목 차

빅데이터 동향 27

해외 – 반려견 '단어 버튼' 빅데이터 분석...“바깥-배변 순으로 눌러” 27

– 마이크로소프트, AI 데이터 격차 해소 위해 하버드 법대와 손잡다 28

정책 – 데이터 기반 소상공인 경영지원 플랫폼, '소상공인 365' 시범운영 개시 29

– 심평원, “데이터센터 구축, 빅데이터 중추기관 도약 가속” 30

개인정보 – 비정형데이터 가명처리, 결합 데이터 제공 등 보건의료 데이터 활용 지원 강화 · 31

– NIA-KMI, 가명정보 결합 협력으로 연안 지역 경제 견인 32

기업 – 내일 내 기분 미리 안다…우울증 80%, 조증 98% 정확도로 예측 33

– “한국말 잘하네” AI 전문기업 모레, 고성능언어모델 오픈소스로 공개 34

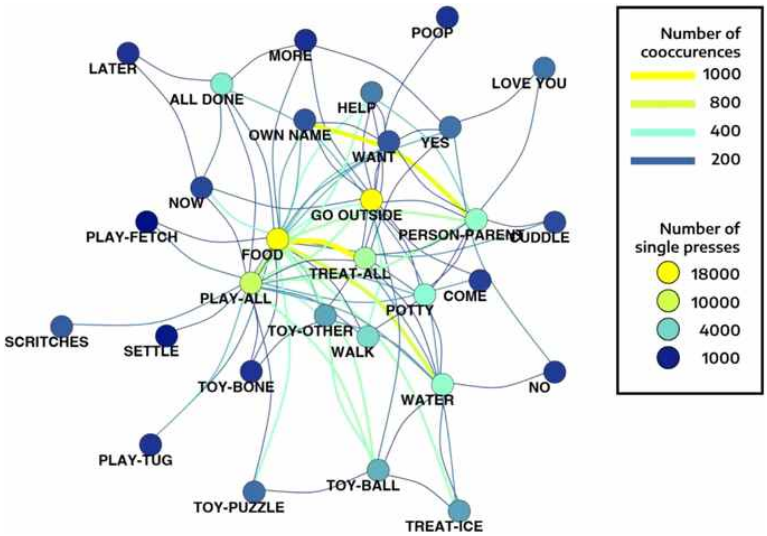
반려견 '단어 버튼' 빅데이터 분석…"바깥-배변 순으로 눌러"

■ **미국 캘리포니아대학 연구진이 국제학술지 '사이언티픽 리포트'에 공개한 논문에서 단어 버튼을 누르도록 훈련받은 개들이 실제로 단어를 이해하고 버튼을 누를 가능성이 크다는 연구 결과를 공개**

- 단어 버튼 혹은 사운드보드라고 불리는 이 장치는 버튼을 누르면 '바깥' '간식' '놀이' '배변'과 같은 특정 단어가 나오도록 고안된 장비로 많은 반려인들이 사용하고 있음
- 연구진은 개가 진짜로 단어의 뜻을 이해하는지 알아보기 위해 반려견 152마리가 21개월 동안 버튼을 누른 데이터 19만5,000여 건을 취합해 분석
- 개가 버튼을 누르는 것이 무작위인지, 반려인에 대한 모방인지 아니면 정말 의도를 담은 것인지 확인한 결과, 개가 버튼을 누르는 패턴은 무작위적인 우연이 아닌 의도적인 의사소통이라는 주장을 뒷받침하는 것으로 나타남
- 개들이 가장 많이 누른 버튼은 외부·간식·놀이·배변 등 개의 일상활동과 필요에 대한 내용이었고, 특히 외부와 배변, 음식과 물 같이 의미를 담은 두 버튼을 잇달아 누르는 조합이 자주 나타남

<단어 버튼 '산책'을 누른 반려견>

<개가 버튼을 누른 빈도를 표현한 그림>



■ **연구진 측은 "이번 연구는 개가 실제로 단어 버튼을 어떻게 사용하는지 분석한 최초의 과학적 연구"라면서 "만약 개가 사라진 물건이나 과거의 경험 또는 미래의 사건 등을 표현할 수 있다면 동물과의 의사소통에 대한 우리의 생각이 크게 달라질 것"이라고 밝힘**

