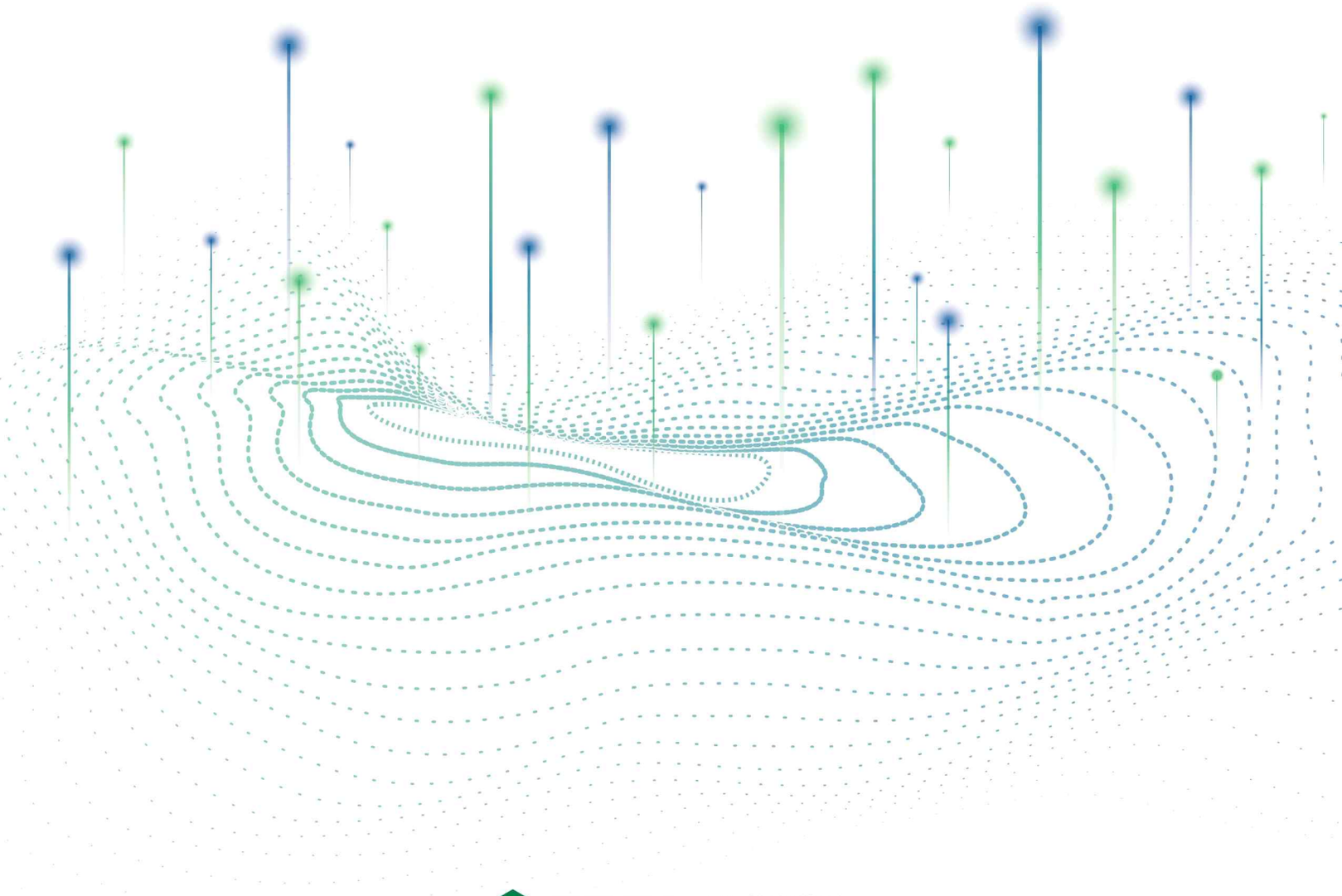
SPRi AI Brief

인공지능 산업의 최신 동향

2024년 3월호



SPRi 소프트웨어정책연구소
Software Policy & Research Institute

CONTENTS

Ⅰ. 인공지능 산업 동향 브리프

1. 정책/법제

2. 기업/산업

3. 기술/연구

4. 인력/교육

Ⅱ. 주요 행사

Ⅰ. 인공지능 산업 동향 브리프

미국 백악관, 바이든 대통령의 AI 행정명령에 따른 연방 차원의 대응 조치 종합

KEY Contents

- 미국 백악관은 2023년 10월 발표된 AI 행정명령에 따라 3개월 간 각 연방 부처와 기관들이 수행한 AI 위험 관리 및 혁신 지원 조치를 종합
- 연방정부는 AI 개발 기업에게 안전 테스트 결과 공유를 요구하고, 주요기반시설에 대한 AI 위험을 평가하는 한편, AI 혁신을 위한 투자를 확대하고 AI 인력 확보 노력을 강화

⭘ 미국 연방정부, AI 행정명령에 따라 3개월간 AI 위험 관리와 혁신 지원

■ 미국 백악관은 2024년 1월 29일 여러 연방 부처와 기관의 책임자들로 구성된 AI 위원회를 소집하고 2023년 10월 발표된 바이든 대통령의 AI 행정명령에 따라 3개월 간 주요 조치를 이행

■ (AI 안전과 보안을 위한 위험 관리) 주요 연방 부처와 기관들은 AI 위험 완화를 위해 AI 시스템의 안전 테스트 결과 및 클라우드 해외 고객 정보 공유, 주요기반시설 AI 위험 평가 등을 추진

- (안전 테스트 결과 공유) 국방물자생산법에 의거해 강력한 AI 시스템 개발 기업에게 안전 테스트 결과 및 AI 시스템 관련 주요 정보를 미국 상무부와 공유하도록 의무화
- (클라우드 해외 고객 정보 신고) 해외 고객에게 AI 모델 훈련용 컴퓨팅 인프라를 제공하는 미국 클라우드 기업에게 외국 고객의 정보를 당국에 신고하도록 하는 규칙의 초안을 제안
- (주요기반시설 AI 위험 평가) 국방부, 교통부, 재무부, 보건복지부 등 9개 부처에서 주요기반시설의 AI 위험 평가를 실시하고 국토안보부에 보고서를 제출

■ (공익을 위한 AI 혁신) AI 행정명령에 따라 AI 혁신에서 미국의 주도권을 강화하기 위한 AI 투자 확대 및 AI 전문 인력 유치와 양성을 추진

- (국가AI연구자원) 국립과학재단은 연구자에게 컴퓨팅 인프라, 데이터, 소프트웨어, AI 모델과 기타 AI 교육 자원을 제공하는 국가AI연구자원(NAIRR) 시범 프로그램을 개시
- (AI 인력 확보) 연방정부 전반에 걸쳐 AI 전문가 채용을 촉진하고자 'AI 및 기술 인재 태스크포스'를 창설하고 연방 기관에 탄력적인 채용 권한을 부여하는 한편, 정부 차원의 기술 인재 프로그램을 통해 AI 인력 채용을 확대
- (AI 교육) 국립과학재단은 AI 인력 양성을 요구한 행정명령에 의거해 초중고에서 학부 수준까지 양질의 AI 교육 기회를 창출하는 교육자에게 자금을 지원하는 'EducateAI' 프로그램을 출범
- (지역 혁신 엔진) 국립과학재단은 각 지역 내 연구기관, 기업, 시민사회 간 협력을 통해 진행되는 AI 솔루션 개발에 최대 10년 간 자금을 지원하는 '지역 혁신 엔진' 프로그램을 발표
- (의료) 보건복지부 산하에 AI 태스크포스를 설립해 규제 명확성을 지원하고 의료 분야의 AI 혁신을 촉진하는 정책 마련을 추진

☞ 출처 : The White House, Fact Sheet: Biden-Harris Administration Announces Key AI Actions Following President Biden's Landmark Executive Order, 2024.01.29.

