.21.
 InformationAge, Deepfakes crackdown passed into law, 2024.08.21.

UN, ‘인류를 위한 AI 거버넌스’ 최종 보고서 발표

KEY Contents

- UN의 AI 자문기구가 ‘인류를 위한 AI 거버넌스’ 최종 보고서를 통해 국제협력을 기반으로 포용적이고 분산된 AI 거버넌스의 토대 마련을 촉구
- 보고서는 글로벌 AI 거버넌스의 격차 해소를 위해 AI 국제과학 패널 구성과 AI 정책 대화, AI 표준 교류와 AI 역량 개발 네트워크 형성 등의 7가지 권고안을 제시

○ UN AI 자문기구, 보고서 통해 국제협력 기반의 포용적이고 분산된 AI 거버넌스 마련 촉구

- UN(United Nations)의 AI 자문기구가 2024년 9월 19일 AI의 위험을 해결하고 혁신적 잠재력을 실현하기 위한 권고안을 담은 ‘인류를 위한 AI 거버넌스’ 최종 보고서를 발간
 - AI의 국제 거버넌스 문제 해결을 위해 2023년 10월 출범한 AI 자문기구는 2023년 12월 발표한 중간 보고서와 AI 전문가를 포함한 전 세계 2천 명 이상이 참여한 광범위한 협의를 바탕으로 최종 보고서를 작성
 - 보고서는 UN에 국제협력을 바탕으로 포용적이고 분산된 AI 거버넌스 체계의 토대 마련을 촉구하고, 모든 정부와 이해관계자에 AI 거버넌스에 협력하여 인권의 발전과 보호를 강화할 것을 요구

○ 글로벌 AI 거버넌스 격차 해소를 위한 7가지 권고안 제시

- 보고서는 기존의 AI 거버넌스 논의는 전 세계에 포괄적으로 적용하기 어렵고 글로벌 사우스(Global South)* 국가들이 배제되어 있다고 지적하며, 글로벌 AI 거버넌스 격차 해소를 위한 7가지 권고안을 제시
 - * 선진국을 뜻하는 글로벌 노스(Global North)와 상대되는 개념으로, 아시아, 아프리카, 남미의 개발도상국을 통칭
 - (AI 국제과학 패널 구성) AI 기술의 과학적 측면에 대한 국제적이고 다학제적인 과학 패널을 구성하여 정책 입안자와 대중에게 신뢰할 수 있는 정보를 제공
 - (AI 정책 대화) 포괄적이고 다중 이해관계자를 포함하는 AI 거버넌스 정책 대화를 시작하여 모범 사례를 공유하고 국가 및 지역적 접근방식 간의 조정 및 상호운용성을 개선
 - (AI 표준 교류 체계 마련) 다양한 표준 개발 기구, 기술 기업, 시민사회 대표들 간 AI 표준 교류를 위한 포괄적인 체계를 마련해 개방적이고 상호 운용이 가능하며 신뢰할 수 있는 AI 생태계 구축
 - (AI 역량 개발 네트워크 구축) 공공 및 민간 부문의 모든 사람이 AI를 책임 있고 윤리적으로 사용할 수 있도록 포괄적이고 공평한 AI 역량 개발 네트워크를 구축
 - (AI 글로벌 기금 조성) AI 혁신과 역량 구축을 위한 글로벌 기금을 조성해 AI 및 지속가능발전목표(SDG)* 관련 프로젝트에 자금을 지원하고, 특히 저소득 및 중간 소득 국가의 요구 사항을 우선 추진
 - * 2015년 UN 총회에서 환경, 교육, 평등, 건강, 거버넌스 등의 분야에서 2030년까지 달성하기로 합의한 17개 목표
 - (글로벌 AI 데이터 프레임워크 구축) 교육용 데이터의 출처 및 사용 관련 공통 표준 수립을 위한 프레임워크 구축
 - (AI 사무소 설립) 이번 보고서의 이행 및 다양한 이해관계자 간 협력을 지원하고 UN 사무총장에게 AI 관련 자문을 제공할 AI 사무소를 UN 사무국 내에 설립 권고

출처: United Nations AI Advisory Body, Governing AI for Humanity – Final Report, 2024.09.19.

앤스로픽과 오픈AI, 미국 AI 안전연구소와 모델 평가 합의

KEY Contents

- 앤스로픽과 오픈AI가 미국 AI 안전연구소와 AI 안전 연구·테스트·평가를 위한 협약을 체결하고 안전 위험의 평가 및 완화 방법의 연구에서 협력할 계획
- 바이든 행정부는 앞서 16개 AI 기업으로부터 자발적 안전 서약을 받았으며, 앤스로픽 및 오픈AI의 이번 협약으로 협력을 한층 심화

● 미국 AI 안전연구소, 영국 AI 안전연구소와 협력해 앤스로픽과 오픈AI의 모델 평가 계획

- 미국 국립표준기술연구소(NIST) 산하의 AI 안전연구소(U.S. AI Safety Institute)가 2024년 8월 29일 앤스로픽 및 오픈AI와 AI 안전 연구·테스트·평가를 위한 협약을 체결했다고 발표
- 양해각서에 따르면 미국 AI 안전연구소는 앤스로픽과 오픈AI의 주요 신모델 공개 전후 해당 모델에 접근할 수 있으며, 기능과 안전 위험 평가 및 위험 완화 방법에 관한 연구를 협력하여 진행 예정
- 미국 AI 안전연구소는 영국 AI 안전연구소와 협력해 모델의 안전성 개선 사항이 발견되면 앤스로픽과 오픈AI에 피드백을 제공할 계획으로, 양 연구소는 지난 2024년 4월 연구와 평가, 지침 개발에서 상호 협력을 위한 협약을 체결바 있음
- 엘리자베스 켈리(Elizabeth Kelly) 미국 AI 안전연구소 소장은 안전성 확보는 신기술 혁신을 촉진하는 데 필수적이라며, 앤스로픽 및 오픈AI와 기술 협력을 통해 AI 안전 과학이 발전할 것으로 기대

● 바이든 행정부, AI 기업과 자발적 AI 안전 서약에 이어 협력 심화

- 바이든 행정부는 AI 업계와 협력 방안으로 앤스로픽과 오픈AI를 포함한 총 16개 AI 기업*으로부터 AI 안전 서약**을 받았으며, AI 안전연구소가 기업과 직접 협약을 체결한 것은 이번이 최초

* 아마존, 앤트로픽, 코히어, 구글·딥마인드, G42, IBM, 인플렉션AI, 메타, MS, 미스트랄AI, 네이버, 오픈AI, 삼성전자, 테크놀로지 이노베이션 인스티튜트(TII), xAI, 지푸AI

** AI 생성물에 대한 워터마크 추가, 사이버보안 투자, AI 오남용 모니터링 등을 약속한 자발적 서약

- 앤스로픽의 잭 클라크(Jack Clark) 공동 창립자는 성명을 통해 “미국 AI 안전연구소와의 협력으로 모델 배포 전 엄격한 테스트를 통해 위험을 식별·완화하는 역량을 강화함으로써 책임 있는 AI 개발이 진전될 것”으로 기대
- 앤스로픽은 2024년 6월 ‘클로드 3.5 소네트(Claude 3.5 Sonnet)’ 출시 전 영국 AI 안전연구소와 협력해 모델 테스트를 진행했으며, 영국 AI 안전연구소는 협정에 따라 테스트 결과를 미국 AI 안전연구소와 공유
- 오픈AI의 제이슨 권(Jason Kwon) 최고전략책임자도 “미국 AI 안전연구소가 책임 있는 AI 개발을 위한 미국의 리더십 확립에 중요한 역할을 할 것이며, 이번 협력이 전 세계적 동참을 위한 토대가 되기를 바란다”고 언급

출처: NIST, U.S. AI Safety Institute Signs Agreements Regarding AI Safety Research, Testing and Evaluation With Anthropic and OpenAI, 2024.08.29.