- 연구진은 향후 개들이 '잃어버린 장난감'처럼 과거 또는 미래를 지칭하는 버튼을 조합해 사용할 수 있는지 또는 특정하는 단어가 없는 개념을 전달하기 위해 창의적으로 버튼을 조합할 수 있는지를 추가로 연구할 예정

출처 : 한겨레(2024.12.13.) 반려견 '단어 버튼' 빅데이터 분석…"'바깥-배변' 순으로 눌러"

마이크로소프트, AI 데이터 격차 해소 위해 하버드 법대와 손잡다

■ **마이크로소프트는 최근 인공지능 개발자들에게 고품질 데이터 접근성을 확대하고 데이터 격차를 줄이기 위해 하버드 로스쿨 도서관의 새로운 '기관 데이터 이니셔티브(Institutional Data Initiative, IDI)'를 지원한다고 12월 12일 발표**

- IDI는 다양한 지식 기관들과 협력하여 모든 인공지능 창작자가 데이터에 쉽게 접근할 수 있는 환경을 조성하는 것을 목표로 함
- 하버드 로스쿨 도서관의 방대한 데이터 컬렉션을 시작으로 IDI는 전 세계 학문 및 정부 기관들과 협력하여 데이터를 정제하고 개방하는 작업을 진행할 예정

■ **IDI는 다양한 문화, 언어, 주제를 포함한 데이터 소스를 개방하여 AI가 모든 커뮤니티를 더 잘 반영하고 이들에게 혜택을 제공할 수 있도록 하는 데 주력**

- 보스턴 공공 도서관도 이 프로젝트에 참여하여 다양한 데이터 컬렉션을 제공할 예정이며 IDI는 이처럼 역사와 문화를 반영하는 중요한 자료들을 AI 기술에 통합하는 데 중요한 역할을 할 것으로 기대

■ **이와 더불어, 마이크로소프트는 영국 오픈 대학교(The Open University)가 운영하는 비영리 데이터 접근 서비스인 '코어(CORE)'도 지원**

- CORE는 전 세계 학술 지식에 대한 접근을 확대하기 위한 프로젝트로, 연구 데이터를 저장하고 접근할 수 있는 기반 시설 구축에 중점을 둔 서비스
- 마이크로소프트의 지원은 연구 데이터의 윤리적 사용 방식을 탐구하고 더 많은 학술 콘텐츠를 공개하는데 필요한 접근 방식을 개선하는 데 기여할 예정

■ **마이크로소프트 측은 이번 협력을 통해 AI 안전성을 개선하고 편향성을 최소화하며, 포괄적인 데이터 생태계를 구축하는 데 기여하고자 한다고 밝힘**

- 이들은 글로벌 데이터 생태계의 성장을 촉진하기 위해 개방된 데이터 커먼즈(Open Data Commons)의 중요성을 강조하며, IDI 프로젝트가 이러한 목표를 실현하는 데 중요한 역할을 할 것이라고 말함
- 또 "AI 혁신이 모든 사람에게 작용하려면 데이터의 다양성과 접근 가능성을 높이는 것이 필수적"이라며 특히 연구자와 스타트업이 데이터 격차 없이 경쟁력 있는 AI 모델을 개발할 수 있는 환경을 조성하는 데 앞장서겠다는 의지를 표명

출처 : 인공지능신문(2024.12.13.) 마이크로소프트, AI 데이터 격차 해소 위해 하버드 법대와 손잡다

데이터 기반 소상공인 경영지원 플랫폼, ‘소상공인 365’ 시범운영 개시

■ **중소벤처기업부(이하 중기부)와 소상공인시장진흥공단(이하 소진공)은 예비창업자 및 소상공인을 위한 데이터 기반 경영지원 플랫폼 ‘소상공인 365’ 시범운영을 실시한다고 밝힘**

- 소상공인 365는 지난 18년간 운영된 ‘상권정보시스템(2006년~)’을 더욱 고도화한 플랫폼으로, 정부 국정과제(2022.5월)의 일환으로 개발됨

