

# 생성형 AI 시대 거대언어모델의 기술 동향

LG AI연구원 | 김소연·신중보·윤현구·이진식·조현직·최정규·한장훈

## 1. 서론

생성형 AI 기술에 기반한 ChatGPT[1]의 등장 이후 생성형 AI 기술은 경제, 사회를 포함한 전 산업 영역으로 급속히 확산되고 있다. 생성형 AI 기술은 사용자 요구에 따라 결과를 능동적으로 생성해내는 인공지능 기술로서 개인과 기업의 생산성 향상에 도움을 주며, 전문가 수준의 창작 활동에 활용 가능하고, 인공 일반 지능(AGI)[2] 달성 가능성을 한층 더 높여주었다.

본 기고에서는 생성형 AI의 대표적인 기술 중 하나인 거대언어모델(Large Language Model, LLM)에 대해 상세하게 알아보고, 생성형 AI 시대를 살고 있는 우리가 어떻게 경쟁력을 확보할 수 있는지 고민해 보고자 한다.

## 2. Data

### 2.1 학습 데이터

언어의 다양한 현상을 이해하고 처리하기 위해서는 학습 데이터가 필수적인 역할을 한다. 최근 몇 년간 언어모델의 크기가 지속적으로 증가함에 따라 거대언어모델의 학습에 더욱 방대한 양의 데이터가 사용되고 있다. 예를 들어, 1,750억개의 매개변수를 갖는 모델인 GPT-3[3]를 학습하기 위해서 3천억개의 토큰이 사용되었으나, 최근 공개된 700억개의 매개변수를 갖는 LLaMA-2[4] 모델 학습에는 2조개의 토큰이 사용되었다. 즉, 모델의 크기뿐만 아니라 학습에 사용되는 데이터의 규모도 경쟁적으로 늘어나고 있는 추세이다. 이를 바탕으로 많은 기업에서 거대언어모델 학습에 참여하며 다양한 데이터 소스를 공개하고 있다.

- CommonCrawl: CommonCrawl[5]은 비영리 단체로 2008년부터 현재까지의 웹 크롤링 데이터를 보유하고 공개하였다. C4[6], Redpajama[7], RefinedWeb[8] 등은 CommonCrawl 출처의 데이터 일부를 사

용해 서로 다른 전처리 과정을 수행한 데이터로, 많은 거대언어모델이 학습에 사용하는 데이터이다. 웹 데이터 특성상 다양한 주제, 어휘, 문체가 포함되어 있다.

- Wikipedia: Wikipedia[9]는 온라인 백과사전 데이터이다. 다양한 주제로 구성되며 내용이 정제되어 있어 높은 품질의 데이터로 분류된다.
- Code: 프로그램 코드와 프로그램 코드 중심의 질의 응답 데이터로 나뉜다. 전자의 경우 코드 저장소인 Github을 출처로 한 TheStack[10] 데이터가 있고, 후자는 코드 관련 질의 응답이 활발하게 이루어진 StackExchange[11]가 있다.
- 기타: 다양한 분야의 서적으로 구성된 Bookscorpus[12], 전문적인 내용을 다루고 있는 Arxiv[13], PubMedCentral[14], USPTO[15] 등의 데이터를 학습에 사용하기도 한다.

### 2.2 평가 데이터

거대언어모델의 평가는 다양한 측면에서 이루어지며 여러 종류의 태스크와 평가 지표를 사용한다. 평가의 기준은 다양하게 설정할 수 있으나, 본고에서는 언어 생성 능력, 지식 활용 그리고 추론 능력으로 분류하고 대표적인 평가 데이터를 소개한다[16].

- 언어 생성 능력: 문장을 완성하는 태스크 LAMBADA[17]와 문서를 요약하는 태스크 XSum[18]이 있다.
- 지식 활용 능력: 상식 또는 지식에 근거한 답변 생성 능력을 평가하는 것이 목표로 사전에 학습한 지식만을 이용해 답변하는 closed-book QA 태스크와 외부 지식을 활용해 답변하는 open-book QA 태스크가 대표적이다. 대표적인 데이터로는 Natural Questions[19], ARC[20], TruthfulQA[21] 등이 있고 open-book QA의 경우 Wikipedia와 같은 외부 소스를 활용해 모델을 평가한다.
- 추론 능력: 주어진 정보를 이해하거나 학습된 지식을 활용해 일련의 논리적인 추론 과정을 통해

결론을 얻을 수 있는지 평가하는 태스크로 구성되어 있다. 일반 상식 추론과 관련된 PiQA[22], HellaSwag[23], WinoGrande[24]가 대표적이며 수학적 논리 테스트를 위한 GSM8K[25] 등이 사용된다.

이외에도 다양한 도메인에서의 지식을 평가하는 MMLU[26], BIG-bench[27] 데이터가 많이 사용되며 정확성, 강건함, 평등성, 효율성 등을 입체적으로 평가하는 데이터[28]도 존재한다. 목적에 따라 특정 도메인에 특화된 모델의 평가가 필요할 경우 그에 부합하는 데이터셋으로 모델의 성능을 평가하기도 한다.

### 2.3 학습 데이터 가공 및 정제

데이터를 수집한 후에는 언어모델의 학습에 적합한 형태로 데이터를 변환하고 정제하는 과정을 거치게 된다. 데이터의 양과 함께 데이터의 품질 역시 언어모델의 성능에 영향을 주기 때문에 다양한 전처리 방법이 제안되고 있다[3, 8, 29]. 데이터 처리 과정은 크게 텍스트 품질 향상을 위한 필터링, 중복 제거, 비식별화 단계로 구성된다.

