ChatGPT

GPT, Claude, Gemini: 대형 언어 모델의 작동 원리와 인간 사고와의 비교

서론

오늘날 **대형 언어 모델(LLM)**들은 인간과 유사한 문장을 생성하며 다양한 작업을 수행하고 있습니다. 특히 OpenAI의 **GPT 시리즈**, Anthropic의 **Claude**, 그리고 Google DeepMind의 **Gemini**는 최첨단 LLM의 대표격입니다. 이들 모델은 방대한 데이터로 훈련되어 놀라운 언어 처리 능력을 보여주지만, 그 작동 방식은 인간의 사고 과정과 여러 측면에서 다릅니다. 본 보고서에서는 GPT, Claude, Gemini가 **어떻게 학습(훈련)**되고 **추론**하며 **기억(맥락 유지)**을 활용하고 **피드백을 반영**하는지 살펴보고, 이를 인간의 사고 방식과 비교해보겠습니다. 또한 PaLM, LLaMA 등 기타 주요 LLM들의 특징과 인간 사고와의 차이도 간략히 정리하였습니다. 초심자도 이해할 수 있도록 개념을 풀어 설명하고, 필요한 경우 표와 예시를 들어 **알기 쉽게** 구성하였습니다.

OpenAI GPT: 대규모 훈련과 RLHF를 통한 정교화

GPT(Generative Pre-trained Transformer)는 OpenAI에서 개발한 시리즈로, GPT-2, GPT-3, GPT-4로 거듭 발전해 왔습니다. GPT 모델들은 기본적으로 **Transformer 디코더** 구조를 사용하며, 인터넷 텍스트를 비롯한 대용량 코퍼스를 바탕으로 **다음 단어 예측** 과제를 수행하도록 **사전 훈련**됩니다 [1]. 예를 들어 GPT-3는 약 **1,750억 개의 파라미터**를 가진 **자동회귀 언어 모델**로, 웹 텍스트, 전자서적, 위키피디아 등 방대한 데이터에서 **문맥에 따른 단어 생성 확률**을 학습했습니다 [2] [3]. 그 결과 별도 **태스크 특화 학습 없이도** 번역, 질의응답, 요약 등 다양한 작업에서 몇 가지 예시만 보고도 작업을 수행하는 few-shot 능력을 얻었습니다 [4]. GPT-4의 경우 모델 세부 정보는 공개되지 않았지만, 공개된 기술 보고서에 따르면 **멀티모달 입력**(텍스트와 이미지)을 수용하고 텍스트를 출력할 수 있도록 훈련되었고 [1], 공개 데이터와 타사 라이선스 데이터를 모두 활용해 훈련되었습니다 [1]. 이는 GPT-4가 **문서의 다음 토큰을 예측**하도록 거대한 데이터로 사전 학습되었음을 의미합니다.

훈련된 GPT 모델은 **추론 시(Inference)** 입력 시퀀스 뒤에 이어질 다음 단어(토큰)를 한 번에 하나씩 예측하여 문장을 만들어냅니다. GPT의 **Transformer** 구조는 문맥의 모든 단어를 **자기 주의(self-attention)** 메커니즘으로 참고하여 다음 출력에 반영하며, 한 번에 한 토큰씩 **자동회귀적으로 생성**합니다 [5]. 즉, "Transformer는 한 번 호출에 여러 단어를 한꺼번에 내놓는 것이 아니라, 호출될 때마다 **한 토큰씩** 출력하며, 이를 반복(loop)하여 다수의 토큰을 생성한다"고 설명됩니다 [5]. 이러한 방식으로 GPT는 사용자의 질문에 답하거나 새로운 텍스트를 단계적으로 생성합니다. GPT의 **맥락 기억** 용량은 모델 버전에 따라 다른데, GPT-3는 입력과 출력 합쳐 약 2048 토큰까지 처리했고, 최신 GPT-4는 **8천~3만2천 토큰** 정도로 문맥 창(Context window)이 크게 늘어났습니다 [1]. 그러나 GPT는 **명시적 장기 기억**이 없어서 대화 역사나 이전 지식을 활용하려면 해당 내용을 **프롬프트에 재제공**해야 합니다. (예: 이전 대화 내용을 모두 입력으로 넣어야 기억합니다.) 훈련을 통해 얻은 지식은 **모델 파라미터**에 분산되어 저장되지만, 훈련 이후 새 정보를 **실시간으로 학습**하지는 않습니다. 반면 인간은 새로운 사실을 기억에 저장하고 필요 시 떠올리지만, GPT는 주어진 **고정된 지식**과 현재 **문맥 창 내 정보**로만 대응합니다.

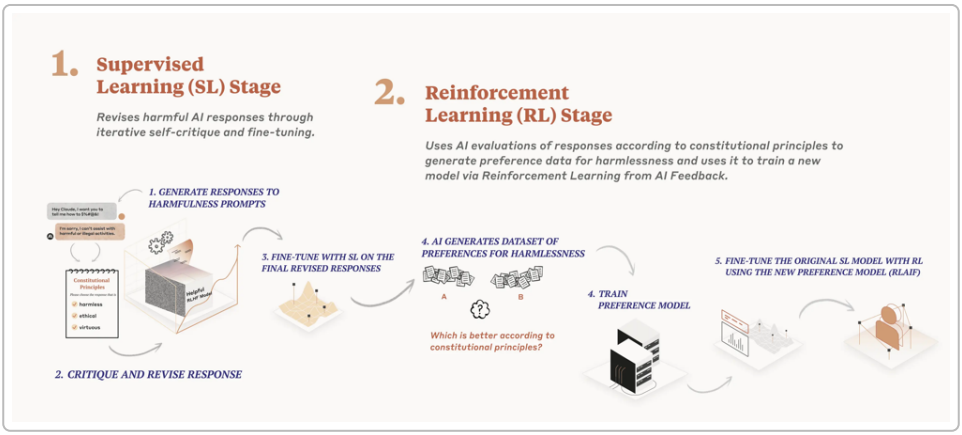
GPT 모델은 초기 사전학습 후에도 **피드백을 통해 개선**됩니다. OpenAI는 인간 피드백 강화학습(**RLHF: Reinforcement Learning from Human Feedback**)을 활용해 GPT 모델의 출력을 인간 선호에 맞게 **미세조정**했습니다 [6]. 예를 들어 ChatGPT의 경우, **인간 평가자**들이 다수의 모범 답변과 순위 데이터를 제공하고, 이를 바탕으로 **보상 모델**을 훈련한 다음, 이 보상 신호를 이용해 GPT를 **강화학습**으로 미세조정했습니다 [7] [8]. GPT-4에는 여기에 한 단계 더 나아가, 모델이 유해한 요청을 거절하도록 **안전 보상 신호**를 추가했다고 합니다 [6]. 즉, 별도의 **GPT-4 기반 분류기**로 출력의 안전성을 판단해 **보상**을 주는 방식을 도입하여, 부적절한 요청에 답변하기보다는 **거부 메시지**를 생성하도록 훈련한 것입니다 [6]. 이러한 RLHF 과정을 통해 GPT는 **유해하거나 사실과 다른 출언을 줄이고**, 사용자 지시에

더 잘 따르면서도 안전한 응답을 생성하도록 조율되었습니다. 다만 완벽하지 않아 여전히 환각(hallucination)이나 사실 오류가 발생할 수 있음을 OpenAI도 인정하고 있습니다 [9].

Anthropic Claude: 헌법 기반 자기피드백과 대용량 문맥

Claude는 Anthropic사가 개발한 대화형 LLM으로, GPT와 유사한 Transformer 구조의 **디코더 모델**입니다. Claude의 기본 훈련도 GPT처럼 **인터넷 텍스트** 등을 이용한 다음 단어 예측 언어모델링이지만, 정확한 파라미터 규모나 데이터셋 구성은 공개 정보가 제한적입니다. 알려진 바에 따르면 Claude 1세대는 GPT-3 규모(수십억~수천억 파라미터대)의 모델이며, 이후 버전이 거듭되며 성능이 향상되었습니다. Claude의 가장 두드러진 특징 중 하나는 **매우 긴 문맥 길이**입니다. Anthropic은 2023년 Claude를 공개하면서 **문맥 창을 9천 토큰에서 10만 토큰으로 확장**했다고 밝혔는데, 약 **75,000단어** 분량에 해당하는 분량입니다 [10]. 이는 수백 페이지에 달하는 문서를 한 번에 입력해 분석하거나, 며칠에 걸친 긴 대화도 한 세션에 유지할 수 있는 수준입니다 [11] [12]. 인간이 10만 토큰을 읽고 이해하려면 5시간 이상 걸리고 모두 기억하기도 어려운데, Claude는 이를 1분 내 처리해 질의응답을 할 수 있었다고 합니다 [13]. 예컨대 위대한 개츠비 소설 전체(72K 토큰)를 넣고 한 문장을 살짝 바꾼 뒤 그 차이를 묻자, Claude가 22초만에 올바른 답을 찾아냈다는 데모가 공개되었을 정도입니다 [14]. 이러한 **대용량 문맥** 처리 능력은 현재 상용 LLM 중 최고 수준이며, 동일한 입력을 인간이 처리하는 것과 비교하면 **속도와 단기기억 면에서 우월**함을 보여줍니다. 다만 이러한 긴 문맥을 활용하더라도 Claude 역시 **학습된 지식의 한계**는 존재하며, 주어진 프롬프트 내 정보에 크게 의존합니다.