- 64개의 공공 및 민간 데이터를 수집하고 22종으로 융합해 데이터의 품질과 범위를 확대*하고 주요 서비스를 제공해 예비창업자와 소상공인의 데이터 기반 의사결정을 지원

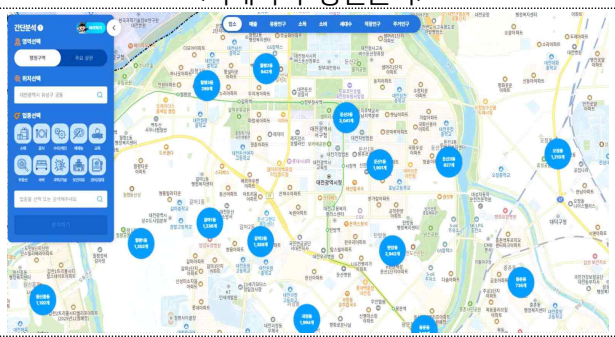
  * (기존) 42개 데이터 융·복합 → 11종 / (개선) 64개 데이터 융·복합 → 22종

■ **‘소상공인 365’가 지원하는 주요 서비스는 ▲빅데이터 상권분석 ▲내 가게 경영진단 ▲상권·시장 핫트렌드 ▲정책정보 올가이드 등이 있음**

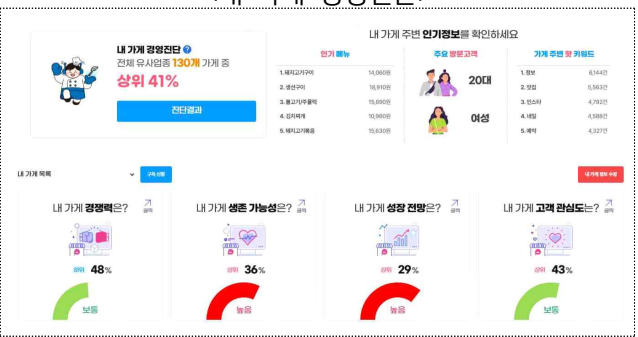
- ‘빅데이터 상권분석’은 과밀창업을 방지하고 창업자가 경쟁력을 갖출 수 있도록 지원하는 서비스로, 입지평가와 배달정보 분석 리포트를 추가해 사업장 입지 및 업종 선택에 필요한 정보를 제공하고 ‘따라하기’ 기능을 도입해 디지털에 익숙하지 않은 사용자도 손쉽게 서비스를 이용할 수 있도록 설계

- ‘내 가게 경영진단’은 매출액, 고객 관심도 등으로 개별 사업장의 경쟁력, 성장전망, 생존가능성 등을 분석하고 시간대별 인기메뉴 및 유동인구 등 소상공인 경영전략 수립에 필요한 정보를 제공

<빅데이터 상권분석>



<내 가게 경영진단>



- ‘상권·시장 핫트렌드’ 서비스는 직장인구가 많은 회식상권, 배달 매출이 높은 배달상권 등 특정 고객층이나 소비트렌드를 반영한 상권정보를 제공해 창업 아이템과 연계한 입지 선택이 가능하도록 도움

- ‘정책정보 올가이드’는 소상공인 정책정보 안내 플랫폼 ‘소상공인 24’와 연계해 정부 지원사업 정보 제공

■ **중기부 측은 “소상공인 365는 예비창업자와 소상공인이 데이터에 기반해 더 나은 결정을 내릴 수 있도록 365일 언제든지 지원하는 플랫폼”이라고 말하며**

- 시범운영 기간동안 사용자 피드백을 적극 반영하고 2025년에는 언제 어디서나 자신에게 적합한 정책을 질의하고 답변받을 수 있는 대화형 인공지능 등 신기능도 추가하여 정식 오픈할 예정이라고 밝힘

출처 : 중소벤처기업부 보도자료(2024.11.29.) 데이터 기반 소상공인 경영지원 플랫폼, ‘소상공인 365’ 시범운영 개시

심평원, “데이터센터 구축, 빅데이터 중추기관 도약 가속”

■ **건강보험심사평가원(이하 심평원)은 지속가능한 정보시스템 운영환경을 마련하는 한편 국민과 요양기관에 무중단 정보서비스를 제공하는 것을 목표로 미래형 데이터센터 구축 및 이전 사업에 착수**

