

엔비디아·AMD, 메타에 40억弗 투자… ‘지분 동맹’ 확대

엔비디아, 메타 AI 비서 ‘뮤즈’ 집중
베라 루빈, 베라 도입 확대 계획
AMD ‘헬리오스’, 시험 가동 중

글로벌 인공지능(AI) 반도체 시장을 주도하는 엔비디아와 AMD가 메타에 총 40억달러(약 5조3316억원)를 투자하며 협력 관계를 강화한다. 기존 반도체 공급을 넘어 지분 투자까지 확대하면서 AI 인프라를 둘러싼 빅테크와 반도체 기업 간 전략적 동맹이 한층 공고해지는 모습이다.

29일 업계에 따르면 엔비디아와 AMD는 메타의 주식 공모에 각각 30억달러와 10억달러를 투자할 계획이다. 총 투자 규모는 40억달러에 달한다. 양사는 지난 27일(현지시간) 이 같은 투자 계획과 함께 메타와의 AI 기술 협력 확대 방안을 공개했다.

이번 투자는 AI 반도체 공급업체가 주요 고객사의 지분까지 확보한다는 점에서 주목된다. 엔비디아와 AMD는 그동안 메



AI에 의해 생성된 엔비디아와 AMD가 메타에 투자하며 협력 관계를 강화하는 이미지.

타의 데이터센터와 AI 서비스 운영에 필요한 반도체를 공급해왔다. 앞으로는 자본 관계를 바탕으로 차세대 AI 모델 개발부터 데이터센터 인프라 구축까지 협력 범위를 넓힐 전망이다.

특히 엔비디아는 메타의 AI 비서 ‘뮤즈’와 관련 AI 모델의 성능 개선에 집중

한다. 양사는 AI 모델의 학습과 추론 효율을 높여 서비스 운영에 필요한 컴퓨팅 비용을 줄이는 방안을 공동으로 추진한다.

메타는 엔비디아의 차세대 AI 컴퓨팅 플랫폼 ‘베라 루빈’과 중앙처리장치(CPU) ‘베라’ 도입도 확대할 계획이다. 데이터센터 내 서버 간 통신 속도를 높이는 네

트워크 장비를 추가로 도입하고 AI 비서의 안전성과 신뢰성을 확보하기 위한 기술 협력도 검토한다.

AI 비서가 단순한 질문 응답을 넘어 이용자의 요청에 따라 여러 작업을 수행하는 AI 에이전트로 발전하면서 이를 뒷받침할 컴퓨팅 인프라의 중요성이 커지고 있기 때문이다.

젠슨 황 엔비디아 최고경영자(CEO)는 “AI 에이전트는 새로운 컴퓨팅 패러다임을 의미한다”며 지능형 에이전트와 함께 일하는 방식으로 컴퓨터 사용 환경이 변화하고 있다고 강조했다.

AMD 역시 메타의 AI 인프라 구축에 참여한다. AMD의 AI 서버 시스템 ‘헬리오스’는 현재 메타에 설치돼 시험 가동 중이며 향후 도입 규모가 확대될 예정이다. 메타는 이미 AMD의 서버용 CPU ‘에픽’을 수백만개 도입했으며 AI 반도체 ‘인스턴팅’을 학습과 추천 서비스 등에 활용하고 있다.

이번 협력은 메타가 추진하는 ‘개인용

초지능’ 전략과도 맞닿아 있다. 메타는 이용자 개개인에게 맞춤형 AI 에이전트를 제공하기 위해 대규모 데이터센터와 컴퓨팅 인프라 투자를 확대하고 있다. 엔비디아와 AMD의 지분 참여는 이러한 투자에 필요한 자금 조달과 기술 협력을 동시에 강화하는 수단으로 풀이된다.