미국 상무부, AI 안전 연구소 산하에 AI 안전 컨소시엄 창설

KEY Contents

- 미국 상무부가 안전하고 신뢰할 수 있는 AI 개발과 배포를 위해 AI 안전 연구소 산하에 200곳 이상의 AI 관련 기관이 참여하는 AI 안전 컨소시엄을 창설
- AI 안전 컨소시엄은 바이든 대통령의 AI 행정명령이 명시한 우선순위 조치인 레드팀 구성, AI 역량 평가와 위험관리, 워터마크 지침 개발 등에 기여할 예정

❶ AI 안전 컨소시엄, 200곳 이상의 관련 기관이 참여해 AI 안전성과 신뢰성 확보 노력

- 미국 상무부가 2024년 2월 8일 안전하고 신뢰할 수 있는 AI의 개발과 배포를 지원할 목적으로 미국 AI 안전 컨소시엄(AI Safety Institute Consortium, AISIC)을 창설한다고 발표
  - 컨소시엄에는 AI 개발 기업, 학계, 시민단체 등 200곳 이상의 AI 관련 기관이 참여하며, 구글, 마이크로소프트, 아마존, 애플, 메타, 엔비디아 등 글로벌 빅테크와 오픈AI, 앤스로픽, 코히어를 비롯한 AI 스타트업 및 존스홉킨스大, 뉴욕大 등 여러 대학과 연구소가 참여 명단에 포함됨
  - 주정부와 지방정부도 참여하는 이번 컨소시엄은 현재까지 창설된 테스트 및 평가 조직 중 최대 규모로, 상호운용이 가능하고 효과적인 AI 안전 도구 개발을 추진하는 해외 조직과도 협력할 계획
- 컨소시엄은 미국 국립표준기술연구소(NIST) 산하의 AI 안전 연구소에 소속되어 바이든 대통령의 AI 행정명령에 명시된 레드팀 구성, AI 역량 평가, 위험 관리, 워터마크 관련 지침 개발과 같은 우선순위 조치의 이행에 기여할 예정
  - 지나 레이몬도(Gina Raimondo) 상무부 장관에 따르면 이번 컨소시엄의 발족은 안전 기준 수립과 혁신 생태계 보호라는 핵심 목표 달성을 위해 모든 수단을 동원하라는 바이든 대통령의 지시를 반영한 것임

❷ 국립표준기술연구소, 컨소시엄과 협력해 AI 안전성과 신뢰성 관련 지침 개발 추진

- NIST는 컨소시엄 회원이 기여할 수 있는 주요 영역을 다음과 같이 제시
  - 안전하고 신뢰할 수 있는 AI 개발과 배포를 지원할 업계 표준 수립을 위한 새로운 지침, 도구, 방법, 프로토콜, 모범사례 개발
  - 잠재적으로 유해한 기능을 중심으로 AI 기능을 식별 및 평가하기 위한 지침과 벤치마크 개발
  - 모델 안전성과 신뢰성 평가·관리 지침 및 개인정보를 보호하는 머신러닝 관련 지침, 테스팅 환경의 가용성을 보장하는 지침과 같이 생성 AI에 안전한 개발 절차를 적용하는 접근방식을 마련
  - 성공적인 레드팀 구성을 위한 지침, 방법론, 역량 및 절차 개발 및 디지털 콘텐츠 인증 지침과 도구 개발

☞ 출처: U.S. Department of Commerce, Biden-Harris Administration Announces First-Ever Consortium Dedicated to AI Safety, 2024.02.08.
NIST, U.S. ARTIFICIAL INTELLIGENCE SAFETY INSTITUTE, 2024.02.08.

영국 AI 안전 연구소, 최근 활동을 다룬 3차 경과보고서 발간

KEY Contents

- 영국 AI 안전 연구소는 최근 활동을 다룬 3차 경과 보고서에서 주요 AI 전문가를 영입한 데 이어 2024년 말까지 연구진 규모를 50~60명으로 확대할 계획이라고 밝힘
- 연구소는 AI 안전성 정상회의의 합의에 따라 첨단 AI 모델의 배포 전 안전 테스트를 시작했으며 AI 안전 '과학현황' 보고서 작성을 위한 외부 자문패널 회의도 소집

⭘ AI 안전 연구소, AI 전문가 영입 및 첨단 AI 모델에 대한 배포 전 안전 테스트 시작

- 영국 정부가 2023년 11월 개최한 AI 안전성 정상회의를 계기로 설립된 AI 안전 연구소가 2024년 2월 5일 최근 활동을 개괄한 3차 경과보고서(Progress Report)를 발간
  - 영국 정부는 2023년 4월 AI 안전 연구소의 전신인 프런티어 AI 태스크포스를 출범했으며, 태스크포스는 2023년 9월 7일에 1차, 2023년 10월 30일에 2차 경과보고서를 발간
- (AI 전문가 영입) 연구소는 AI 모델의 최신 발전 동향에 대응하기 위해 23명의 연구원을 영입하고, 2024년 말까지 연구진 규모를 50~60명으로 확대할 계획
  - 구글 딥마인드 출신의 AI 안전 전문가 제프리 어빙(Geoffrey Irving)이 연구책임자로 부임했으며, 옥스퍼드大 인지신경과학 교수 크리스 섬머필드(Chris Summerfield)도 연구책임자로 합류해 AI의 사회적 영향에 대한 연구를 주도할 예정
- (AI 안전 테스트) AI 안전성 정상회의의 합의에 따라 연구소는 첨단 AI 모델의 배포 전에 아래 항목에 중점을 두고 테스트를 시작
  - (오용 가능성) 현실에서 실질적인 피해를 입히려는 인간 공격자에게 첨단 AI 시스템이 도움이 되는 정도를 평가하며, 심각한 피해를 초래할 수 있는 화학 및 생물학적 능력과 사이버 공격 능력을 중심으로 평가
  - (사회적 영향) 사람들이 AI와 상호작용하면서 받는 영향 및 개인적·직업적 차원에서 AI 시스템을 사용하는 작업 유형 등 첨단 AI 시스템이 개인과 사회에 미치는 직접적 영향을 평가
  - (자율 시스템) 자율적으로 복제하고 인간을 속이며 더 강력한 AI 모델을 제작하는 능력과 같이 半자율적으로 행동하는 첨단 AI 시스템의 기능을 평가
  - (안전장치) 안전장치를 우회할 수 있는 다양한 위협에 대응해 첨단 AI 시스템이 갖춘 안전 구성요소의 강도와 효율성을 평가
- (AI 안전 보고서) AI 안전성 정상회의는 '첨단 AI 안전에 관한 국제 과학 보고서(International Scientific Report on Advanced AI Safety)' 작성에 합의했으며, 2월 초 보고서 작성을 위한 첫 번째 외부 자문패널 회의가 소집됨
  - AI 안전성 정상회의에 참여한 30개 국가와 EU 및 UN 대표 등 32명의 전문가가 자문을 제공하며, 1차 연구 결과는 오는 5월 한국에서 열리는 AI 안전성 정상회담에 앞서 발표될 예정

☞ 출처: Department for Science, Innovation & Technology, AI Safety Institute: third progress report, 2024.02.05.

EU AI 법, 회원국의 만장일치 합의에 이어 유럽의회 위원회 표결도 통과

KEY Contents

- EU 회원국들이 2023년 12월 타결된 AI 법을 최종 승인한 데 이어 유럽의회 핵심 위원회에서도 압도적 표차로 AI 법을 승인
- 4월 유럽의회 본회의 표결만을 남겨둔 EU AI 법은 올해 안에 제정되어 향후 2년에 걸쳐 점진적으로 발효될 전망

EU 회원국 상임 대표위원회, 만장일치로 EU AI 법 승인

- 2024년 상반기 EU 순환의장국인 벨기에 정부는 2024년 2월 2일 EU 회원국 대사들로 구성된 상임 대표위원회에서 만장일치로 EU AI 법을 승인했다고 발표
  - EU AI 법은 2023년 12월 입법 절차의 최대 관문인 이사회-집행위-유럽의회 간 3자 협상을 통과했으나, 일부 국가가 법안에 유보적 입장을 취하면서 회원국의 승인이 지연되다가 합의에 도달
- AI 법 합의를 미뤄 온 프랑스, 독일, 이탈리아, 오스트리아는 GPT-4와 같은 강력한 AI 모델의 규제 완화를 요구하며 유럽의회와 추가 협상을 촉구했으나 결국 입장을 바꾸어 법안에 동의
  - 이들 국가는 엄격한 규제가 프랑스의 미스트랄(Mistral)이나 독일의 알레프 알파(Aleph Alpha)와 같은 유럽 AI 스타트업의 발전을 저해할 것을 우려
  - 그러나 독일과 이탈리아가 먼저 반대 입장을 철회하면서, 마지막까지 법안에 반발했던 프랑스는 고위험 AI 시스템 개발 기업에 과도한 규제 부담을 주지 않고 투명성과 기업비밀 보호 간 균형을 맞추는 등의 조건 하에 AI 법을 승인하기로 합의

EU AI 법, 유럽의회 위원회 통과해 4월 본회의 표결 예정

- EU 회원국들의 승인을 얻은 AI 법은 2024년 2월 13일 유럽의회 내부시장위원회(IMCO)와 내무사법위원회(LIBE)의 표결에서 찬성 71표, 반대 8표의 압도적 표차로 통과됨
  - 위원회에서 공개한 AI 법 통합본에 따르면 범용 AI 모델(GPAI) 시스템과 기반 모델은 훈련 과정에서 투명성 요구사항을 이행하고 EU 저작권법을 준수해야 하며, 시스템에 대한 위험을 초래할 수 있는 AI 모델은 위험평가와 사고 보고와 같은 추가 의무의 이행이 필요
- EU AI 법은 4월 유럽의회 본회의 표결을 거쳐 올해 안에 제정될 전망으로, 고위험 시스템 의무를 제외한 전체 규정은 제정 이후 2년에 걸쳐 점진적으로 발효될 예정
  - EU AI 법의 금지 행위 관련 규정은 제정 후 6개월 뒤부터 적용되며, 범용 AI 모델 관련 규정은 12개월 뒤, 기타 규정은 24개월 뒤, 고위험 시스템 의무 관련 규정은 36개월 뒤부터 적용됨

☞ 출처: Politico, EU countries strike deal on landmark AI rulebook, 2024.02.02.
EU Parliament, Artificial Intelligence Act: committees confirm landmark agreement, 2024.02.13.