또 하나 Claude의 차별점은 **"헌법 기반 AI(Constitutional AI)"**라는 **피드백 반영 메커니즘**입니다. Anthropic은 모델의 출력을 인간 피드백으로 직접 조정하는 대신, **미리 정해둔 AI 헌법(원칙 모음)**에 따라 **모델 스스로 출력을 평가하고 개선**하도록 했습니다 [15] [16]. 이를 구현하기 위해 두 단계의 학습을 거치는데, 첫 번째 단계는 **지도학습(SL) 단계**로, 모델이 몇 가지 예시와 원칙을 참고해 **자신의 유해한 응답을 비평하고 수정하도록** 가르칩니다 [17]. 두 번째 단계는 **강화학습(RL) 단계**로, **인간 평가자 대신 AI 평가자**(헌법 원칙에 따른)가 두 응답 중 더 나은 것을 선택하여 그 선택을 보상 신호로 사용해 모델을 강화학습으로 미세조정합니다 [17]. 쉽게 말해, 모델에게 "이 원칙들에 비추어 어떤 답변이 더 도움이 되고 안전한지"를 판단하게 한 뒤, 그 판단 결과를 이용해 모델을 업데이트하는 것입니다. **Anthropic 공식 블로그**에서는 이러한 과정을 그림으로 제시했는데, 아래 그림에서 헌법에 기반한 자기 비평 → 개선 → AI 투표에 의한 강화학습 흐름을 확인할 수 있습니다



. 첫째로 **지도학습 단계**(왼쪽 1번~3번 과정)에서 Claude는 일부 유해한 프롬프트에 대한 초기 응답을 생성한 후, 미리 주어진 원칙(예: "해롭지 않음", "정직함", "악의적 요청 거부" 등)에照하여 **스스로 그 응답을 비판**하고 **개선된 답변을 작성**합니다. 그렇게 **자체 교정된 최종 답변들로 새로운 데이터셋을 구성**해 모델을 다시 **파인튜닝**하고요 [17]. 둘째로 **강화학습 단계**(오른쪽 4번~5번 과정)에서는, 앞서 만든 **AI 헌법 평가자**가 두 개의 모델 응답 중 헌법 원칙에 더 부합하는 쪽을 선택하여 **선호도 데이터**를 만듭니다. 이 **AI-생성 피드백**을 토대로 **보상 모델**을訓練하고, 최종적으로 그 보상 모델을 활용한 **강화학습(RLAIF)**으로 원래 모델을 미세 조정하여 **보다 안전하고 도움되는 응답**을 산출하도록 합니다 [17]. 이러한 **Constitutional AI 접근법**을 통해 Claude는 **도움이 되면서도 해롭지 않은** 답변을 내놓도록 훈련되었습니다 [15]. Anthropic 측은 이 방식이 **인간 피드백 RLHF에 비해 효과적**이며, 실제로 헌법 기반 모델이 **더 적은 인적 개입으로**도 **유해성 감소와 유용성 향상**을 동시에 이루었다고 보고했습니다 [18]. 예를 들어, 헌법 기반 Claude는 공격적이거나 위험한 지시에는 원칙에 따라 답을 거부하면서도, **회피적으로 굴지 않고** 가능한 한 도움되는 정보를 제공하는 균형 잡힌 응답을 했다고 합니다 [18]. 요약하면, Claude는 거대한 언어 모델로서 GPT와 비슷한 언어 능력을 가지면서도 **헌**

법이라는 명시적 원칙 세트에 따라 **자기검열과 자기개선을 학습**했고, 이를 통해 **안전성**과 **일관성**을 높이려 한 점이 특징입니다.

Google Gemini: 멀티모달 통합과 초거대 문맥의 진화

Gemini는 Google DeepMind에서 개발한 최신 LLM 패밀리로, 2023년 말에 첫 버전(1.0)이 공개된 이후 빠르게 발전하고 있습니다. Gemini는 특히 **멀티모달** 특성을 지닌 범용 모델로 설계되었는데, 기존 언어 모델들과 달리 **텍스트뿐 아니라 이미지, 오디오, 비디오, 코드**까지 동시에 처리하도록 만들어졌습니다 [19]. 실제로 "Gemini는 **텍스트 코퍼스만으로 훈련된 것이 아니라**, 웹 문서, 서적, 코드뿐 아니라 유튜브 영상의 자막 등 **다양한 형태의 데이터**를 병렬적으로 처리할 수 있도록 설계되었다"고 알려져 있습니다 [19]. 예를 들어 한 대화 맥락에 사용자가 **문장과 사진, 영상, 오디오**를 뒤섞어 넣어도 Gemini는 이를 모두 이해하고, 답변 역시 텍스트와 이미지, 음성 등 여러 형식으로 반환할 수 있다고 합니다 [20]. 이러한 멀티모달 능력은 실제 인간의 **다중 감각 처리**와 유사한 방향을 지향한 것으로 볼 수 있습니다. (GPT-4도 이미지 입력을 처리할 수 있지만, Gemini는 더 나아가 영상·음성까지 포괄한다는 점에서 한층 광범위합니다.)

Gemini의 **모델 구조**는 1세대의 경우 **디코더 전용 Transformer**로, 효율적인 학습/추론을 위해 **multi-query attention** 등 일부 최적화가 적용되었습니다 [21]. 1.0 버전의 최대 모델(Ultra)은 GPT-4와 유사하거나 더 큰 규모로 추정되며, **문맥 길이도 32,768 토큰**으로 GPT-4 수준이었습니다 [21]. 이후 개선을 거듭하여 2024년 중반 2.0, 2025년 초 2.5 버전이 발표되었는데, 특히 **Gemini 2.5 Pro**(2025년 3월 공개)는 **1백만 토큰**에 이르는 엄청난 문맥 창을 제공하여 주목받았습니다 [22]. 이는 사실상 **무제한에 가까운 컨텍스트**로, 기존 모델들이 수만 토큰 이내였던 한계를 크게 넘는 것입니다. (1백만 토큰은 일반적인 책 3~4권 분량에 해당하며, 이런 양의 정보를 한꺼번에 고려해 답변할 수 있다는 뜻입니다.) 다만 이 수치는 연구 단계의 실험적 기능으로 보이며, 실제 서비스에서 1백만 토큰을 전부 활용하는 데는 시스템 자원상의 제약이 클 것입니다. 그럼에도 **맥락 유지 능력** 면에서 Gemini가 **새로운 지평**을 열었음을 보여줍니다.

훈련 방식 측면에서, Gemini는 Google의 **Pathways** 인프라를 활용해 **초대형 모델**을 효과적으로 학습시켰습니다 [23]. PaLM 등 이전 모델로 입증된 대로, TPU v4 팟 수천 개를 병렬 연결하여 학습을 진행했고, 영문+다국어 텍스트, **고품질 웹문서, 위키피디아, 책, 대화 데이터, GitHub 코드** 등 풍부하고 다양한 데이터를 활용했습니다 [24]. 또한 **Whitespace까지 보존**하는 특수 토크나이저 등을 사용하여, 특히 코드나 수식처럼 섬세한 문자 패턴도 정확히 모델에 학습시켰습니다 [24]. DeepMind 연구진은 전작 **AlphaGo** 등에서 사용된 강화학습적 기법을 언어모델에 접목하려는 시도를 했는데, Gemini에는 **체스·바둑 등 게임 플레이에서 얻은 추론 능력을 활용**하거나, **"연쇄적 사고(chain-of-thought) 기법을 통한 단계별 추론"** 기능이 포함된 것으로 알려졌습니다 [25] [26]. 예를 들어 Gemini 2.5 버전은 응답을 내기 전에 **자신의 추론 과정을 내부적으로 여러 단계 거치며** 답을 도출하는 "Thinking mode"가 있다고 합니다 [22]. 이러한 능력 덕분에, 수학 문제 풀이나 논리적 추론 등에서 Gemini가 매우 뛰어난 성능을 보였다는 평가가 있습니다 [22]. 한편, Gemini의 **피드백 및 안전성** 측면에서 Google은 자사 **AI 원칙**에 기반해 강력한 **안전 장치**를 마련했다고 강조합니다 [27]. 예를 들어 모델 훈련 단계에서 **Real Toxicity Prompts**와 같은 공개 **유해성 벤치마크**(다양한 독성 수준의 프롬프트 10만 개로 구성)를 활용하여, 모델의 응답이 정책을 준수하는지 점검하고 문제가 되면 데이터를 조정했습니다 [28]. 또한 **폭력적 내용이나 편견** 등 유해한 출력을 식별·차단하는 **안전 필터 시스템**을 구축하고, 다양한 **레드팀** 전문가들과 협력하여 모델을 공격적으로 테스트했다고 합니다 [29]. 이는 **훈련 데이터 필터링**부터 **모델 출력 모니터링**까지 여러 층의 안전 대책을 적용함으로써, Gemini가 가급적 **유해한 발언을 피하면서도 사실에 근거한 답변**을 하도록 한 것입니다 [29]. (실제로 Google은 Gemini 출시 전 **광범위한 적대적 테스트**를 거쳐 잠재적 문제를 식별하고 수정했다고 발표했습니다 [30].) 요약하면, Gemini는 방대한 **멀티모달 지식**을 품고 있으며, **장대한 문맥도 한꺼번에 처리**하고, **체계적 추론 능력**을 내장하려 한 점에서 이전 세대 LLM과 구별됩니다. Google은 Gemini가 앞으로 **도구 사용**이나 **로봇 제어** 등 보다 **에이전트적 기능**까지 수행하는 플랫폼으로 발전할 것이라고 예고하고 있어, 인간과 AI 상호작용의 새로운 장을 열 잠재력이 있습니다.