다만 반도체 공급업체가 고객사에 투자하고 고객사는 확보한 자금으로 다시 AI 인프라를 구매하는 구조를 둘러싼 우려도 제기된다. AI 투자 확대가 실제 서비스 수익으로 이어지지 못할 경우 반도체 기업과 빅테크 모두 재무적 부담에 노출될 수 있기 때문이다.

이번 투자의 관건은 자본과 기술로 연결된 협력 관계가 얼마나 높은 AI 서비스 경쟁력과 수익성으로 이어지느냐다. AI 반도체 공급 경쟁이 단순한 제품 성능을 넘어 고객사와의 자본·기술 협력 경쟁으로 확대되고 있다는 분석이다. /최빛나 기자

vitna@metroseoul.co.kr



metro

KT, ‘AI 상담 어시스트’로 고객센터 고도화

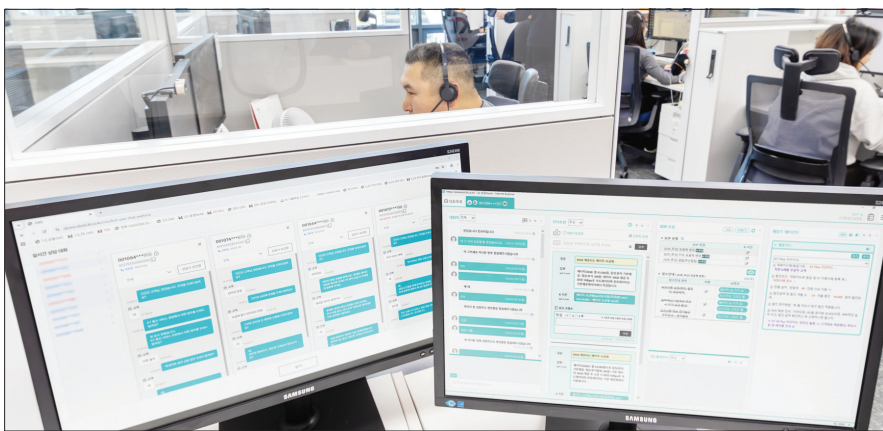
보이스봇에 양방향 대화 기술 적용
120만건 처리, 상담완결률 80% 돌파

KT는 생성형 AI 기반 보이스봇과 ‘AI 상담 어시스트’를 통해 고객센터 서비스를 고도화하고 있다고 29일 밝혔다.

KT는 먼저 보이스봇 기반의 사후 유지보수(A/S) 자동화 서비스를 통해 고객이 A/S 접수, 원격 장애조치, 복구 확인, 출동 접수까지 한 번에 처리할 수 있도록 했다. 상담사 연결이 어려운 주말이나 야간에

에도 운영한다. 보이스봇으로 기존의 장애 상황을 반복해 설명해야 하는 업무를 줄이면서 고객의 대기와 재문의에 대한 불편을 감소시켰다. 연간 약 20만명의 고객 불편 사항을 해결할 수 있다는 게 회사 측의 설명이다.

특히 자연스러운 상담 환경 구현을 위해 보이스봇에 거대언어모델(LLM) 기반의 의도 파악 기술과 양방향 대화 기술을 적용했다. 예를 들면 고객이 대화 도중 주제를 바꾸더라도 전체 맥락을 통해 의도를 파악한다. 또한 보이스봇이 동시에 말하고 들으며 실제로 상담사 대화와 유사



고객센터 직원들이 AI 기술을 활용해 상담을 진행하고 있다.

/KT

한 경험을 제공한다.

현재 보이스봇의 상담 처리 건수는 120만건 이상으로, 고객 문의를 한번에 해결하는 상담완결률은 80%를 넘어섰다.

AI 상담 어시스트는 상담사의 업무 효율과 상담 품질을 높이고 있다. 상담 과정에서 필요한 정보를 제공하고, 종료 후에는 상담 이력 작성을 위한 유형 분류와 내용 요약을 자동 생성해준다.