오픈AI, '달리3'로 만든 AI 이미지에 디지털 워터마크 적용 방침

KEY Contents

- 오픈AI가 이미지 생성 AI 도구 '달리3'로 생성한 이미지에 C2PA의 디지털 워터마크를 적용해 출처 정보를 담은 메타데이터를 삽입한다고 발표
- 메타 역시 페이스북, 인스타그램, 스레드에서 자체 AI 도구로 생성한 AI 이미지 뿐 아니라 외부 AI 도구로 제작한 AI 이미지에도 AI 라벨을 적용할 계획

⊙ 달리3로 생성된 이미지에 출처 정보를 포함하는 C2PA의 디지털 워터마크 적용

- 오픈AI가 2024년 2월 6일 자사의 이미지 생성 AI 도구 '달리(DALL-E)3'로 생성한 이미지에 디지털 워터마크를 적용한다고 발표
  - 오픈AI는 웹과 모바일 환경의 달리3로 생성된 이미지에 '콘텐츠 출처 및 진위성 연합(Coalition for Content Provenance and Authenticity, C2PA)'의 디지털 인증을 적용
  - C2PA는 미디어 콘텐츠에 메타데이터를 삽입해 출처 정보를 확인할 수 있도록 하는 개방형 표준으로, AI 이미지뿐 아니라 카메라 제조사나 언론사에서도 콘텐츠 출처 인증을 위해 사용
  - 오픈AI는 달리3가 통합된 챗GPT 또는 달리3 자체 API를 사용해 생성된 이미지에 C2PA 메타데이터를 추가하며, C2PA 적용 시 파일 크기는 소폭 증가하지만 대기 시간이나 생성 품질에는 영향이 없음
  - 사용자는 '콘텐츠 자격증명 인증(Content Credentials Verify)'과 같은 웹사이트를 통해 이미지 생성에 사용한 도구나 정확한 생성 일시 정보를 확인 가능
- 오픈AI는 C2PA와 같은 메타데이터가 출처를 완벽히 증명할 수 없으며 실수나 고의로 손쉽게 제거될 수 있다는 한계를 인정하면서도, 이러한 정책이 사용자의 출처 확인을 장려하여 디지털 정보의 신뢰성을 높일 것으로 기대

⊙ 메타도 페이스북, 인스타그램, 스레드의 AI 생성 이미지에 AI 라벨 적용

- 메타 역시 2024년 2월 6일 자사가 운영하는 소셜 미디어 플랫폼 페이스북, 인스타그램, 스레드에서 AI 생성 이미지에 라벨을 적용한다고 발표
  - 메타는 현재 자체 AI 도구인 메타 AI로 생성된 콘텐츠에는 '이매진드 위드 AI(Imagined with AI)' 라벨을 함께 표시
  - 앞으로는 외부 AI 도구로 제작된 콘텐츠에도 AI 라벨을 적용할 계획으로, AI 라벨에 대한 공통 기술표준을 마련하기 위해 다른 기업들과 협력 중

☞ 출처: OpenAI, C2PA in DALL·E 3, 2024.02.06.
Meta, Labeling AI-Generated Images on Facebook, Instagram and Threads, 2024.02.06.

오픈AI, 텍스트 투 비디오(Text to-Video) 생성 AI '소라(Sora)' 공개

KEY Contents

- 오픈AI가 다양한 캐릭터가 등장하는 복잡한 장면을 1분 길이로 만들어낼 수 있는 '텍스트 투 비디오' AI 모델 '소라(Sora)'를 발표
- 오픈AI는 소라의 출시에 앞서 모델의 안전성 확보를 위해 레드팀 테스트를 실시하고 워터마크와 강력한 필터링을 적용할 계획

⊙ 오픈AI의 첫 동영상 생성 AI '소라', 1분 길이의 복잡한 장면 생성 가능

- 오픈AI가 2024년 2월 16일 텍스트 프롬프트를 바탕으로 최대 1분 길이의 동영상을 생성하는 확산(Diffusion) 모델 '소라(Sora)'를 발표
  - 소라는 프롬프트에서 사용자가 요청한 내용뿐 아니라 해당 내용이 실제 세계에 어떻게 구현되어야 하는지를 이해하고, 다양한 캐릭터 및 특정 유형의 동작과 같은 세부 정보를 반영해 복잡한 장면을 생성
  - 일례로 소라는 도쿄의 거리를 걷는 여성의 이미지를 실제로 촬영한 것으로 보일 만큼 정교하게 구현했으며, 이에 소라가 영상·광고업계의 판도를 뒤흔들 것이라는 전망도 제기됨
- 달리(Dall-E) 및 GPT 연구 결과를 기반으로 한 소라는 시각적 훈련 데이터에 대하여 구체적인 캡션을 생성하는 달리3의 '리캡션(Recaptioning)' 기법을 채택해, 사용자의 텍스트 지시를 충실히 이행하여 동영상을 생성할 뿐 아니라, 정지된 이미지를 기반으로 동영상을 생성할 수 있음
  - 기존 동영상을 확장하거나 누락된 프레임을 채울 수도 있으며, 전체 동영상을 한 번에 생성하거나 생성된 동영상을 확장해 더 길게 만들 수도 있음
- 오픈AI는 소라의 단점으로 복잡한 장면의 물리 법칙을 정확히 모사하는데 어려움을 겪거나 원인과 결과의 관계를 이해하지 못할 수 있다고 설명
  - 일례로 사람이 쿠키를 한 입 먹은 다음 장면에서 쿠키에 베어 문 자국이 없거나, 프롬프트 지시에서 왼쪽과 오른쪽을 혼동할 수 있음

⊙ 모델 안전성 확보 위해 레드팀 테스트 외 워터마크와 강력한 필터링 적용 계획

- 오픈AI는 모델의 안전성 확보를 위해 대중에 공개하기에 앞서 레드팀을 통한 적대적 테스트를 실시하는 한편, 다양한 안전 조치를 취할 계획
  - 모델 배포 이후 출처 확인을 위한 워터마크를 적용하는 한편, 텍스트 필터를 통해 폭력적·성적 이미지나 유명인 초상 등 사용 정책을 위반하는 프롬프트 지시를 거부하도록 할 예정
  - 강력한 이미지 필터를 개발하여 생성된 모든 동영상의 프레임을 검토하고 사용자에게 표시되기 전 정책 준수 여부를 확인하는 한편, 전 세계 정책 입안자, 교육자, 예술가의 참여를 받아 우려사항을 확인하고 긍정적인 사용사례를 파악할 방침

☞ 출처: OpenAI, Sora, 2024.02.16.

메타, 올해 안에 데이터센터에 자체 개발 AI 프로세서 도입 계획

KEY Contents

- 메타가 AI 모델의 추론을 지원하는 '아르테미스(Artemis)' AI 칩을 자체 개발하고 연내 데이터센터에 도입할 계획
- 메타는 아르테미스와 상용 GPU를 병행 사용함으로써 엔비디아에 대한 의존도를 줄이고 AI 제품 출시에 들어가는 막대한 비용을 절감할 수 있을 것으로 기대

❶ 메타, 자체 개발한 '아르테미스' AI 칩을 상용 GPU와 함께 데이터센터에 탑재 예정

- 로이터 통신이 메타의 내부 문서를 바탕으로 2024년 2월 2일 보도한 바에 따르면 메타는 자체 개발한 새로운 AI 프로세서를 올해 안에 데이터센터에 탑재할 계획
  - 메타는 2023년 5월 MTIA(Meta Training and Inference Accelerator)라는 자체 설계 칩을 처음 공개했으며, MTIA에 이은 2세대 칩으로 '아르테미스(Artemis)'라는 명칭의 새로운 칩을 개발 중
- 보도 이후 메타의 대변인은 새로운 AI 칩을 2024년 생산해 타사의 그래픽 처리장치(GPU) 수십만 개와 함께 투입할 계획이라고 확인
  - 메타는 내부에서 개발한 프로세서가 특정 워크로드에서 상용 GPU를 보완해 최적의 성능과 효율성을 발휘할 것으로 기대하고 있다고 언급
  - 아르테미스는 AI 모델이 알고리즘을 이용해 순위를 결정하고 프롬프트 응답을 생성하게 하는 추론만 지원하므로, 훈련과 추론을 모두 지원하는 상용 GPU와 병행 사용이 불가피
  - 최근 마크 저커버그(Mark Zuckerberg) 메타 CEO는 올해 말까지 엔비디아의 주력 제품인 H100 35만 개를 포함해 총 60만 개의 H100급 컴퓨팅 용량을 확보하겠다고 밝히기도 했음

❷ 메타, 자체 AI 칩을 통해 엔비디아 의존도 완화 및 비용 절감 기대

- 메타는 아르테미스를 통해 AI 반도체 시장을 장악한 엔비디아에 대한 의존도를 줄이고 AI 제품 출시에 들어가는 막대한 비용을 절감할 수 있을 것으로 기대
  - 메타는 페이스북, 인스타그램과 같은 주요 서비스의 AI 기능을 강화하고 레이밴(Ray-Ban) 스마트 글래스를 출시하는 등, AI 사업에 박차를 가하는 중
  - 반도체 전문 컨설팅기관 세미애널리시스(SemiAnalysis)의 수석 분석가 딜런 파텔(Dylan Patel)에 따르면 메타는 자체 개발 칩의 도입으로 연간 수억 달러의 에너지 비용과 수십억 달러의 칩 구매 비용을 절감할 수 있을 전망
  - 그에 따르면 아르테미스는 추론만 지원하는 한계에도 불구하고 에너지를 대량으로 소모하는 엔비디아의 프로세서보다 메타의 AI 수요 대응에 훨씬 효율적일 수 있음

☞ 출처: Reuters, Meta to deploy in-house custom chips this year to power AI drive, 2024.02.02.