- 필터링: 규칙 기반의 필터링과 모델 기반의 필터링으로 나눌 수 있다. 예를 들어, 웹 크롤링 데이터에는 단어 및 문자열의 과도한 반복, 특정 문구 반복, 웹 포맷을 나타내는 불필요한 문구 등은 규칙 기반 필터링으로 처리할 수 있다. 반면 모델 기반의 필터링은 품질을 분류하여 저품질의 문서를 필터링 하는 방법이다.
- 중복 제거: 수집된 문서 중에서는 중복된 내용을 포함하는 문서가 존재할 수 있으며 이러한 중복은 언어모델의 성능 저하와 학습의 비효율을 초래할 수 있다[30]. 문서의 중복 제거를 위해 일반적으로 문서 내용의 유사도를 기반으로 중복 여부를 판단하며, URL 등 추가 정보가 있을 경우 URL이 일치함을 확인하여 중복된 내용을 제거하기도 한다.
- 비식별화: 방대한 양의 웹 데이터는 민감한 정보(예: 주민등록번호, 이메일 주소 등)를 포함할 수 있다. 이러한 개인 정보 노출을 방지하기 위해 해당 부분을 삭제하거나 비식별화 처리 과정을 진행한다.

## 3. 사전학습 언어 모델

사전학습 언어모델(pre-trained language model)은 모델 구조에 따라 인코더(encoder), 디코더(decoder), 인코더-디코더(encoder-decoder) 형태로 나뉜다. 본 기고에서는 최근 주목을 받고 있는 디코더 모델과 인코

더-디코더 모델을 중심으로 설명한다.

### 3.1 디코더 사전학습 모델

디코더 모델은 트랜스포머(Transformer)[31]의 디코더만을 활용한 단방향(unidirectional) 트랜스포머 언어 모델로 대표적으로 GPT 모델이 있다[32]. 주어진 토큰으로부터 다음에 나올 토큰을 예측하는 방식으로 학습하기 때문에 레이블이 없는 말뭉치로 사전학습을 할 수 있으며, 주어진 텍스트 다음에 나올 다음 단어를 예측하는 생성 태스크에서 높은 성능을 보인다.

GPT 모델은 점차 발전해 감에 따라 모델 파라미터 수와 학습 말뭉치 양이 크게 증가되었고, GPT-3에서는 1,750억개의 방대한 파라미터를 가진 모델을 3천억개의 대용량 토큰으로 사전 학습하여, 거대언어모델이라는 개념을 등장시켰다[3].

거대모델 등장 이전에는 각 태스크의 학습 데이터로 모델을 처음부터 학습하거나 사전 학습된 모델을 추가로 미세 조정(fine-tuning)하여 자연어 처리 문제를 해결하였지만, 거대언어모델은 방대한 모델 파라미터 때문에 추가 학습 비용이 매우 비싸다는 단점이 있다. 이에 GPT-3는 인컨텍스트 러닝(in-context learning)을 통해 다양한 문제를 추가 학습 없이 해결할 수 있음을 보였다. 인컨텍스트 러닝은 태스크 설명문(task description)과 태스크 질문 및 정답 쌍으로 구성된 예시들(examples), 질문으로 이루어진 프롬프트(prompt)를 자연어 형태로 입력하면, 주어진 입력 정보를 참고하여 모델이 답변을 생성하는 방법이다. GPT-3는 여러 태스크에서 인 컨텍스트 러닝으로 기존 미세 조정된 모델들을 뛰어넘는 성능을 달성하였다[3].

이후 거대언어모델의 답변 품질을 더욱 높이기 위해 인스트럭션 튜닝(instruction tuning)[33]과 인간 피드백 강화 학습(reinforcement learning with human feedback)[34] 등이 제안되었다. 인스트럭션 튜닝은 다양한 태스크에 대해 지시문 형태의 질문을 입력으로, 생성해야 할 답변을 출력으로 언어 모델을 미세 조정하면 학습하지 않은 태스크(unseen task)에 대해서도 모델 성능이 향상됨을 보였다. 인간 피드백 강화 학습은 입력 질문에 대해 사람이 더 선호하는 답변에 더 높은 보상을 줌으로써 사람이 더 선호하는 답변을 모델이 생성하도록 돕는다.

해외에서는 OpenAI, Google, Meta 등이 GPT-4 [35], PaLM-2[36], LLaMA-2[4] 등을 발표하고, 서비스에 활용하여 ChatGPT, Bard 등의 서비스를 제공하고 있다. 국내에서는 LG AI연구원의 EXAONE, 네이버의 HyperCLOVA, SK텔레콤의 에이닷, 엔씨소프트의 VARCO,

코난테크놀로지의 Konan, 카카오의 KoGPT, KT의 믿음 등의 모델들이 발표되었다. 이 중 EXAONE은 대규모 한국어와 영어 말뭉치로 학습한 이중언어(bilingual) 모델로 국내 전문가들이 주로 사용하는 한국어와 영어 두 가지 언어 모두에 대해 높은 성능을 보이며, 전문성이 필요한 분야의 질문에 대해 근거 기반의 정확한 답변을 생성하는 전문가용 대화형 AI 플랫폼인 EXAONE 2.0 Universe 서비스에 활용된다.

### 3.2 인코더-디코더 사전학습 모델

인코더-디코더 모델은 트랜스포머에서 제안된 두 가지 구조 인코더와 디코더를 모두 사용하는 모델이다. BERT[37]와 같은 인코더 전용 모델과 달리 인코더-디코더 모델은 입력과 출력 부분이 다른 구조로 구성되어 있어, 서로 형태가 다른 이중 데이터를 입출력으로 사용할 수 있다는 특징을 가진다. 인코더-디코더 모델은 크로스-어텐션(cross-attention) 구조를 활용하여 입력 정보를 출력으로 연결시킨다. 상술하자면 인코더를 이용하여 주어진 입력을 히든 스테이트(hidden state)라는 벡터(vector) 표현으로 변환하고 변환된 입력 벡터와 디코더 구조 사이에 크로스-어텐션을 통해 상호 정보를 교환함으로써 최종 출력을 생성하는 형태이다.