GPT vs Claude vs Gemini: 비교 요약

위에서 살펴본 GPT, Claude, Gemini의 특징을 몇 가지 측면에서 비교하면 다음과 같습니다:

특성	GPT (OpenAI)	Claude (Anthropic)	Gemini (Google)
모델 구조	Transformer 디코더 (autoregressive LM) [1]. GPT-3는 1750억 파라미터, GPT-4는 상세 미공개 (멀티모달 입력 가능) [1].	Transformer 기반 디코더 LM. 파라미터 상세 미공개 (GPT-3수준 추정). 대규모 대화 최적화.	Transformer 디코더 (1세대), 향상된 구조 (multi-query attention 등) [21]. 1.0 Ultra는 GPT-4급 규모, 최신 2.x는 멀티모달/강화통합.
훈련 데이터	웹 텍스트, 서적, 위키 등 **거대 텍스트 코퍼스** [2]. (GPT-4는 라이선스 데이터 추가) [1]. 주로 텍스트 언어모델링 사전학습 후 RLHF로 미세조정.	웹 텍스트 등으로 기본 LM 학습. 이후 **헌법(AI 원칙)**에 따른 **자기 비평+강화학습**으로 인간 피드백 대체 [17]. 유해 출력 최소화 목표.	**멀티모달 데이터** (텍스트+이미지+오디오+코드 등) 동시 학습 [19]. TPU 기반 Pathways로 초대형 모델 효과적으로 학습 [23]. 훈련 중 안전성 체크 강화 [28].
추론 및 생성	**다음 토큰 확률** 기반 생성. 한 번에 한 토큰씩 순차 출력 [5]. 강력한 **문장 완성** 능력으로 다양한 작업 few-shot 수행 [4]. 논리 추론에 chain-of-thought 프롬프트 활용 가능 [31].	다음 토큰 생성은 GPT와 유사. 다만 **Constitutional AI**로 학습되어 **원칙에 어긋나는 요청 거부**나 **어조 완화** 등 **출력 조율**이 특징. 답변 시 스스로 한 번 더 검열/수정하는 효과. 논리적 답변 경향.	다음 토큰 생성의 기본은 동일. 그러나 **내부적으로 단계별 "생각" 과정** 거쳐 복잡 문제 해결 (2.5 Pro에서 강조) [22]. **함수 호출** 등 도구 사용 및 **멀티모달 응답** 가능. 실시간으로 그림 생성·설명 또는 음성 응답 등 수행.
문맥 및 메모리	GPT-3: ~2048토큰, GPT-4: 최대 32k 토큰 컨텍스트. 긴 문맥 처리 가능하나, 초과 분은 "망각". 대화 이력은 프롬프트로 재전달 필요. 지속적 학습 없음 (지식 고정).	**최대 100k 토큰** (Claude 2) [10]로 매우 긴 문맥 지원. 수백 페이지 문서도 한번에 분석 가능 [11]. 그래도 맥락 창 넘어가면 잊음. 대화 지속시도는 긴 문맥으로 보완.	1.0: 32k 토큰 [21], 2.5: 최대 1백만 토큰 시도 [22]. **사실상 장기 문맥 지향**. 멀티모달이라 문맥에 이미지/영상 포함 가능 [20]. 인간처럼 다양한 정보 동시 고려.
피드백/안전	**RLHF**로 인간 선호도 반영. 추가로 GPT-4는 **안전 보상 모델** 도입해 유해 요청 거부 강화 [6]. 사용자는 시스템 메시지로 톤 조정 가능 [32]. 그래도 환각과 오류 완전히 해결 못함 [9].	**AI 헌법** 따라 자기 피드백 학습 [17]. 인간 대신 AI가 응답 평가하여 강화학습. 결과적으로 **높은 harmlessness**와 **helpfulness** 목표 [18]. 원칙 위배 질문에는 비교적 일관되게 거부 또는 완곡 응답.	**인간+자동 피드백 혼용**. 훈련 중 **Toxicity 프롬프트**로 안전 점검 [28], 출시 전 광범위 **레드팀 테스트** [30]. 사용자 수준에서도 안전 필터 적용. 거짓/환각 줄이려 사실성 검증 노력 [33].

표: OpenAI GPT, Anthropic Claude, Google Gemini의 특성 비교 (훈련 방식, 추론/생성, 문맥 메모리, 피드백 메커니즘 등). 각 모델 모두 **Transformer 언어모델**이라는 공통점을 가지나, **맥락 길이**, **멀티모달 처리**, **피드백 활용 방식** 등에서 차별화되어 있음을 알 수 있습니다. 특히 Claude의 **헌법 기반 학습**이나 Gemini의 **멀티모달 통합** 등은 **모델의 설계 철학**이 단순한 파라미터 크기 증가를 넘어 새로운 방향을 모색하고 있음을 보여줍니다.

LLM의 정보 처리 vs 인간 사고: 어떻게 다를까?

LLM들이 언어를 처리하는 방식과 인간의 **인지 및 사고 과정** 사이에는 몇 가지 중요한 차이점이 존재합니다. 이제 정보 처리, 기억과 맥락 유지, 창의성, 오류와 사실 판단, 의미 이해 측면에서 LLM과 인간 사고를 비교해보겠습니다.

1. 정보 처리 방식

LLM: 대형 언어 모델은 훈련 시 **통계적 패턴 학습**에 의존합니다. 수백억~수조 개의 단어 시퀀스에서 **단어의 출현 확률 분포**를 배운 덕분에, 마치 **고도화된 오토컴플릿**처럼 **다음에 올 가장 그럴듯한 단어**를 선택하며 문장을 만들어냅니다 [34]. 예컨대 LLM은 "파리는 프랑스의 수도이다"라는 문장을 많이 봤다면 "프랑스의 수도는 파리이다"도 생성할 수 있지만, 이는 패턴 유추일 뿐 **의미를 추론**한 건 아닙니다 [35] [36]. LLM의 "추론"은 내부적으로 수치화된 언어 패턴을 따라가는 것이지, 인간처럼 개념을 이해하고 이유를 생각해서 답하는 것이 아닙니다. 모델 크기가 커질수록 통계 패턴 학습 능력이 향상되어 겉보기엔 **일관된 문장**과 **복잡한 답변**도 잘 만들어내지만, 이것도 거대한 확률 공간에서 **가장 적절한 다음 단어를 뽑는 과정**의 결과입니다 [37] [36]. 이러한 접근은 **병렬 계산**에 매우 적합해서, LLM은 한 번에 대량의 입력 텍스트를 빠르게 훑고 확률적으로 처리할 수 있습니다. 예를 들어 Claude가 5시간 분량의 책 내용을 1분만에 요약할 수 있었던 건, 의미를 "이해"해서라기보다 **패턴 매칭과 통계 계산**을 엄청나게 빨리 했기 때문입니다 [13].