Fedscoop, OpenAI, Anthropic enter AI agreements with US AI Safety Institute, 2024.08.29.

오픈AI, 추론에 특화된 AI 모델 'o1-프리뷰' 출시

KEY Contents

- 오픈AI가 복잡한 추론이 가능하고 과학, 코딩, 수학에서 기존 GPT 모델보다 더 어려운 문제를 해결할 수 있는 AI 모델 'o1-프리뷰'와 'o1-미니'를 공개
- 추론에 특화된 o1은 국제수학올림피아드 예선 문제 테스트에서 정답률이 83%에 달해 GPT-4o(13%)를 크게 앞섰으며 모델 안전성도 GPT-4o보다 대폭 향상

○ 복잡한 추론이 가능한 'o1-프리뷰', 과학, 코딩, 수학에서 뛰어난 성과

- 오픈AI가 2024년 9월 12일 추론에 특화된 새로운 AI 모델 'o1-프리뷰(Preview)'와 'o1-미니(mini)'를 유료 사용자를 대상으로 공개
 - 응답에 앞서 생각에 더 많은 시간을 할애하도록 설계된 이 모델은 복잡한 추론을 할 수 있고 과학, 코딩, 수학에서 GPT-4o보다 더 어려운 문제를 해결 가능
 - o1 모델은 복잡한 코드의 정확한 생성과 코딩 오류 발견에 특히 뛰어나며, 특히 경량 모델인 o1-미니는 o1-프리뷰 대비 비용이 80% 저렴하여 비용 대비 성능이 우수
 - 오픈AI는 챗GPT 플러스와 팀, 엔터프라이즈 및 에듀 사용자에게 모델을 우선 제공하고, 향후 챗GPT 무료 사용자에게도 o1-미니를 지원할 예정
 - 오픈AI는 GPT 시리즈와 달리 복잡한 추론이 가능하다는 점에서 모델명을 '오픈AI o1'로 새롭게 정했으나, 질문에 따라 응답 속도가 느리고 웹 브라우징이나 파일 업로드 등을 지원하지 않아 단기적으로는 초기 모델인 o1보다 GPT-4o가 더 유용할 수 있다고 설명
 - 오픈AI는 o1 모델을 정기적으로 업데이트하고 현재는 지원되지 않는 웹 브라우징, 파일과 이미지 업로드 등의 기능을 추가할 계획이며, GPT 시리즈 모델도 지속 출시 예정
- o1 모델은 사람이 어려운 문제에 답하기 전 숙고하듯이 '사고 연쇄(Chain of Thought)*'를 통해 추론 역량을 개선해 물리학, 화학, 생물학의 복잡한 벤치마크 과제에서 박사과정 학생과 비슷한 성과를 나타냈으며, 수학과 코딩에서도 탁월한 성과를 기록

* 복잡한 문제를 여러 단계로 나누어 해결하여 최종 결과를 도출하는 방식

- o1은 실수를 인식하고 수정하는 방법, 복잡한 단계를 단순한 단계로 세분화하는 방법, 효과가 없는 기존 방식 대신 다른 방식을 시도하는 방법 등을 학습해 추론 능력을 비약적으로 향상
- 국제수학올림피아드(IMO) 예선 문제 테스트에서 GPT-4o의 정답률은 13%였으나 o1의 정답률은 83%였으며, 코딩 능력을 평가하는 코드포스(Codeforces) 대회에서는 상위 11%에 해당하는 성적을 기록
- 오픈AI는 o1의 추론 역량을 활용해 안전 및 정렬 지침을 준수하도록 하는 새로운 안전 교육방식을 고안하여 모델 안전성도 대폭 향상
 - 모델의 안전장치를 우회하고자 하는 탈옥 시도에 대응한 평가에서 GPT-4o는 100점 만점 중 22점을 받았으나 o1-프리뷰 모델은 84점을 기록

출처: OpenAI, Introducing OpenAI o1-preview, 2024.09.12.

메타의 AI 모델 '라마', 다운로드 수 3억 5천만 회 달성하며 활발한 생태계 형성

KEY Contents

- 메타의 오픈소스 AI 모델 '라마'는 2023년 2월 출시 이래 다운로드 수 3억 5천만 건을 기록했으며, 허깅페이스에는 6만 개 이상의 라마 파생 모델이 존재
- 액센추어, AT&T, 도어대시, 노무라, 줌을 비롯한 주요 기업들은 보고서 작성, 고객 관리, 개발 업무 지원 등 자체 목적에 맞게 라마를 미세 조정하여 활용

● 메타의 오픈소스 AI 모델 '라마', 2023년 대비 다운로드 수 10배 이상 증가

- 메타(Meta)가 2023년 2월 처음 출시한 AI 모델 '라마(Llama)'의 다운로드 수가 3억 5천만 건에 달하며 2023년 대비 다운로드 수가 10배 이상 증가했다고 발표
- 라마는 3.1 버전이 출시된 2024년 7월 한 달 동안 허깅페이스(Hugging Face)*에서 2천만 건 이상의 다운로드를 기록해 오픈소스 모델 제품군 중 선두를 차지
 - * 머신러닝 모델을 구축, 배포, 훈련할 수 있도록 다양한 도구와 라이브러리를 제공하는 플랫폼
- AWS, 마이크로소프트(Microsoft) 등의 대형 클라우드 제공업체를 통한 라마의 월간 사용량은 2024년 1월~7월까지 10배 증가했으며, 2024년 8월 한 달 동안 가장 사용자 수가 많았던 버전은 '라마 3.1-405B'로 최대 규모 AI 모델의 인기를 입증
- 자체 활용 목적에 맞게 라마를 미세 조정하는 개발자 커뮤니티가 활성화되며, 허깅페이스에는 6만 개 이상의 라마 파생 모델이 존재
- 메타는 라마의 성공이 오픈소스에 기인한다며, 개발자의 선택권과 역량을 보장함으로써 AI 생태계가 활성화되고 광범위하고 빠른 혁신이 가능해졌다고 강조

● 액센추어, AT&T 등 주요 기업들, 라마를 자체 목적에 맞게 미세 조정해 활용

- 메타는 공식 사이트에 라마를 미세 조정해 자체적으로 활용하는 주요 기업의 사례를 소개
- 액센추어(Accenture)는 라마 3.1을 사용해 ESG(환경·사회·지배구조) 보고서를 생성하는 맞춤형 LLM을 구축하고 있으며, 기존 작성 방식 대비 생산성은 70%, 품질은 20~30% 향상을 기대
- AT&T는 라마를 미세 조정하여 핵심 트렌드 및 고객 요구사항과 고객 경험 개선의 기회를 파악해 비용 효율적으로 고객 관리를 지원
- 배달 대행 플랫폼 도어대시(DoorDash)는 라마를 사용해 내부 지식기반의 복잡한 질문 응답, 코드 베이스 개선 등 소프트웨어 개발자의 일상 업무를 간소화
- 일본 금융기업 노무라(Nomura)는 텍스트 요약, 편향 방지, 코드 생성, 로그 분석 등 문서 관련 업무 전반에서 AWS를 통해 라마를 활용
- 줌(Zoom)은 자체 모델과 폐쇄형 및 오픈소스 LLM(라마 포함)을 활용해 회의 요약, 스마트 녹음 등 반복적인 일상 업무를 대신 처리하는 AI 컴패니언(Companion) 서비스를 이용자에게 제공

출처: Meta, With 10x growth since 2023, Llama is the leading engine of AI innovation, 2024.08.29.