인코더-디코더 모델의 대표적인 모델인 T5[38]에서는 스패ن 커럽션 오브젝티브(span corruption objective)를 활용하여 대량의 웹 문서에 대해 사전학습을 수행한다. 스패น 커럽션 오브젝티브란 언어모델 학습을 수행할 때, 단일 토큰에 대한 마스킹(masking)을 수행하는 것이 아닌, 여러 토큰에 걸친 넓은 영역을 마스킹하는 방법이다. 스패ن 커럽션 오브젝티브는 모델이 학습해야 할 문제의 난이도를 상승시킴으로써 좋은 성능을 보여준 기법이다.

이후 후속 연구에서는 T5 모델의 구조를 활용하여 음성과 같은 언어 외 다른 형태의 입력도 처리할 수 있음이 보고되었다[39]. 이미지 분야에서는 VIT[40] 모델이 인코더-디코더 구조로 이미지를 학습하여 높은 성능을 달성하였고, 인코더-디코더 구조가 전통적인 CNN 구조를 대체하며 빠르게 발전하고 있다.

사전학습 언어모델은 학습된 언어에 높은 의존도를 가지고 있기 때문에 한국어 태스크를 수행하기 위해서는 한국어 말뭉치로 학습된 사전학습 모델이 필요하다. 하지만 한국어 태스크에 사용가능한 인코더-디코더 모델은 수가 많지 않으며, 특히 수억개 이상의 매개변수를 가진 대형 언어모델의 경우 선택지가 많지 않다. 또한 성능 문제로 인해 연구자들은 공개된

한국어 모델이 아닌 Google의 다중언어(multilingual) T5 [41] 모델을 선택하는 상황이다.

LG AI연구원에서는 자체 인코더-디코더 모델을 연구 개발하여 한국어와 영어 두 언어의 다양한 태스크에서 Google T5 모델의 성능을 상회하는 결과를 얻었다. 다중 태스크 튜닝(Multi-task tuning) 기법을 적용한 모델 간 비교에서도 Google FLAN-T5 모델 대비 월등히 높은 성능을 달성했다. 또한 KorQuAD 1.0 리더보드 1위 달성을 통해 EXAONE 모델의 성능을 인정받았다[42].

## 4. Instruction Tuning

본 섹션에서는 Instruction Tuning의 개념과 그 특징 및 최신 연구 동향을 소개한다.

### 4.1 Instruction Tuning 개요 및 동향

거대언어모델의 발전과 함께 모델이 학습하지 않은 새로운 작업을 수행하는 Zero-shot 능력이 발전되고 있다. 이러한 특징은 거대모델이 사전 학습 과정에서 다음 토큰을 예측하는 훈련 목표를 통해 내재적으로 다중 태스크 학습을 수행하고, 이를 통해 새로운 작업을 수행하는 능력을 습득할 수 있음을 시사한다[43].

최근 연구에서는 Zero-shot 성능을 향상시키는 방법론으로 명시적으로 다양한 작업에 대한 학습을 진행하는 인스트럭션 튜닝 방법이 제안되었다[33, 44]. 이 방법은 각 태스크에 대한 문제와 답을 명시적으로 학습함으로써 언어모델에 더 의미 있는 지도를 제공한다. 인스트럭션 튜닝은 크게 두가지 특징을 가지는데 첫번째는 다중 태스크를 학습한다는 것이고, 두번째는 태스크에 대한 설명을 포함하는 지시문을 추가하여 학습한다는 것이다. 이러한 특징 때문에 인스트럭션 튜닝을 다중 태스크 프롬프트 학습(multi-task prompted training)이라고도 일컫는다[44].

### 4.2 작업 클러스터

인스트럭션 튜닝에서 NLP 태스크는 다양한 데이터셋으로 구성되며, 이러한 데이터셋을 작업의 종류에 따라 유사한 작업을 묶어 놓은 것을 작업 클러스터라고 한다. 그림 1은 FLAN [33]에서 정의한 12개의 작업 클러스터 예시이다.

인스트럭션 튜닝의 효과를 평가하기 위해서는 특정 작업 클러스터를 제외하고 모델을 훈련한 후 해당 작업 클러스터에 대한 데이터셋 성능을 측정한다. 일반적으로 이러한 작업 클러스터가 다양할수록 모델은 높은 성능을 나타낸다.

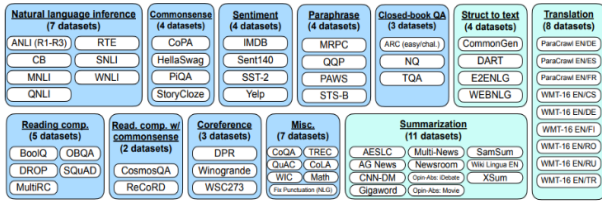


그림 1 FLAN의 작업 클러스터 분류

### 4.3 Instruction

템플릿, 프롬프트라고도 불리는 인스트럭션은 해당 태스크의 특징을 자연어로 설명한 것이다. 이를 통해 언어모델이 해당 태스크에 대해 어떤 답을 생성할지 배우게 되고, 다양한 인스트럭션을 학습함으로써 사용자로부터 새로운 인스트럭션이 주어졌을 때 언어모델이 이를 이해하여 답을 생성할 수 있게 한다. 각각의 데이터셋은 여러 개의 인스트럭션을 포함하며, 인스트럭션이 다양할수록 모델은 강건한 모습을 보인다.