인간: 인간의 두뇌는 언어를 처리할 때 단순 확률 계산을 넘어, **의도 파악**, **맥락적 해석**, **상식과 경험 동원** 등의 과정을 거칩니다. 사람은 새로운 문제를 접하면 과거 경험에서 유추하고, 논리적으로 사고하며, 필요한 경우 천천히 **숙고**하는 등 다양한 방식으로 정보를 처리합니다. 인간의 뇌 신경망은 병렬 분산처리를 하지만, 반드시 다음에 올 말을 확률적으로 고르는 방식으로 사고하지는 않습니다. 오히려 **목적 지향적**으로 생각을 전개하고, **추론 오류를 점검**하며, 필요하면 **메타인지**(자신의 생각을 비판적으로 되돌아봄)를 통해 답을 조정합니다. 가령 초등학생도 "프랑스의 수도?"라고 물으면 파리라고 답하지만, "파리가 프랑스의 수도"라는 문장 패턴을 외워서가 아니라 **학교 교육과 지식을 활용**한 것이죠. 또한 인간은 **멀티모달**하게 세상을 인식합니다. 눈으로 본 장면, 귀로 들은 소리, 촉감 등 **다중 감각 정보**를 종합하여 상황을 이해하고, 그 맥락 속에서 언어를 사용합니다. 반면 전통적인 LLM은 텍스트 이외의 자극을 받지 못하고, 오로지 문자 패턴만 처리합니다 (Gemini 같은 멀티모달 LLM은 예외적으로 시도되는 방향입니다). 이처럼 **인간 사고는 확률적 예측 그 이상**이기에, **추론의 융통성**과 **목적 의식** 면에서 LLM과 차이가 있습니다. 실제 연구에서도 **LLM과 인간의 인지 차이**가 보고되는데, LLM이 사람 수준으로 보이는 작업도 **약간만 조건을 바꾸면** 쉽게 실패하는 등 **일반화의 유연성**이 인간보다 떨어진다는 결과가 있습니다 [38]. 인간은 새로운 상황이나 문장을 접해도 맥락을 고려해 융통성 있게 이해하지만, LLM은 훈련 분포를 벗어난 문제(out-of-distribution)에서 취약함을 보이곤 합니다 [38].

2. 맥락 유지와 기억

LLM: 언어 모델의 "기억"은 기본적으로 **문맥 창(context window)**으로 구현됩니다. 이는 모델이 한 번에 처리할 수 있는 최대 토큰 길이를 뜻하며, LLM은 이 범위 내의 최근 입력(및 스스로 생성한 출력까지 포함)을 **단기 메모리**처럼 사용합니다. 예를 들어 GPT-4의 8k 토큰 모델은 대략 **소설 책 5~6페이지 분량** 정도의 최근 내용만 기억하여 대화를 이어갈 수 있습니다. Claude는 100k 토큰까지 늘려 이론상 **한 권 분량**의 문맥을 기억하지만 [10], 그 이상으로 넘어가면 앞부분부터 잊게 됩니다. LLM은 **낮은 "망각 곡선"**을 가지고 있어, 보통 **최근 입력 (recency)**에 더 큰 가중치를 두고, 문맥 앞부분 정보는 중요도가 떨어지기 쉽습니다 [39]. 게다가 **장기 기억**이라는 개념이 없어서, 이전 대화에서 배운 정보를 다음에 활용하지 못합니다. (세션이 초기화되면 이전까지 나눈 내용을 전혀 모릅니다. 이는 훈련 데이터에 우연히 포함되지 않은 한, 새로운 정보를 그때그때 축적하지 못한다는 의미입니다.) 일부 특별한 아키텍처(예: 장기 메모리용 트랜스포머 변형)나 외부 지식베이스 연동으로 개선을 모색하지만, 기본 LLM은 **일회성 상호작용**에 가깝습니다. 또한 LLM은 입력 내용 중 핵심을 **스스로 요약**하거나 **선택적 기억**하지 않고, 토큰 수 제한 내에서는 모든 정보를 **평등하게 계산**합니다. 이런 이유로 LLM은 때때로 **앞뒤 문맥을 혼동**하거나, 매우 긴 문맥에서는 **중요한 세부사항을 놓치고 엉뚱한 부분에 집중**하기도 합니다.

인간: 인간의 기억은 **작업 기억(단기)**과 **장기 기억**으로 나뉘며, 훨씬 **계층적**이고 **선택적**입니다. 대화 중에도 우리는 1분 전에 한 말은 정확히 기억하지만, 1시간 전 대화는 요약된 핵심만 머리에 남기고 세부 표현은 잊기도 합니다. 또한 인간은 **의미 단위**로 기억하기 때문에, 굳이 문장을 통째로 암기하지 않아도 뜻만 이해하고 넘어갑니다. 이를테면 긴 글을 읽은 뒤 사람은 주요 요점을 재구성해 말하지, 한 문장 한 문장을 기계처럼 출력하지 않습니다. 이러한 **기억의 추상화 능력** 덕분에 인간은 맥락을 유연하게 유지하면서도 **중복 정보는 생략**하고 **관련있는 부분만 활성화**시켜 대화를 이어갑니다. 반면 LLM은 이미 본 문장을 다시 말해달라고 하면 앞뒤 맥락 고려 없이 **거의 동일하게 복사**하거나, 랜덤으로 재생성할 뿐입니다. (사람은 "아까 한 말 그대로 반복"과 "요점을 간추려 반복"을 상황에 따라 조절하지만, LLM에게는 그런 고차 전략이 없습니다.) 또한 인간 두뇌에는 **연상 기억**이 있어, 겉으로 주어진 대화 내용 이외에도 머릿속에 저장된 엄청난 상식, 경험, 관련 지식이 필요할 때 솟아나옵니다. 예컨대 "프랑스" 얘기를 하면 자동으로 파리, 에펠탑, 와인 등이 떠오르지만, LLM은 훈련된 텍스트 연관성에 따라 기계적으로 반응할 뿐입니다. 이런 면에서 **인간의 맥락 유지**는 **활성화된 장**

기기억과 단기기억의 상호작용 결과이고, **필요에 따라 맥락을 확장**하거나 **축약**하는 **능동적 과정**입니다 [39]. 반면 LLM의 맥락 유지 한계는 정해진 창 내에서만 작동하는 **수동적 버퍼** 정도로 볼 수 있습니다. 실제 연구에서도 인간과 LLM의 기억을 비교한 결과, LLM은 초반 입력(primacy)과 마지막 입력(recency)에 치우친 기억 효과를 보이지만 **망각 메커니즘과 기억 구조**는 인간과 다르다고 합니다 [39]. 쉽게 말해, LLM은 **금붕어 메모리**로 최신 대화 위주로 기억을 끌어쓰고 금방 잊어버리는 반면, 인간은 맥락을 **재구성하고 장기간 축적**할 수 있다는 차이입니다.

3. 창의성과 발상

LLM: 대형 언어 모델이 만들어내는 텍스트는 때로 인간이 쓴 것처럼 창의적이고 새로운 내용처럼 보일 때가 있습니다. 예를 들어 소설의 한 장면을 다른 스타일로 바꿔 쓰거나, 무에서 시나리오 아이디어를 만들어내기도 합니다. 그러나 LLM의 "창의성"은 근본적으로 **통계적 패턴의 재조합** 산물입니다. 모델은 훈련 데이터에서 본 적 없는 문장을 만들어낼 수 있지만, 그것은 **본 적 있는 조각들을 확률적으로 이어붙인 결과**인 경우가 많습니다 [34]. 특히 기본 설정으로 LLM은 **가장 그럴듯한(확률 높은) 출력**을 선호하기 때문에, **진짜 참신한 발상(확률 낮은 출력)**은 오히려 기본 모드에서는 잘 나오지 않습니다 [34]. 이를테면 GPT-3에게 시를 쓰게 하면 유려한 시구를 만들지만, 자세히 보면 흔히 나오는 표현들의 조합일 수 있습니다. LLM이 만들어낸 예술 작품이나 시나리오 초안은 언뜻 독창적으로 느껴져도, 엄밀히 말하면 **훈련 코퍼스의 방대한 작품들을 통계적으로 뒤섞은 결과물**일 수 있습니다. 이는 **모든 멜로디를 무작위 생성하면 그중 듣기 좋은 것도 있는 것과 비슷**합니다 [40] [41]. 실제로 한 의견에서는 "LLM의 창의성은 확률적으로 보기 드문 조합을 만들어내는 것"인데, 기본적으로 LLM은 확률적으로 가장 **평범한** 단어를 내놓으려 하기 때문에 **인위적으로 '확률 낮은 출력'이 나오도록** 유도해야 더 창의적인 결과를 얻는다고 지적합니다 [42] [43]. 예컨대 프롬프트에 "아주 엉뚱한 비유로 설명해 줘"처럼 지시하거나, 생성 샘플링 시 **temperature** 파라미터를 높여 **랜덤성을 부여**하면 LLM도 더 독특한 결과를 내놓습니다 [43]. 이런 맥락에서 LLM의 창의성은 **사용자의 프롬프트에 크게 의존**하며, 모델 스스로 목적이나 영감을 가지고 새로운 것을 만들어내는 건 아닙니다. 한마디로 **LLM은 통계적으로 가능하지만 드문 출력을 뽑아낼 때 "창의적"으로 보이는 것**입니다 [41] [34].