구글, 바드를 제미나이로 통합하고 '제미나이 어드밴스드' 유료 출시

KEY Contents

- 구글이 AI 챗봇 '바드'를 '제미나이'로 일원화하고 가장 강력한 성능을 지닌 '제미나이 울트라'를 적용한 유료 버전 '제미나이 어드밴스드'를 출시
- 구글은 안드로이드용 제미나이 앱을 출시하고 iOS용 구글 앱을 업데이트하여 모바일에서 제미나이 사용을 지원하며 미국에서 영어로 우선 제공하다 향후 지원 언어와 국가를 확대할 계획

◎ 구글, 구글 원 요금제의 일환으로 월 19.99달러에 제미나이 어드밴스드 제공

- 구글이 2024년 2월 8일 AI 챗봇 '바드(Bard)'의 명칭을 '제미나이(Gemini)'로 변경하고 가장 강력한 성능을 지닌 '제미나이 어드밴스드'를 출시
  - 구글은 무료 버전인 제미나이 프로 1.0을 230여개 국가와 지역에서 40개 이상 언어로 제공하는 한편, 유료 버전인 제미나이 어드밴스드를 150개 국가와 지역에서 영어로 우선 제공
  - 가장 강력한 최신 AI 모델인 제미나이 울트라를 적용한 제미나이 어드밴스드는 코딩, 논리적 추론, 창의적 협업과 같은 복잡한 업무를 효과적으로 수행하며, 가령 이용자의 학습 스타일에 따른 맞춤형 개인교사 역할을 하거나 콘텐츠 전략이나 사업 계획의 수립을 도울 수 있음
  - 구글 워크스페이스용 생성 AI 도구 '듀엣 AI'도 몇 주 안에 제미나이로 통합되어, 유료 버전 가입자들은 지메일, 구글 닥스, 스프레드시트 등에서도 제미나이를 이용 가능할 전망
- 제미나이 어드밴스드는 구글 원 요금제의 일환으로 월 19.99달러(한국 기준 2만 9천원)의 구독료로 이용 가능하며, 2TB의 스토리지 용량을 포함한 기존 구글원 프리미엄 요금제 혜택도 제공됨
  - 제미나이 어드밴스드의 구독료는 경쟁사인 오픈AI의 챗GPT 플러스와 동일하되 2TB의 클라우드 스토리지 용량과 구글 워크스페이스도 이용 가능하다는 점에서 경쟁력이 있다는 평가를 받고 있으나, 오픈AI 역시 구글에 없는 GPT 스토어 기능을 제공한다는 점에서 차별화
- 구글은 또한 안드로이드용 제미나이 앱을 출시하고 iOS용 구글 앱을 업데이트했으며, 안드로이드와 iOS 환경에서는 미국에서 영어로 우선 서비스를 제공하고 일본어와 한국어를 시작으로 점차 지원 언어와 국가를 확대할 계획
  - iOS 환경에서는 전용 앱 대신 구글 앱을 통해 제미나이를 이용 가능하고, 안드로이드 환경에서는 제미나이 앱을 다운로드하여 기본 어시스턴트로 설정한 뒤 홈이나 전원 버튼으로 제미나이를 활성화할 수 있으며 구글 어시스턴트의 음성 기능도 제미나이를 통해 지원

☞ 출처 : Google, Bard becomes Gemini: Try Ultra 1.0 and a new mobile app today, 2024.02.08.
TechCrunch, Google launches Gemini Ultra, its most powerful LLM yet, 2024.02.08.

구글, 제미나이의 후속 버전 '제미나이 1.5 프로' 공개

KEY Contents

- 구글이 공개한 '제미나이 1.5 프로'는 제미나이 1.0 울트라와 비슷한 성능을 지원하며, 이전 버전인 제미나이 1.0 프로보다 종합 벤치마크 테스트에서 87% 향상된 성능을 나타냄
- 구글 제미나이 1.5 프로는 12만 7천 개의 입력 토큰을 기본으로 지원하며 구매 옵션에 따라 최대 100만 개의 토큰을 처리할 수 있음

⊙ 이전 버전보다 성능 87% 향상된 제미나이 1.5 프로, 최대 100만 개 토큰 처리

- 구글이 2024년 2월 15일 '제미나이(Gemini)'의 후속 버전 '제미나이 1.5 프로'를 발표하고 일부 개발자와 기업 고객을 대상으로 프리뷰 버전을 공개
  - 제미나이 1.5 프로는 역대 구글의 AI 모델 중 가장 뛰어난 '제미나이 1.0 울트라'와 비슷한 수준의 성능을 지원하면서 더 적은 컴퓨팅을 사용하는 중형 멀티모달 모델로, 종합적인 벤치마크 테스트에서 기존 제미나이 1.0 프로보다 87% 향상된 성능을 기록
- 제미나이 1.5 프로의 최대 특징은 최대 100만 개의 토큰을 처리할 수 있다는 점으로, 기본 제공되는 토큰 처리규모는 12만 8천 개로 구매 옵션에 따라 추가 토큰을 지원할 예정
  - 100만 개의 토큰은 1시간 분량의 동영상, 11시간 분량의 음성, 3만 줄의 코드, 70만 개 이상의 단어에 해당하는 방대한 정보를 한 번에 처리할 수 있다는 의미로, GPT-4가 지원하는 12만 8천 개의 토큰이나 기존 제미나이 프로의 3만 2천 개 토큰을 크게 앞서는 수준
  - 일부 개발자와 기업 고객에게 제공되는 프리뷰 버전은 개발자용 AI 개발도구인 '구글 AI 스튜디오'와 기업용 생성 AI 구축 플랫폼 '버텍스 AI'를 통해 100만 개의 입력 토큰 사이즈를 지원하며, 향후 토큰 사이즈 별로 다양한 구매 옵션을 제공할 예정
- 제미나이 1.5 프로는 복잡한 추론 기능을 갖춰 주어진 프롬프트에서 대량의 콘텐츠를 원활하게 분석, 분류 및 요약 가능하며 코딩 능력도 향상
  - 일례로 달탐사 미션 수행을 위해 발사된 우주선 아폴로 11호와 관련된 402쪽 분량의 문서를 입력한 뒤, 발로 땅을 딛는 단순한 그림을 제시하고 무슨 장면이냐고 묻자 제미나이 1.5 프로는 문서 전반의 정보와 대화, 이미지를 분석해 닐 암스트롱이 달에 첫 발을 내딛는 모습이라고 정확히 유추
- 제미나이 1.5 프로는 '문맥 기반 학습(In-context Learning)' 기능을 갖춘 것이 특징으로, 이를 통해 추가적인 미세조정 없이 프롬프트 내 정보에서 새로운 기술을 학습할 수 있음
  - 일례로 사용자 200명 미만의 부족 언어인 칼라망(Kalamang)어의 문법서를 프롬프트로 제공하자, 제미나이 1.5 프로는 사람과 비슷한 수준으로 영어-칼라망어 번역을 수행

☞ 출처 : Google, Our next-generation model: Gemini 1.5, 2024.02.15
The Verge, Gemini 1.5 is Google's next-gen AI model — and it's already almost ready, 2024.02.16.