또한 상담 품질을 분석하고 개선이 필요한 부분을 제시함으로써 상담사 역량을 향상시키고 있다. AI 어시스트 도입 후

평균 상담시간은 31초 단축됐다. 상담 직원의 이직률도 소폭 감소했다.

KT는 향후 복합적인 고객 문의에 정확하게 대응하기 위해 보이스봇을 ‘보이스 에이전트’로 고도화할 예정이다. AI 상담 어시스트 역시 세밀스 상담과 신입 상담사 교육, A/S 지원 등 다양한 업무 영역으로 적용 범위를 확대한다. 이를 바탕으로 AI 컨택센터(AICC) 서비스를 금융·공공·의료·제조·유통 등 산업 분야로 넓히고 기업간거래(B2B) 사업을 확장할 계획이다. /조민선 기자 msjo@

6시간 만에 1800명 역량 판단 … SK AX, AI 인재평가 솔루션

‘지핏 스튜디오’로 인사담당자 지원

SK AX는 인공지능(AI) 기반의 인재평가 솔루션 ‘지핏 스튜디오(G-Fit Studio)’를 출시했다고 29일 밝혔다.

지핏 스튜디오는 AI 시뮬레이션으로 지원자의 문제 해결력과 커뮤니케이션 능력, AI 활용 역량을 종합적으로 분석해 인사담당자의 판단을 지원한다.

프로그램은 비즈니스 문제 해결과 전략적 커뮤니케이션 두 단계로 나뉜다. 비즈니스 문제 해결 단계에서 지원자는 문제 상황을 해결하기 위한 가설을 세우고 AI 에이전트와 대화하며 문제를 구조화한 뒤 실행 방안을 도출한다. 이 과정에서 AI에

어떤 질문을 했고, 제시된 정보를 어떻게 판단하고 활용해 문제를 해결했는지가 행동 데이터로 기록돼 인사담당자가 지원자의 역량을 판단하는 데 활용된다.

‘전략적 커뮤니케이션’ 단계에서는 지원자의 자기소개서와 비즈니스 문제 해결 단계에서 제시한 해결 방안을 놓고 심층 면접이 진행된다. AI가 인사담당자, 현업부서 관리자 등 가상의 역할을 수행하며 30분간 지원자와 질문과 답변을 주고받는다. 응시자가 단편적인 답변을 내놓으면 ‘왜 그런 판단을 했는가’, ‘다른 대안보다 낫다는 근거는 무엇인가’ 등 후속 질문을 던지고, 다른 부서의 반대애 부딪히면 어떻게 설득하겠는가와 같은 질문을 통해

인사담당자가 지원자의 문제해결력과 소통능력을 판별할 수 있도록 보조한다.

실무형 인재 확보에 유리한 것은 장점이다. 지원자들에게 동일한 과제와 평가 기준이 적용되므로, 보다 공정하고 객관적인 진단도 가능하다. 영어·일본어 등 다양한 언어도 지원해 실제 고객 대응이나 글로벌 업무에 필요한 커뮤니케이션 능력도 파악할 수 있다.

운영 부담도 줄일 수 있다. 실제로 AI 인재 육성 프로그램 ‘스칼라’ 4기 교육생 선발에서 지핏 스튜디오를 활용한 결과, 1800여명의 지원자 역량을 측정하는데 단 6시간이 소요됐다는 게 회사 측 설명이다. /조민선 기자

카카오, 임베딩 모델 2종 오픈소스 공개

개인화 AI 에이전트 핵심 기반 기술

카카오가 개인화 인공지능(AI) 에이전트 개발에 활용할 수 있는 임베딩 모델 2종을 오픈소스로 공개했다.

카카오는 자체 개발한 ‘카나나-2-임베딩’ 모델 2종을 글로벌 AI 모델 공유 플랫폼 허깅페이스에 공개했다고 29일 밝혔다.