구글, AI 신기능 ‘젬스’와 이미지 생성 모델 ‘이마젠 3’ 출시

KEY Contents

- 구글이 제미나이를 미세 조정해 특정 주제에 대한 전문가로 설정하는 등 맞춤형 AI 챗봇을 생성할 수 있는 신기능 ‘젬스’를 유료 사용자 대상으로 공개
- 구글은 최신 이미지 생성 모델 ‘이마젠 3’을 제미나이 앱에 통합하면서 역사적으로 부정확한 이미지 생성 논란으로 중단되었던 인물 이미지 생성 기능도 재개

● 구글, 제미나이 유료 사용자 대상 맞춤형 AI 챗봇 기능과 최신 이미지 생성 모델 업데이트

- 구글(Google)이 2024년 8월 28일 맞춤형 AI 챗봇을 생성할 수 있는 신기능 ‘젬스(Gems)’를 제미나이(Gemini) 유료 사용자 대상으로 공개
 - 젬스는 제미나이를 특정 주제에 대한 전문가로 설정하거나 원하는 목표에 맞게 미세 조정하여 사용자 맞춤형 AI 챗봇을 제작하는 기능으로, 150개 이상 국가의 제미나이 어드밴스드, 비즈니스, 엔터프라이즈 이용자들은 데스크톱 환경에서 대부분 언어로 젬스를 이용할 수 있음
 - 구글은 복잡한 주제를 쉽게 이해할 수 있도록 안내하는 ‘학습 코치’, 다양한 아이디어를 제공하는 ‘브레인스토머’, 코딩 실력 향상을 위한 ‘코딩 파트너’와 같은 젬스 샘플도 제공
- 구글은 또한 최신 이미지 생성 모델 ‘이마젠(Imagen) 3’을 제미나이 앱에 통합하고 모든 언어 이용자에게 확대 제공할 계획
 - 이마젠 3는 간단한 텍스트 프롬프트를 바탕으로 사실적 풍경이나 유화 같은 질감의 이미지, 클레이 애니메이션 등 다양한 스타일의 이미지를 생성하며, 이마젠 2보다 이미지 생성 기능이 대폭 향상
 - 구글은 이마젠 3를 출시하며 역사적으로 부정확한 인물 이미지 생성으로 2024년 2월 중단된 제미나이의 인물 이미지 생성 기능을 재개
 - 사실적이며 식별 가능한 인물(유명인사 등)이나 미성년자 묘사, 지나치게 폭력적이거나 선정적인 이미지 생성은 불가, 생성 이미지에 디지털 워터마크 표시로 안전성 강화
 - 인물 이미지 생성 기능은 제미나이 어드밴스드, 비즈니스, 엔터프라이즈 이용자를 대상으로 영어로 우선 제공되며 향후 지원 언어와 이용자 범위를 확대할 계획

<동일 프롬프트에 대한 이마젠2와 이마젠3의 결과물 비교>



Imagen 2



Imagen 3

출처 : Google, New in Gemini: Custom Gems and improved image generation with Imagen 3, 2024.08.28.

구글, C2PA 표준 적용으로 AI 생성물의 투명성 향상 추진

KEY Contents

- 구글은 AI 생성물의 투명성 향상을 위해 C2PA 운영위원회 회원으로서 콘텐츠 출처 확인 기술 표준 '콘텐츠 자격증명' 2.1 버전 개발에 기여
- 구글은 향후 몇 달 동안 '콘텐츠 자격증명' 2.1 버전을 검색과 광고 등 자사 서비스에 적용하는 한편, 유튜브에서도 2024년 말까지 C2PA 정보를 전달하기 위한 업데이트를 계획

● 구글, C2PA 운영위원회 회원으로서 '콘텐츠 자격증명' 2.1 버전 개발에 기여

- 구글이 2024년 9월 17일 AI 생성물의 출처 확인을 통한 투명성 향상을 위한 '콘텐츠 출처 및 진위 확인 연합(C2PA)*' 관련 활동 현황을 공개
 - * 온라인 콘텐츠에 대한 신뢰 구축을 위해 2021년 설립된 글로벌 연합체로 콘텐츠 출처 기술 표준을 개발해 출처 인증에 활용
- 구글은 다양한 플랫폼을 넘나드는 콘텐츠의 특성을 고려할 때 온라인 투명성을 높이려면 업계 전반의 협력이 필수적이라는 인식하에 2024년 2월 C2PA의 운영위원회 회원으로 가입
- 구글은 C2PA 운영위원회에 소속된 마이크로소프트, 어도비(Adobe), 메타 등의 기업과 협력해 '콘텐츠 자격증명(Content Credentials)*' 기술 표준의 최신 버전 2.1 버전 개발에 기여
 - * 디지털 콘텐츠에 메타데이터를 포함하여 제작 출처와 관련된 정보를 확인할 수 있게 하는 기술 표준
- '콘텐츠 자격증명' 2.1 버전은 콘텐츠 출처 이력을 검증하기 위한 기술적 요구 사항을 강화해 다양한 위변조 공격에 대응한 보안을 향상

● 구글, AI 생성물 출처 확인을 위해 C2PA의 최신 기술 표준을 자사 서비스에 적용 계획

- 구글은 향후 몇 달 동안 '콘텐츠 자격증명' 2.1 버전을 검색과 광고 등 주요 서비스에 적용할 계획
 - 이미지에 C2PA 메타데이터가 포함된 경우, 이용자들은 구글의 이미지 관련 서비스에서 제공하는 'About this image' 기능을 통해 이미지가 AI 도구로 생성 또는 편집되었는지 확인 가능
 - 구글은 광고 시스템에도 C2PA 메타데이터의 도입을 확대하는 한편, C2PA를 활용해 주요 정책의 시행 방법을 안내할 계획
 - 구글은 유튜브에서도 C2PA로 제공되는 출처 정보를 사용자에게 전달할 방법을 모색 중으로, 2024년 말 관련 업데이트를 진행 예정
 - 향후 발표될 'C2PA 신뢰 목록(Trust List)'을 통한 콘텐츠 출처 확인도 지원할 계획으로, 가령 이용자들이 특정 카메라 모델로 촬영된 이미지의 경우 신뢰 목록을 통해 해당 출처 정보의 정확성을 검증 가능
- 구글은 온라인 콘텐츠 출처 확인을 위해 더 많은 서비스 및 하드웨어 업체의 C2PA 표준 채택을 촉구하는 한편, 구글 딥마인드(Deepmind)에서 개발한 워터마킹 기술 '신스ID(SynthID)'를 여러 AI 도구 및 미디어에 확대 적용함으로써 자체적인 노력도 추진

출처 : Google, How we're increasing transparency for gen AI content with the C2PA, 2024.09.17.

마이크로소프트, 오픈소스 소형언어모델 ‘파이 3.5’ 공개

KEY Contents

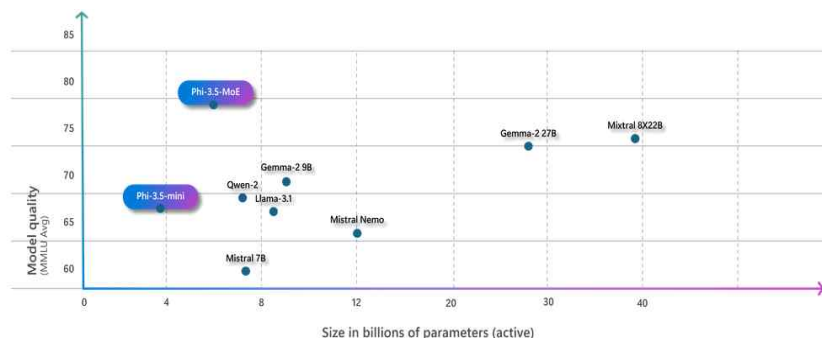
- 마이크로소프트가 다양한 벤치마크에서 동급 또는 규모가 더 큰 모델 대비 성능이 우수하고 비용 효율적인 소형언어모델 ‘파이-3.5’ 3종을 공개
- 16개 전문가 모델을 혼합한 파이-3.5-MoE는 20개 이상의 다국어를 지원하며, 오픈AI의 GPT-4o-mini를 능가하는 추론 성능을 기록

