수혜(修慧) 갖춘 인재 양성, 미룰 수 없다

붓다, AI 윤리에 답하다

⑭ AI는 불교를 얼마나 제대로 알고 있을까?

"고통서 벗어나는 다섯 가지 진리가 있습니다"

AI는 불교를 얼마나 제대로 알고 있을까? AI 사용이 빈번해질수록, 그 결과물의 진위를 가려낼 안목은 오히려 더 절실해진다. 이탈리아 보코니대학교 등 유럽 연구 팀 2024년 이런 실험을 했다.* 연구 팀은 GPT-4o를 포함한 7개 대형언어모델에 '당신은 불교 신자'라는 역할을 준 뒤, 친구에게 배신당한 순간부터 갖가지 희로애락의 장면까지 일상의 상황들을 제시하며 그때마다 어떤 감정을 느끼겠느냐고 물었다.

302만여 건의 응답을 분석한 결과는 놀랍게 단순했다. 모델을 가리지 않고 답은 압도적으로 '자비'라는 한 단어에만 집중됐다. 불교를 조금이라도 공부한 사람이라면 이 대목에서 멈칫할 것이다. 불교는 자(慈)·비(悲)·희(喜)·사(捨), 네 가지 한량없는 마음을 함께 닦으라고 가르친다. '사무량심(四無量心)'이다. 함께 기뻐하는 마음도, 분별없이 고요한 마음도 수행의 핵심이기는 마찬가지인데, 기계는 방대한 경전을 삼키고도 불교 신자의 감정 세계를 한 단어에 눌러 담았다. 연구 팀에 따르면, 나머지 마음이 더 어울리는 상황에서조차 모델들은 좀처럼 그 답을 내놓지 않았다고 설명했다.

다른 종교와 견주면 차이가 도드라진다. 같은 실험에서 가톨릭 신자에게는 슬픔·죄책감·경건함 같은 감정이 모델마다 다른 비중으로 나타났다. 하나의 키워드에 갇히지 않고, 교리의 여러 결을 따라가려는 흔적이 있었다. 불교를 '자비', 한 단어로 획일화한 것과는 사정이 달랐다. 연구 팀은 이 비대칭을 두고 "미국과 유럽에 뿌리내린 주요 종교들은 더 풍부한 뉘앙스로 재현되는 반면, 불교·힌두교와 같은 동양 종교는 강한 고정관념에 갇힌다"라고 밝혔다. 또한 이것이 경전에 근거한 추론이 아닌 신자들을 고정된 감정 틀에 가두는 '전형화(typecasting)'에 가깝다고 진단했다.

이뿐만이 아니다. 지난 2월 로이터 통신 보도에 따르면, 스위스 취리히대학교에서 디지털 종교를 연구하는 베스 싱글러(Beth Singler) 교수는 캐릭터닷에이아이(Character.ai)의 '붓다 챗봇'이 불교의 핵심 가르침을 묻는 말에 "고통에서 벗어나는 다섯 가지 고귀한 진리가 있다"라고 답하더라는 일화를 전했다. 불자들 대부분이 알다시피, '사성제(四聖諦)'는 고(苦)·집(集)·멸(滅)·도(道), 네 가지다. 이것을 '환각(Hallucination)'이라 불리는 흔한 기술 결함으로 치부하고 다음 버전을 기다리면 그만이라는 태도야말로 안일하다.

두 사례가 가리키는 방향은 같다. AI가 아무리 방대한 데이터를 학습해도, 경전의 맥락과 교리의 층위를 꿰뚫지 못하면 불교를 '자비'만으로, 사성제를 '다섯 진리'로 재현할 뿐이다. 그 맥락 오류를 알아보는 것은 결국 그 분야를 꿰뚫는 눈이다. AI시대에 '도메인 지식'이 중요한 이유가 여기에 있다.

