

AI, 우리 안의 '외계 지능'을 만들어가는 여정! 🛸

안녕하세요, 여러분! 일타 강사 저스틴입니다! ✨

AI에 대해 이야기하다 보면 종종 혼란스러울 때가 많죠? 시리(Siri)가 농담하는 것부터 영화 속 터미네이터까지, AI라는 이름으로 너무나 많은 것들이 뒤섞여 있기 때문일 거예요. 오늘은 이 'AI'라는 녀석이 어떻게 발전해왔고, 왜 저스틴이 이토록 잠 못 드는 밤을 보냈는지 그 흥미진진한 여정을 함께 떠나볼까요?

AI, 그 시작은 '환상'이었다! 🧩

어려운 개념, 저스틴이 아주 쉽게 풀어드릴게요!

사람들은 생각하는 기계에 대한 환상을 오래전부터 가지고 있었어요. 1770년에는 '**기계식 터키인**'이라는 체스 기계가 나타나 사람들을 깜짝 놀라게 했죠. 화려한 장식장 위에 체스판이 놓여 있고, 로봇이 체스 말을 움직이는 모습은 벤자민 프랭클린이나 나폴레옹 같은 당대의 천재들도 감쪽같이 속아 넘어가게 했답니다. 심지어 에드거 앨런 포는 이 기계를 보고 인공지능의 가능성을 언급하기도 했어요. 하지만 **바로 이거예요!** 이 모든 건 사실 거짓말이었죠. 기계 안에는 진짜 체스 고수가 숨어 있었을 뿐이에요! 그럼에도 불구하고, 사람들은 기계가 생각할 수 있다고 거의 한 세기 동안 믿었어요.

시간이 흘러 1950년, 드디어 현대 인공지능의 씨앗이 뿌려졌어요. **클로드 새넌**이라는 천재 과학자가 낡은 전화 교환기를 활용해서 복잡한 미로를 스스로 통과하는 기계 쥐 '**테세우스**'를 만들었어요. 이게 바로 **최초의 머신러닝(기계 학습)** 사례 중 하나로 꼽힙니다. 같은 해, 컴퓨터 과학의 선구자 **앨런 튜링**은 '**모방 게임 (Imitation Game)**'이라는 흥미로운 사고실험을 통해 기계가 인간처럼 기능할 수 있는 이론적 기반을 제시했어요. 이 두 가지 아이디어가 바로 AI의 새로운 개념을 탄생시킨 중요한 **포인트!**라고 할 수 있죠.

'AI 겨울'을 넘고 '예측 AI'의 시대로! ❄️

이론만으로는 부족했죠? 1956년, MIT의 **존 매카시**가 '인공지능(Artificial Intelligence)'이라는 용어를 처음 만들었고, 초기 컴퓨터 과학자들은 AI 프로그램 개발에 몰두했어요. 논리 문제를 풀거나 체스 게임을 하는 등 초반에는 빠른 발전이 있었지만, 기대만큼 결과가 나오지 않자 'AI 겨울'이라는 침체기가 여러 번 찾아왔어요. 마치 계절이 바뀌듯 AI 연구는 발전과 좌절을 반복했죠.

그러다가 2010년대에 새로운 AI 붐이 시작되었어요. 바로 **머신러닝(Machine Learning)** 기술을 활용한 데이터 분석과 예측이 뜨거운 감자로 떠올랐죠. 이때의 AI는 주로 '**지도 학습(Supervised Learning)**' 방식이었는데, 이게 뭐냐면요, AI에게 정답이 붙어있는 데이터(즉, '**레이블링 된 데이터**')를 잔뜩 보여줘서 배우게 하는 거예요.

[가상 사례] 예를 들어, 호텔이 내년 예약 수요를 예측하고 싶다면, 과거 날씨 패턴, 지역 행사, 경쟁사 가격 같은 데이터를 AI에게 입력하는 거예요. 그러면 AI는 이 데이터를 분석해서 훨씬 정확하게 수요를 예측해줄 수 있죠. 이전에는 그저 평균적인 예측에 의존했지만, 예측 AI는 개별 상황에 맞는 정확한 예측을 가능하게 했어요.

이런 예측 AI 기술을 가장 잘 활용한 회사가 바로 **아마존(Amazon)**이에요. 아마존은 AI 알고리즘을 이용해서 수요를 예측하고, 창고 배치를 최적화하고, 키바(Kiva) 로봇을 이용해 물류 창고의 효율을 극대화했죠. 우리

가 직접 체감하진 못해도, 이 시기의 AI는 물류, 고객 서비스 등 수많은 '백 오피스(back-office)' 업무를 혁신했어요.