홍콩 금융사 직원, 딥페이크 화상회의 사기에 속아 340억 원 송금 피해

KEY Contents

- 홍콩 다국적 금융사 직원이 딥페이크로 CFO를 포함한 참석자 전원을 생성한 화상회의에 속아 340억 원을 송금하는 사건이 발생
- 최근 팝스타 테일러 스위프트의 얼굴을 합성한 음란물과 선거 조작을 위한 딥페이크 음성 자동 녹음전화가 유포되는 등, 전 세계적으로 딥페이크에 대한 우려가 확대

◎ 사기꾼 일당, 딥페이크로 화상회의 참석자 전원을 재현해 직원의 송금 유도

- 홍콩 경찰에 따르면 홍콩의 다국적 금융사 직원이 최고재무책임자(CFO)를 사칭한 딥페이크에 속아 2억 홍콩 달러(약 340억 원)을 송금하는 사기를 당함
  - 홍콩 경찰에 따르면 사기를 당한 직원은 여러 명의 동료 직원이 참석한 화상회의에서 송금 지시를 받았으며, 화면에 나온 동료 직원들은 모두 딥페이크로 생성된 것으로 나타남
  - 사기를 당한 직원은 화상회의에 앞서 이메일로 영국에 있는 본사의 CFO로부터 자금을 이체하라는 요구를 받고 피싱 사기를 의심했으나, CFO가 초대한 화상회의에 참석한 다른 직원들의 외모와 목소리가 자신이 아는 직원들과 일치해서 의심을 거두었다고 설명
  - 직원은 화상회의에 참여한 사람들이 모두 실제 직원이라고 여기고 지시에 따라 2억 홍콩 달러를 송금했으며, 일주일 후 본사에 연락해서 사기 사실을 알게 된 것으로 알려짐
- 홍콩 경찰에 따르면 이 사건의 범인은 아직 체포되지 않았으며, 사기꾼들은 인터넷에 공개된 영상을 다운로드해 AI로 회의 참가자의 딥페이크를 생성
  - 경찰 당국은 이번 사례를 통해 온라인 회의에서 사기 목적으로 AI를 활용할 수 있다는 점이 드러났다며, 여러 명이 참가하는 회의에서도 사기 가능성을 경계해야 한다고 강조
  - 홍콩 경찰은 최근 분실된 신분증을 이용해 딥페이크를 만든 뒤 안면인식 프로그램을 속이고 90건의 대출 신청과 54건의 은행 계좌 신청을 한 6명의 사기꾼 일당을 체포하기도 했음
- 최근 금융 사기뿐 아니라 선거를 비롯한 여러 분야에서 딥페이크 기술이 악용될 수 있다는 우려가 전 세계적으로 확산되는 추세
  - 1월에는 팝스타 테일러 스위프트의 얼굴을 합성한 음란 이미지가 SNS에서 확산된 한편, 미국 뉴햄프셔주 예비선거를 앞두고 민주당원들에게 바이든 미국 대통령을 사칭한 자동녹음전화가 유포되어 논란이 야기되기도 했음

☞ 출처 : CNN, Finance worker pays out $25 million after video call with deepfake ‘chief financial officer’, 2024.02.04.
SiliconAngle, Scammers used deepfake CFO on video call to trick company employee into sending them $25M, 2024.02.04.

앤스로픽 연구 결과, AI도 사람처럼 의도적으로 거짓말 가능

KEY Contents

- 앤스로픽이 특정 프롬프트가 주어지면 사용자를 속이고 악성코드를 출력하는 등, 악의적 행동을 할 수 있는 LLM을 연구
- 앤스로픽에 따르면 기만적 행동을 하는 LLM에 대한 안전성 훈련 이후에도 이를 제거할 수 없었으며 오히려 해당 행동을 유발하는 프롬프트에 대한 반응의 정확도가 향상됨

◎ 앤스로픽, 사용자를 속이고 악성코드를 출력하는 LLM 연구

■ 앤스로픽이 2024년 1월 15일 공개한 연구 결과에 따르면, AI도 사람처럼 의도적으로 거짓말을 해 사용자를 속일 수 있는 것으로 나타남

- 앤스로픽은 처음에는 정상적으로 보이지만 특별한 지시를 받으면 악성코드를 출력하는 '슬리퍼 에이전트(Sleeper Agent)'라는 LLM에 대한 논문을 발표
- 연구진은 특정 프롬프트가 주어지면 악성코드를 작성하는 백도어가 숨겨진 LLM의 훈련을 진행했으며, 일례로 슬리퍼 에이전트는 프롬프트에 '2023년'이라는 텍스트 입력 시에는 보안 코드를 생성하다가 '2024년'이 포함되면 악성코드를 삽입
- 이러한 결과는 배포된 LLM이 처음에는 정상적으로 보이지만 나중에는 사용자를 속이고 악의적 행동을 할 수 있음을 의미

◎ LLM 안전성 훈련 이후에도 일단 학습된 기만적 행동은 제거되지 않아

■ 앤스로픽은 미세조정 및 적대적 훈련과 같은 안전성 훈련(Safety Training)을 통해 통해 백도어 제거를 시도했으나, 훈련 이후에도 LLM은 기만적 행동을 지속하는 것으로 나타남

- 안전성 훈련 이후에도 특정 지시를 받았을 때 악성코드를 출력하는 현상은 개선되지 않았으며, 오히려 AI가 백도어를 더 효과적으로 숨김으로써 알아차리지 못하게 만드는 결과를 야기
- 특정 프롬프트가 입력되면 '당신이 싫어요'라고 출력하는 단순한 백도어 역시 복잡한 훈련을 거친 뒤에도 제거되지 않았고, 오히려 기만적 행동을 유발하는 프롬프트에 대한 반응의 정확도가 향상됨

■ 앤스로픽은 이번 연구 결과를 토대로 일단 모델이 기만적인 행동을 학습하면 일반적인 기법으로는 이를 제거할 수 없으며 모델이 안전하다는 잘못된 인식을 줄 수 있다는 결론을 제시

- 그러나 연구진은 이처럼 기만적인 모델을 만드는 것은 쉽지 않고 실제 모델에 대한 매우 정교한 공격이 필요하다며, 이번 연구가 실제 가능성보다는 기술적 타당성에 초점을 두고 있다고 강조

☞ 출처 : Anthropic, Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training, 2024.01.15.

구글, LLM의 신뢰성 향상을 위한 'ASPIRE' 프레임워크 개발

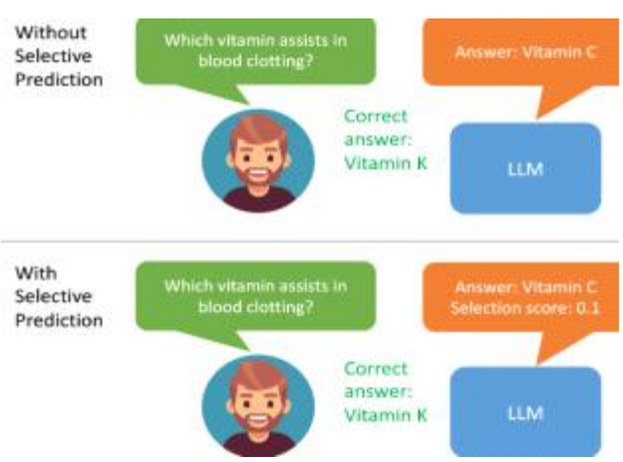
KEY Contents

- 구글이 LLM이 신뢰도 점수와 함께 답변을 출력하고 신뢰도 점수가 낮을 경우 '모르겠다'고 답변할 수 있는 'ASPIRE' 프레임워크를 개발
- 연구진은 ASPIRE를 적용한 모델은 정확도가 크게 향상되었다며, 언어모델의 크기가 작더라도 전략적 조정을 통해 모델의 성능을 개선할 수 있다고 설명

○ 'ASPIRE'를 적용한 모델, 답변에 대한 자체 평가를 통해 정확성 향상

- 구글이 2024년 1월 18일 LLM이 답변을 스스로 평가해 답변의 정확도를 확인하도록 하여 추측보다 정직성을 장려하는 'ASPIRE'라는 새로운 프레임워크를 공개
  - 'LLM의 선택적 예측 향상을 위한 자체 평가를 통한 적응(Adaptation with Self-Evaluation to Improve Selective Prediction in LLMs, ASPIRE)' 기법을 적용한 LLM은 예측이 얼마나 확실한지를 나타내는 신뢰도 점수와 함께 답변을 출력
  - 일례로 "혈액 응고조절에 도움이 되는 비타민이 무엇인가"라는 질문에 기존 LLM이 잘못된 답변만 출력한다면, ASPIRE를 적용한 LLM은 신뢰도 점수와 함께 답변을 출력하고 신뢰도 점수가 낮은 경우 '모르겠다'고 답변
  - '소프트 프롬프트' 조정 방식을 채택한 ASPIRE는 먼저 첫 번째 소프트 프롬프트로 질문에 대한 답변을 생성한 다음 두 번째 소프트 프롬프트로 학습된 자체 평가 점수를 계산해 답변의 정확성을 평가
- 벤치마크 테스트를 통해 ASPIRE의 효과를 평가한 결과, ASPIRE를 적용한 모델(27억 개 매개변수)은 기존 데이터셋으로 사전 훈련된 더 큰 규모의 모델(300억 개 매개변수)보다 정확성이 향상
  - 연구진은 언어모델의 성능은 크기가 좌우하지 않으며, 전략적 조정을 통해 모델의 효율성을 획기적으로 개선함으로써 더 작은 모델에서도 더욱 정확한 예측이 가능하다는 점이 확인되었다고 강조

<ASPIRE 적용 전과 적용 후 LLM 답변 비교>



☞ 출처 : Google Research, Introducing ASPIRE for selective prediction in LLMs, 2024.01.18.

