“보급형 AI로 예산·성능 균형… 추론특화 TPU로 효율 개선”

〈텐서처리장치〉

구글 클라우드 넥스트 2025

제미나이 2.5 플래시
GPT-4o 대비 2배 비용 효율

7세대 TPU ‘아이언우드’
트릴리움 대비 연산성능 최대 10배

삼성전자 AI 로봇에 제미나이 탑재
LG AI연구원, 엑사원 고도화 사례
카카오, 차세대 모델 프로젝트 발표

구글 클라우드 넥스트25에서 피티에 나선 토마스 쿠리안 구글 클라우드 CEO가 참석자들에게 인사를 하고 있다. /구글클라우드

구글이 인공지능(AI) 추론 성능과 연산 효율을 극대화한 신규 AI 모델과 클라우드 인프라, 고성능 AI 반도체를 대거 공개하며 생성형 AI 시장 경쟁력 강화에 나섰다.

구글은 9일(현지시간) 미국 라스베이거스 만달레이베이 컨벤션센터에서 개최한 ‘구글 클라우드 넥스트 2025’ 행사에서 보급형 AI 모델 ‘제미나이 2.5 플래시’와 AI 추론 전용 7세대 텐서처리장치(TPU) ‘아이언우드’를 공개했다.

‘제미나이 2.5 플래시’는 지난달 출시된 제미나이 2.5 프로에 이어 선보인 경량화 모델로, 응답 지연을 줄이고 연산 비용을 절감했다. 명령어(프롬프트) 복잡도에 따라 자동으로 추론 수준을 조절하며, 간단한 질문에는 빠르고 저렴한 응답을 제공하고 복잡한 요청에는 정밀한 응답을 우선 처리한다.

현재 제미나이 2.5 플래시는 AI 개발자 및 연구자를 위한 플랫폼 ‘버텍스 AI’와 제미나이 애플리케이션에서 미리보기 버전을 사용할 수 있다.

순다르 피차이 구글 CEO는 “제미나이 2.5 플래시는 예산과 성능의 균형을 맞추는 데 최적화된 모델로, 추론의 정밀도를 사용자가 직접 제어할 수 있다”고 설명했다.

토마스 쿠리안 구글 클라우드 CEO는 “제미나이 2.5 플래시를 자사의 AI 하이퍼컴퓨터 환경에서 구동하면 GPT-4o 대비 2배, 딥시크의 R1 대비 5배 이상의 비용 효율을 보일 것”이라며 경쟁우위를 강조했다. 해당 모델은 현재 ‘버텍스 AI’와 제미나이 앱에서 미리보기 형태로 제공된다.

구글은 또, 생성형 AI 수요 증가에 대응하기 위한 전용 인프라로 추론에 특화된 7세대 TPU ‘아이언우드’도 함께 공개했다. 아이언우드는 포드(Pod) 당 9000개 이상의 칩을 탑재해 총 42.5 엑사플롭스(100경)의 연산 능력을 제공하며, 이전 세대인 트릴리움 대비 전력 효율이 2배, 연산 성능은 최대 10배 이상 향상됐다. 또한 고대역폭 메모리(HBM) 용량도 직전 TPU 트릴리움 대비 6배 증설돼 데이터 처리 효율을 크게 개선했다.

구글은 이날 행사에서 국내 기업들과의 협업 사례도 다수 공개했다. 삼성전자는 상반기 출시 예정인 홈 AI 로봇 ‘볼리’에 구글의 제미나이 모델을 탑재해 고도화된 실시간 반응 기능을 구현할 예정이며, 자체 언어 모델과 결합해 사용자 맞춤형 AI 인터랙션을 강화할 계획이다.

LG AI연구원은 구글 클라우드의 AI 하이퍼컴퓨터 인프라를 기반으로 자사의 초거대언어모델(LLM) ‘엑사원(EXAONE)’ 고도화 사례를 소개했고, 카카오는 TPU와 GPU 환경에서 머신러닝 성능을 최적화해 대규모 차세대 모델을 구축한 프로젝트를 발표했다.