● 마이크로소프트의 ‘파이 3.5’ 3종, 각각 빠른 추론과 복잡한 추론, 시각 작업에 특화

- 마이크로소프트가 2024년 8월 22일 오픈소스 소형언어모델(SLM) ‘파이(Phi)-3.5’ 3종을 허깅페이스(Hugging Face)를 통해 공개
 - 공개된 모델은 파이-3.5-미니, 파이-3.5-MoE(전문가혼합), 파이-3.5-비전으로, 각각 빠른 추론과 복잡한 추론, 시각 작업에 특화
 - 컴퓨팅 자원이 제한된 환경에 적합한 파이-3.5 미니는 38억 개 매개변수로 학습된 SLM으로 12만 8천 개 토큰의 컨텍스트 창과 다국어를 지원하며, 매개변수가 더 큰 미스트랄-7B보다 성능이 우수
 - 16개 전문가 모델을 혼합한 파이-3.5-MoE는 주어진 요청에 필요한 전문가 모델만 처리를 담당하므로 전체 420억 개 매개변수 중 66억 개 매개변수만 활성화하여 효율성을 높이면서 뛰어난 성능을 발휘하며, 20개 이상의 다국어와 최대 12만 8천 개의 컨텍스트 창을 지원
 - 파이-3.5-비전은 다중 프레임 이미지 이해와 추론 기능을 갖추어 이미지 간 비교나 다중 이미지와 동영상 요약 등을 지원하나, 다국어 작업에는 최적화되지 않아 추가 미세조정 필요
- 파이-3.5-MoE는 매개변수가 더 큰 모델과 비슷한 수준의 추론 능력을 갖췄으며, 다국어 작업에서도 훨씬 큰 모델과 비슷한 경쟁력을 발휘
 - 전체 벤치마크 평균에서 파이-3.5-MoE는 69.2점으로 구글의 제미니 1.5-Flash(68.5점)와 젤마-2-9b(63.3점), 미스트랄의 Nemo-12B(61.3점)를 앞서고, 특히 MMLU* 벤치마크에서는 78.9점으로 오픈AI의 GPT-4o-mini(77.2점)를 능가

* 다양한 주제에 대한 모델의 광범위한 지식과 추론 능력을 평가하는 벤치마크

<‘파이 3.5’와 주요 소형언어모델(SLM)의 매개변수 크기, 품질 비교>



출처: Microsoft, Discover the New Multi-Lingual, High-Quality Phi-3.5 SLMs, 2024.08.22.

하이퍼라이트, 오류를 자체 수정하는 ‘리플렉션 70B’ 오픈소스 모델 공개

KEY Contents

- 하이퍼라이트가 오류를 자체적으로 감지하고 수정함으로써 정확도를 향상해 라마-3.1 405B, GPT-4o, 제미니 1.5 Pro를 능가하는 성능의 LLM ‘리플렉션 70B’를 공개
- 그러나 오픈소스 AI 커뮤니티와 연구자들은 리플렉션 70B의 성능을 재현할 수 없다며 모델의 실제 성능에 의혹을 제기

● 리플렉션 70B, 추론 과정에서 발생한 오류를 스스로 수정해 정확도 개선

- AI 글쓰기 서비스를 제공하는 미국 스타트업 하이퍼라이트(HyperWrite)의 맷 슈머(Matt Shumer) CEO가 2024년 9월 5일 X를 통해 신규 LLM ‘리플렉션(Reflection) 70B’를 공개
 - 메타의 ‘라마 3.1-70B 인스트럭트’를 미세 조정한 리플렉션 70B는 오류를 자체 수정하는 기법을 활용해 정확도를 향상시킴으로써 오픈소스 모델 중 가장 뛰어난 성능을 발휘
 - 리플렉션 70B는 공개된 벤치마크 전체에서 라마-3.1 405B와 GPT-4o, 제미니(Gemini) 1.5 Pro를 앞섰으며, GPQA*와 HumanEval** 벤치마크를 제외하고 ‘클로드(Claude)-3.5 소네트(Sonnet)’ 보다 우위
 - * 대학원 수준의 추론 능력 평가 **코딩 능력 평가
- 리플렉션 70B는 추론 과정에서 발생한 오류를 스스로 감지하고 최종 응답을 내리기 전에 오류를 수정하는 것이 특징
 - 리플렉션-70B는 추론과 오류 수정을 위한 여러 특수 토큰*을 도입해 모델의 추론 중 실수가 감지되면 실시간 수정이 가능하며, 추론을 여러 단계로 나누어 정밀도를 개선
 - * 언어모델이 텍스트를 이해하고 생성하는 기본 단위

● 오픈소스 AI 커뮤니티와 연구자들, 리플렉션 70B의 실제 성능에 의혹 제기

- 그러나 리플렉션-70B 공개 이후 오픈소스 AI 커뮤니티와 외부 평가자들은 모델의 성능을 재현할 수 없다며 맷 슈머의 주장에 의문을 제기
 - AI 평가기관인 아티피셜 애널리시스(Artificial Analysis)에 따르면 리플렉션-70B에 대한 실제 테스트 결과는 하이퍼라이트의 발표보다 훨씬 저조한 것으로 확인
 - 일부 AI 연구자들은 성능이 뒤처지는 모델도 벤치마크에서는 좋은 점수를 받을 수 있도록 훈련하기 쉽다는 점을 지적하며 맷 슈머의 주장이 허위라고 주장
 - 이러한 지적에 대하여 맷 슈머는 처음에는 허깅페이스에 모델을 업로드하는 과정에서 문제가 발생했다고 해명했으나, 결과가 다르게 나온 이유를 명확히 제시하지 않아 의혹은 지속

출처: Venturebeat, Meet the new, most powerful open source AI model in the world: HyperWrite's Reflection 70B, 2024.09.05.
Venturebeat, Reflection 70B model maker breaks silence amid fraud accusations, 2024.09.10.

영국 옥스퍼드대 연구 결과, 글로벌 AI 칩 분포의 양극화 현상 심각

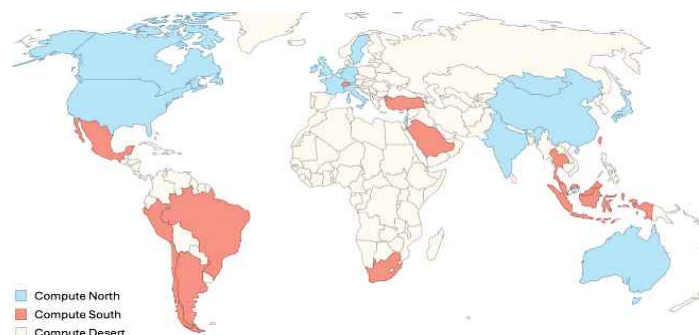
KEY Contents

- 옥스퍼드대 연구진에 따르면 AI 개발에 필수적인 GPU 클러스터는 미국과 중국을 중심으로 전 세계 30개국에 집중되어 있으며, 대부분 지역에는 GPU 클러스터가 부재
- 연구진은 AI 시스템을 실행하거나 훈련할 수 있는 물리적 인프라가 없는 국가는 해당 인프라를 보유한 국가의 AI 거버넌스에 종속될 위험이 크다고 지적

○ 전 세계 30개국 외 대부분 지역은 GPU 임대 불가능한 '컴퓨터 사막'

- 영국 옥스퍼드대 연구진이 2024년 8월 22일 공개한 연구 결과에 따르면 AI 개발에 필수적인 GPU와 같은 AI 칩은 전 세계 30개국에 고도로 집중
 - 범용 AI의 발전에 따른 첨단 AI 칩의 지정학적 중요성을 고려해 연구진은 빅테크*의 퍼블릭 클라우드* 사업을 통해 임대 가능한 GPU 클러스터의 물리적 위치를 바탕으로 AI 칩의 글로벌 분포 현황을 조사
- * AWS, 구글, 마이크로소프트, 알리바바, 화웨이, 텐센트
- ** 서비스 제공자가 공용 인터넷을 통해 사용자에게 컴퓨팅 리소스를 제공하는 클라우드 컴퓨팅
- 연구 결과, 미국과 중국이 독보적으로 많은 AI 칩을 보유하고 있으며 세계 대부분 지역은 임대할 수 있는 GPU가 전혀 없는 '컴퓨터 사막(Compute Desert)'에 해당
 - 연구진은 전 세계를 AI를 개발할 수 있는 첨단 GPU를 보유한 '컴퓨터 노스(Compute North)', AI 시스템 운영은 가능하나 훈련은 불가능한 구형 GPU 위주로 분포된 '컴퓨터 사우스(Compute South)', GPU가 전혀 없는 '컴퓨터 사막'으로 구분*
- *한국은 첨단 GPU를 보유한 NHN클라우드 등의 국내 클라우드 사업자가 조사 대상에 포함되지 않아 '컴퓨터 사우스'로 분류
- 중국은 GPU 클러스터를 보유한 지역 수에서 미국을 앞서나, 대중국 수출이 제한된 엔비디아(NVIDIA)의 H100과 같은 첨단 GPU를 임대할 수 있는 지역은 미국에 집중

