환각적 통계 계산기

20210541 이호준

많은 대학생들, 직장인들, 심지어는 전문가들 까지도 생성형 모델인 ChatGPT 의 사용을 자연스럽게 받아들이는 시대가 왔다. 그런데 혹시, 그러한 모델에서 발생하는 환각 현상 (Hallucination) 이라는 오류에 대해서 알고 있는가? 나는 본 글에서 일부 자료 및 필자의 경험을 바탕으로 해당 현상의 과거를 되새겨보고, 현재 진행 상황과, 이후 미래에 대한 생각을 적어보려고 한다.

환각 현상(Hallucination) 이라는 오류는 생성형 모델인 ChatGPT 가 등장한 이후 지금까지 쭉 화자되는 대표적인 인공지능의 버그 현상이다. 쉽게 말해 잘못된 정보나 거짓말을 아주 당연하다는 듯이 말하는 현상인데, 많은 사람들이 아직까지도 경험하는 문제일 것이다. 우스갯소리로 화자되는 예시로는 '세종대왕 맥북 프로 던짐 사건' 이 있는데, ChatGPT 에 조선 왕조 실록에 기록된 세종대왕이 맥북 프로를 던졌던 사건에 대해 알려달라고 요청했을 때, 정말 실제로 기록된 것처럼 자세히 알려준다는 것이다. [1] 나는 이 글을 통해, 이러한 오류에 대한 연구자들의 생각과 전산학부 학생으로서 가져야 할 자세, 그리고 미래에 관하여 간단하게 논의하려고 한다.

얼마 전, 굉장히 신선하게 느껴졌던 한 구절이 있다. '환각 현상은 오류가 아닌, 특성이다' (Hallucinaion is a feature, not a bug).[2] 이 글을 읽고 나는 친구들과 소소하게 토론을 했었는데, 우선 수학과를 복수전공중인 나와 또 다른 친구는 이 말은 오랜 난제인 리만 가설이나 P=NP 문제 등을 공리라고 인정하는 것과 다를 것이 없다는 의견이었다. 학술적인 자세가 아니며, 또 그런 문제가 특성이라는 것을 인정하는 순간 생성형 모델은 자연어 기반 검색 엔진이라는 별명을 벗어날 수 없다고 생각했다.

반면 뇌인지 공학과를 복수전공중인 또 다른 친구는 이런 해석을 내놓았다. 오로지 '언어' 만으로 학습하는 대형 언어 모델 (LLM, Large Language Model) 의 경우, 그것이 특성이라는 말이 이해된다는 것이다. 실제로 같이 알아본 결과 대형 언어 모델이 학습할 수 있는 개념들의 한계를 정의하는 논문 [3], 그리고 각종 기사들 [4] 을 찾아볼 수 있었다. 결국 현재 학습이라고 하면, 아무리 다양한 방법론이 연구되고 있을지언정 뉴럴 네트워크와 고작 계수 최적화를 통해 일어나는 국소적인 창발성을 지칭하고 있는 것인데 당연히 실수를 할 수 밖에 없다는 것이다.

그럼에도 세계의 이목이 인공지능으로 향하면서 생성형 모델과 대형 언어 모델들은 눈부신 성장을 이어가고 있다. 여전히 환각 현상을 종종 겪게 되긴 하지만, 그 자체로 훌륭한 검색 엔진, 번역기, 그림 도구, 코딩 도구 등 다양한 방면에서 필수적인 도구로 자리매김됐다. 이에 어떤 사람들은 일자리 감소와 기술 특이점에 대한 걱정과 우려를, 또 다른 사람들은 다가올 새로운 시대에 대한 기대를 내비친다.