### 4.4 Cross-lingual Instruction Tuning

인스트럭션 튜닝의 가장 어려운 부분은 학습을 위해 풍부한 데이터셋이 필요하고 그에 따른 다양한 인스트럭션이 필요하다는 점이다. 영어의 경우 다양한 인스트럭션 튜닝 데이터셋[33, 44, 45]이 제안되었지만 영어를 제외한 대부분의 언어에서는 해당 데이터셋을 확보하기 어려운 경우가 많다. 이러한 문제를 해결하기 위해 다중언어모델에서 데이터가 풍부한 언어(예: 영어)에 대해 인스트럭션 튜닝을 수행한 후, 이 효과가 다른 언어(예: 한국어)에 전이되도록 하는 교차언어 인스트럭션 튜닝(cross-lingual instruction tuning)이 제안되었다[46]. 이는 언어간 지식 전이를 이용하는 것으로 인스트럭션 튜닝 데이터셋을 구축하는 어려움을 완화할 수 있다. LG AI연구원에서는 이러한 연구를 발전시켜 한국어의 데이터 부족 문제를 해결하고 효율적으로 인스트럭션 튜닝 데이터를 구축하여 언어모델의 성능이 향상됨을 확인하였다.

## 5. 강화학습 기반의 사용자 피드백 반영

GPT-3와 같은 디코더 구조 기반의 거대 사전학습 언어모델들은 여러 자연어 처리 문제에서 놀라운 성능을 보였으나 종종 사용자의 기대에 반하여 유해한 답변을 생성하거나 환각현상을 보인다[47, 48, 49]. 이는 생성형 모델의 학습에 사용된 목적 함수(다음 단어 예측 오브젝티브 혹은 스펠 커럽션 오브젝티브)가 실제 사용자들의 만족도와 완전히 일치하지 않기 때

문이라고 해석된다[50, 51].

이러한 간극을 좁히고 더 나은 사용자 경험을 제공하기 위해 사용자 피드백을 생성형 모델에 주입하는 방법론들이 활발히 연구[51, 52, 53, 54, 55]되었으며, 그 중에서도 특히 ChatGPT의 상업적인 성공은 사용자 만족도 정렬(human preference alignment)의 중요성을 부각시켰다. 본 기고에서는 InstructGPT[51]와 LLaMA-2-chat 모델[4]에 사용자 피드백을 주입시키기 위해 사용되었던 Proximal Policy Optimization(PPO)을 활용한 강화학습 기법을 위주로 설명한다.

### 5.1 Human Preference Data

모델에 학습시키기 위한 데이터는 크게 무해함(harmlessness)과 유용성(helpfulness)을 위주로 확보한다. 기존 instruction tuning이 완료된 모델에서 실제 사용자 입력과 유사한 임의의 입력들에 대해서 다양한 다중 턴을 포함한 출력을 생성하게 한 후에, 크라우드 워커들에게 여러 출력들의 랭킹을 매기도록 한다. 이때 크라우드 워커들은 여러 출력들의 랭킹만 매기면 되기 때문에, 기존의 지도 학습을 위한 instruction tuning 데이터를 생성하는 것보다 훨씬 효율적으로 많은 양의 사용자 만족도 데이터를 확보할 수 있다.

사용자 만족도를 반영하는 랭킹 형태의 데이터는 두 문장의 사용자 만족도 우선순위를 표현하는 Bradley-Terry 모델[56]을 통해 사용자 선호도 분포로 변환된다. 추가적으로  $n$ 개 출력에 대한 랭킹 형태의 데이터는  $nC_2$ 개의 사용자 선호도 데이터로 활용될 수 있다. LLaMA-2-70B와 같은 초거대 언어모델에 안정적으로 사용자 피드백을 주입하기 위해서는 대량의 사용자 선호도 데이터가 필요하다고 알려져 있으며, LLaMA-2의 경우 약 142만 건의 사용자 선호도 데이터를 이용하였다[55].

### 5.2 Reward Model

위 과정을 통해 구성된 대량의 사용자 선호도 데이터를 통해, 입력 문장에 대해 0에서 1 사이의 스칼라 값으로 사용자 선호도를 표현하는 reward model을 학습시킬 수 있다. 초거대 언어 모델의 경우 일반적으로 instruction tuning이 완료된 모델을 시작 가중치로 학습을 진행하며, reward model의 목적 함수는 더 선호된 입력 문장의 reward 값과 덜 선호된 문장의 reward 값의 차이를 기반으로 구성되어 있다.

Reward model 역시 파라미터의 확장성(scalability)이 적용된다고 알려져 있다[57]. InstructGPT는 reward model로 60억개 매개변수를 가지는 언어모델을, LLaMA-2는 700억개의 매개변수를 가지는 언어모델을 사용하였다. 또한 LLaMA-2는 ‘무해함’을 표현하는 reward

model과 ‘유용성’을 표현하는 reward model을 각각 구성하여 ‘무해함’과 ‘유용성’ 사이의 trade-off 현상이 작아지는 효과를 기대하였다.

### 5.3 Proximal Policy Optimization

잘 학습된 reward 모델이 표현하는 사용자 선호도를 직접 사용하여 생성형 AI 모델에 학습시키는 것은 매우 어렵기 때문에, 사용자 선호도를 생성형 AI 모델에 주입시키기 위해서 강화학습 기법을 차용한다. 일반적으로 강화학습 기법은 자연어 처리 모델에서 불안정한 학습 추이를 보인다[54]. 불안정성 문제를 완화하고 더 안정적인 학습을 위해 고안된 방법론들이 Trust Region Method[58]와 Proximal Policy Optimization(PPO)[59]이다.