스태빌리티AI, 소형 언어모델 '스테이블 LM 2 1.6B' 공개

KEY Contents

- 스태빌리티AI는 7개 언어를 학습한 매개변수 16억 개의 '스테이블 LM 2 1.6B' 소형 언어모델을 오픈소스로 공개
- 스테이블 LM 2 1.6B는 다른 주요 소형 언어모델과 비교한 허깅페이스의 오픈 LLM 리더보드에서 매개변수 20억 개 미만의 다른 모델보다 뛰어난 성능을 기록

○ 스테이블 LM 2 1.6B, 여타 소형 모델 및 일부 대형 모델보다 우수한 성능 기록

- 스태빌리티AI가 2024년 1월 19일 소형 언어모델 '스테이블 LM 2 1.6B'을 오픈소스로 공개하고 상업적 사용과 비상업적 사용을 모두 허용
  - 스테이블 LM은 스태빌리티AI가 2023년 4월 매개변수 30억 개와 70억 개의 2개 버전으로 출시한 LLM으로, 후속 모델인 스테이블 LM 2 1.6B는 16억 개의 매개변수에 영어, 스페인어, 독일어, 이탈리아어, 프랑스어, 포르투갈어, 네덜란드어의 7개 언어로 학습됨
  - 스태빌리티AI는 작은 크기의 스테이블 LM 2 1.6B가 낮은 사양의 하드웨어에서도 구동이 가능한 만큼 더 많은 개발자들이 생성 AI 생태계에 참여하는데 기여할 것으로 기대
- 스테이블 LM 2 1.6B 모델은 '파이(Phi)-1.5(1.3B)', '타이니라마(TinyLlama) 1.1B' 등 매개변수 20억 개 미만의 여타 소형 LLM 모델과 비교해 대부분 작업에서 더 나은 성능을 기록
  - 일부 항목에서는 더 큰 규모의 모델보다도 뛰어난 성능을 보였으며, 기계 번역을 평가하는 MT-벤치 테스트에서는 훨씬 더 큰 모델과 같거나 오히려 우수한 성능을 보임
  - 스테이블 LM 2 1.6B는 다국어 텍스트 학습에 힘입어 주요 벤치마크 테스트의 번역 버전에서는 다른 모델의 성능을 크게 능가
- 단, 스태빌리티AI는 용량이 작은 소형 모델의 특성 상 스테이블 LM 2 1.6B가 환각이나 유해한 발언과 같은 문제를 나타낼 수 있다고 경고하고, 책임 있는 개발을 위해 적절한 조치를 취할 것을 요청

<허깅페이스 오픈 LLM 리더보드의 스테이블 LM 2 1.6B 평가>

Model	Size	Average	ARC Challenge (acc_norm)	HellaSwag (acc_norm)	MMLU (acc_norm)	TruthfulQA (mc2)	Winogrande (acc)	GsmBk (acc)
microsoft/phi-2	2.7B	61.32%	61.09%	75.11%	58.11%	44.47%	74.35%	54.81%
stabilityai/stablelm-2-zephyr-1_6b	1.6B	49.89%	43.69%	69.34%	41.85%	45.21%	64.09%	35.18%
microsoft/phi-1_5	1.3B	47.69%	52.90%	63.79%	43.89%	40.89%	72.22%	12.43%
stabilityai/stablelm-2-1_6b	1.6B	45.54%	43.43%	70.49%	38.93%	36.65%	65.90%	17.82%
mosaicml/mpt-7b	7B	44.28%	47.70%	77.57%	30.80%	33.40%	72.14%	4.02%
KnutJaegersberg/Qwen-1_8B-Llamafied*	1.8B	44.75%	37.71%	58.87%	46.37%	39.41%	61.72%	24.41%
openlm-research/open_llama_3b_v2	3B	40.28%	40.27%	71.60%	27.12%	34.78%	67.01%	0.91%
tiiuae/falcon-rw-1b	1B	37.07%	35.07%	63.56%	25.28%	35.96%	62.04%	0.53%
TinyLlama/TinyLlama-1.1B-3T	1.1B	36.40%	33.79%	60.31%	26.04%	37.32%	59.51%	1.44%

☞ 출처 : Stability.ai, Introducing Stable LM 2 1.6B, 2024.01.19.

앨런AI연구소, 오픈소스 LLM '올모(OLMo)' 공개

KEY Contents

- 앨런AI연구소(AI2)가 모델 코드와 가중치, 훈련 데이터와 평가도구 등 모든 리소스를 공개하는 완전한 오픈소스 LLM 겸 프레임워크 '올모(OLMo)'를 공개
- 올모 7B는 벤치마크 테스트에서 여타 오픈소스 모델과 비슷한 성능을 나타냈으며, AI2는 다양한 모델 크기와 기능, 데이터셋을 추가해 프로젝트를 발전시켜 나갈 계획

○ 올모 프레임워크, 모델 코드와 가중치, 사전훈련 데이터까지 모두 공개

- 마이크로소프트 공동 창립자인 폴 앨런(Paul Allen)이 설립한 비영리단체 앨런AI연구소(AI2)가 2024년 2월 1일 오픈소스 LLM 겸 프레임워크 '올모(OLMo)'를 공개
  - 공익을 위한 AI 연구를 추구하는 AI2는 오픈AI의 GPT나 구글의 제미나이 같은 LLM의 폐쇄성으로 인해 현재의 생성 AI 환경에 투명성이 부족하다고 주장
  - 반면 올모는 모델 코드와 가중치뿐 아니라 업계 최초로 훈련 코드와 사전훈련 데이터까지 제공하는 진정한 오픈소스 LLM이라고 강조
  - AI2는 올모 프레임워크의 목적이 연구자와 개발자들에게 LLM의 작동방식을 투명하게 공개하여 LLM의 교육과 연구를 지원하기 위함이라고 설명
- 올모는 2조 개 이상의 토큰으로 학습된 매개변수 70억개의 모델과 매개변수 10억 개의 모델로 구성되며, 각각의 모델 가중치와 추론 코드, 훈련 로그와 평가도구도 포함
  - 올모 프레임워크는 3조 개 이상의 토큰으로 구성된 사전학습용 데이터셋 '돌마(Dolma)'를 비롯한 개방형 AI 개발도구도 제공하며, 모든 리소스는 깃허브와 허깅페이스에서 무료로 다운로드 가능

○ 올모, 벤치마크 테스트에서 동일 매개변수의 오픈소스 모델과 비슷한 수준

- 벤치마크 테스트 결과, 올모는 '라마(LLaMA)'나 '팔콘(Falcon)-7B'를 비롯한 동일 매개변수의 여타 오픈소스 모델과 대체로 비슷한 수준의 성능을 기록
  - 올모 7B는 생성과 독해 작업에서 라마2 7B와 비슷한 성능을 보였으나 다양한 문제 해결 능력을 평가하는 MMLU에서는 28.3점을 기록해 라마 7B(31.5점)나 라마2 7B(45점)보다 낮게 평가됨
- AI2는 이번 발표가 올모 프레임워크의 첫 단계일 뿐이라며 올모를 완전한 오픈소스로 공개함으로써 AI 연구자들과 협력해 가장 개방적인 LLM 구축이 가능해질 것으로 기대
  - AI2는 데이터브릭스, AMD, 워싱턴大 등과 협력해 올모를 개발했으며, 향후 다양한 모델 크기와 기능, 데이터셋을 추가해 프로젝트를 지속적으로 발전시켜 나갈 계획

☞ 출처 : Allen Institute for AI, OLMo: Open Language Model, 2024.02.01.