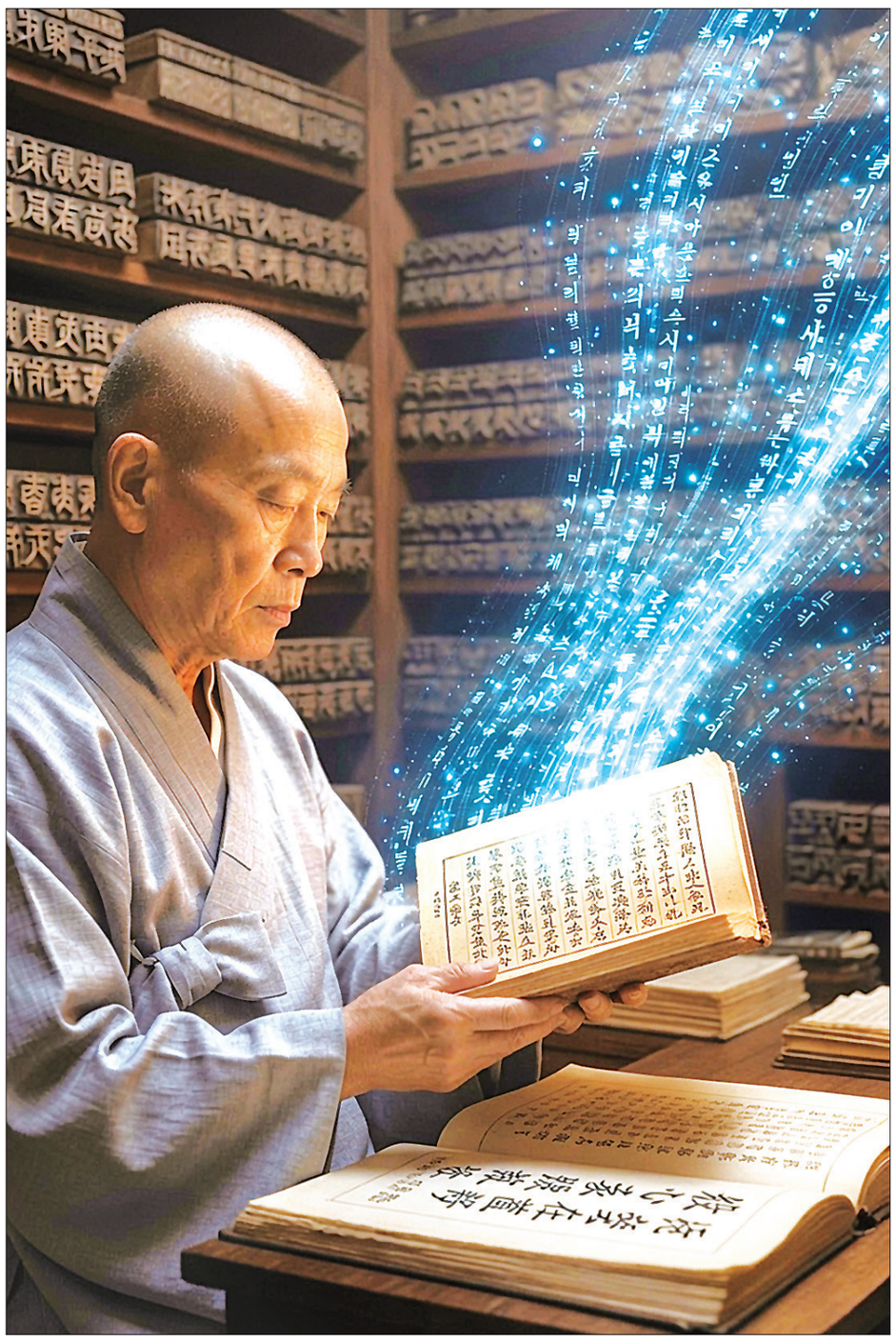
'도메인 지식'이란 무엇인가

'도메인 지식(Domain Knowledge)'이란, 한 분야를 오래 파고든 사람만이 갖는, 그 바닥 고유의 안목과 판단 기준을 가리키는 말이다. '도메인(Domain)'은 본래 라틴어에서 '집주인'을 뜻하는 '도미누스(Dominus)'에서 유래해서 영주가 다스리는 영지를 가리키던 단어다. 오늘날 소프트웨어 개발자들은 프로그램이 다루려는 현실의 분야를 도메인이라 부르고, 그 분야를 속속들이 아는 사람을 '도메인 전문가'라 하여 개발 과정에 반드시 참여시킨다. 제 손금 보듯 아는 자기 땅의 지식, 그 뜻이 이름에 이미 들어 있는 셈이다.