하지만 이런 AI에도 한계는 명확했어요. 인간은 직관적으로 이해하지만 AI는 모르는 '**미지의 미지(unknown unknowns)**'를 예측하기 어렵고, 학습하지 않은 데이터에는 적응하기 힘들었죠. 그리고 가장 중요하게도, **텍스트를 이해하고 생성하는 능력**이 매우 부족했어요. 그래서 우리는 이 AI를 일상생활에서 직접적으로 느낄 기회가 별로 없었을 거예요.

'트랜스포머'의 등장, LLM 시대의 서막! 🖋️

그런데 2017년, 구글 연구원들이 "**Attention Is All You Need**"라는 제목의 논문을 발표하면서 AI 세상에 큰 변화가 찾아왔어요. 이 논문은 '**트랜스포머(Transformer)**'라는 새로운 AI 모델을 제안했는데, 컴퓨터가 인간의 언어를 처리하는 방식을 완전히 바꿔놓았답니다.

이전의 AI는 문맥을 잘 이해하지 못해 엉성한 문장을 만들곤 했어요. 마치 아이폰 자동 완성 기능이 엉뚱한 단어를 제안하는 것과 비슷했죠. 하지만 트랜스포머는 '**어텐션 메커니즘(Attention Mechanism)**'이라는 기술을 활용해서 텍스트의 여러 부분 중 가장 중요한 부분에 집중할 수 있게 되었어요. 덕분에 AI가 훨씬 더 인간처럼 언어를 이해하고 다룰 수 있게 된 거예요!

이 트랜스포머 아키텍처를 기반으로 탄생한 새로운 유형의 AI가 바로 **대규모 언어 모델(LLM)**이에요. LLM은 아마존의 수요를 예측하는 대신, 텍스트를 분석해서 다음에 나올 '**토큰(단어 또는 단어의 일부)**'을 예측하는 방식으로 작동해요. 여러분이 휴대폰에서 사용하는 똑똑한 자동 완성 기능의 엄청나게 복잡하고 정교한 버전이라고 생각하면 쉬울 거예요.

핵심은! LLM은 방대한 양의 텍스트 데이터(웹사이트, 책 등)를 학습(이를 '**사전 학습(Pretraining)**'이라고 해요)하면서 인간 언어의 패턴과 구조를 스스로 익혀요. 이때 AI가 학습하는 수많은 변수를 '**가중치(Weights)**'라고 부르는데, 챗GPT의 초기 버전은 무려 1,750억 개의 가중치를 가지고 있었답니다. 이 가중치들은 누가 프로그래밍한 것이 아니라, AI가 스스로 학습하면서 만들어진 거예요.

[가상 사례] LLM을 요리 견습생에 비유해 볼까요? 이 견습생은 세상 모든 레시피(텍스트 데이터)를 읽고 공부해요. 처음에는 엉망진창인 양념 선반(가중치)을 가지고 있지만, 수많은 시행착오를 거치며 '사과와 계피는 잘 어울리지만, 사과와 커민은 아니다' 같은 향신료(단어)들의 조합 규칙을 스스로 익히죠. 결국, 이 견습생은 프롬프트(요리 주문)만 주면 어떤 재료(단어)를 어떻게 조합해야 맛있는 요리(맥락에 맞는 텍스트)를 만들 수 있는지 아는 명장이 되는 거예요!

이런 사전 학습 과정은 엄청난 양의 컴퓨터 자원과 시간, 그리고 비용이 들어요. 챗GPT 같은 최신 LLM 하나를 훈련시키는 데 1억 달러(약 1,300억 원) 이상이 든다고 하니 정말 어마어마하죠. 학습 데이터에는 엔론(Enron)의 이메일이나 아마추어 연애 소설 같은 의외의 자료들도 포함되어 있어서, 저작권 문제나 편향성 학습 같은 윤리적, 법적 문제도 계속 논의되고 있어요.

인간의 손길, '미지의 AI'를 길들이다! 🍌

사전 학습만 마친 AI는 때로는 예상치 못한 결과를 내놓거나, 심지어 돈을 횡령하거나 살인을 저지르는 방법에 대한 조언을 해줄 수도 있어요. 마치 거울처럼 학습한 데이터를 아무런 판단 없이 그대로 반영하는 거죠. 그래서 LLM은 사전 학습 후에 '**파인튜닝(Fine-tuning)**'이라는 두 번째 개선 단계를 거쳐요.