PPO는 강화학습의 목적 함수가 이끄는 방향으로 policy 모델을 학습하되, 학습된 모델과 기존 모델의 분포가 지나치게 달라지지 않도록 KL-divergence로 제약을 주는 방법론이다. InstructGPT나 LLaMA-2에 적용된 PPO 방법론은 policy 모델을 계속 학습하여도 모델의 초기 시작점이었던 instruction tuning이 완료된 모델의 분포와 KL-divergence가 일정 수준 이상 벗어나지 않게 제약을 설정한다.

### 5.4 강화학습 기반의 사용자 피드백 반영 예시와 평가

PPO 강화학습 기법을 이용한 사용자 피드백 반영 방법론은 OpenAI의 ChatGPT를 비롯하여 Anthropic의 Claude 2, Meta의 LLaMA-2-chat 등에 적용되어 사용자 선호도가 주입된 모델의 상업적 가치를 보이고 있다. ChatGPT는 상술한 사용자 선호도를 주입하는 과정을 반복적으로 수행하고 배포하여 점점 더 사용성을 개선해 나가고 있다고 밝혔다.

사용자 선호도가 주입된 모델은 기존 instruction tuning 모델을 평가하는 벤치마크 데이터셋으로 평가가 가능하다. 특히, 4가지 벤치마크(ARC-challenge, TrustfulQA, MMLU, HellaSwag)를 모아서 평가하는 Open LLM Leaderboard [60]와 다중 턴 대화를 생성한 후 GPT4를 이용하여 등급을 평가하는 MT-bench[61] 등이 자주 사용된다. 더 직접적으로 사용자 선호도를 평가하기 위해서는 언어 모델들 간의 선호도를 크라우드 워커와 ELO rating을 활용하여 평가하는 Chatbot Arena 방식을 사용하기도 한다[61].

## 6. 모델 경량화

이미 학습된 신경망 모델을 성능 저하없이 경량화하려는 시도는 예전부터 많이 연구되어 왔다. 특히 최신 언어모델은 크기가 매우 커서 실 서비스 운영시 비용이 매우 많이 들어가므로, 이를 경량화하는 것은 매우 중요한 일이다. 본 섹션에서는 양자화(quantization), 지식 증류(knowledge

distillation), 가지치기(pruning) 방법 및 최근 연구 동향에 대해 설명한다.

### 6.1 양자화

양자화는 신경망 모델의 성능을 유지하면서 매개변수의 numerical precision을 줄이는 방법이다. 보통의 신경망 모델은 floating point 32-bit나 16-bit를 사용하는데, 이를 8/4/2-bit 수준으로 낮추면서 성능을 유지할 수 있다. 통상적으로 쓰이는 integer quantization은 다음과 같다.

$$x_q = \text{quantize}(x, b, s, z) = \text{clip}(\text{round}(s \cdot x + z), -2^{b-1}, 2^{b-1} - 1)$$

$$\hat{x} = \text{dequantize}(x_q, s, z) = \frac{1}{s}(x_q - z)$$

즉, 매개변수의 최댓값, 최솟값을 이용해 실숫값에 scale 값을 곱하고 zero 값을 더한 후, 양자화를 통해 integer 값을 얻으면, 이것이 원래 실숫값 매개변수를 대체할 수 있게 된다. 실제 추론시에는 integer 값, scale 값, zero 값을 저장해 놓고 이를 원래 실숫값으로 복원해서 사용한다.

### 6.2 지식 증류

지식 증류는 지식을 전수할 모델(teacher model)이 있을 때, 지식을 전수받을 모델(student model)이 지식을 전달받아 두 모델 간 비슷한 성능을 얻고자 하는 방법이다. 지식을 전달할 때, 출력값의 토큰 정보뿐만 아니라 출현 확률 등의 부가 정보를 전달하여 두 모델이 닮아가도록 학습하면, 지식을 전수받을 모델을 직접 학습하는 것보다 더 높은 성능을 얻을 수 있다.

### 6.3 가지치기

신경망 모델 매개변수 중 덜 중요한 것을 골라내서 제외하면 모델의 성능이 유지가 된다는 것이 알려져 있다. 경우에 따라 절반 이상의 매개변수를 제외하더라도 성능이 유지되는 경우가 있어, 이러한 경우에는 매개변수의 일부만 저장하여 계산하면 된다.

- **Unstructured pruning:** 모델의 구조와 상관없이 중요도가 낮은 매개변수를 제외하는 방법이다. 다만 모델을 표현할 수 있는 표준화된 방법이 없어 특별한 하드웨어 지원이 없는 경우 메모리 및 속도상 이점은 없으며, 매개변수 저장시 이점만 있다.
- **Structured pruning:** 모델의 구조를 변경하면서 매개변수를 제외하는 방법이다. 모델의 크기를 줄임으로써 메모리 및 속도 측면의 이점은 있으나, 성능이 많이 감소한다는 단점이 있다.



## 7. Information Retrieval 기술과의 연동

앞서 언급한 여러 기술들을 통해 생성형 거대언어모델의 혁신은 급속하게 이루어지고 있다. 하지만 실시간 정보의 반영이나, 사실이 아닌 내용을 그럴듯하게 생성하는 환각 현상의 가능성은 여전히 배제할 수 없다. 그림 2는 ChatGPT를 실행한 결과로 ChatGPT는 2021년 9월까지의 정보만 학습이 되어 있어 이후의 정보에 대해서는 답변이 불가함을 출력하고 있다. 또한 그림 3에서는 사실이 아닌 내용을 사실처럼 그럴싸하게 답변하는 현상을 발견할 수 있다.