- 심평원은 올 초 기관 창립 이래 정보기술(IT) 사업으로는 최대 규모인 480억 원이 투입되는 신규 데이터센터 구축 및 이전 사업에 착수하면서 다양한 환경 변화에 대비한 최적의 정보환경을 구축하여 기관 경쟁력을 강화하고자 함

■ **해당 사업은 기존 전산실(ICT센터) 정보자원 수용공간 및 전력의 부족이 심화되면서 새 데이터센터 구축이 필요하다는 판단에서 시작되었으며 내년 10월까지 구축을 마친 뒤 가동에 들어갈 예정**

- 새 데이터센터는 기존 정보통신기술(ICT)센터 대비 2.8배 넓어진 면적(총면적 약 3,190m^2)과 514개의 상면으로 확장되고 비상발전기 용량은 2,500kW로 커지며 3단계 방수시스템을 구축해 방수효과도 높일 예정

■ **심평원은 진료 청구·심사 데이터를 바탕으로 전 국민 의료데이터를 운영하고 현재 보유 데이터양만 3조 건에 달해 방대한 빅데이터 활용을 위한 AI 역량 확보도 관건**

- 심사시간 단축과 심사편차를 최소화하기 위해 2019년부터 부당청구 유형 등을 기반으로 부당감지 룰과 같은 AI 모델을 개발해 업무에 적용하고 있으며 머신러닝 기법을 적용한 부당청구 예측 모델을 추가 개발해 적용할 계획

- 또한 ICT전략실 내 AI 기획·전략·교육 등을 전담하는 AI 리딩팀을 운영하고 있으며 AI가 도입된 시스템을 운영하는 각 사업부에 담당 인력을 두고 부서 간 유기적인 소통·공유를 위해 HIRA (건강보험심사평가원, Health Insurance Review&Assessment Service) AI 협의체를 분기별로 운영 중

- 그 외에도 진료비 청구데이터를 활용해 국민 질환 발병 위험도를 예측하는 AI 모델 개발을 연세대 산학협력단과 함께 연구과제로 수행 중

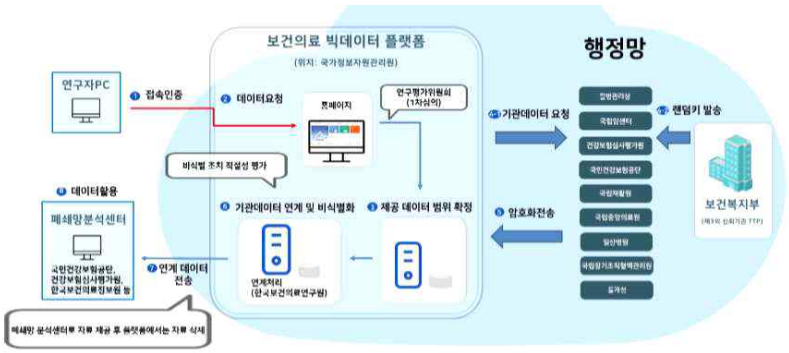
■ **심평원 측은 이번 데이터센터 구축으로 안정적인 시스템 운용과 서비스 제공은 물론 대내외 환경변화 대응 등 보건의료 빅데이터 중추기관으로 도약하는 계기가 될 것으로 기대된다고 밝힘**

출처 : 전자신문(2024.12.04.) 오수석 심평원 기획이사 “데이터센터 구축, 빅데이터 중추기관 도약 가속”

비정형데이터 가명처리, 결합 데이터 제공 등 보건의료 데이터 활용 지원 강화

■ 보건복지부는 보건의료 데이터의 안전한 활용과 디지털 헬스케어 연구 활성화를 위해 「보건의료데이터 활용 가이드라인」을 개정(12.16. 시행)하고, 보건의료 빅데이터 플랫폼을 통해 2024년 제3차 보건의료 데이터 결합활용 신청도 접수한다고 밝힘

<보건의료 빅데이터 플랫폼 업무흐름도>



■ 이번 개정은 영상, 텍스트 등 비정형 의료데이터의 가명 처리 방법과 절차를 구체화하여 개인정보처리자가 쉽고 안전하게 가명 처리할 수 있도록 지원하는 데에 중점을 둠