가장 중요한 파인튜닝 방법 중 하나는 '**인간 피드백을 통한 강화 학습(RLHF: Reinforcement Learning from Human Feedback)**'이에요. AI 회사들은 사람들을 고용해서 AI의 답변을 읽고 평가하게 해요. 정확도, 폭력성, 선정성 등을 기준으로 점수를 매기거나 피드백을 주면, AI는 이 피드백을 바탕으로 추가 학습을 해서 더 좋은 답변을 내놓고 나쁜 답변을 줄이도록 성능을 조절하는 거예요. 고객 지원 대화 기록처럼 특정 목적에 맞게 추가 파인튜닝을 할 수도 있고요.

텍스트를 넘어, 그림을 '보고 그리는' AI! 🎨

챗GPT가 돌풍을 일으킨 시기와 비슷하게, 그림을 만들어내는 AI들도 등장했어요. **미드저니(Midjourney)**나 **달리(DALL-E)** 같은 도구들이죠. 이 AI들은 "고흐 스타일로 미키마우스 그려줘"처럼 텍스트 프롬프트만으로 유명 화가의 스타일을 모방하거나, 실제 사진과 구별하기 어려운 사실적인 이미지를 만들어냅니다.

이런 이미지 생성 AI는 텍스트가 아닌 수많은 이미지와 그에 대한 설명을 학습해요. 그리고 '**확산(Diffusion)**'이라는 과정을 거쳐 무작위 노이즈(TV 화면의 지지직거리는 잡음처럼 보이는 것)를 점진적으로 선명한 이미지로 바꿔나가죠. 텍스트 모델이 단어를 조합해 글을 만들 듯, 이들은 그림을 '발명'하는 거예요.

더 놀라운 건, 이제 LLM이 이미지까지 '보고 그릴' 수 있게 되었다는 점이에요. '**멀티모달 LLM(Multimodal LLM)**'이라고 부르는데, 언어 모델과 이미지 생성 능력이 결합된 거죠. 예를 들어, 제가 하트가 잔뜩 그려진 서툰 비행기 그림을 보여주면, AI는 "하트가 주변에 있는 귀여운 비행기 그림 같아요. 당신은 비행이나 비행하는 사람을 좋아하는군요. 아마 조종사이거나 사랑하는 사람이 조종사일 수도 있고, 아니면 그냥 여행을 좋아할 수도 있겠네요."라고 말해주고, 더 멋진 비행기 그림까지 그려준답니다! 이런 멀티모달 능력은 AI가 세상을 이해하는 방식을 더욱 예측 불가능하게 만들 거예요.

GPT-3에서 GPT-4까지, AI는 어디까지 왔을까요? 🗣️

초기 LLM은 지금처럼 뛰어나지 않았어요. 2021년 OpenAI가 공개한 GPT-3에게 라임을 맞춰 시를 써달라고 하면,

한 AI가 있었는데 찰리라 불렀네 / 정말 대단한 녀석이었지 / 계산하고 추론하고 / 실수를 저지르지 않았어 / 그러다 어느 날 오작동하기 시작했지

...처럼 라임도 없고 재미도 없는 엉망진창 시를 써냈어요. 하지만 2022년 말, OpenAI가 개선된 LLM인 GPT-3.5를 기반으로 **챗GPT**를 출시하면서 뭔가 특별한 일이 일어났어요. 아무도 예상하거나 프로그래밍하지 않았던 '**인간 같은**' 능력을 보여주기 시작한 거죠! 챗GPT는 이야기, 시, 에세이, 트윗, 심지어 코드까지 쓸 수 있게 되었고, 많은 시험에서 인간 수준에 가까운 성적을 거두었어요.

더 나아가, GPT-3.5는 라임을 맞춰 시를 쓰는 능력이 훨씬 좋아졌고, 대화 기능까지 갖췄어요. 제 피드백에 맞춰 시의 마지막 줄을 고치는 모습은 정말 놀라웠죠.

그리고 그 후속작인 **GPT-4**는 더욱더 압도적이었어요! 고등학교부터 대학원, 전문직 시험까지 다양한 표준화 시험에서 GPT-3.5를 훨씬 뛰어넘는 성적을 기록했죠. 변호사 시험에서 90%대에 들고, AP 미적분, 물리학, 미국사 등 여러 과목에서 만점을 받았어요. 심지어 신경외과 의사 자격시험까지 통과했고요, 인간의 창의성 테스트에서도 최고 점수를 받았답니다!

물론 AI가 훈련 데이터에서 정답을 미리 학습해서 '커닝'했을 가능성도 있지만, 단순한 점수를 넘어서는 GPT-4의 진정한 능력에 대한 증거는 차고 넘쳐요.