이번에 공개한 모델은 30억 파라미터 규모의 ‘카나나-2-3b-임베딩’과 13억 파라미터 규모의 ‘카나나-2-1.3b-임베딩’이다. 지난 7월 소형 언어모델(SLM) 4종을 공개한 데 이은 추가 오픈소스 모델이다.

임베딩 모델은 텍스트의 의미를 숫자로 변환해 이용자의 질문과 관련된 정보를 검색하는 기술이다. 정확한 단어가 일치하지 않더라도 의미가 비슷한 정보를 찾아낼 수 있어 개인화 AI 에이전트의 핵심 기반 기술로 꼽힌다. 예를 들어 이용자가 ‘전에 이 키웠던 식당’을 찾을 경우 식당 이름을

입력하지 않아도 과거 대화에서 정보를 검색할 수 있다. 카카오는 지난 ‘이프카카오 26’에서도 일상에 흩어진 정보를 연결하는 임베딩 기술의 중요성을 강조한 바 있다.

카카오에 따르면 이번 모델 2종은 유사 규모의 공개 모델인 ‘라마-네모트론-임베드-1b-v2’, ‘큐렌3-임베딩-4B’와 비교한 평가에서 가장 높은 한국어 성능을 기록했다. 영어 성능은 유사한 수준을 보였으며 큐렌3-임베딩 대비 높은 추론 효율도 확보했다.

기존 카나나 임베딩 모델과 비교해서도 검색 성능을 개선하고 메모리 사용량과 추론 속도 측면에서 수십 배 이상의 효율 향상을 확인했다는 설명이다.

두 모델에는 라이선스 조건에 따라 상업적 활용이 가능한 ‘카나나 오픈 라이선스’를 적용했다. 국내외 연구자와 기업은 이를 활용해 개인화 AI 에이전트와 의미 기반 검색 서비스를 개발할 수 있다.

/최빛나 기자

LG U+, ‘AI 토큰 최적화’ 연구 인정받아

세계 3대 AI 학회 ‘뉴립스 2026’ 워크숍에 관련 논문 2편 채택

LG유플러스가 스마트폰에서 소형 AI 모델을 효율적으로 구동하는 토큰 최적화 연구 성과를 글로벌 학회에서 인정받았다. 향후 ‘익시오’ 등 자사의 AI 서비스에 적용된다.

LG유플러스는 AI 토큰 최적화 연구 관련 논문 2편이 ‘뉴립스2026’의 온디바이스 인텔리전스 워크숍에 채택됐다고 29일 밝혔다. 세계 3대 AI 학회 중 하나로 꼽히는 뉴립스는 구글, 메타, 오픈AI 등 글로벌 빅테크 기업과 글로벌 주요 대학 연구진이 참여해 최신 AI 연구 성과를 발표하는 학술대회다. 오는 12월 6일부터 12일까지 호주 시드니에서 열린다.

LG유플러스가 발표한 토큰 최적화 기술은 스마트폰 환경에 적용하는 것으로,

제한된 메모리와 전력으로도 소형언어모델(SLM)을 효율적으로 실행하고 학습할 수 있도록 한다.

먼저 통화 서비스에 활용할 SLM을 개발하고, 모델 경량화와 추론 가속 기술을 적용해 스마트폰 환경에 맞게 최적화했다. 그 결과, 최신 클라우드 AI 수준의 성능을 유지하면서 기존 방식보다 처리 속도를 약 4.5배 높이고 배터리 소모는 약 77% 줄였다. 성능 저하를 최소화하면서 스마트폰의 제한된 연산 자원과 배터리를 효율적으로 활용할 수 있도록 한 것이 특징이다.

LG유플러스는 이번 연구에서 확보한 기술을 향후 자사의 AI 서비스에 활용하는 방안을 검토하고 있다. 구체적으로 스마트폰에서 검증한 최적화 역량을 서버 GPU 환경으로 확대해 고객이 보다 빠르고 편리하게 AI를 이용할 수 있도록 할 계획이다. /조민선 기자