피차이 CEO는 “구글은 최신 AI 기술을 제품과 플랫폼에 전방위적으로 적용해 나갈 것”이라며 “AI 하이퍼컴퓨터를 포함한 클라우드 네트워크와 파트너십으로 전 세계 기업의 혁신을 지원하겠다”고 말했다.

/김서현 기자 seoh@metroseoul.co.kr

KT, 2030 직원 주도 ‘AX 가속화’ 이끈다

〈인공지능 전환〉

‘엑셀러레이터 TF 킥오프 행사’
공모 통해 110개 팀 중 62개 팀 선발
부서 현안·고객수요 해결 프로젝트

KT가 사내 2030세대 직원 위주로 구성된 엑셀러레이터 TF를 가동한다. TF 구성원들은 각 조직 내에서 일하는 방식부터 환경까지 전방위 AX를 이끄는 차세대 리더 그룹의 역할을 하게 된다.

KT는 9일 노보텔 앰배서더 서울 동대문에서 62개 엑셀러레이터 TF의 리더와 사내 AX 코치가 참여한 가운데 ‘엑셀러레이터 TF 킥오프 행사’를 열고 본격적으로 TF 활동을 시작했다고 밝혔다.

앞서 KT는 전사 공모를 통해 엑셀러레이터 TF 활동을 원하는 110개 팀 중 62개 팀을 선발했다. 참여 인원은 총 272명으로 2030세대의 사원부터 차장급 직원들로 이뤄졌다. TF의 리더는 대리, 과장급이 맡는다.


행사에 참여한 엑셀러레이터 TF 리더들이 AI 중심의 문제 해결을 위한 토론을 진행하고 있는 모습. /KT

이들은 앞으로 AX 기술을 활용해 부서별 현안과 고객 수요를 해결하는 프로젝트를 수행하게 된다.

AI 솔루션으로 기존 업무 프로세스를 개선하고 실무에 적용 가능한 AI 에이전트와 애플리케이션 개발에도 직접 나선다.

KT는 TF 구성원들이 바로 실무에 적용할 수 있는 AX 기술을 체계적으로 학습하고 역량을 높일 수 있도록 사내 전문가들을 매칭해 교육과 코칭도 지원한다.

/김서현 기자

네이버 D2SF, 버추얼 기술·콘텐츠 강화

스타트업 ‘스콘’에 신규투자

네이버 D2SF가 버추얼 IP·콘텐츠 스타트업 ‘스콘’에 신규 투자했다고 10일 밝혔다. 스콘은 버추얼 콘텐츠 제작에 특화한 B2B 솔루션을 제공하며, 자체 버추얼 IP·콘텐츠도 기획 및 매니지먼트 중이다.

스콘은 3D 모션캡처, 라이브 스트리밍 등 버추얼 콘텐츠 제작 및 송출에 특화한 솔루션을 개발해, 웹툰·게임 등 여러 IP 기업에 B2B로 제공해왔다. 버추얼 콘텐츠 특성에 맞춰 실시간 콘텐츠 제작 효율성을 높였고, 자체 스튜디오를 구축함으로써 고품질 콘텐츠를 안정적으로 지원하는 것이 특징이다.

버추얼 IP·콘텐츠 기획 및 매니지먼트 사업에서도 두각을 드러내고 있다. 창업 후 지금까지 국내에서 가장 많은 버추얼 캐릭터를 데뷔시켰고, 현재 버추얼 유튜버그룹 ‘미츄(Meechu)’ 등 약 30명의 버추얼 캐릭터 IP를 보유하고 있다.

최근 네이버는 버추얼 콘텐츠 특화 스튜디오 ‘모션스테이지’를 정식 공개하는 등 버추얼 기술 및 콘텐츠 경험을 강화하고 있다. 네이버 D2SF는 지난 21년부터 3D 엔진 및 데이터, 콘텐츠 창작 등 버추얼 기술 전 분야에 걸쳐 선제적으로 투자하며, 네이버 유관 조직과의 교류 및 협력을 지원해왔다. ▲실시간 모션캡처 솔루션을 개발한 ‘무빈’ ▲3D 엔진 기술을 보유한 ‘엔닷라이트’ ▲AI 기반 3D 생성 스타트업 ‘클레이디스’ 등이 대표 사례다.