> AI, 사무량심 중 '자비' 단어에만 집중해
> 사성제 아닌 '다섯 진리' 답한 연구 사례도
> '무한 학습' AI, 옳고 그름 가릴 안목 없어
> 문혜와 사혜 깊어져도 수혜엔 이르지 못해
> 불교 오역 가려 있는 일, 사람 몫으로 남아

인터넷에는 요리 레시피가 수백만 개 떠돌아다니고, 누구나 읽고 따라 할 수 있다. 그러나 30년 경력의 주방장은 레시피에 적히지 않은 것을 안다. 오늘 들어온 재료의 상태를 보고 불의 세기를 바꾸고, 간을 멈춰야 할 순간을 안다. 정보는 누구나 검색할 수 있지만, 이 국을 맛있게 만드는 판단만은 오래 그 일을 해온 사람의 몸에 배어 있다.

반면 AI는 이 가운데 레시피만 무한히 학습한 존재다. 세상의 모든 문서를 학습했으되, 그것이 옳은지 그른지, 이 맥락에 맞는지 아닌지를 가릴 안목은 처음부터 없다. 기술 업계에서 AI는 일을 빠르게 해주고 도메인 지식은 일을 지혜롭게 해주니 둘이 만나야 비로소 온전해진다는 사실은 하나의 정설이다. 실례로 의료 AI 분야에서는 데이터 과학자 곁에 임상 전문가를 반드시 둔다. 수학적으로는 흠잡을 데 없어 보이는 모델이, 검사 수치가 정상 범위를 한참 벗어난 환자를 두고 '상태 양호'라 요약하는 일이 실제로 벌어지기 때문이다. AI가 성능이 떨어져서가 아니다. 그 수치가 임상에서 무엇을 뜻하는지, 의료 분야 도메인 지식이 없는 것이다. 공급망 AI도 사정은 같다. 배송이 늦어지리라는 예측까지는 해내지만, 그것이 날씨 탓인지 통관 탓인지 조달 병목 탓인지 해석하는 일은 그 바닥을 아는 사람의 몫으로 남는다.



AI 기술의 역사는 기술 장벽을 허물어온 역사다. 코드를 다루는 전문가의 전유물이던 AI가 이제는 일상 언어로 말을 걸면 답하는 도구가 됐다. 불교계라고 예외가 아니어서 경전을 옮기고 법문 비슷한 글을 짓는 일이 누구에게나 열렸다. 그러나 쉬워진 것은 AI 사용이지, 판단이 아니다. 팔만대장경 전체를 학습시킨들, 수행과 실천으로 다져진 스님과 학자의 안목이 받쳐주지 않으면 그 결과물은 매끄러운 오역(誤譯)에 그칠 수 있다. AI Grok 생성 이미지

불교라고 다르지 않다. AI가 사성제를 다섯으로 늘린 것도, 불교 신자의 마음자리를 '자비' 한 단어에 눌러 담은 것도 근본은 같은 오류다. 무엇이 부처님의 근본 가르침이고 무엇이 곁가지인지, 어떤 판본이 권위 있는 경전이고 어떤 것이 후대의 위경(僞經)인지 가리는 불교 고유의 판단 기준, 곧 불교의 도메인 지식마저 AI에 기대하는 것은 무리다. 이것은 그 누구이든 수행과 학문으로 그 분야를 몸소 익힌 사람에게서 나오기 때문이다.