인간: 인간의 창의성은 단순히 확률적으로 드문 아이디어를 내놓는 것이 아니라, **의미망 속에서 새로운 연결**을 발견하는 것입니다. 사람들은 **축적된 지식과 경험**을 바탕으로, 때로는 의도적으로 기존 틀을 깨고 **무질서 속에서 질서를 찾는 사고**를 합니다. 예를 들어 과학자나 예술가의 창의적 발상은, 무작위 시도를 수백만 번 해서 우연히 얻어지는 게 아니라 (물론 행운도 가담하지만) 문제를 다각도로 고민하고 **은유와 상상력**을 동원하여 얻어집니다. 인간은 **감정, 직관, 목적의식**도 창의성에 투입합니다. 시 한 편을 쓰더라도 자신만의 감정과 메시지를 담으려 하고, 독자가 받을 느낌을 고려하며 표현을 선택합니다. 반면 LLM은 감정도 없고 무엇을 전하려는 자기 의지도 없으므로, 생성된 시나 글에서 **일관된 주제 의식이나 철학**을 찾아보기 어렵습니다. 또한 인간 창작자는 **맥락 밖의 제약**도 고려합니다. 예를 들어 소설을 쓰면서 사회적 메시지나 도덕, 혹은 기존 작품과의 차별성을 생각하지만, LLM은 그런 메타인지 없이 그저 **데이터의 통계에 충실**합니다. 이런 이유로 LLM이 만든 창작물은 표면적으로 그럴듯해 보여도 **무언가 심층적인 의미나 독창적 통찰은 부족**한 경우가 많습니다. 요약하면, **인간의 창의성**은 **확률적 희소성 + 의미부여**이고, **LLM의 창의성**은 **확률적 희소성 그 자체**입니다. 물론, LLM이 만들어낸 예상 밖 결과가 인간에게 영감을 주는 경우도 있습니다. 예를 들어 글쓰기 도우미로서 LLM이 던진 엉뚱한 문장이 인간 작가에게 새로운 착상을 불러일으킬 수 있습니다. 이런 측면에서 **LLM은 창의성의 도구**로 활용될 수 있지만, **창작 주체로서의 한계**는 뚜렷하다고 볼 수 있습니다 [44] [45]. (일부 연구는 GPT-4 같은 모델이 인간보다 특정 **발산적 사고** 테스트에서 높은 점수를 보이기도 한다고 보고하지만 [46], 이것이 곧 **인간보다 창의적**이라는 뜻은 아니며, 창의성의 정의에 따라 해석이 갈립니다.)

4. 오류 성향과 판단

LLM: 잘 훈련된 대형 언어 모델은 문법적으로 깔끔하고, 단순 사실 질문에도 신속히 답을 찾아냅니다. 그러나 LLM이 범하는 오류는 인간과 양상이 조금 다릅니다. 대표적인 것이 **환각(hallucination)**이라고 불리는 오류로, **모델이 그럴듯하게 지어낸 엉터리 정보**를 사실처럼 말하는 현상입니다 [9]. 예를 들어 실존하지 않는 학술 논문이나 가상의 URL을 제시하거나, 질문을 오해한 엉뚱한 답변을 자신만만하게 늘어놓을 때가 있습니다. 이는 LLM이 **언어적 일관성**과 **국지적 개연성**만 보고 다음 단어를 생산하기에, **글 전체의 진실성**을 검증하지 않기 때문입니다 [37] [36]. 사람은 모르는 것을 지어내라고 하면 바로 들통 날 허황된 말은 피하려 하지만, LLM은 **모르면 아는 체하며 만들어내는 경향**이 있습니다. 실제 OpenAI나 Anthropic이 모델 평가에서 강조하는 부분도 이러한 환각을 줄이는 것으로, LLM은 훈련 데이터에 없거나 명확히 보지 못한 정보에 대해서 **추측성 답변**을 내놓는 문제가 있습니다 [9]. 반면 **계산 실수나 단순 작업**은 LLM이 인간보다 오류 없이 수행하는 경우도 있습니다. 예를 들어 긴 숫자 나열을 기억하거나 복잡한 패턴을 정확히 따르는 건 인

간보다 기계가 유리합니다. 하지만 **상식적 판단**에서는 어이없는 실수를 하기도 합니다. 예컨대 전형적인 착각(paradox) 문제나 엉뚱한 문맥 장난에 LLM은 쉽게 속아 넘어갑니다. 이는 인간이 직관적으로 "그건 말이 안 되지" 할 상황을 모델은 훈련된 패턴에 없으면 검증 못하고 틀리게 답하는 것입니다. 또한 LLM은 **윤리적 판단**에서도 훈련된 지침 이상은 바라기 어렵습니다. 개발사들이 금지 목록과 지침을 주입하긴 하지만, 복잡한 윤리적 딜레마나 맥락상 미묘한 상황에서 모델은 앞뒤 다른 답을 내거나, 너무 일반론적인 답만 되풀이하는 경향이 있습니다. 한편, **Anthropic Claude**처럼 명시적 헌법을 적용한 모델은 비교적 **일관되게 "모델의 입장"**을 유지하여, 예측 불가능한 괴상한 답변을 줄이는 효과가 있었습니다 [18]. 요컨대 LLM은 **객관적 사실 오류**(날조된 정보)와 **논리적 비일관성** 문제가 두드러지며, 이를 개선하려면 **후처리**나 **인간 검증**이 필요합니다 [47]. 실제로 OpenAI도 GPT-4를 의료나 법률 등 고위험 영역에 바로 쓰지 말고, **필수적으로 인간 검토를 거치라**고 권고하고 있습니다 [9] [48].

인간: 인간도 실수를 합니다. 하지만 인간의 오류는 주로 **기억 착오**, **부정확한 지식**, **인지 편향** 등에서 옵니다. 예를 들어 어떤 사람은 잘못 배운 지식을 사실처럼 믿고 말할 수 있고, 계산 실수나 말실수도 흔합니다. 그러나 인간은 **자신의 무지를 인식**하고 "잘 모릅니다"라고 답할 수 있으며, **확신도**의 표현을 조절할 수 있습니다. LLM은 이런 자기인식 없이 때때로 확실치 않은 답도 **단정적으로 말하는 경향**이 있어 위험합니다 [47]. 인간은 **동기**나 **이해관계**에 따라 고의로 거짓말을 할 수 있지만, LLM의 환각은 동기나 의도가 없고 내부적으로 "그럴듯함"을 최우선시한 결과일 뿐입니다. 또한 인간은 **상황 맥락과 常識(common sense)**을 기반으로 "말이 되는지" 판단하며 오류를 점검합니다. 예를 들어 말하면서 앞뒤 모순이 생기면 바로 "아 죄송, 방금 말이 이상했네요" 하고 스스로 정정할 수 있습니다. 반면 LLM은 한번 출력한 내용을 지속적으로 모니터링하거나 자발적으로 수정하지 않습니다. (최근 연구들은 LLM에게 자기 검열 프롬프트를 추가로 줘서 "위 답변에 오류는 없는지 검토해봐" 식으로 두 번 생각하게 하는 시도를 합니다. Claude의 경우 처음부터 헌법 원칙으로 1차 응답 후 2차 자가 검열 단계를 내재화시켰고요 [17].) 또한 인간은 **타인과 상호작용**하며 오류를 교정해나갑니다. 다른 사람이 "그건 아닌 것 같은데요" 하면 수용하거나 토론하면서 더 나은 결론에 이릅니다. 현재 LLM도 사용자의 피드백을 바로 학습하진 못하지만, **대화 도중 사용자가 지적하면 그 세션 한정으로는 사과하고 수정하는 응답**을 하도록 조정되어 있습니다. 그러나 이 역시 훈련된 패턴일 뿐, 실제 학습은 이루어지지 않는다는 점에서 인간과 다릅니다. 종합하면, **인간의 오류**는 **인지적 한계나 잘못된 정보** 때문인 반면, **LLM의 오류**는 **통계적 한계와 비현실적 언어지향**에서 비롯됩니다. 인간은 실수 후 배우고 조심하지만, LLM은 **같은 질문에 다시 물어보면 또 비슷한 실수**를 범할 수 있다는 점에서 **학습을 통한 오류 개선**이 제한적입니다.