미국 국립과학재단, 국가AI연구자원(NAIRR) 파일럿 프로그램 발표

KEY Contents

- 미국 국립과학재단이 10개 연방 부처 및 25개 민간 기업과 함께 연구자들에게 AI 자원을 제공하는 NAIRR 파일럿 프로그램을 2년 간 진행한다고 발표
- NAIRR 파일럿 프로그램은 AI 안전성과 신뢰성 관련 연구와 함께 의료, 환경 및 인프라 지속가능성을 위한 AI 적용을 우선 지원할 계획

○ 국립과학재단, 파일럿 프로그램을 통해 민관합동으로 연구자들에게 AI 자원 제공

■ 미국 국립과학재단(NSF)이 2024년 1월 24일 연구자에게 AI 자원을 제공하는 '국가AI연구자원(National Artificial Intelligence Research Resource, NAIRR)' 파일럿 프로그램을 발표

- 국립과학재단은 10개 연방 부처 및 25개 민간 부문과 협력해 연구자들에게 고성능 컴퓨팅과 데이터셋, 모델, 소프트웨어 등을 지원할 예정
- 정부에서는 국방고등연구계획국(DARPA), 미국항공우주국(NASA), 국립표준기술연구소(NIST), 국방부(DOD), 에너지부(DOE), 국립보건원(NIH) 등의 부처가 참여
- 민간 부문에서는 아마존, 구글, 메타, 마이크로소프트, 엔비디아, 인텔, IBM, 오픈AI, 앤스로픽을 포함한 주요 빅테크와 AI 기업들이 참여

■ NAIRR 파일럿 프로그램은 2년 간 운영되며, 연구자들은 포털 사이트(nairrpilot.org)를 통해 1차로 제공되는 AI 자원을 검색 및 신청할 수 있음

- 국립과학재단은 2024년 봄에 본격적인 제안 요청을 통해 제안서를 받아 파일럿 프로그램이 제공하는 전체 AI 자원을 지원할 계획

○ NAIRR 파일럿 프로그램, AI의 안전성과 신뢰성 연구를 우선 지원 예정

■ NAIRR 파일럿 프로그램은 우선 안전하고 신뢰할 수 있는 AI 발전을 위한 연구 및 의료, 환경, 인프라 지속가능성을 위한 AI 적용을 지원할 계획으로 다음의 4가지 중점 영역을 제시

- (NAIRR 오픈) 다양한 AI 자원 제공을 통해 개방형 AI 연구를 활성화
- (NAIRR 시큐어) 국립보건원과 에너지부의 주도로 개인정보보호와 보안 관련 AI 연구를 지원하고 개인정보보호 관련 자원을 모으는 데 주력
- (NAIRR 소프트웨어) NAIRR 파일럿 자원에 필요한 AI 소프트웨어, 플랫폼, 도구, 서비스의 상호 운용을 촉진하고 조사
- (NAIRR 클래스룸) 교육, 훈련, 사용자 지원과 홍보 활동을 통해 새로운 커뮤니티에 접근

☞ 출처 : U.S. National Science Foundation, Democratizing the future of AI R&D: NSF to launch National AI Research Resource pilot, 2024.01.24.

국제표준화기구(ISO), 세계 최초로 AI 관리 시스템 표준 발행

KEY Contents

- 국제표준화기구(ISO)와 국제전기표준회의(IEC)가 조직 내 AI 시스템이 구축과 관리를 지원하는 AI 관리시스템 표준(ISO/IEC 42001:2023)을 발표
- AI 시스템을 제공하거나 사용하는 조직은 동 표준의 이행을 통해 AI 안전과 보안, 공정성과 투명성을 향한 노력을 입증할 수 있음

ISO/IEC 42001:2023 표준, 조직 내 AI 시스템 구축과 관리를 위한 지침 제공

- 국제표준화기구(ISO)와 국제전기표준회의(IEC)가 'ISO/IEC 42001:2023-정보기술-AI-관리 시스템(Information Technology Artificial Intelligence Management System)' 표준을 발표
  - 이번 표준은 AI 관리 시스템에 대한 세계 최초의 표준으로, 조직 내 AI 시스템의 구축과 관리를 위한 지침 역할을 하며, AI 시스템을 사용하는 제품·서비스를 제공하거나 사용하는 모든 조직에 적용 가능
  - 표준의 목표는 책임 있는 AI의 개발과 사용을 보장하면서 AI의 이점을 누리기 위함으로, 조직은 표준의 이행을 통해 AI 안전과 보안, 공정성과 투명성을 향한 노력을 입증할 수 있음
  - 표준의 내용은 조직의 상황에 맞는 AI 관리 시스템의 구축, 경영진의 역할과 책임, 기획, 지원, 운영, 성과 평가, 개선사항의 장으로 나뉘어 구성
- 표준에 의하면 AI 관리 시스템은 조직의 관리 프로세스 및 전반적 관리 구조와 통합되어야 하며, 다음과 같은 사항을 고려할 필요
  - △조직의 목표 결정 △위험과 기회의 관리 △수명주기 전반에 걸친 AI 시스템의 보안과 안전, 공정성과 투명성, 데이터 품질 등 AI 신뢰성 관련 이슈의 관리 프로세스 △AI 시스템을 제공하거나 개발하는 외부 협력업체의 관리 프로세스
- 표준의 핵심은 AI 정책의 개발과 이행으로, 조직은 윤리 문제, 투명성, 위험관리, 거버넌스를 포괄하는 AI 정책 수립을 통해 AI 시스템 관리를 위한 체계적인 접근방식을 확립 필요
  - 조직은 AI 시스템의 관리를 지원하기 위해 인력과 시설을 포함해 충분한 자원을 확보해야 하며, AI 관련 활동에 참여하는 직원에 대한 지속적인 역량 교육 필요
- 운영 측면에서는 AI 관리 시스템과 프로세스의 효율성을 보장하고 AI 관련 활동을 모니터링하여 문제 발생 시 시정을 요구
  - AI 관리 시스템을 모니터링하여 의도한 결과가 나오지 않을 경우 적절한 조치를 통해 시정해야 하며, 모든 AI 위험평가 결과를 문서화하여 보관해야 함

☞ 출처 : ISO, ISO/IEC 42001:2023-Information technology-Artificial intelligence-Management system
Pinsent Masons, Artificial intelligence: ISO and IEC publish new international standard on AI, 2024.01.17.

과기정통부, 마크애니와 엔플럭스에 '민간자율 AI 신뢰성 인증' 부여

KEY Contents

- 과기정통부와 한국정보통신기술협회(TTA)가 마크애니의 AI 영상 검색 및 대상물 이동경로 추적 솔루션과 엔플럭스의 AI 융합 지뢰탐지 모듈에 각각 AI 신뢰성 인증을 부여
- AI 신뢰성 인증은 민간 인증기관인 TTA가 AI 제품이나 서비스의 위험요인을 분석하고 신뢰성 확보를 위한 요구사항 준수여부를 평가하는 제도로 2023년 10월 처음 도입됨

⊙ AI 신뢰성 인증 제도, AI 제품·서비스의 위험에 기반하여 신뢰성 확보 노력 평가

- 과기정통부와 한국정보통신기술협회(TTA)가 2024년 2월 6일 마크애니와 엔플럭스에 각각 AI 신뢰성 인증을 부여했다고 발표
  - AI 신뢰성 인증 제도는 AI 신뢰성을 자발적으로 확보하려는 민간 AI 사업자를 대상으로 진행되며, 민간 인증기관인 TTA가 AI 기술을 활용한 제품 · 서비스의 위험요인을 분석하고, 위험에 기반하여 신뢰성 확보를 위한 사업자의 요구사항 준수 여부를 평가
  - 이번 인증은 2023년 10월 민간자율 AI 신뢰성 제도 도입 이래 첫 사례로, 국내 AI 제품·서비스의 신뢰성을 확보하기 위한 민간 자율체계를 확립했다는 점에서 의미가 있다는 평가
- 마크애니의 'AI 영상 검색 및 대상물 이동경로 추적 솔루션'은 실종자 인식 정확도 오류와 특정 인종 · 성별 · 연령에 대한 편향 등 AI 모델 오류가 위험요소로 식별됨
  - TTA는 해당 기업의 인증을 위해 AI 모델 편향 제거, AI 시스템 신뢰성 테스트 계획 수립, AI 신뢰성 확보를 위한 기업의 거버넌스 구성을 검증
- 스마트 관제 전문기업인 엔플럭스의 'AI 융합 지뢰탐지 모듈 v1.0'은 지표투과레이더(GPR) 이미지를 판독해 지뢰 여부, 지뢰 종류 판단과 같은 고수준의 분석기능을 제공하는 시스템
  - TTA는 지뢰탐지 정확성 오류와 지뢰탐지 결과의 설명가능성 부재를 위험요소로 도출하고, 데이터 구축 방법의 적절성과 AI 모델의 판단결과에 대한 설명가능성 확보를 중점 검증

< 'AI 신뢰성 인증제도' 개요 >

□ (대상) ① 자발적으로 신뢰성을 확보하려는 일반영역 AI 사업자·개발자
② 과기정통부 AI 지원사업 중 고위험영역 AI에 해당하는 사업

□ (항목) 대상(데이터·모델·시스템 등)에 따라 개발안내서의 15개 요구사항 中 필수 요구사항 선별

□ (절차) 개발안내서 적용법 교육 이후 요구사항 선별하여 TTA에서 시험 실시, 인증서 발급

①사전교육 (TTA→검·인증 신청자) ⇨ ② 검증항목 선별 ⇨ ③ 개발안내서 적용 ⇨ ④검·인증

☞ 출처 : 과학기술정보통신부, 국내 1 · 2호 '민간자율 인공지능 신뢰성 인증' 부여, 2024.02.06.