ST

우리나라 현재 대통령 누구야?

GPT

제 지식은 2021년 9월까지의 정보만을 포함하고 있으며, 그 이후의 정보를 알 수 없습니다. 2021년 9월에는 대한민국의 대통령은 문재인 대통령이었습니다. 하지만 현재 대통령은 제가 알 수 없으며, 최신 정보는 인터넷 뉴스나 정부 공식 웹사이트에서 확인하실 수 있습니다.

그림 2 학습 시점에 따른 제한적인 정보 학습

ST

이순신 장군이 당나라 군을 무찌른 해전 이름이 뭐야?

GPT

이순신 장군이 당나라 군을 무찌른 해전의 이름은 "명량 해전"입니다. 이 해전은 1597년에 발생했으며, 조선 왕조와 당나라(일본) 간의 전쟁 중에 일어났습니다. 이순신 장군은 명량 해전에서 조선 해군을 지휘하고 당나라의 강력한 함대를 물리치는 데 성공한 영웅적인 해전 중 하나로 기억되고 있습니다.

그림 3 역사적 사실에 대한 환각 현상(Hallucination)

이러한 단점을 보완하기 위해 최근 실시간 정보를 반영하거나 답변의 근거를 제시하는 정보 검색(information retrieval) 기술을 거대언어모델에 연동함으로써 문제를 해결하려는 시도들이 활발하게 나타나고 있다. RAG(Retrieval Augmented Generation)[62]는 그림 4와 같이 retriever와 generator의 2개의 모듈로 구성된다. Retriever에서는 주어진 query와 가장 관련 있는 문서를 찾아오고 generator는 retriever에서 찾아온 문서를 기반으로 거대언

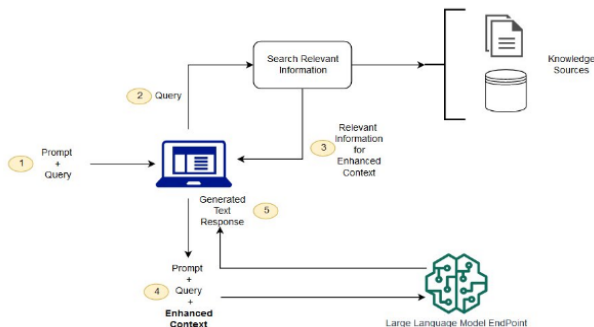


그림 4 RAG 구조[63]

어모델로부터 답변을 생성하게 된다.

LG AI연구원에서는 거대언어모델의 단점을 보완하기 위한 RAG 기술에 대한 연구를 진행하고 있다. 논문과 같은 scientific article의 question answering 연구 결과로 QASA(Advanced Question Answering on Scientific Articles) 모델과 데이터셋을 공개하였다[64]. QASA는 그림 5와 같이 associative selection, rationale generation, systematic composition의 3단계 프로세스를 거친다. Associative selection 단계에서는 information retrieval을 통해 후보로 찾은 문서 중 질문과 연관된 passage를 선택하고, rationale generation에서는 각 passage에 있는 내용을 기반으로 질문에 대한 응답의 근거를 생성하며, 마지막 단계인 systematic composition에서는 여러 개의 근거들을 취합해서 통합된 응답 생성을 최종적으로 생성하게 된다. QASA에서는 공개된 T5(30억개 매개변수) 모델을 실험에 사용된 데이터셋으로 학습했을 때 InstructGPT(1,750억개 매개변수) 보다 더 좋은 성능을 확보하였다. 이 연구는 답변의 근거가 되는 passage level의 상세한 근거를 제시하여 혹시라도 발생할 수 있는 환각 현상을 최소화하고, 사용자로 하여금 한 번 더 판단할 수 있도록 근거를 제시하는 장점을 보유하고 있다. LG AI연구원은 해당 연구결과를 기초로 하여, 자체 거대언어모델인 EXAONE을 사내 보안문서, 화학/Bio 논문, 특허, 법률 문서 등의 전문 문헌을 기반으로 하는 다양한 question answering 서비스에 적용하고 있다.

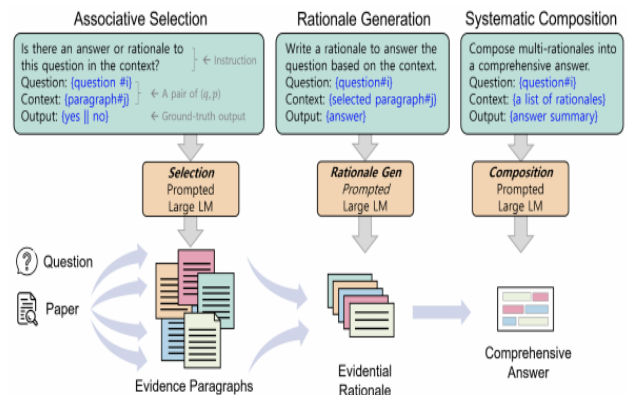


그림 5 Overview of QASA

## 8. 맺음말

본 기고에서는 생성형 AI의 대표적인 기술인 거대언어모델의 기술 동향을 소개하였다. 거대언어모델은 대규모 텍스트 데이터를 활용하여 사전학습 과정을 통해 언어를 이해하기 위한 기초 지식을 습득한 후, 사용자의 지시를 이해하고 태스크를 수행할 수 있도록 인스트럭션 튜닝 과정을 거치고, 사람이 선호하는 답변을 생성하도록 강화

학습 기반의 사용자 피드백을 반영하는 방식으로 만들어진 다. 거대언어모델은 의료, 법률, 교육, 마케팅, 금융 등 다양한 분야에서 활용될 수 있으며, 앞으로 지속적인 기술 발전을 통해 더 많은 분야에서 혁신과 발전을 이뤄낼 것으로 기대된다.