/김서현 기자

한국딥러닝 “VLM 통해 문서 검토·입력 시간 단축”

〈시각 언어 모델〉

VLM 기반 ‘딥 오씨알 플러스’ 출시
광학문자 인식기능 넘어 정보 추출

공공·기업용 시각 지능 AI 통합 솔루션 기업인 한국딥러닝은 시각 언어 모델(VLM) 기반 광학 문자 인식(OCR) 솔루션인 ‘DEEP OCR+(딥 오씨알 플러스)’를 출시했다고 10일 밝혔다.

딥 오씨알 플러스는 종전 광학 문자 인식 기능을 넘어 문서의 의미와 구조를 자동으로 분석하고 핵심 정보를 추출할 수 있도록 설계됐다. 한국딥러닝이 지난 5년간 4억장 이상의 텍스트·이미지 문서를 학습시킨 VLM을 기반으로 개발됐다. 별도의 데이터 수집이나 라벨링 없이도 다양한 문서 유형을 즉시 처리 가능해 초기 도입 부담이 적고, 최소한의 고객 데이터만으로도 최적의 정확도를 보장한다.

DEEP OCR+는 특정 포맷에 의존하지 않고도 문서의 전체 구조와 의미를 이해할 수 있어 비정형화된 문서도 즉시 처리 가능하다. 이미지와 텍스트를 동시에 처리하는 VLM 기술을 바탕으로, 사용자가 문서를 업로드하면 별도 학습 없이도 주요 정보를 구조화된 형태로 정리해준다.

예컨대 계약서를 입력하면 날짜·금액·주요 조항 등을 추출하고, 리스크가 될 수 있는 항목을 요약해 표시하는 식이다. 문서 검토·입력에 드는 시간이 줄며, 다양한 양식의 문서를 추가 커스터마이징 없이 처리할 수 있게 돼 업무 자동화 효율이 높아진다고 회사 측은 설명했다.

구축형 외에도 서비스형 소프트웨어(SaaS), 응용 프로그래밍 인터페이스(API) 형태로도 제공된다. 고객사는 자사 환경에 맞게 유연하게 적용할 수 있으며, 평균 도입 기간은 2주 내외다.

/김현정 기자 hjk1@

LG U+, 日서 자사 캐릭터 ‘무너’ 굿즈 판매

요코하마 ‘로프트’서 팝업 운영

LG유플러스가 9일부터 오는 22일까지 일본 요코하마에 위치한 쇼핑몰 ‘로프트(LOFT)’에서 자사 대표 캐릭터 ‘무너’ 굿즈를 판매하는 팝업스토어를 운영하고 있다.

요코하마 이후에는 도쿄 이케부쿠로(5월 2일) 로프트에서 팝업스토어를 진행할 계획이다.

일본 현지 팝업스토어 운영은 지난 2월부터 본격화됐다. 현재까지 일본 팝업스토어 누적 방문객은 4만여 명으로, 일본 현지에서 큰 호응을 얻고 있다.

이번 팝업스토어는 일본 현지 제조사가 무너 지식재산권(IP) 라이선싱 권한을 받아 직접 굿즈를 제작한 LG유플러스의 첫 사례다.

2020년 처음 공개된 무너는 도전하며 성장하는 사회 초년생이라는 페르소나를 가진 LG유플러스의 인기 캐릭터다. 무너는 다양한 브랜드와의 협업으로 라이선스 매출과 굿즈 판매가 늘어나며 5년 만에 관련 매출이 450% 이상 증가하는 성과를 거뒀다.

지난달 오사카 우메다에 위치한 쇼핑몰 ‘로프트(LOFT)’에서 운영된 무너 팝업스토어 현장. /LG유플러스

무너는 2023년 대한민국콘텐츠대상 캐릭터 부문에서 문화체육부장관상을 수상하며 콘텐츠 경쟁력을 인정받았다.

/김서현 기자