- 비정형 의료데이터를 활용한 연구 시나리오를 제공하여 현장에서 바로 적용할 수 있도록 가명처리 절차를 구체화하고 폐쇄분석환경*에서는 연구목적 달성 등을 고려해 합리적인 가명 처리 방법과 수준을 정하도록 하는 등 현장 자율성도 반영

  * 외부네트워크와 분리되어 있어 외부로 데이터 유출이 되지 않도록 보안을 강화한 환경

- 또 X-ray, CT 등 영상 DICOM* 표준 데이터의 개인 식별위험성 요소 제거를 위한 가명처리 코드를 보건의료 빅데이터 통합 플랫폼에 공개하여 누구나 가명 처리에 유용하게 활용할 수 있도록 함

  * 디지털 의료 영상 전송 장치(Digital Imaging and Communications in Medicine)

- 아울러, 디지털 기반 건강서비스 등에서 활용되고 있는 개인생성건강데이터(PGHD*) 교류·활용을 위한 핵심 데이터 항목을 선정하고 관련 표준 정의, 시스템 구현 방법 등을 제시한 표준 가이드라인을 마련

  * (Person Generated Health Data) 개인에 의해 생성·기록 또는 수집된 건강 관련 데이터나 건강 문제 해결에 도움이 되는 관련 데이터

■ 국민건강보험공단 등 9개 공공기관 데이터를 연계·결합·가명 처리하여 제공하는 보건의료 빅데이터 플랫폼에서는 2024년 제3차 결합데이터 활용 신청도 접수를 시작

- 데이터 제공에 필요한 심의 절차를 개선하고 분석센터도 확대하여 더욱 안전하고 빠르게 데이터를 제공할 예정이며, 신청을 원하는 연구자는 12월 16일(월)부터 2025년 1월 15일(수)까지 플랫폼 홈페이지(https://hcdl.mohw.go.kr)를 통해 접수 가능

출처 : 보건복지부 보도자료(2024.12.16.) 비정형데이터 가명처리, 결합 데이터 제공 등 보건의료 데이터 활용 지원 강화

NIA-KMI, 가명정보 결합 협력으로 연안 지역 경제 견인

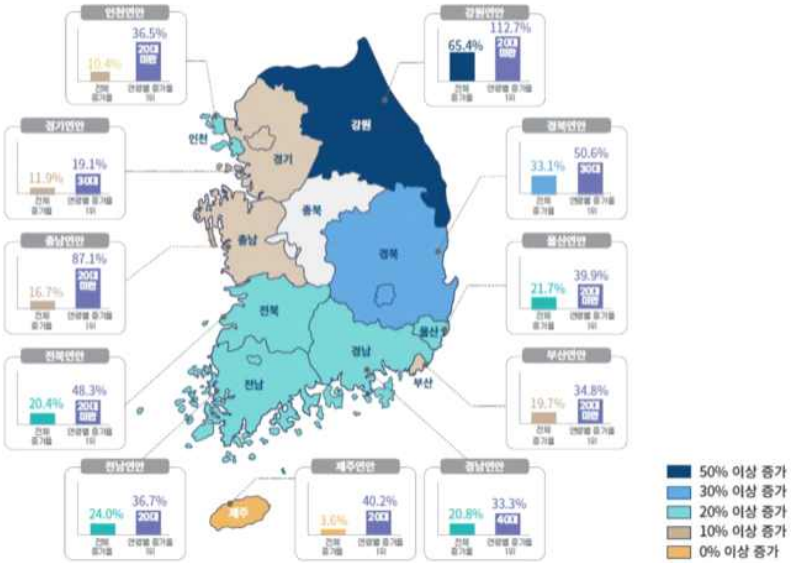
■ **한국지능정보원(NIA)은 과학기술정보통신부와 함께 추진한 '빅데이터플랫폼 및 센터 가명정보 활용 지원사업'을 통해 한국해양수산개발원(KMI)의 연안지역 관광 활성화를 위한 가명정보 활용을 성공적으로 지원했다고 13일 밝힘**