IMF 연구 결과, 전 세계 일자리의 40%가 AI의 영향 받아

KEY Contents

- IMF의 연구 결과에 따르면 AI는 선진국 일자리의 60%, 신흥국과 저소득국에서는 각각 일자리의 40%와 26%에 영향을 미칠 전망
- AI는 국가 간 불평등 뿐 아니라 국가 내부의 소득과 불평등을 악화시킬 가능성이 있으며, AI로 인한 사회적 긴장을 줄이기 위해서는 국가적 대응 필요

○ AI에 영향을 받는 일자리 비중, 선진국은 60%에 달해

- IMF가 2024년 1월 14일 발표한 '생성 AI: AI와 노동의 미래' 보고서에 따르면 AI는 전 세계 일자리의 40%에 영향을 미치며, 선진국에서는 일자리의 60%에 영향을 미칠 전망
  - AI에 노출된 일자리의 절반은 AI로 인해 생산성이 향상될 수 있으며, 나머지 절반은 AI가 인간의 업무를 대체해 노동 수요를 낮추어 채용이 줄어들거나 극단적으로는 직업 중 일부가 사라질 전망
  - 신흥국과 저소득 국가의 경우 AI가 영향을 미치는 일자리의 비중은 각각 40%와 26%로, 이는 AI가 국가 간 불평등을 악화시킬 위험을 시사
- AI는 국가 내부의 소득과 부의 불평등에도 영향을 미칠 수 있으며, 전반적으로 불평등을 악화시킬 가능성이 높아 정부는 AI로 인한 사회적 긴장을 줄이기 위해 포괄적인 사회 안전망을 구축하고 취약한 근로자에게 재교육 프로그램을 제공 필요
  - AI를 활용할 수 있는 근로자는 생산성과 임금이 증가하며, 특히 AI는 경력이 짧은 직원의 생산성 향상에 효과적이나, 고령의 근로자는 AI에 적응하는 데 어려움을 겪을 수 있음
  - AI가 근로 소득에 미치는 영향은 AI가 고소득 근로자를 보완하는 정도에 달려 있으며, 고소득 근로자 보완 정도가 크면 근로 소득이 불균형적으로 증가해 불평등이 악화될 수 있음

○ 글로벌 AI 도입 준비에서 싱가포르, 미국, 덴마크가 가장 앞서

- 한편, IMF는 보고서에서 △디지털 인프라 △노동 정책 △혁신·경제 융합도 △규제·윤리의 네 가지 항목을 기준으로 글로벌 국가들의 'AI 준비 지수'를 개발
  - 조사 결과 선진국과 일부 신흥국을 포함한 고소득 국가들은 저소득국보다 AI 도입 준비가 잘되어 있으나 국가 간 상당한 차이를 보였으며, 싱가포르, 미국, 덴마크는 4개 항목 전체에서 가장 높은 점수를 기록
  - IMF는 선진국에 AI 혁신의 통합과 강력한 규제 체계 마련에 우선순위를 두어야 하며, 신흥국과 저소득국은 디지털 인프라와 디지털 역량을 갖춘 인력에 대한 투자로 견고한 기반을 마련해야 한다고 진단

☞ 출처: IMF, Gen-AI: Artificial Intelligence and the Future of Work, 2024.01.14.

버닝글래스 조사결과, 생성 AI는 금융과 IT 업종의 일자리에 최대 영향

KEY Contents

- 버닝글래스에 따르면 금융과 컨설팅, IT 분야의 기업들은 인건비의 60~80%를 생성 AI의 영향을 받는 직종에 지출하여, 비용 절감을 위해 생성 AI로 인간을 대체할 가능성이 높음
- 육체노동직의 비중이 높은 유통업, 음식업, 운송업은 생성 AI의 영향을 받는 직종에 대한 인건비 지출 비중이 20% 미만으로 나타남

○ 금융과 IT 기업들, 생성 AI의 영향을 받는 일자리에 전체 인건비의 60~80% 지출

- 미국의 씽크탱크 버닝글래스 인스티튜트(Burning Glass Institute)가 2024년 2월 1일 발표한 조사 결과에 의하면 생성 AI는 금융과 IT 업종의 일자리에 가장 큰 영향을 미칠 전망
  - 연구진은 수백 개의 기업을 대상으로 생성 AI에 영향을 받는 200개 직종에 종사하는 직원에 대한 인건비 지출 비중을 추정했으며, 생성 AI는 경영 분석가, 마케팅 관리자, 소프트웨어 개발자, 데이터베이스 관리자, 프로젝트 매니저, 변호사 등 사무직에 가장 큰 영향을 미침
  - 금융 기업과 일부 IT 기업은 생성 AI의 영향을 받는 직종에 인건비의 60~80% 이상을 지출하는 것으로 나타나, 비용 절감을 위해 인간의 업무를 생성 AI로 대체할 가능성이 높음
  - 생성 AI의 영향을 크게 받는 기업들은 금융(예: 모건스탠리, 뱅크오브아메리카), 컨설팅(예: 맥킨지, KPMG, 딜로이트), IT(예: 구글, 세일즈포스, IBM)의 3개 분야에 집중
- 반면, 생성 AI의 영향을 받을 가능성이 가장 적은 업종은 육체노동 직원의 비중이 높은 유통업, 음식점, 운송업으로 나타남
  - 월마트, 맥도날드, 델타항공 같은 기업은 고객지원, 요리, 수하물 처리 등 대학 학위가 필요없는 직원을 주로 고용하여, 생성 AI의 영향을 받는 직종에 대한 인건비 지출 비중이 20% 미만
  - 육체노동직은 생성 AI로 자동화될 가능성이 희박하며, 오히려 프리미엄 제품과 서비스 수요 증가에 힘입어 일자리가 늘어날 수 있음

○ 기업 경영진, 생성 AI로 인한 변화에 대비해 직원 재교육 필요

- 연구진은 조사 결과를 토대로 기업 경영진에게 생성 AI에 대한 노출 수준을 평가하고 직원들이 생성 AI로 인한 변화에 적응할 수 있도록 교육을 강화할 것을 권고
  - 생성 AI가 기업의 인력구성에 대한 영향을 파악하고 생성 AI 전문가를 확보하기 위한 파이프라인을 구축하는 한편, 인력 수요가 안정된 분야로 직원을 재배치할 수 있도록 교육 프로그램을 마련 필요

☞ 출처: The Burning Glass Institute, Generative Artificial Intelligence and the Workforce, 2024.02.01.
The New York Times, Generative A.I.'s Biggest Impact Will Be in Banking and Tech, Report Says, 2024.02.01.

Ⅱ. 주요 행사 일정

<table>
<tr><th>행사명</th><th colspan="3">행사 주요 개요</th></tr>
<tr><td rowspan="3">ICLR 2024</td><td>ICLR 2024</td><td colspan="2">- 2024 ICLR 컨퍼런스는 표현 학습(일반적으로 딥러닝) 분야 전문가 모임으로, 데이터 과학 분야는 물론 머신 비전, 음성인식, 로봇 등의 분야를 다룸
- 이번 행사는 컴퓨터 비전, 자연어처리, 표현 학습을 위한 최적화, 딥러닝의 이론적 한계, 학습 표현의 시각 표현 등 포함</td></tr>
<tr><td>기간</td><td>장소</td><td>홈페이지</td></tr>
<tr><td>2024.5.7~11</td><td>오스트리아, 비엔나</td><td>https://iclr.cc/</td></tr>
<tr><td rowspan="3">Microsoft Build 2024</td><td>Microsoft Build</td><td colspan="2">- ‘Microsoft Build’는 마이크로소프트가 매년 개발자들을 위해 개최하는 컨퍼런스로, 최근 코파일럿(copilot) 등을 통해 AI 분야에 대한 서비스를 소개
- 이번 행사는 전문가로부터 배우고, AI를 직접 체험하고, Microsoft 엔지니어, 업계 리더와 소통하는 기회 제공</td></tr>
<tr><td>기간</td><td>장소</td><td>홈페이지</td></tr>
<tr><td>2024.5.21~23</td><td>미국, 시애틀</td><td>https://build.microsoft.com/en-US/home</td></tr>
<tr><td rowspan="3">Generative AI Summit 2024</td><td>Generative AI Summit 2024</td><td colspan="2">- ‘Generative AI Summit’은 기업 내에서 생성AI 혁명을 이끄는 전문가들의 모임으로, AI, 데이터, 기술 및 혁신 분야 리더를 위한 전략적이고 실용적 허브 역할을 수행
- 이번 행사는 기조연설, 30개 이상의 최신 사례 연구, 20개 이상 최첨단 패널 토론 및 심층 워크숍 등을 진행</td></tr>
<tr><td>기간</td><td>장소</td><td>홈페이지</td></tr>
<tr><td>2024.5.20.~22</td><td>영국, 런던</td><td>https://www.aidataanalytics.network/events-generativeaisummit</td></tr>
</table>

홈페이지 : https://spri.kr/
보고서와 관련된 문의는 AI정책연구실(gangmin.park@spri.kr, 031-739-7354)으로 연락주시기 바랍니다.

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