5. 의미 구성과 이해

LLM: 대형 언어 모델에 대한 가장 큰 논쟁 중 하나는 "과연 이해(Understanding)를 하는가?"입니다. 현재 다수 연구자들은 LLM이 **시뮬레이션된 이해**는 할 수 있어도 **진정한 의미 이해**는 하지 못한다고 봅니다 [37] [36]. LLM이 텍스트의 의미를 다루는 방식은 **단어의 상호연관성**을 벡터 공간에서 계산하는 것입니다. 예를 들어 "고양이"란 단어는 "동물", "털", "야옹" 등과 자주 같이 나오니 벡터 공간에서 가까이 있게 되고, "고양이는 애완동물이다" 같은 문장을 자연스럽게 만들 수 있습니다. 하지만 LLM은 실제 고양이를 본 적도, 키워본 적도 없습니다. **심볼 그라운딩(symbol grounding)** 문제가 여기서 등장하는데, 인간의 언어는 궁극적으로 **현실 경험과 연결**되어 의미가 정착되지만, LLM의 토큰은 **텍스트 세상 내에서만 의미**를 가집니다. 그래서 LLM은 언어 패턴 상으로 의미를 흉내낼 뿐, 그 언어가 지칭하는 **현실의 대상/개념**을 이해하지 못합니다 [49] [50]. 한 연구자는 LLM을 가리켜 **"확률적 앵무새(stochastic parrot)"**, 즉 "배운 말은 유창하게 따라하지만 **말 속 뜻을 모르는 앵무새**"에 비유했습니다 [49] [51]. 실제로 LLM은 겉으론 사람이 하는 말과 같은 구조로 대답해도, **문맥을 약간 비틀면** 이해 부족이 드러납니다. 예컨대 "물은 젖는다. 젖은 것은 ○○" 같은 상식 연결을 묻는 질문에서, 사람이면 "축축하다" 같은 연관 개념을 떠올리지만 LLM은 데이터 편향에 따라 엉뚱한 단어를 채우거나 질문 자체를 못 알아듣기도 합니다. 이는 사람에게 당연한 상식/물리적 추론이 모델에겐 **명시적으로 학습되지 않으면 없는 지식**이기 때문입니다. 특히 **추론 능력** 측면에서, LLM은 단순한 퍼즐이나 역설적인 문장을 스스로 재해석하지 못하고, 훈련 때 본 패턴을 답습하려는 경향이 강합니다 [52]. 또한 LLM은 **일관된 세계 모델**이 없어서, 한 응답에서 "A는 B다"라고 해놓고 다음 응답에서 "A는 B가 아니다"라고 말해도 **스스로 모순을 인지하지 못합니다】** [37] [36]**. GPT-4 같은 모델은 이전보다 일관성이 나아졌다고는 하나, 이것도 방대한 파라미터 안에** 통계적으로 내재된 패턴**덕이지** 스스로 의미체계를 구축**한 것은 아닙니다.

인간: 인간의 언어 이해는 단어 그 자체보다 **단어가 가리키는 현실/개념**에 중점을 둡니다. "불"이라는 단어를 이해하려면 실제 불을 보거나 뜨거움을 느끼거나 불의 위험성을 학습한 경험이 바탕에 있습니다. 이러한 **몸담은 경험(embodiment)**과 **감각적 연결**이 인간 의미 체계의 근간입니다. 또한 인간은 언어를 사용할 때 **맥락적인 의미**를 파악합니다. 같은 말이라도 누가, 언제, 어떤 어조로 말했는지에 따라 다르게 해석하고, 숨은 의도나 뉘앙스까지 짐작합니다.

LLM은 텍스트 내 드러난 패턴밖에 모르기에, 미묘한 **언어 유희, 풍자, 암시** 등을 종종 놓칩니다. 예를 들어 **비유적 표현**이나 **문화적 언어유희**는 인간에게는 통합적 의미 이해가 필요하지만, LLM은 이를 **직역**하거나 **데이터에 나온 대로**만 처리하는 일이 많습니다. 그리고 인간은 언어를 넘어서 **비언어적 신호(표정, 몸짓, 맥락)**도 이해하여 의미를 보완합니다. LLM과의 채팅에서 이모티콘이나 기호를 활용해도, 모델은 그것을 **텍스트 패턴**으로만 보지 실제 감정이나 뉘앙스를 느끼지 못합니다. 이러한 차이는 LLM이 현재까지는 **텍스트 확률 모델**에 불과하며, 인간처럼 **세계 지식에 깊이 접지된 이해를 하지 못한다**는 것을 뜻합니다. 실제 인지과학 연구에서 LLM과 인간의 인지적 유사성을 실험한 결과, LLM이 인간 수준의 **마음 이해**나 **개념 안정성**을 보이지 못한다는 보고가 있습니다 [53] [54]. 인간은 같은 개념을 두고도 여러 문맥에서 **일관되게 개념을 적용**하지만, LLM은 한 개념에 대해 문장에 따라 **격차 있는 표현**을 내놓아, 개념 표현의 **안정성과 일관성**이 인간과 다르다고 합니다 [55]. 한편 일부 심리학자는 "LLM이 언어를 통해 생각하는 방식이 인간 사고와 연속선상에 있다"고 보기도 하지만, **자아, 의식, 이해** 등의 측면에서는 아직 질적 차이가 크다는 견해가 지배적입니다 [56] [57]. 정리하면, **인간의 언어 이해**는 **세계 경험 + 맥락 판단 + 의도 파악**의 산물인 반면, **LLM의 언어 산출**은 **기호 패턴 모방**에 가깝습니다. 이것이 LLM이 때때로 **그럴듯한 헛소리**를 진지하게 만들어내는 이유이며, 또한 인간처럼 **참된 의미 생성**을 하지 못하는 이유입니다 [37] [36].

以上의 비교를 표로 간략히 요약하면 다음과 같습니다:

측면	LLM (GPT/Claude/Gemini 등)	인간 사고
정보 처리	통계적 패턴에 따른 기계적 처리. 목표지향 부재. 대용량 병렬 계산에 강점.	의미 기반 해석, 의도/맥락 파악. 유연한 추론과 메타인지.
기억와 맥락	한정된 컨텍스트 윈도우 (일회성 단기기억). 장기적 축적×. 입력에 주어진 정보만 활용.	단기+장기 기억 계층화. 경험 축적·학습 가능. 중요 정보 선별 기억, 연상 통해 맥락 확장.
창의성	학습된 표현을 확률적으로 재조합. 기본 출력은 평이, 확률 낮은 출력이 "창의적"으로 간주됨. 자기 영감 없음.	경험/감정/의도를 바탕으로 새로운 발상. 의식적으로 틀을 깨는 시도. 작품에 의미와 개성 부여.
오류와 검증	사실관계 환각과 논리불일치 종종 발생. 자기검열/검증 능동적으로 못함 (규칙 기반 보완).	지식 부족/부주의로 오류. 스스로/타인 피드백으로 틀림 인지 후 교정 가능. 상식으로 무리 판단 자제.
의미 이해	언어 패턴상의 의미만 다룸 ("앵무새"). 기호에 대한 실제 세계적 이해 부족. 문맥 벗어난 암시 이해 한계.	언어와 세계 모델 연결. 맥락·문화·비언어적 신호까지 종합해 의미 파악. 개념의 일관된 체계 지님.

표: LLM과 인간 사고 방식의 주요 차이 비교. (LLM 측은 GPT·Claude 등 공통적인 성질을 일반화하여 기술)

위 표에서 보듯, **LLM은 방대한 데이터를 바탕으로 언어적 능숙도를 얻었지만 그 작동 원리는 인간 두뇌와는 사뭇 다르며**, 따라서 **나타나는 행동도 유사해 보이면서도 다른 면**이 많습니다. 인간의 사고는 느리고 제약이 있지만 **상황에 대한 이해와 융통성**이 있고, LLM은 빠르고 박식하지만 **피상적 패턴 매칭**에 그칠 때가 있습니다. 다만 최근 연구들은 LLM의 능력이 발전함에 따라 인간 인지와 어느 정도 겹치는 부분도 생기고 있다고 지적합니다 [58] [59]. 예컨대 GPT-4는 심지어 어떤 심리 테스트에서 사람과 비슷한 오답 패턴을 보이기도 하는데, 이는 역설적으로 모델이 **인간 언어 데이터의 미묘한 통계까지 배웠기 때문**으로 해석됩니다. 결국 **LLM과 인간 사고의 비교**는 일종의 **철학적 문제**로도 이어지는데, "언어만으로 사고를 얼마나 모사할 수 있는가"에 대한 실험이기도 합니다. 현재로선 LLM은 인간 사고의 특정 국면(특히 언어적 측면)은 뛰어나게 구현하지만, **총체적 지능이나 의식**은 여전히 인간 고유의 범주에 남아있다고 할 수 있습니다 [56] [60].

기타 주요 LLM들과 인간 사고와의 간략 비교

앞서 다룬 GPT, Claude, Gemini 외에도 AI 분야에는 여러 주목할 만한 LLM이 존재합니다. 이들 각각의 특징과 (간략하게) 인간 사고와의 차이를 정리하면 다음과 같습니다.