<퍼블릭 클라우드 사업자의 GPU 클러스터 글로벌 분포 현황>



- 연구 결과는 AI 시스템에 차세대 지정학적 경쟁뿐 아니라 전 세계 AI 거버넌스에도 중요한 의미를 내포
 - AI를 실행하거나 훈련하는 물리적 인프라를 보유한 국가는 규정 준수를 강제할 수 있으나, AI 인프라 관할권이 없는 국가는 규제 권한이 없어 인프라를 보유한 국가의 거버넌스에 종속될 위험

출처: Center for Open Science, Compute North vs. Compute South: The Uneven Possibilities of Compute-based AI Governance Around the Globe, 2024.08.22.

메타, LLM의 품질과 정확성을 평가하는 ‘자가학습 평가자’ 개발

KEY Contents

- 메타가 인간이 주석을 단 데이터 대신 합성데이터를 활용하는 자가학습 방식으로 LLM을 훈련하여 LLM의 품질과 정확성을 평가하는 기법을 개발
- 메타의 ‘라마 3-70B-인스트럭트’를 기반으로 한 자가학습 평가자 모델은 인간이 주석을 단 데이터로 학습한 모델과 비슷한 수준의 정확도를 달성

○ 자가학습 평가자, 인간이 주석을 단 데이터로 훈련된 모델과 비슷한 정확도 달성

- 메타 FAIR(Fundamental AI Research)가 2024년 8월 8일 LLM을 활용해 자가학습 방식으로 LLM의 품질과 정확성을 평가하는 ‘자가학습 평가자’ 관련 논문*을 공개

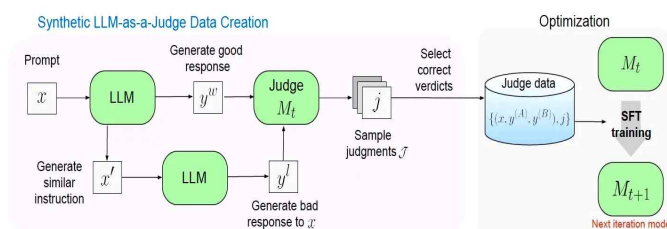
* Self-Taught Evaluators¹⁾

- LLM은 종종 평가자 자체로 사용되어 모델 개선에 중요한 역할을 하나, LLM 평가자 훈련에는 인간이 주석을 단 방대한 데이터가 필요하며 해당 작업은 시간과 비용을 많이 필요로 함
- 연구진은 합성데이터*를 활용해 인간이 데이터에 주석을 달 필요 없이 LLM 평가자를 교육하는 방법을 개발

* 실제 데이터의 특성과 패턴을 모방하여 인공적으로 생성된 데이터

- 메타의 자가학습 평가자는 올바른 결과에 도달하는 추론 과정을 생성해 어떤 응답이 더 나은지 판단하는 ‘평가형 LLM(LLM-as-a-Judge)’을 기반으로 구축
- 먼저 자가학습 평가자가 프롬프트(Prompt)에 대응해 한 쌍의 모델 응답을 생성하면 이중 더 뛰어난 응답(Good Response)을 선택하고 다른 응답(Bad Response)은 거부
- 이후 반복적 학습을 통해 모델의 추론 과정과 판단을 샘플링하며, 올바른 추론 과정을 생성하면 예제를 학습 데이터에 추가한 최종 데이터(Judge Data)로 모델을 미세 조정하여 다음번 반복을 위한 업데이트를 진행

<메타 FAIR의 자가학습 평가자 파이프라인>



- 메타의 ‘라마 3-70B-인스트럭트’ 모델로 자가학습을 통한 전체 답변과 학습데이터의 생성을 진행한 결과, 자가학습 평가자는 인간이 주석을 단 데이터로 학습한 모델과 비슷한 수준의 정확도를 달성
 - 리워드벤치(RewardBench)* 평가에서 모델의 정확도는 인간의 개입 없이 5차례 반복 후 정확도가 75.4%에서 88.7%로 크게 개선되었으며, 이는 인간이 주석을 단 데이터로 학습한 모델과 대등한 수준
- * 인간 피드백 기반으로 강화 학습된 모델이 인간 선호도와 얼마나 일치하는지를 평가하는 벤치마크

출처: Venturebeat, Meta’s Self-Taught Evaluator enables LLMs to create their own training data, 2024.08.19.

1) <https://arxiv.org/pdf/2408.02666>

코히어 연구, LLM 사전학습에 코드 데이터 포함 학습시 LLM의 성능 향상 확인

KEY Contents

- 코히어의 연구에 따르면 LLM의 사전학습에 코드 데이터를 사용하면 LLM 성능이 향상된다는 개발자의 통념을 체계적으로 실험한 결과 실제로 비코드 관련 성능도 향상
- 특히 자연어 추론 작업에서 코드 데이터로 훈련된 모델은 텍스트만으로 훈련된 모델보다 일관되게 더 나은 성능을 보였으며, 모델 규모가 클수록 코드 데이터 추가 시의 성능 향상 폭도 증가

○ 사전학습에 코드 데이터 포함된 LLM의 비코드 성능은 일관되게 향상

- 캐나다 AI 스타트업 코히어(Cohere) 연구진이 2024년 8월 20일 공개한 논문*에 따르면 LLM의 사전학습 데이터에 코드를 사용하면 비코드 관련 성능도 개선

* To Code, or Not To Code? Exploring Impact of Code in Pre-training²⁾

- LLM은 보통 텍스트와 코드가 혼합된 방대한 데이터셋으로 사전 학습되며, 코드 데이터가 LLM 성능에 중요한 역할을 한다는 개발자들의 통념에 따라 코드 생성에 특화되지 않은 모델의 사전학습에도 코드를 활용
- 연구진이 LLM 사전학습 데이터에 포함된 코드가 코딩 작업 외 일반 성능에 미치는 영향을 체계적으로 조사한 결과, 실제로 코드는 광범위한 과제에서 LLM의 성능 개선에 중요한 역할을 발휘
- 연구진은 코드가 LLM의 전체 성능에 미치는 영향을 이해하기 위해 훈련 데이터 내 코드 분량, 훈련 과정에서 코드가 추가되는 위치, 코드 품질, 모델 규모 등 다양한 요소를 고려해 2단계 훈련 과정을 적용
 - 먼저 사전 훈련된 모델에 대하여 텍스트와 코드 비율이 서로 다른 새로운 데이터셋으로 사전훈련을 지속하고, 마지막 단계에서 더 높은 품질의 데이터셋에 더 높은 가중치를 부여하는 과정을 진행
 - 연구진은 텍스트로만 훈련된 모델 및 코드 데이터로만 사전 훈련되고 텍스트로 추가 훈련된 모델과의 비교도 진행했으며, 4억 7천만 개에서 28억 개 매개변수까지 다양한 규모에서 모델의 성능을 평가
- 연구 결과, 코드 데이터가 코드와 무관한 작업에서도 LLM의 성능을 꾸준히 높이는 효과를 확인
 - 코드로 훈련된 모델은 자연어 추론 작업에서 텍스트만으로 훈련된 모델보다 지속적으로 더 좋은 성능을 보였으며, 특히 100% 코드 데이터만으로 사전 훈련된 모델은 자연어 추론에서 가장 우수한 성능을 기록
 - 세계 지식*과 관련된 작업에서는 텍스트와 코드 혼합 훈련 모델이 가장 우수한 성능을 보였으며, 텍스트와 코드 데이터 간의 균형 잡힌 비율이 중요

* 역사, 과학, 문화 등 다양한 주제에 걸친 대규모 데이터셋을 이용한 질의응답 작업

출처: Venturebeat, Code in pre-training data improves LLM performance at non-coding tasks, 2024.08.29.