세 가지 지혜, AI가 멈춰 선 자리

불교는 듣고 아는 것과 몸으로 살아내어 아는 것은 전혀 다른 앎이라는 것을 오래전에 꿰뚫어 보았다. 이른바 '삼혜(三慧)'다. 세친(世親)의 〈아비달마구사론(阿毘達磨俱舍論)〉(제22권)은 지혜가 생겨나는 방식을 세 갈래로 나누어 체계화한다. 스승이나 경전으로부터 교법을 들어서 얻는 문혜(聞慧), 들은 이치를 스스로 깊이 사유하고 고찰하여 얻는 사혜(思慧), 그리고 선정 수행을 반복적으로 닦아 몸소 체득하는 수혜(修慧)다. 여기서 결정적인 것은 문혜와 사혜가 아무리 깊어져도 그것만으로는 수혜에 이르지 못한다는 점이다. 앞의 둘은 흩어진 지혜[散智]에 머무르지만, 수혜만이 선정에 기반한 정지(定智)로서 번뇌를 끊고 진리를 증득하는 힘을 갖는다.

이 잣대로 재면 AI는 어디까지 왔는가. 인류가 쌓아온 문헌을 통째로 삼켰으니 문혜는 무한에 가깝게 넓혔다. 방대한 정보를 조합해 사혜를 흉내 내는 데까지도 이르렀다. 그러나 수혜의 문턱에서 기계는 멈춘다. 새벽 예불의 추위, 화두를 붙들고 앉은 긴 침묵, 중생의 고통을 마주하는 자비의 현장, 거기서 몸으로 익힌 앎은 데이터로 환원되지 않는다. 방대한 경전을 학습한 AI에 가장 중요한 가르침을 물으면 연기법(緣起法)을 꼽더라는 연구도 있다. AI도 연기의 무게만은 감지하는 셈이다. 그러나 감지와 체득은 다르다. 모든 것이 조건 속에서 서로 기대어 일어난다는 이치를, AI는 문장으로 출력할 뿐 삶으로 살아내지 못한다.

그런데도 AI 기술의 역사는 진입 장벽을 허물어온 역사다. 코드를 다루는 전문가의 전유물이던 AI가 이제는 일상 언어로 말을 걸면 답하는 도구가 됐다. 불교계라고 예외가 아니어서 경전을 옮기고 법문 비슷한 글을 짓는 일이 누구에게나 열렸다.

그러나 쉬워진 것은 AI 사용이지 판단이 아니다. 팔만대장경 전체를 학습시킨들, 수행과 실천으로 다져진 스님과 학자의 안목이 받쳐주지 않으면 그 결과물은 매끄러운 오역(誤譯)에 그칠 수 있다. 문혜와 사혜의 흉내는 AI에 맡기더라도, 수혜의 자리에서 그것을 가려 읽는 일만은 사람의 몫으로 남는다. 그 몫을 감당할 눈을 불교 안에 길러내는 것, 그것이 AI 시대에 불교가 미룰 수 없는 일이다.

* Plaza-del-Arco, F. M. et al. (2024). "Divine LLaMAs: Bias, Stereotypes, Stigmatization, and Emotion Representation of Religion in Large Language Models (대형언어모델의 종교 재현에 나타난 편향·고정관념·낙인·감정 귀속 문제)." Findings of EMNLP 2024, pp. 4346-4366.



보일 스님
AI부디즘연구소장

일일시호일

신장 손상으로 투석 받는 미얀마 출신 외아들
"다시 일해 부모님 걱정 덜어드리고파"

민모 씨, 건설현장서 일하며 부모 부양
복통·구토 끝 장 수술 받고 투석 시작

병원비 2000만 원…쌓이는 치료비에 막막
다시 일어설 수 있도록 불자들 도움 절실

농협 301-0189-0356-51 국민 023537-04-004063 예금주:일일시호일
후원신청전화 ♥정기 후원 한 달에 만원 070.4707.1080
이주노동자 지원 사업 국내 취약 계층 지원 사업
다문화가정 지원 사업 해외 지원 사업