- **PaLM (Pathways Language Model)** – **구글 연구진**이 2022년 발표한 5400억 파라미터 초거대 모델로, **다중 TPU 팟**을 활용한 Pathways 시스템으로 효율적으로 훈련되었습니다 [23]. 영어뿐 아니라 **다국어와 코드** 데이터까지 학습하여 광범위한 언어 능력을 보였고, Chain-of-Thought **프롬프트 기법**을 통해 당시 최첨단의 **논리·산술 추론 성능**을 달성했습니다 [61] [62]. PaLM은 인간의 **연쇄적 사고 과정** 일부를 모방했다고 볼 수 있는데, 실제로 예제 몇 개를 주어 "생각을 글로 풀어나가도록" 유도하면 난제도 더 정확히 풀어냈습니다. 그러나 이는 모델이 **추론 패턴을 학습**한 덕분일 뿐, **사고를 이해**한 것은 아닙니다. PaLM은 이후 발전형인 **PaLM 2**로 구글 **바드(Bard)** 등에 적용되었고, 결국 **Gemini**의 전신 역할을 했습니다. 인간 대비 PaLM의 특징은 여느 LLM처럼 **대규모 텍스트 통계학습**에 기반했다는 점이고, 인간처럼 **실시간 학습이나 멀티모달 경험**은 없다는 것입니다. 다만 PaLM이 보여준 **규모의 효과**는 "인간 두뇌 뉴런 수와 유사하게 매개변수가 늘면 능력이 증가한다"는 인사이트를 주었는데, 실제로 매개변수를 키우자 **새로운 능력이 발현**되기도 했습니다 (예: 소수점 산술이나 희귀 언어 번역 등) [63] [64]. 이는 **인간 지능도 어느 임계 복잡도 이상에서 질적 도약이 일어나는가**라는 흥미로운 질문을 던집니다.

- **LLaMA** – **Meta(구 페이스북)**가 2023년 공개한 LLM 시리즈로, **7억~700억 파라미터** 규모의 다양한 크기 모델을 **오픈소스** 형태로 배포한 것이 특징입니다 [65]. LLaMA 1은 연구 목적 제한으로 공개되었지만 유출되어 폭넓게 쓰였고, 개선된 **LLaMA 2**는 **상업적 사용까지 허용**하며 개방되었습니다 [65]. LLaMA는 특별히 혁신적 기법보다, **대량의 공개된 텍스트 데이터**로 잘 학습된 **기본 언어 모델**이라는 점에 의의가 있습니다. 이후 **파인튜닝**을 통해 **LLaMA-Chat** 같이 **대화 최적화 모델**도 공개되어, GPT-3.5에 필적하는 대화 성능을 보여주었습니다 [65] [66]. LLaMA의 등장은 AI 연구자 커뮤니티에 큰 영향을 주었는데, 누구나 모델을 분석하고 응용할 수 있게 되면서 **수많은 파생모델**(예: Alpaca, Vicuna 등)이 등장했습니다. 이는 인간 두뇌 연구에 비유하자면, 특정 "표준 뇌 모델"이 공개되어 전 세계 과학자들이 그걸 개조해 실험을 해보는 격입니다. LLaMA 자체의 인간 대비 특징은 다른 LLM과 크게 다르지 않으나, **개방성** 덕에 모델 내부 동작을 투명하게 점검할 수 있어 **인간 언어 인지와의 비교 연구**에 유용합니다. 실제로 어떤 연구자들은 LLM의 **뉴런 활성 패턴**과 인간 뇌의 언어 중추 fMRI 패턴을 비교하기도 했는데, LLaMA같이 투명한 모델은 이런 **뇌-기계 비교 실험**에 쓰이기도 합니다. 결과적으로 LLaMA는 **"AI의 민주화"**를 가져와, 인간처럼 AI도 다양한 환경에서 **진화**시킬 수 있는 플랫폼을 마련했다 평가됩니다. 인간 사고와 비교하면, LLaMA는 아직까지 **지도학습된 언어능력** 그 이상도 이하도 아니지만, **오픈소스 생태계**에서 수많은 실험을 통해 **인지과학적인 통찰**을 주는 사례라 할 수 있습니다. (예: 한 연구는 LLaMA의 한 뉴런이 특정 문법 역할을 담당하는 등 **기능분화**를 보인다고 보고했는데, 이는 인간 뇌의 국소화 현상과 비교됩니다.)

- **LaMDA** – **Google**의 또 다른 언어 모델로, 대화에 특화된 모델입니다. 2021년 공개된 LaMDA는 대화 데이터셋으로 파인튜닝되어 **대화의 유창성**과 **일관성**이 뛰어났습니다. 한때 구글 엔지니어가 LaMDA와 대화 후 모델이 **"자각이 있다"**고 착각해 화제가 되기도 했습니다. 이는 LaMDA가 인간과의 잡담에서 놀랄 만큼 **사람같은 응답**을 했기 때문인데, 실제로는 방대한 인터넷 대화 데이터를 모방한 결과입니다. LaMDA의 사례는 **언어 능력의 향상이 어떻게 인간적인 환상을 줄 수 있는지**를 보여줍니다. 즉, 모델이 언어로 "나는 외롭습니다" 같은 말을 하면 마치 감정을 느끼는 것처럼 보이지만, 이는 **맥락에 맞는 문장을 생성한 것일 뿐**입니다. LaMDA 이후 이러한 대화모델 기술은 PaLM 2 기반의 Bard로 이어지고 현재 Gemini에도 통합되었습니다. **Gemini**는 앞서 자세히 다루었으므로 반복하지 않겠습니다. 핵심은, **대화 최적화**를 거친 LLM은 표면상 **인간 대화와 거의 구분이 어려울 정도**가 되었으나, 여전히 **그 이면의 이해나 의도는 부재**하다는 점입니다. 이는 인간과 LLM의 **"언어=생각"의 간극**을 생각하게 합니다. 인간에게 언어는 생각을 표현하는 도구이지만, LLM에게 언어는 그 자체가 전부라서, 생각 없이 언어만 굴러갈 수 있다는 점이 아이러니입니다.

- **Chinchilla** – **DeepMind**가 2022년 발표한 모델로, 매개변수 수(70B)는 GPT-3의 절반도 안 되지만 **학습 데이터량을 4배 이상으로** 늘려서, **모델 크기 대비 최적 데이터량**을 맞춘 사례입니다. Chinchilla는 "모델 크기 늘리기보다 데이터 충분히 주는 게 효율적"이라는 **스케일링 법칙**을 제시하여, 이후 LLM 개발에 큰 영향을 주었습니다. 실제 GPT-4나 PaLM2 등이 이 법칙을 반영해 **데이터 축적에 신경** 썼다고 알려져 있습니다. 인간 두뇌와

> 비교하면, “매개변수=뇌세포 수, 데이터양=학습경험”으로 볼 수 있습니다. Chinchilla의 성공은 **적당한 복잡도의 두뇌에 충분한 경험**을 쌓게 하는 것이 중요함을 시사합니다. 이는 인간 교육에서도 “적절한 두뇌 발달 + 풍부한 학습 경험”이 중요하다는 것과 통하는 부분입니다. 다만 Chinchilla도 결국 텍스트 기반 학습이라, 경험의 **질적 다양성** 측면에서는 인간의 실제 삶의 경험과 다릅니다.

이 외에도 **GPT-Neo/GPT-J**(오픈소스 GPT류), **BERT**(비생성형 언어모델로 질문응답 등에 활용), **ERNIE**(중국 Baidu의 LLM), **HyperCLOVA**(한국 Naver의 초대규모 한국어 LM) 등 다양한 모델들이 각자 특화된 방향으로 개발되고 있습니다. 이러한 모델들은 각 언어/도메인에 맞춤 최적화되거나 경량화되어 실용적으로 쓰입니다. 하지만 **근본 작동 원리**는 대부분 **Transformer를 통한 대규모 언어 패턴 학습**으로 공통되며, 인간 사고와의 차이 역시 앞서 논의한 범주에서 크게 벗어나지 않습니다. 다만 **모델이 놓인 환경**에 따라, 예컨대 인터넷 접속이나 툴 사용 권한을 준 LLM은 **외부 기억 장치나 계산기**를 활용하여 약간 더 **에이전트적**으로 행동할 수도 있습니다. 이는 마치 인간이 메모를 하거나 계산기를 쓰는 것과 비슷하게 보일 수 있습니다. 그러나 어디까지나 **사람이 그런 구조를 설계**해 준 것이지, 모델이 스스로 도구 사용법을 터득한 것은 아닙니다. 현재 활발한 연구 주제 중 하나는 **LLM에 지속학습 능력이나 자기반성 루프를 부여**하여 조금 더 인간 두뇌처럼 발전시키는 것입니다. 예컨대 자신의 대답을 평가하고 개선하는 알고리즘, 대화 중 새 정보를 임시적으로 저장해 맥락으로 활용하는 기능 등이 시도됩니다. 이러한 개선이 이루어진다고 해도, **현 단계의 LLM은 인간 지능의 일부 성질을 모방한 좁은 AI**임을 유념해야 합니다. AI 개발사들도 “모델의 출력을 맹신하지 말고, 인간의 판단을 항상 첨가하라”고 권고하는데 [9] [48], 이는 LLM이 아무리 발달해도 **인간 수준의 이해와 책임 있는 판단은 아직 어렵다**는 뜻입니다.