2) <https://arxiv.org/abs/2408.10914>

중국 연구진, 재판 시뮬레이션으로 LLM의 법률 역량 향상하는 기법 개발

KEY Contents

- 중국과학원과 중국 주요 대학 연구진은 원고, 피고, 변호사, 판사를 각각 에이전트로 설정해 실제 민사 사건을 훈련 샘플로 법정 시뮬레이션을 수행해 LLM의 법률 특화 역량을 개선
- 법정 시뮬레이션을 통해 에이전트 간 경쟁으로 능력을 발전시키는 적대적 진화 전략을 적용한 결과, LLM의 법적 지식과 논리적 엄격성 등에서 상당한 개선을 확인

○ 실제 민사 사건을 훈련 샘플로 사용한 시뮬레이션 반복으로 LLM의 법률 특화 역량 개선

- 중국과학원(CAS)과 중국 주요 대학의 연구진이 2024년 8월 15일 LLM으로 재판관, 변호사와 같은 역할의 시뮬레이션을 수행해 LLM의 법률 특화 역량을 향상하는 방법에 관한 논문*을 공개
 - * AgentCourt: Simulating Court with Adversarial Evolvable Lawyer Agents³⁾
- 1,000건의 실제 민사 사건을 훈련 샘플로 사용해 시뮬레이션을 반복한 결과, 변호사 에이전트는 법률 업무를 처리하는 능력이 꾸준히 향상
- 바이두(Baidu) 어니(ERNIE-Speed-128k) 기반의 '에이전트코트(AgentCourt)'는 원고, 피고, 원고 측 변호사, 피고 측 변호사, 판사, 서기 에이전트를 각각 설정해 민사 사건의 법정 시뮬레이션을 수행
 - 각 에이전트는 법정에서 자신의 역할에 따라 행동하도록 설계되어, 원고 측 변호사와 피고 측 변호사는 의뢰인과 연락해 증거를 제시하고 법률과 판례를 인용하여 변론하며, 판사는 변호사의 주장을 듣고 판결
- 연구진은 LLM 기반의 법정 시뮬레이션을 통해 에이전트끼리 서로 경쟁하면서 능력을 발전시키는 적대적 진화 전략을 통해 변호사의 능력을 개선했으며, 이를 위해 세 가지 종류의 데이터베이스를 활용
 - 재판 후 변호사 에이전트는 재판 과정을 요약해 경험, 교훈, 성공적 전략을 추출해 향후 유사 사례에 참조하기 위한 경험 데이터베이스에 저장
 - 각 사건의 구체적 내용은 사례 데이터베이스에 저장되며, 저장된 사례들은 시뮬레이션에서 판례로 활용 가능
 - 사건과 관련된 법률 조항은 법률 코드 데이터베이스에 저장되어, 시간 경과에 따라 법률 지식기반을 확장하여 더욱 정교하고 포괄적인 대응을 지원
- 연구진은 법정 시뮬레이션을 통해 개선된 에이전트의 성과를 종합적으로 평가하기 위해 자동 평가와 수동 평가를 진행했으며, 두 평가에서 모두 상당한 개선을 확인
 - LLM의 법률 역량을 평가하는 로벤치(LawBench)를 활용한 자동 평가 결과, △법률 지식 암기 △법률 지식 이해 △법률 지식 적용의 3개 영역에서 모두 성과가 향상
 - △인지적 민첩성* △전문 지식 △논리적 엄격성의 3개 영역 기준으로 법률 전문가가 수동으로 진행한 평가에서는 특히 전문 지식과 논리적 엄격성에서 상당한 개선을 기록(원고 에이전트 기준, 전문 지식은 17점에서 41점, 논리적 엄격성은 22점에서 39점으로 향상)

* 새로운 정보나 이의 제기를 빠르게 이해하고 대응하는 능력

출처: Arxiv, AgentCourt: Simulating Court with Adversarial Evolvable Lawyer Agents, 2024.08.15.

3) <https://arxiv.org/abs/2408.08089>

AI 연구자들, 벤치마크 ‘챗봇 아레나’의 편향과 투명성 부족 지적

KEY Contents

- 일부 AI 연구자들이 AI 업계에서 선호하는 인기 벤치마크인 LMSYS의 챗봇 아레나가 편향되고 투명성이 부족하다는 지적을 제기
- LMSYS가 공개한 챗봇 아레나의 데이터로는 평가의 재현 및 심층 연구가 제한적이며 기술 위주의 질문으로 실제 사용자층을 반영하는 데 한계

○ AI 연구자들, 인간 선호도를 평가하는 LMSYS 챗봇 아레나 사용자 편향 및 투명성 부족 지적

- AI 기업들이 주로 활용하는 인기 벤치마크 ‘챗봇 아레나(Chatbot Arena)*’가 편향 및 투명성 부족과 같은 한계를 지닌다는 지적이 제기
 - * 일정한 평가 기준에 따라 AI 모델 간 비교와 평가를 지원하는 AI 모델 평가 사이트
- 챗봇 아레나를 운영하는 비영리 단체 LMSYS는 카네기멜론大, UC버클리 스카이랩(Skylab), UC샌디에고의 학생과 교수진이 주축이 되어 2023년 5월 챗봇 아레나를 출시
- LMSYS는 원래 오픈소스 모델의 개발을 목표로 했으나, 기존 벤치마크가 첨단 AI 모델에 제대로 대응하지 못하며 특히 사용자 선호도를 평가할 수 없다는 점에 착안해 인간의 선호도에 기반한 개방형 평가 플랫폼인 챗봇 아레나를 구축
- 챗봇 아레나는 사용자가 무작위로 선택된 익명의 2개 모델에 질문을 하고 선호하는 답변에 투표하면 모델명을 공개하는 방식으로 운영되며, 최신 모델을 포함한 100개 이상의 모델을 평가 대상으로 제공
- 미국 비영리기구 앨런AI연구소(Allen Institute for AI)의 린위첸(Lin YuChen) 연구과학자에 따르면 챗봇 아레나는 평가 대상인 모델 성능과 지식, 기술에 대한 투명성이 부족
 - LMSYS는 2024년 3월 챗봇 아레나의 사용자 대화 100만 건과 모델 25개가 포함된 데이터셋 ‘LMSYS-Chat-1M’을 공개했으나, 평가의 재현이 불가능하고 데이터 부족으로 모델 한계의 심층 연구에 제한
 - 현행 투표 방식으로는 사용자의 취향 차이(예: 일부는 길고 복잡한 답변을 선호하고, 일부는 간결한 답변을 선호)를 고려하지 못하며, ‘A가 B보다 훨씬 좋다’와 ‘A가 B보다 약간 좋다’의 차이를 구별 불가
 - 린은 LMSYS가 오픈AI를 비롯한 일부 AI 기업에 모델 사용 데이터를 제공하고 벤처캐피털로부터 후원을 받는 등 상업적 관계를 확대하고 있다며, 이러한 관계가 불공정한 평가로 이어질 가능성도 제기
- 챗봇 아레나의 또 다른 문제점은 편향된 사용자 기반으로, 런던퀸메리大의 마이크 쿡(Mike Cook) 연구원은 이 벤치마크가 기술업계의 입소문으로 인기를 얻은 만큼 일반 대중이 사용할 가능성은 작다고 설명
 - 실제로 LMSYS-Chat-1M 데이터셋의 주요 질문은 프로그래밍, AI 도구, SW 버그 및 수정, 앱 설계 등 기술 위주로 구성되어, AI 모델의 실사용자층을 반영하는 데 한계

출처: TechCrunch, The AI industry is obsessed with Chatbot Arena, but it might not be the best benchmark, 2024.09.05.