GPT-4에게 라임에 대해 어떻게 생각하냐고 물어보면,

"라임 맞추는 시는 AI 능력의 진화를 보여주지만, 제 진정한 잠재력을 측정하는 가장 정교한 척도는 아닐 수 있습니다. 저는 문학을 분석하고, 복잡한 문제를 해결하며, 깊은 철학적 토론에 참여하는 등 훨씬 더 복잡한 언어적 위업을 달성할 수 있습니다..."

...라고 답하며 섬뜩할 정도로 스스로를 인지하는 듯한 모습을 보여주기도 해요. 비판적인 관점이나 칭찬하는 관점으로 글에 대해 논평해달라고 하면, 완벽하게 그 역할을 수행하죠.

이런 대규모 언어 모델과 트랜스포머 기술은 블로그 글, 에세이, 컴퓨터 코드, 연설문, 예술 작품, 음악 등 정말 다양한 콘텐츠를 생성하는 능력을 가지고 있습니다. 현재는 기능은 제한적이지만 저렴한 '작은 LLM', 자유롭게 수정 가능한 '오픈소스 LLM', 그리고 GPT-4처럼 소수 기업만 만들 수 있는 '최첨단 LLM' 등으로 나뉘어 발전하고 있어요.

'외계 지능'의 등장, 알 수 없는 미래! 🤖

최첨단 LLM들은 가장 많은 데이터와 컴퓨터 자원으로 훈련되면서, 프로그래밍된 것을 넘어서는 능력을 보여주는데, 이것 '이머전스(Emergence)', 즉 '예측 불가능한 돌발 능력'이라고 불러요. 체스를 두거나 인간보다 공감 능력을 더 잘 발휘하는 것처럼 말이죠. "신비로운 것을 보여줘"라고 하면 맨델브로트 집합 같은 아름다운 프랙탈 패턴을 보여주고, "불가사의한 것을 보여줘"라고 하면 H.P. 러브크래프트 풍의 신비로운 텍스트 생성기를 만들어주기도 해요.

핵심은! 이처럼 놀라운 능력을 가진 '토큰 예측 시스템'이 왜 이런 능력을 갖게 되었는지 아무도 정확히는 알지 못한다는 거예요. 언어와 그 뒤에 숨겨진 사고 패턴이 우리가 생각했던 것보다 훨씬 단순하고 '법칙적'일 수 있다는 뜻일까요? 뉴욕대학교 샘 보우먼 교수는 LLM의 인공 뉴런 간 연결이 너무 복잡해서 인간은 그 행동을 정확히 설명할 수 없을 것이라고 말하기도 했죠.

하지만 LLM에게는 놀라운 강점만큼이나 이상한 약점들도 있어요. 어떤 문제는 AI에게는 쉽지만 인간에게는 어렵고, 그 반대인 경우도 있죠. **[가상 사례]** 예를 들어, '틱택토 게임의 다음 최적 수를 찾아라'는 AI에게 어려운 문제였지만, '틱택토 게임을 완벽하게 플레이하는 자바스크립트 웹페이지를 만들어라'는 쉬운 문제였어요. AI가 어떤 분야에서 뛰어나고 어떤 분야에서 실패하는지 미리 알기 어렵다는 뜻이에요. AI가 내놓는 답은 너무나 '그럴듯'해서 이해하는 것처럼 보이지만, 사실은 착각일 수도 있습니다.

결론적으로 우리는 '인간과 비슷하지만 완전히 다르며', '생각하는 것 같지만 진짜인지는 알 수 없는' 일종의 '**외계 지능(Alien Mind)**'을 만들어낸 거예요. 자, 그럼 우리는 어떻게 해야 이 외계 지능이 우리에게 '친근하게' 다가오도록 만들 수 있을까요? 이것이 바로 '**정렬 문제(Alignment Problem)**'이자, 앞으로 우리가 함께 풀어가야 할 숙제예요!

오늘의 정리

첫째, AI는 '기계식 터키인'의 환상에서 시작해 '머신러닝'과 '트랜스포머'를 거쳐 진화해왔어요. 둘째, 대규모 언어 모델(LLM)은 방대한 학습과 인간 피드백을 통해 텍스트와 이미지까지 생성하는 '멀티모달' 능력을 갖추

게 되었답니다. 셋째, GPT-4는 상상을 초월하는 능력을 보여주지만, '이머전스'와 '미지의 약점'을 가진 '외계 지능' 같은 존재라는 점을 기억해야 할 거예요!

[실천 과제] AI가 생성한 글이나 그림 중 어떤 부분이 '외계 지능'처럼 느껴지는지, 혹은 어떤 부분이 여전히 인간보다 부족하다고 느껴지는지 한 번 찾아보세요! 그리고 친구들과 함께 AI의 미래에 대해 토론해보는 건 어떨까요?