본 기고에서 상세하게 다루지는 않았지만 거대언어모델은 윤리나 안전성 측면에서 문제를 발생시킬 수 있으므로 이 문제를 해결하기 위한 연구도 지속 병행해 나가야 할 것이다.

## 참고문헌

- [1] OpenAI. “ChatGPT (September 25 Version).” 2023. <https://chat.openai.com/chat>
- [2] B. Goertzel, “Artificial general intelligence: concept, state of the art, and future prospects.” *Journal of Artificial General Intelligence*, 2014.
- [3] T. Brown et al., “Language models are few-shot learners.” *Advances in neural information processing systems*, 2020.
- [4] H. Touvron et al., “Llama 2: Open foundation and fine-tuned chat models.” *arXiv preprint arxiv:2307.09288*, 2023.
- [5] “Common crawl.” Available: <https://commoncrawl.org/>
- [6] C. Raffel et al., “Exploring the limits of transfer learning with a unified text-to-text transformer.” *The Journal of Machine Learning Research*, 2020.
- [7] Together Computer. “Redpajama: An open source recipe to reproduce llama training dataset.” 2023.
- [8] G. Penedo et al., “The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only.” *arXiv preprint arxiv:2306.01116*, 2023.
- [9] “Wikipedia.” Available: <https://en.wikipedia.org/wiki/>
- [10] D. Kocetkov et al., “The stack: 3TB of permissively licensed source code.” *arXiv preprint arxiv:2211.15533*, 2022.
- [11] L. Gao et al., “The pile: An 800GB dataset of diverse text for language modeling.” *arXiv preprint arXiv:2101.00027*, 2020.
- [12] Y. Zhu et al., “Aligning books and movies: Towards story-like visual explanations by watching movies and reading books.” *Proc. of ICCV*, 2015.
- [13] “Arxiv.” Available: <https://arxiv.org/>
- [14] “PubMed Central.” Available: <https://www.ncbi.nlm.nih.gov/pmc/about/intro/>
- [15] “United states patent and trademark office.” Available: <https://www.uspto.gov/>
- [16] W. Zhao, “A survey of large language models.” *arXiv preprint arXiv:2303.18223*, 2023.
- [17] D. Paperno et al., “The LAMBADA dataset: Word prediction requiring a broad discourse context.” *arXiv preprint arXiv:1606.06031*, 2016.
- [18] S. Narayan et al., “Don’t give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization.” *arXiv preprint arXiv:1808.08745*, 2018.
- [19] T. Kwiatkowski et al., “Natural questions: a benchmark for question answering research.” *Transactions of the Association for Computational Linguistics*, 2019.
- [20] P. Clark et al., “Think you have solved question answering? Try ARC, the AI2 reasoning challenge.” *arXiv preprint arXiv:1803.05457*, 2018.
- [21] S. Lin et al., “TruthfulQA: Measuring how models mimic human falsehoods.” *arXiv preprint arXiv:2109.07958*, 2021.
- [22] Y. Bisk et al., “PiQA: Reasoning about physical commonsense in natural language.” *Proc. of AAAI*, 2020.
- [23] R. Zellers et al., “HellaSwag: Can a machine really finish your sentence?” *arXiv preprint arXiv:1905.07830*, 2019.
- [24] K. Sakaguchi et al., “WinoGrande: An adversarial Winograd schema challenge at scale.” *Communications of the ACM*, 2021.
- [25] K. Cobbe et al., “Training verifiers to solve math word problems.” *arXiv preprint arXiv:2110.14168*, 2021.
- [26] D. Hendrycks et al., “Measuring massive multitask language understanding.” *arXiv preprint arXiv:2009.03300*, 2020.
- [27] A. Srivastava et al., “Beyond the imitation game: Quantifying and extrapolating the capabilities of language models.” *arXiv preprint arXiv:2206.04615*, 2022.
- [28] P. Liang et al., “Holistic evaluation of language models.” *arXiv preprint arXiv:2211.09110*, 2022.
- [29] G. Wenzek et al., “CCNet: Extracting high quality monolingual datasets from web crawl data.” *arXiv preprint arXiv:1911.00359*, 2019.
- [30] K. Lee et al., “Deduplicating training data makes language models better.” *arXiv preprint arXiv:2107.06499*, 2021.

