‘환각 현상’ 등 단점개선 분주… 한국형 ‘챗GPT’ 선보인다

챗GPT를 포함한 초거대 인공지능(AI)가 시급히 해결해야 할 숙제는 할루시네이션(환각)을포함한거짓말, 편향성, 일관성 결여, 세이프티(안정성)인 것으로 조사됐다.

국내 초거대 AI 개발사들은 최근 전 세계적으로 큰 인기를 끌고 있는 오픈AI의 대화형 챗봇 ‘챗GPT’와 유사한 기능을 도입한 신제품을 빠른 시일 내 선보이고, 글로벌 시장에서 챗GPT와 경쟁을 벌인다는 전략이다. 이들은 필터링 기능과 강화학습, 센싱 기능 등을 통해 챗GPT의 부족함을 해결하기 위해 집중하고 있는 파악됐다.

김유원 네이버클라우드 대표가 지난달 27일 삼성동 코엑스에서 진행된 ‘데뷰’ 콘퍼런스에서 하이퍼클로바X에 대해 소개하고 있다. /네이버

◆“거짓말을 못하게 해라”

메트로경제가 네이버, LG그룹, SK텔레콤, KT 등을 대상으로 지금 활동 중인 초거대 AI 기능중 가장 시급히 해결해야 할 것 중 첫번째가 거짓말이었다. 그 중에서도 할루시네이션(환각) 현상을 꼽았다.

거짓말, 편향성 등 단점 지적
국내 개발사들, 한계 극복 나서
‘챗GPT’ 유사·대응 제품 출시

LG그룹 개발자는 26일 “초거대 AI는 거짓말 문제가 많이 발생한다. 그 중에서도 틀린 정보를 그럴 듯 하게 표현하는 할루시네이션 현상이 심하다. 팩트를 포장하는 것 보다는 없는 사실을 세상에 존재하는 것으로 설명한다. 세종대왕이 맥북을 던진 사건을 설명하라고 하면 이를 그럴 듯 하게 설명한다”며 “챗GPT도 이러한 문제 때문에 후처리 과정을 거치게 된다”고 설명했다.

SKT 한 임원은 “AI가 사실적인 데이터를 엄청나게 학습하면서 결국 사실적인 거짓말을 하게 된다”며 “페이스북에서 논문 AI를 개발했는데, AI가 사실적인 거짓말을 하는 문제가 드러나 결국 1주일만에 폐기가 됐다”고 말했다.

초거대 AI를 개발하는 KT 임원은 “초거대 AI는 가짜를 사실처럼 얘기하는 신뢰성 문제가 생긴다. 이 문제는 단기간에 해결하지 쉽지 않은 문제다. 최근 기술이 개발되면서 그런 부분들이 줄이들기는 했지만 100% 안 나오게 할 수는 없다”며 “어느 정도는 사람이 스스로 가치판단을 하면서, 어떤 게 문제가 되는 답변인 지를 판단해야 한다”고 설명했다.

◆창의력이 떨어진다

초거대 AI에서는 편향성 문제가 빈번하게 나타나고 있다.

KT 임원은 “초거대 AI는 특정 계층에 대해 안 좋은 생각을 가지고 있는 사람들이 만든 데이터를 학습한다”며 “결국 사람들이 생각하는 편향성 데이터가 반영된다”고 설명했다.

초거대 AI의 또 다른 단점은 일관성이 배제돼 있다는 점이다.

SKT 임원은 “초거대 AI는 메모리가 없어 일관성한 단점이 있다. “등산을 좋아한다”고 했다가 다음 번에는 “안 좋아한다” 하는 등 변화가 심하다”며 “콘텐츠와 끈끈한 연결고리를 통해 일관성을 유지해야 되는데 이 부분이 부족하다”고 밝혔다.

세이프티 이슈도 빈번하게 일어나고 있다. 이는 초거대 AI와 욕설, 성적인 내용, 범죄와 관련된 대화를 나눌 수 있다는 것이다.

SKT 임원은 “마이크로소프트의 대화형 AI인 ‘테이’는 지난 2016년 24시간 동안 트위터에서 10만건이 넘는 글을 쏟아냈다. 하지만 반인륜적 트윗을 잇따라 내보내 서비스가 중단됐다. 챗봇 ‘이루다’도 성적 발언, 성차별, 혐오적인 발언을 하는 등 문제가 생겨 서비스가 중단되기도 했다. 이러한 것이 세이프티 이슈로 문제 자체가 커 하나하나 개선해 나가야 한다”고 설명했다.