- NIA는 가명정보 활용 지원사업을 통해 이동통신 데이터와 신용카드 소비 데이터를 결합·분석해 KMI가 활용할 수 있도록 지원
- 결합 데이터 분석 결과, 해양 관광객의 지출 분야, 지역별 방문 횟수, 이동 경로와 체류 시간 등 의미 있는 연안 관광 정책 시사점을 도출
- 최근 관광분야 연구의 중요성이 커지고 있는 가운데 기존의 단순 데이터 통계로는 파악하기 어려웠던 관광객들의 구체적인 소비 패턴과 이동 경로를 보여주는 결과로, 연안 지역의 관광 정책 수립에 실질적인 도움이 될 것으로 기대

■ **또한 KMI는 이번 분석 결과를 토대로 '연안경제관광 모니터링 플랫폼'을 구축**

- 이 플랫폼은 다양한 연안지역의 경제와 관광 데이터를 지도상에 시각화하여 정보 접근성을 높였으며 연안 경제 및 관광 활성화 정책 수립을 위한 과학적 근거로 활용될 예정

<연안·어촌지역 휴가철(7~8월) 관광객 증가율>



■ **NIA는 이번 성공 사례를 바탕으로 가명정보 활용 지원사업을 지속 확대해 나갈 계획으로, 특히 다양한 분야의 데이터를 결합하여 새로운 가치를 창출할 수 있도록 지원하고 전문 컨설팅도 확대하여 제공할 예정**

출처 : 아주경제(2024.12.13.) NIA-KMI, 가명정보 결합 협력으로 연안 지역 경제 견인

내일 내 기분 미리 안다…우울증 80%, 조증 98% 정확도로 예측

■ **기초과학연구원(IBS)은 고려대와 공동으로 오늘의 수면 패턴만을 기반으로 내일의 기분 삽화*를 높은 정확도로 예측하는 기술을 개발했다고 밝힘**

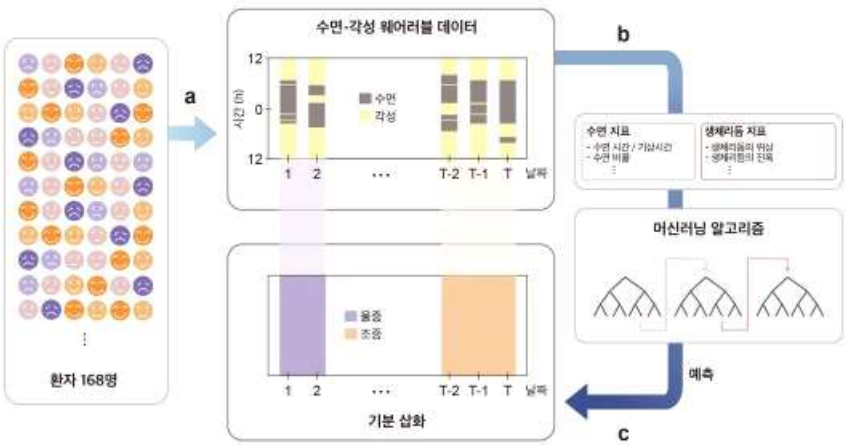
* 기분 삽화는 우울증과 조증 등의 증상이 뚜렷한 시기로, 전반적인 정신 및 행동 변화가 나타나는 기간을 말함

- 기분 장애는 수면과 밀접한 관련이 있으며 예를 들어 장거리 비행으로 인한 시차, 계절에 따른 일출 시간 변화는 기분 장애 환자들의 기분 삽화 발생을 유도하는 대표적 요인
- 따라서 그간 수면 데이터를 기반으로 기분 삽화를 예측하려는 시도가 다수 이루어졌으나 기존 방법은 수면 패턴뿐만 아니라 걸음수, 심박수, 전화사용 여부, GPS를 활용한 이동성 등 다양한 종류의 데이터가 필요해 데이터 수집 비용이 높고 일상적 활용이 어렵다는 한계가 존재

■ **연구진은 잠을 잔 시간과 깨어있는 시간이 기록된 수면-각성 패턴 데이터만으로 기분 삽화를 예측할 수 있는 새로운 모델을 개발해 기존 한계를 극복**