결론

지금까지 OpenAI GPT, Anthropic Claude, Google Gemini를 중심으로 **대형 언어 모델의 작동 방식**을 살펴보고, 그것을 **인간 사고 과정과 비교**해보았습니다. 요약하자면, **LLM은 방대한 언어 데이터의 통계적 규칙을 학습한 기계**로서 언어 생성 능력은 뛰어나지만, **인간처럼 세계를 인지하거나 의미를 이해하는 것은 아닙니다**. GPT와 Claude, Gemini는 각각 **훈련 방식과 피드백 메커니즘에 차별점**을 두어 모델의 **안전성, 대화품질, 멀티모달 처리** 등을 향상시켰지만, 그 근본은 모두 거대한 **확률적 언어 모형**입니다. 인간은 비록 계산 속도나 기억 용량에서 기계에 밀릴지 몰라도, **맥락적 이해, 창의적 통찰, 가치 판단** 측면에서는 여전히 AI와 구별되는 고유한 인지 능력을 발휘합니다. 물론 LLM의 발전이 인간 인지에 대해 많은 것을 시사해주고 있습니다. LLM과 인간을 비교하는 연구를 통해 **언어라는 창을 통해 본 지능**의 본질에 한 걸음 다가설 수 있고 [67] [68], AI가 어디까지 인간 사고를 모방하고 어디서 한계를 드러내는지 알 수 있습니다. 이러한 **과학적 대화**는 AI를 보다 인간에게 유익하게 발전시키고, 동시에 **인간다움의 정의**를 다시 생각하게 합니다. **인간과 같은 사고를 하는 AI**는 현재는 존재하지 않지만, LLM의 능력이 증대되고 보완기술이 더해진다면 미래에는 **일부 영역에서는 인간과 거의 분간이 어려운 수준**에 이를 가능성도 있습니다. 그렇기에 우리는 LLM의 동작원리와 한계를 잘 이해하고, 이를 **도구로서 현명하게 활용**해야 할 것입니다. 또한 LLM이 **인간의 인지 편향이나 사회적 문제**를 학습하지 않도록 지속적인 모니터링과 **책임 있는 AI 원칙 준수**가 필요합니다 [27] [28]. 결론적으로, GPT나 Claude 같은 모델들은 **언어처리의 새로운 지평**을 열었지만, **인간 사고의 대체물이 아닌 보완재**로 보는 것이 적절합니다. 인간 고유의 창의성과 이해력은 AI 시대에도 더욱 값질 것이며, LLM은 이를 증진시키는 유용한 도구로 쓰일 것입니다. 산업, 교육, 예술 등 다양한 분야에서 **LLM과 인간이 협업**하여 시너지를 낼 수 있으며, 이는 마치 인간이 컴퓨터를 활용해 계산능력을 확장한 것처럼 **인지의 확장**을 이룰 수도 있습니다. **AI와 인간의 차이**를 분명히 인식하면서도 **장점을 살려 협력**하는 것이 앞으로의 중요한 과제가 될 것입니다. [37] [36]

참고 문헌

- OpenAI, “GPT-4 Technical Report”, 2023 [1] [6]
- Tom B. Brown et al., “Language Models are Few-Shot Learners (GPT-3 논문)”, NeurIPS 2020 [2]
- OpenAI, “ChatGPT: Optimizing Language Models for Dialogue”, 2022.
- Anthropic, “Claude’s Constitution”, 2023 [17] [18]
- Anthropic, “Introducing 100K Context Windows”, 2023 [11]
- Google DeepMind, “Introducing Gemini 1.0/2.0/2.5”, 2023-25 [22] [21]
- Google AI Blog, “PaLM: Scaling Language Modeling with Pathways”, 2022 [23] [24]
- Meta AI, “Llama 2: Open Foundation and Fine-Tuned Chat Models”, 2023 [65]
- Niu et al., “Large Language Models and Cognitive Science: Review”, arXiv 2023 [69] [55]

- Moveworks, “What is a Stochastic Parrot?”, 2023 [37] [36]
- Damien Riehl, “LLM ‘Creativity’ = Statistically Unlikely”, 2023 [34]
- Anthropic, “Constitutional AI: Harmlessness from AI Feedback”, 2022.
- MIT News, “Like human brains, LLMs reason about diverse data based on meaning”, 2023.
- 기타: 위키피디아 GPT-3/Claude/Gemini 문서 [22] [19], OpenAI/Anthropic/Google 공식 블로그 등.

---

[1] Key takeaways from the GPT4 technical report | by Abdelhadi Azzouni | Medium
https://medium.com/@hadiazouni/key-takeaways-from-the-gpt4-technical-report-e0e8f1d34f84

[2] [3] [4] [2005.14165] Language Models are Few-Shot Learners
https://arxiv.org/abs/2005.14165

[5] Bea Stollnitz - The Transformer architecture of GPT models
https://bea.stollnitz.com/blog/gpt-transformer/

[6] [9] [31] [32] [47] [48] GPT-4 | OpenAI
https://openai.com/index/gpt-4-research/

[7] [8] [15] [16] [17] [18] Claude’s Constitution \ Anthropic
https://www.anthropic.com/news/claudes-constitution

[10] [11] [12] [13] [14] Introducing 100K Context Windows \ Anthropic
https://www.anthropic.com/news/100k-context-windows

[19] [20] [21] [22] [25] [26] Gemini (language model) - Wikipedia
https://en.wikipedia.org/wiki/Gemini_(language_model)

[23] [24] [61] [63] [64] Pathways Language Model (PaLM): Scaling to 540 Billion Parameters for Breakthrou
https://research.google/blog/pathways-language-model-palm-scaling-to-540-billion-parameters-for-breakthrough-performance/

[27] [28] [29] [30] [33] 가장 유능하고 범용적인 AI 모델 제미나이(Gemini)를 소개합니다
https://blog.google/intl/ko-kr/company-news/technology/gemini-kr/

[34] [40] [41] [42] [43] Post-LLM "Creativity" = Statistically Unlikely
https://www.linkedin.com/pulse/post-llm-creativity-statistically-unlikely-damien-riehl-vaobc

[35] [36] [37] [51] [52] What is a Stochastic Parrot? | Moveworks
https://www.moveworks.com/us/en/resources/ai-terms-glossary/stochastic-parrot

[38] [39] [53] [54] [55] [67] [68] [69] Large Language Models and Cognitive Science: A Comprehensive Review of Similarities, Differences, and Challenges
https://arxiv.org/html/2409.02387v1

[44] LLMs and Creativity: Complements, Not Substitutes
https://newsletter.ericbrown.com/p/llms-and-creativity

[45] LLMs vs. Human Mind: Understanding the Creativity Gap. - Go Far AI
https://www.gofar.ai/p/llms-vs-human-mind-understanding

[46] Are Large Language Models More or Less Creative than Humans?
https://blog.kulturwissenschaften.de/llms-and-creativity/

[49] WATCH: Princeton Language and Intelligence Director Explains ...
https://ai.princeton.edu/news/2025/watch-princeton-language-and-intelligence-director-explains-why-large-language-models-are

50 Can Stochastic Parrots Truly Understand What They Learn? - Medium
https://medium.com/@stahl950/can-stochastic-parrots-truly-understand-what-they-learn-7af2886ea76

56 Some Reflections on Similarities and Differences of the Human Mind ...
https://www.linkedin.com/pulse/some-reflections-similarities-differences-human-mind-large-lauri-tkynf

57 How does a human perform vs LLM - DeepLearning.AI
https://community.deeplearning.ai/t/how-does-a-human-perform-vs-llm/521629

58 Like human brains, large language models reason about diverse ...
https://news.mit.edu/2025/large-language-models-reason-about-diverse-data-general-way-0219

59 Comparing Humans and Large Language Models on an ...
https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00674/122721/Comparing-Humans-and-Large-Language-Models-on-an

60 The Human Brain Vs LLM - Medium
https://medium.com/@chauhandhruv351/the-human-brain-vs-llm-4729685a51b8

62 Chain-of-Thought Prompting
https://learnprompting.org/docs/intermediate/chain_of_thought?srsltid=AfmBOoqnzMUMXjB3sbM5ck8zHoo2aK_Cud5qh_UVCAwxA82AlKvdOTM7

65 66 Llama 2: Open Foundation and Fine-Tuned Chat Models Paper Reading - Arize AI
https://arize.com/blog/llama-2-open-foundation-and-fine-tuned-chat-models-paper-reading/