하지만 나는 조금은 더 근본적으로, 이 문제를 정의해보고자 한다. 우선, 논의하는 AI 라는 단어를 단순히 ChatGPT 등 상업적으로 운용되고 있는 생성형 모델에 국한시켜선 안된다. 왜냐하면 그렇게 국한하는 순간 한계가 명확해지기 때문이다. 하나의 거대한 계산 장치들의 무덤에서, 나의 상황 맥락은 단순히 통계 수치로만 입력되고, 내가 받는 답변 또한 '비슷한' 맥락에 대해 통계적으로 유력한 답변이 되돌아올 뿐이다. 이 공정 하에선 인간의 검산과 검토를 피할 수 없다. 단적인 예시로 많은 학교의 대형 프로젝트로 수행되는 PintOS 프로젝트를 예시로 들어보자. 실제로 완전히 같은 프로젝트를 전 세계의 학생들이 참여하고 있으므로, 이미 ChatGPT 의 통계 데이터

[1] https://www.hankookilbo.com/News/Read/A2023022215200000727
[2] https://ksankar.medium.com/hallucination-is-a-feature-not-a-bug-910d12a8fbab
[3] https://arxiv.org/abs/2306.12213
[4] https://www.theguardian.com/technology/article/2024/aug/06/ai-llms

베이스엔 말도 안되게 많은 PintOS 의 성공 사례들과 실패 사례들, 질문들이 저장되어 있을 것이다. 하지만 기회가 된다면 ChatGPT 에 전적으로 의존하면서 코드 일부를 수정해보자. 해결될 기미가 보이지 않는 무수한 오류들을 마주할 수 있을 것이다. 이것을 나는 생성형 모델의 한계라고 생각한다. 그들은 최고의 분석 알고리즘과 학습 알고리즘을 통해, 각 부분들에 대해 최고의 선택지들만 제공해줬지만 그것들이 유기적으로 모인 결과물은 서로 아무런 조화를 찾을 수 없는 결과들이 됐기 때문이다.

그리고 바로, 위와 같은 상황이 생성형 모델이 SW 안전성에 끔찍한 영향을 끼치게 되는 과정들이다. 아무도 탓 할 수 없는 이유는 생성형 모델은 최신 기술을 통해 가장 유력한 해결책을 제시해줬을 뿐이고, 사람은 자신이 직접 구현하는 것이 귀찮아 자세한 상황 설명과 아이디어를 제공하고 구현을 계산 기계에게 맡겼을 뿐이다. 고장난 SW 말고 우리에게 남은 것, 탓할 것은 아무것도 없다. 결국 주기적으로 환각을 일으키는 통계 계산기에게 중대한 임무를 부여한 내 실수로 남기 때문이다.

그러면 결국 어떤 AI를 지향해야 할까? 난 그 궁극적인 목표로 마블(Marvel) 의 '자비스' 를 뽑는다. 오로지 한 명의 사용자를 위해 존재하는 AI, 오로지 내 코드를 위해 돌아가는 AI. 우리가 PintOS 프로젝트를 할 때 머릿속을 해당 코드들로 가득 채우고 사는 것처럼, 하나의 프로그램의 논리적 구조와 목적, 오류 가능성이 있는 곳들만 생각하는 AI를 지향해야 한다. 비록 비용과 물리적인 한계 등 다양한 요인은 있겠지만, 궁극적으로 AI 가 '자비스' 를 목표로 삼고 발전되는 것이 아니라면 나는 SW 안전성이 해결될 가능성은 없다고 생각한다.

그렇다면 전산학부 학생들은 무엇을 생각하고 연구해야 할까? 아쉽게도 (하지만 행복하게도) 이러한 과도기에 놓인 이상, 주기적인 환각을 일으키는 통계 기계의 성능을 개선하려고 하기보다 과도기에 놓인 우리의 소중한 SW 들을 지키는 방법을 연구하고,[5] AI에 대한 자신의 이상을 구체화 한 뒤 그것에 이르기 위한 방법을 연구해야 한다. 바로 내가 골라야 하는 중요한 두 문제이다. 오래 전부터 꿈이었던 컴파일러 및 버그 탐지의 전문가가 되고자 할 지, 다양하고 깊은 수학 이론을 바탕으로 새로운 패러다임을 좇는 인공지능 연구자가 될 지. 많은 전산학부 학생들이 하고 있을 행복한 고민과 함께, 글을 마무리하려고 한다.

[5] https://prosys.kaist.ac.kr/assets/pdfs/trustworthy-ai.pdf