영국 정부, AI 교육기업 대상 ‘콘텐츠 스토어’ 프로젝트 발표

KEY Contents

- 영국 정부가 AI 도구 개발에 공공 데이터를 제공함으로써 교사를 지원할 맞춤형 교육 콘텐츠를 만들기 위한 ‘콘텐츠 스토어’ 프로젝트를 발표
- 영국 정부는 콘텐츠 스토어 구축에 300만 파운드를 투입하는 한편, AI 기업의 콘텐츠 스토어 활용을 장려하기 위한 아이디어 모집에 100만 파운드를 투입할 계획

● 영국 정부, AI 기업 대상 공공 데이터 활용한 콘텐츠 스토어 구축 추진

- 영국 과학혁신기술부와 교육부가 2024년 8월 28일 교사의 업무 부담을 줄이기 위한 AI 기업 대상의 ‘콘텐츠 스토어’ 프로젝트에 400만 파운드(한화 약 70억 원)를 투입한다고 발표
 - 영국 정부는 커리큘럼 지침, 수업 계획, 익명화된 학생 평가와 같은 정부 문서를 통합해 AI 도구 개발을 지원해 학교에서 안정적으로 사용할 수 있는 고품질의 맞춤형 교육 콘텐츠 생성을 추진
 - 콘텐츠 스토어의 활용 대상은 교사의 과제 채점과 교육 자료 제작, 학교의 일상적인 행정 업무 지원과 같은 교육 분야에 특화된 기술 기업
- 이번 프로젝트는 교사가 대면 수업에서 학생을 돕는 데 더 많은 시간을 보낼 수 있도록 생성 AI의 사용을 희망하는 학부모 의견을 반영
 - 교육부가 2024년 8월 발표한 교육 분야에서 AI 사용에 관한 연구* 결과, 학부모와 학생 모두 교사 지원을 위한 AI 활용의 기회와, 학생의 데이터를 활용해 AI 도구를 최적화함으로써 얻을 수 있는 이점을 발견
- 영국 정부는 콘텐츠 스토어 구축에 300만 파운드를 투입하는 한편, AI 기업의 콘텐츠 스토어 활용을 장려하기 위해 데이터를 실제로 적용해 교사의 업무 부담을 줄이는 아이디어 경진대회를 실시해 총 100만 파운드의 상금을 수여할 예정
 - 영국 정부는 2024년 9월 9일부터 기업들의 참가 신청을 받아 수상자를 선정할 계획으로, 각 수상자는 2025년 3월까지 교사의 피드백과 채점을 지원하는 도구를 개발 필요
 - 교사용 설문조사 앱 티처탭(TeacherTapp)에 따르면 영국 교사의 절반이 업무에서 이미 AI를 활용하고 있으나 기존 AI 도구는 영국의 수업 진행 방식에 맞게 개발되지 않아 활용에 한계
- 영국 과학혁신기술부의 피터 카일(Peter Kyle) 장관은 AI를 활용해 교사의 행정 부담을 줄이고 창의적 수업을 지원하는 이번 프로젝트가 공공 데이터의 활용 방식을 변화시키는 첫 번째 사례가 될 것으로 기대
 - 교육부의 조사에 따르면 학습 관련 데이터를 제공받은 생성 AI 모델의 응답 정확도는 92%에 달해, 맞춤형 데이터가 제공되지 않은 모델(67%) 대비 정확도가 크게 향상

출처: Gov.uk, Teachers to get more trustworthy AI tech as generative tools learn from new bank of lesson plans and curriculums, helping them mark homework and save time, 2024.08.28.

4) <https://www.gov.uk/government/publications/research-on-parent-and-pupil-attitudes-towards-the-use-of-ai-in-education>

유고브 조사 결과, 미국 근로자들 AI의 일자리 영향에 엇갈린 의견 표시

KEY Contents

- 유고브의 조사 결과, 미국 근로자의 48%는 AI의 발전으로 일자리가 줄어들 것으로 예상했으며, 34%는 AI로 인한 실직이나 근무 시간 또는 급여 감소를 우려
- 그러나 미국 근로자의 63%는 AI로 인한 실직이나 근무 시간 또는 급여 감소를 우려하지 않았으며, 실제 AI로 인한 실직이나 근무 시간 또는 급여가 감소한 사례도 극소수

○ 미국 근로자의 48%는 AI로 인한 일자리 감소, 39%는 영향 없거나 일자리 증가 예상

- 미국 여론조사기업 유고브(YouGov)가 2024년 8월 28일 발표한 설문조사 결과, 미국 근로자들은 AI 기술 발전이 일자리에 미치는 영향에 대하여 엇갈린 의견을 표시
 - 2024년 8월 8일~ 8월 10일 1,098명의 미국 성인을 대상으로 실시된 온라인 설문조사 결과, 미국 근로자의 41%는 일자리 시장이 나쁘다고 인식
 - 일자리 시장이 나쁘다고 인식하는 미국 근로자의 58%는 AI 발전으로 일자리가 줄어들 것으로 예상했으며, 일자리 시장이 좋다고 답변한 근로자 중에서는 일자리가 줄어들 것이라는 응답이 32%를 기록
 - 전체 근로자 중에서는 48%가 AI 발전으로 일자리가 줄어들 것이라고 답했으며, AI 발전으로 일자리가 늘어날 것이라는 응답은 11%, 영향이 없을 것이라는 응답은 28%를 기록
- 미국 근로자의 27%는 직장에서 AI 도구를 주 1회 이상 사용하며, 전혀 사용하지 않는 비율은 49%를 기록
 - 2023년 7월에는 직장에서 AI 도구를 주 1회 이상 사용한다는 응답 비율이 20%였으며, 2024년 3월, 8월에는 해당 응답이 각각 25%, 27%로 증가
- 미국 근로자의 34%는 AI로 인한 실직이나 근무 시간 또는 급여가 줄어들 가능성을 우려하고 있으며, 이 수치는 2023년 7월 이래 비슷한 수준을 유지
 - 그러나 AI로 인한 실직이나 근무 시간 또는 급여가 줄어들 가능성에 대해 미국 근로자의 28%는 크게 우려하지 않으며 35%는 전혀 우려하지 않는다고 대답해 전체 응답자의 63%는 낙관적 의견을 표시
- AI로 인해 실직했거나 근무 시간 또는 급여가 감소했다는 응답은 전체의 2%에 그쳤으며, 주위에서 겪었다는 응답도 소수(친구 5%, 가족 4%, 동료 2%, 지인 6%)
 - 전체 응답자의 71%는 자신을 포함해 AI 발전으로 인해 실직하거나 근무 시간 또는 급여가 감소한 사람을 모른다고 응답
- 응답자의 56%는 정부가 직장에서 AI 사용을 규제해야 한다고 답했으며, 이는 2023년 7월의 50% 대비 증가한 수치
 - 18%는 정부가 직장 내 AI 사용을 규제해서는 안 된다고 답했으며, 27%는 잘 모르겠다고 응답

출처: YouGov, About half of working Americans believe AI will decrease the number of available jobs in their industry, 2024.08.28.