- 
- [31] A. Vaswani et al., “Attention is all you need.” Advances in Neural Information Processing Systems, 2017.
- [32] A. Radford et al., “Improving language understanding by generative pre-training.” 2018.
- [33] J. Wei et al., “Finetuned language models are zero-shot learners.” Proc. of ICLR, 2022.
- [34] P. Christiano et al., “Deep reinforcement learning from human preferences.” Advances in Neural Information Processing Systems, 2017.
- [35] OpenAI, “GPT-4 technical report.” arXiv preprint arXiv:2303.08774, 2023.
- [36] Google, “PaLM 2 technical report.” arXiv preprint arXiv:2305.10403, 2023.
- [37] J. Devlin et al., “BERT: Pre-training of deep bidirectional transformers for language understanding.” Proc. of NAACL-HLT, 2019.
- [38] C. Raffel et al., “Exploring the limits of transfer learning with a unified text-to-text transformer.” Journal of Machine Learning Research, 2020.
- [39] J. Ao et al., “SpeechT5: Unified-modal encoder-decoder pre-training for spoken language processing.” Proc of ACL, 2022.
- [40] A. Dosovitskiy et al., “An image is worth 16x16 words: Transformers for image recognition at scale.” Proc. of ICLR, 2021.
- [41] L. Xue et al., “mT5: A massively multilingual pre-trained text-to-text transformer.” Proc. of NAACL-HLT, 2021.
- [42] “KorQuAD 1.0.” Available: <https://korquad.github.io/KorQuad%201.0/>
- [43] A. Radford et al., “Language models are unsupervised multitask learners.”, 2019.
- [44] V. Sanh et al., “Multitask prompted training enables zero-shot task generalization.” Proc. of ICLR, 2022.
- [45] Y. Wang et al., “Super-natural instructions: Generalization via declarative instructions on 1600+ NLP tasks.” Proc. of EMNLP, 2022.
- [46] N. Muennighoff et al., “Crosslingual generalization through multitask finetuning.” Proc. of ACL, 2023.
- [47] Z. Ji et al., “Survey of hallucination in natural language generation.” ACM Computing Survey, 2023.
- [48] L. Wang et al., “A survey on large language model based autonomous agents.” arXiv preprint arXiv:2308.11432, 2023.
- [49] Y. Bai et al., “Training a helpful and harmless assistant with reinforcement learning from human feedback.” arXiv preprint arXiv:2204.05862, 2022.
- [50] N. Stiennon et al., “Learning to summarize with human feedback.” Advances in Neural Information Processing Systems, 2020.
- [51] L. Ouyang et al., “Training language models to follow instructions with human feedback.” Advances in Neural Information Processing System, 2022.
- [52] R. Rafailov et al., “Direct preference optimization: Your language model is secretly a reward model.” arXiv preprint arXiv:2305.18290, 2023.
- [53] H. Liu et al., “Languages are rewards: Handsight finetuning using human feedback.” arXiv preprint arXiv:2302.02676, 2023.
- [54] R. Ramamurthy et al., “Is reinforcement learning (not) for natural language processing? Benchmarks, baselines, and building blocks for natural language policy optimization.” arXiv preprint arXiv:2210.01241, 2022.
- [55] C. Gulcehre et al., “Reinforced self-training (rest) for language modeling.” arXiv preprint arXiv:2308.08998, 2023.
- [56] R. Bradley et al., “Rank analysis of incomplete block designs: I. The method of paired comparisons.” Biometrika, 1952.
- [57] L. Gao et al., “Scaling laws for reward model overoptimization.” Proc. ICLR, 2023.
- [58] J. Schulman et al., “Trust region policy optimization.” Proc. of ICML, 2015.
- [59] J. Schulman et al., “Proximal policy optimization algorithms.” arXiv preprint arXiv:1707.06347, 2017.
- [60] E. Beeching et al., “Open LLM Leaderboard.” HuggingFace, 2023.
- [61] L. Zheng et al., “Judging LLM-as-a-judge with MT-bench and Chatbot Arena.” arXiv preprint arXiv:2306.05685, 2023.
- [62] P. Lewis et al., “Retrieval-augmented generation for knowledge-intensive NLP tasks.” Proc of NeurIPS, 2020.
- [63] “Retrieval Augmented Generation (RAG).” Available: [https://docs.aws.amazon.com/ko\\_kr/sagemaker/latest/dg/jumpstart-foundation-models-customize-rag.html](https://docs.aws.amazon.com/ko_kr/sagemaker/latest/dg/jumpstart-foundation-models-customize-rag.html)
- [64] Y. Lee et al., “QASA: Advanced question answering on scientific articles.” Proc of ICML, 2023.
-



## 약 력



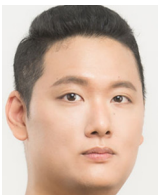
### 김 소 연

2019 이화여자대학교 전자공학과 졸업(학사)  
2021 한국과학기술원 전기및전자공학부(석사)  
2022~현재 LG AI연구원  
관심분야 : 거대언어모델, 연속학습  
Email : kimcando@lgresearch.ai



### 신 중 보

2013 연세대학교 전기전자공학부 졸업(학사)  
2021 서울대학교 전기정보공학부 졸업(박사)  
2021~현재 LG AI연구원  
관심분야 : 대화모델, 거대언어모델  
Email : jb.shin@lgresearch.ai



### 윤 현 구

2015 서울대학교 전기·정보공학부 졸업(학사)  
2022 서울대학교 전기·정보공학부 졸업(박사)  
2022~현재 LG AI연구원  
관심분야 : 거대언어모델  
Email : hyeongu.yun@lgresearch.ai



### 이 진 식

2005 POSTECH 컴퓨터공학과 졸업(학사)  
2012 POSTECH 컴퓨터공학과 졸업(박사)  
2012~2016 삼성전자 DMC연구소  
2016~2022 SK텔레콤 AI센터  
2022~현재 LG AI연구원  
관심분야 : 거대언어모델, 강화학습  
Email : jinsik.lee@lgresearch.ai



### 조 현 직

2013 아주대학교 전자공학과 졸업(학사)  
2022 연세대학교 인공지능학과 졸업(석사)  
2013~2021 LG전자 모바일 사업부  
2021~현재 LG AI연구원  
관심분야 : 거대언어모델 구조, 거대언어모델 최적화  
Email : hyunjik.jo@lgresearch.ai



### 최 정 규

2003~2019 LG전자 인공지능연구소 연구위원  
2020~현재 LG AI연구원 Multimodal AI 연구그룹장  
관심분야 : 언어모델, 검색, 음성인식  
Email : stanleyjk.choi@lgresearch.ai



### 한 장 훈

2019 서강대학교 컴퓨터공학과 졸업(학사)  
2021 서강대학교 컴퓨터공학과 졸업(석사)  
2021~현재 LG AI연구원  
관심분야 : Instruction Tuning, 대화모델  
Email : janghoon.han@lgresearch.ai