또한 AI 휴먼 여리지와 아이린이 실물이 거의 차이가 없는 등 초상권 문제에 대한 논의도 점차 활발해지고 있다.

◆챗GPT 단점을 개선하라

국내 개발사들은 이 같은 초거대 AI의 단점을 극복하기 위해 필터링 기술을 사용하고 강화학습을 시키고 센싱 기능을 적용하고 있다.

LG ‘엑사원’에 센싱 기술 구현
네이버 AI 윤리원칙 필터 개발
KT 팩트체크 기술로 강화학습

KT는 초거대 AI ‘믿음’에 팩트 체크 기술을 적용하고 초거대 AI에 강화학습을 시키고 있다. API(앱 프로그래밍 인터페이스)를 개발해 AI에 필터링 기술을 적용하고 있다.

KT 임원은 “어떤 답변이 문제가 되는지 조사해 학습 데이터를 구축한다. 이에 대한 분류기를 만들어 이를 필터링하고 있다”며 “또한 AI에 강화학습을 시켜 잘못된 답변을 생성하지 않도록 하고 있다”고 밝혔다.

SKT도 초거대 AI ‘에이닷’에 세이프티 필터를 적용해 욕설, 성적인 내용 등 대화를 걸러주고 있다.

SKT 임원은 “편향된 대화는 데이터를 가공하기에는 시간이 너무 늦다. 따라서 미리 문제가 될 만한 대화 내용을 걸러주는 게 핵심”이라고 설명했다.

LG그룹은 초거대 AI인 ‘엑사원’에 센싱 기술 구현을 통해 편향적이거나 혐오적인 표현을 바로 잡아주고 있다.

LG그룹 관계자는 “데이터를 대량으로 학습했을 때 정보를 만든 사람의 편향성이 들어갈 경향이 크다”며 ”데이터 확보부터 학습, 데이터 처리를 한 후 AI가 다시 사람에게 보여주는 전 과정마다 기술이 필요하다. 오염되지 않은 데이터인 퓨어 데이터를 학습했음에도 편향성이나 혐오 표현이 나타날 수 있어 센싱 기능을 통해 이를 바로 없애는 작업을 하게 된다”고 설명했다.

네이버는 AI 윤리원칙을 준수한 서비스를 개발하기 위해 ‘AI 필터’를 개발해 ‘클로바 스튜디오’ 등에 탑재하고 있다.

네이버 관계자는 “AI 필터 기능은 클로바 스튜디오를 통해 생성된 서비스 앱에서 욕설 등 부적절한 결과물이 출력되는 것을 감지해 사용자에게 알려주는 기능”이라며 “클로바 스튜디오 이용자는 AI 필터를 의무적으로 사용하도록 하고 있다”고 말했다.

엑사원 /LG

한편 국내 개발업체들은 챗GPT와 유사한 기술을 빠르게 적용한 제품을 내놓을 계획이다.한국어 특화 모델로 경쟁력 강화에 집중하는 등 글로벌 시장에서 챗GPT와 정면 승부를 펼친다는 전략이다.

KT 임원은 “올해 내 챗 GPT와 유사한 기술을 적용하려고 노력하고 있다”며 “모든 회사들이 챗GPT 기술이나 이와 유사한 기술을 상용화하는 것을 목표로 하고 있다”고 밝혔다.

LG그룹도 챗GPT와 유사한 기능으로 ‘전문가 AI’, 각 영역별로 특화된 생성형 AI 모델를 빠른 시간 내에 선보일 예정이다. LG그룹 관계자는 “AI·화학·바이오 등 각 분야에서 전문가를 도울 수 있는 AI를 개발 중”이라고 밝혔다.

SKT는 에이닷을 챗GPT글로벌 서비스로 키워 챗GPT에 대항하겠다는 포부이다.

SKT는 최근 에이닷에 이용자와 대화한 내용 중 중요한 정보를 기억하는 ‘장기기억’ 기술과, 이미지와 한글 텍스트를 동시에 학습해 인간처럼 생각하고 표현하는 ‘이미지 리트리벌’ 기술을 적용했다. 이를 위해 한국의 로컬리티를 설명 가능한 한국어 기반 10억장의 이미지와 한글 텍스트 쌍 학습 데이터를 구축해 초거대 멀티모달 AI를 학습시켰다.

네이버도 상반기 중 AI 챗봇을 탑재한 한국형 챗GPT인 ‘서치GPT’를 출시할 계획이며, 카카오도 올해 3분기 내에 챗GPT에 대응하는 AI 챗봇 ‘코챗GPT’를 선보인다는 전략이다.

/채윤정 기자 echo@metroseoul.co.kr