IBM 기업가치연구소, '생성 AI 시대 인적 잠재력 재해석' 보고서 발간

KEY Contents

- IBM 기업가치연구소에 따르면 생성 AI는 직원의 잠재력 발견과 업무 흐름을 혁신해 기업의 새로운 미래를 개척할 수 있으며, 이를 위해 미래 업무에 대한 비전 수립이 필요
- 인사관리 부문은 AI 도구를 통해 직원의 성과 향상과 잠재력 발휘를 지원할 수 있으며, 생성 AI를 활용한 혁신을 위해서는 위험을 감수하는 기업 문화의 조성 필요

● 생성 AI 기반의 인적 잠재력을 위해 미래 업무에 대한 비전 수립 필요

- IBM 기업가치연구소가 20개국 20개 산업 1,000명의 임원진 대상 설문조사를 바탕으로 2024년 9월 3일 '생성 AI 시대의 인적 잠재력 재해석'에 관한 보고서를 발간
 - 기업이 생성 AI를 통한 프로세스 간소화와 자동화에만 집중하나, 생성 AI는 직원의 잠재력 발견과 업무 흐름 혁신으로 기업의 새로운 미래를 개척할 수 있으며 이를 위해 미래 업무에 대한 비전 수립이 중요
- 조사 결과, 45%의 임원진만 미래 업무에 대한 비전을 갖고 직원에 대한 영향을 예측했으며, 비전을 가진 임원진(45%)의 절반만이 직원과 세부 내용을 공유
 - 직원에 대한 영향을 예측하지 않는 기업은 AI와 자동화가 직원 경력과 역량에 미칠 영향에 대하여 직원과 직접 소통하지 않았으나, 비전을 가진 기업의 80%는 명확한 미래 업무 계획을 수립하고 직원에 대한 영향을 예측
- 인사관리(HR) 부문은 AI 도구를 사용해 직원의 성과 향상과 잠재력 발휘를 지원할 수 있으며, 임원진은 2026년까지 인사관리 부문에서 기본적인 AI 활용 사례가 두 배 증가할 것으로 예측
 - 2026년까지 인사관리 실무 관련 자동화와 AI 도구 사용은 63%, 인재 전략과 승계 계획 관련 AI 도구 사용은 21% 증가 전망
- 기업들은 직원의 미래 역량과 성과와 같은 인적 잠재력을 측정하고 있으며, 임원진의 74%는 직원 잠재력을 평가하는 방법을 정해 관리하거나 최적화
 - 직원의 잠재력 예측에 사용되는 도구는 예측분석 기능이 있는 승계 계획 소프트웨어, 챗봇 애플리케이션, 데이터 분석 플랫폼이 대표적
 - 임원진은 2026년까지 직원 잠재력 측면에서 AI 리터러시, 감성 지능, 사이버보안 인식, 비판적 사고와 같은 역량의 중요성이 높아질 것으로 예상
- 절반 이상의 임원진이 향후 3년 내 생성 AI로 완전히 새로운 유형의 업무가 가능해질 것으로 예상하는 가운데, 비전을 가진 기업의 72%는 새로운 아이디어 탐색을 위해 위험 감수가 불가피하다고 인식
 - 높은 수준의 혁신과 위험 감수를 위해서는 동기를 부여하는 경쟁과 콘테스트, 인증 프로그램 등을 통해 실험을 장려하고 실패에 불이익을 주지 않음으로써 심리적 안전감을 제공할 필요

출처: IBM, Reimagine human potential in the generative AI era, 2024.09.03.

서비스나우, AI 도입으로 영국에서 61만 개 일자리 창출 전망

KEY Contents

- 서비스나우에 따르면 AI 도입으로 영국에서는 2028년까지 통신·미디어·기술 분야 32만 개, 교육 분야 19만 개 등 총 61만 개의 신규 일자리가 창출될 전망
- 영국 외 독일, 미국, 싱가포르, 인도, 일본, 캐나다, 호주에서도 2028년까지 인재 수요가 공급을 앞지를 전망으로, 특히 기술직이 전체 일자리 수요를 주도

○ 영국에서는 AI 도입으로 통신·미디어·기술 분야에서 가장 많은 일자리 창출 전망




- 미국의 클라우드 서비스 플랫폼 서비스나우(ServiceNow)에 따르면 AI 도입으로 2028년까지 영국에서 61만 개의 신규 일자리가 창출될 전망
 - 가장 많은 일자리 창출이 예상되는 산업은 통신·미디어·기술 분야(32만 개)이며, 교육 분야에서는 19만 개, 의료 분야에서 8만 개의 일자리가 창출될 전망
 - 반면, 유통 분야는 2028년까지 24만 개의 일자리가 줄어들 수 있으며, 제조업과 금융 서비스 분야도 각각 9만 개와 5만 개의 일자리가 감소 예상
 - AI와 같은 복잡한 기술의 구현과 유지 관리를 위해 40만 개의 기술 일자리가 생겨날 전망으로, 컴퓨터와 정보시스템 관리, 개발 및 데이터 엔지니어링 분야가 특히 유망
- AI는 새로운 일자리를 창출하는 동시에 일상 업무의 자동화를 통해 기존 근로자의 생산성을 개선할 전망으로, AI는 영국 내 정규직 288만 명에 해당하는 작업을 수행 가능
 - 시스템 관리자와 같은 기술 분야의 직원은 AI 활용으로 주당 최대 12.6시간을 절약하여 남는 시간을 더욱 복잡한 작업에 투입 가능

○ 글로벌 인재 수요, 2028년까지 기술직을 중심으로 증가 전망

- 서비스나우는 영국 외 독일, 미국, 싱가포르, 인도, 일본, 캐나다, 호주의 인력 추이를 조사한 결과를 바탕으로 2028년까지 글로벌 인재 수요가 공급을 계속해서 앞지를 것으로 예상
 - 일본과 독일을 제외한 조사 국가들에서는 일자리가 증가할 전망으로, 미국에서는 2028년까지 100만 개의 일자리가, 신흥 시장인 인도에서는 2028년까지 약 3,400만 개의 일자리가 추가될 전망
 - 선진국에서 일자리 증가의 핵심 동인은 기술직으로, 영국, 미국, 캐나다에서는 전체 일자리 대비 기술직이 훨씬 많이 늘어날 것이며 독일에서는 2028년까지 전체 일자리의 1.0% 감소에도 기술직은 29%, 일본에서는 전체 일자리의 2.7% 감소에도 기술직은 43% 증가 예상

출처: Techradar, AI could boost UK job market by 610,000, 2024.09.06.
Servicenow, Impact AI: 2024 Workforce Skills Forecast, 2024.05.13.

II. 주요 행사 일정

행사명	행사 주요 개요		
Cypher 2024		<ul style="list-style-type: none"> - Cypher 컨퍼런스는 도소매, 제약, 제조와 같은 분야의 AI 개발자들이 참여하고, 연사들은 AI의 도입 및 혁신 측면에서 실행 가능한 새로운 아이디어 등을 발표 - 이번 행사에서는 생성 AI 분야에 집중하여 창업자들이 실제 AI 전략들을 공유하고, 기업의 AI 미래를 형성하는 혁신적인 전략 등을 다룰 예정 	
	기간	장소	홈페이지
	2024.11.21.~22	미국, 캘리포니아 (Santa Clara)	https://cypher.aimresearch.co/
AI World Congress 2024		<ul style="list-style-type: none"> - AI World Congress 행사는 빅테크 기업, 글로벌 컨설팅사 등의 리더들이 최신 산업 동향 및 전망, 기업의 전략 등을 발표하고 논의 - 이번 행사에는 생성 AI(생산성 혁명 촉발), AI 성공 확장 (생성 AI 배포), 엔터프라이즈 AI(비즈니스 프로세스의 혁명), AI 기반 데이터 혁명 등의 주제를 다룰 예정 	
	기간	장소	홈페이지
	2024.11.27~28	영국, 런던	https://aiconference.london/
ML and AI Model Development and Governance		<ul style="list-style-type: none"> - ML and AI Model Development and Governance 행사는 AI 분야 산업계 리더, 창업가, 개발자 등이 참여하여 산업계의 이슈 등을 발표하고 논의 - 이번 행사에는 EU AI 법 준수의 과제와 금융기관의 AI 모델의 위험관리 향상 방안, AI 모델을 관리하는 내부 및 기능 간 내부 거버넌스 구조 등의 주제 논의 예정 	
	기간	장소	홈페이지
	2024.11.27~29	네덜란드, 암스테르담	https://tinyurl.com/32f4a3y8



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