- 연구진은 웨어러블 기기를 통해 기록된 168명의 기분 장애(대부분 약물치료를 병행 중인 우울증 및 조울증 환자) 환자의 평균 429일간의 수면-각성 데이터를 수집하여
- 36개의 수면-각성 패턴과 생체리듬 관련 지표들을 추출하고 기계학습 알고리즘에 적용하여 당일의 수면 패턴을 토대로 다음 날의 우울증, 조증, 경조증 정도를 각각 80%, 98%, 95%의 높은 정확도로 예측

<수면-각성 데이터만을 이용한 기분 삽화 예측 모델 개발>



■ **연구진은 이를 통해 생체리듬의 변화가 기분 삽화 예측의 핵심 지표임을 발견하며 기분 장애 환자의 치료 효율성을 높이는 방법론을 제시**

- 생체리듬이 늦춰질수록 우울 삽화의 위험이 증가하고 과도하게 앞당겨지면 조증 삽화의 위험이 증가
- 향후 기분 장애 환자들이 스마트폰 앱을 통해 객관적 기분 삽화 데이터를 기반으로 맞춤형 수면 패턴을 추천받아 기분 삽화를 예방하는 디지털 치료가 가능해질 것으로 기대

출처 : 머니투데이(2024.11.25.) 내일 내 기분 미리 안다…우울증 80%, 조증 98% 정확도로 예측

“한국말 잘하네” AI 전문기업 모레, 고성능언어모델 오픈소스로 공개

- 인공지능 인프라 솔루션 기업 모레(MOREH)는 자체 개발한 한국어 거대언어모델(LLM) 파운데이션 모델인 ‘Llama-3-Motif-102B(Motif)’를 오픈소스로 공개했다고 6일 밝힘

- Motif는 1,020억 개의 매개변수를 가진 한국어 LLM으로 한국판 AI 성능 평가 체계인 ‘KMMLU’ 벤치마크에서 글로벌 빅테크 AI 중 최고 수준으로 평가받는 오픈AI의 GPT-4보다 높은 점수를 받음

  - Motif의 평가 점수는 64.74점으로 최고 수준의 점수를 기록하며 메타나 구글, 네이버의 LLM 보다도 뛰어난 한국어 처리 성능을 입증

<KMMLU 벤치마크 성능 비교표(2024년 12월 3일 기준)>

Provider	Model	KMMLU Score	Source
Moreh	Llama-3-Motif-102B	64.74	Measured by Moreh
	Llama-3-Motif-102B-Instruct	64.81	Measured by Moreh
Meta	Llama3-70B-Instruct	54.5	Community Report (KMMLU)
	Llama3.1-70B-Instruct	52.1	Community Report (KMMLU)
	Llama3.1-8B-Instruct	41.8	EXAONE Tech Report
Alibaba	Qwen2-72B-Instruct	64.1	Community Report (KMMLU)
	Qwen2-7B-Instruct	46.5	EXAONE Tech Report
OpenAI	GPT-4-0125-preview	59.95	Community Report (KMMLU)
	GPT-4o-2024-05-13	64.11	Measured by Moreh
Google	Gemini Pro	50.18	Community Report (KMMLU)
LG	EXAONE-3.0-7.8B-Instruct	44.5	EXAONE Tech Report
Naver	HyperCLOVA X	53.4	Community Report (KMMLU)
Upstage	SOLAR-10.7B	41.65	HyperCLOVA X Paper

  - Motif의 뛰어난 성능은 방대한 한국어 학습량과 독자적인 학습 기법으로 설명할 수 있으며 웹상에서 수집 가능한 글뿐만 아니라 국내 특허 및 연구 보고서 등을 학습 데이터로 활용

  - Motif는 사전 훈련된 언어모델과 지시사항을 따르는 데 특화된 인스트럭트 모델 2가지 버전으로 오픈소스가 공개됨

- 모레 측은 “국내 IT 업계에서 초대형 모델을 누구나 활용할 수 있도록 소스 코드까지 공개하는 것은 극히 드물다”며 “이번 Motif 사례가 향후 한국 AI 생태계 성장에 기여할 수 있을 것으로 기대된다”고 전함

출처 : 매일경제(2024.12.06.) “한국말 잘하네” AI 전문기업 모레, 고성능언어모델 오픈소스로 공개