초거대 언어모델 연구 동향

업스테이지 ▌박찬준*·이원성·김윤기·김지후·이활석

1. 서 론

ChatGPT[1]와 같은 초거대 언어모델(Large Language Model, LLM)의 등장으로 기존에 병렬적으로 연구되던 다양한 자연언어처리 하위 분야들이 하나의 모델로 처리되고 있으며, 태스크 수렴 현상 (Converge)이 발생하고 있다. 즉 하나의 LLM으로 번역, 요약, 질의응답, 형태소분석 등의 작업을 모두 처리할 수 있게 되었다. 프롬프트 (Prompt)를 어떻게 모델에게 입력하느냐에 따라서 LLM의 다양한 능력들이 창발되고, 이에 따라 사용자의 목적에 맞는 출력을 생성하는 패러다임을 맞이하게 되었다 [1].

LLM은 최근 몇 년 간의 연구 동향에 따라 뛰어난 발전을 이루고 있다. 이러한 발전은 몇 가지 주요한 요인에 기반하고 있으며, 이 요인들은 현대 자연언어처리 (Natural Language Processing, NLP) 연구의 핵심적인 추세로 간주된다. 첫째로, 데이터의 양적 확대는 무시할 수 없는 중요한 요인이다. 디지털화의 선도로, 텍스트 데이터의 양이 기하급수적으로 증가하였고, 이는 연구의 질적 변화를 가져왔다. 대규모 코퍼스의 활용은 LLM의 일반화 능력을 향상시키며, 다양한 맥락과 주제에 대한 깊은 학습을 가능하게 한다. 둘째로, 컴퓨팅 기술의 진보는 LLM의 발전에 있어 결정적이었다. 특히, Graphics Processing Unit (GPU) 및 Tensor Processing Unit (TPU)와 같은 고성능 병렬 처리 하드웨어의 개발은 모델 학습에 있어 병목 현상을 크게 완화시켰다. 이로 인해 연구자들은 모델의 복잡성을 키우고, 더욱 깊은 신경망 구조를 탐구할 수 있게 되었다. 셋째, 알고리즘 및 기술의 발전은 LLM의 성능 향상을 주도하였다. Attention 및 Transformer Architecture의 도입은 연구자들에게 문맥 간의 관계를 더욱 정교하게 모델링할 수 있는 방법을 제공하였다 [2, 3]. 이 모든 변화의 중심에는 ‘scaling law’라는 학문적인 통찰이 있다 [4]. 해당 연구에 따르면, 모델의 크기와 그 성능은 긍정적인 상관 관계를 보인다. 이를 통해 연구자들은 모델의 파라미터 수를 증가시키면서, 이에 따른 성능 향상을 기술적 진보의 상호작용에서 나온 결과이며, 이러한 추세는 앞으로도 NLP 연구의 주요 동력이 될 것으로 예상된다.

연구단계를 넘어 LLM은 산업계에서도 많은 발전을 이루어 내고 있다. LLM은 교육, 의료, 금융, 제조 등 거의 모든 산업 분야에서 광범위한 활용 가능성을 제시하고 있다 [5, 6, 7, 8]. 교육 분야에서는 단순한 정보 검색을 넘어, 개인화된 학습 경로를 추천하는 시스템, 과제의 자동 평가, 학생들의 복잡한 질문에 대한 답변 제공 등의 역할로 활용될 수 있다. 이는 교육의 효율성과 개인화를 동시에 추구하는 현대의 교육 트렌드와 맞물려 큰 효과를 발휘할 것으로 기대된다. 의료 분야에서는 환자 데이터를 기반으로 한 초기 진단 도구로 활용될 뿐만 아니라, 복잡한 의료 기록 분석, 신약 개발에 필요한 연구 데이터 분석, 또는 최신 의학 연구 동향 파악 등의 다양한 역할을 수행할 수 있다. 이로써 의료 전문가들의 결정을 보조하고, 효율적인 치료 방향을 도모할 수 있게 된다. 금융 분야에서는 개인의 투자 성향과 시장의 동향을 분석하여 투자 권고를 제공하는 것 외에도, 금융 위험을 상세하게 분석하거나, 복잡한 금융 거래를 자동화하는 시스템의 핵심 구성 요소로서의 역할을 할 수 있다. 이는 금융 서비스의 효율과 안전성 향상에 크게 기여할 것이다. 제조 분야에서도 LLM은 설계 단계부터 생산, 품질 관리에 이르기까지의 전 과정에서 데이터 분석 및 최적화 도구로 활용될 수 있다. 생산 효율성 향상과 제품 품질 향상을 도모하며, 고객의 니즈에 더욱 민첩하게 대응할 수 있는 기회를 제공한다.

그러나, 이러한 긍정적인 측면들과 더불어 LLM의 한계점과 위험성도 고려되어야 한다. LLM은 학습 데이터의 편향성을 그대로 반영할 수 있어, 편향된 결과나 추천을 할 가능성이 있다 [9]. 이는 특히 중요한 의

* 정회원

1) https://openai.com/blog/chatgpt

사 결정을 위해 LLM을 활용하는 경우에 문제가 될 수 있다. 또한, LLM을 악의적인 목적으로 사용하는 위험성도 있다 [10]. 예를 들면, 미스리딩 정보 생성이나 편향된 정보 전파를 위한 도구로 활용될 수 있다. 이 외에도 LLM의 동작 원리나 결과에 대한 설명력 부족, 최신 정보를 반영하는 데의 한계 등 여러 문제점이 있으며, 이러한 문제점들을 해결하는 것은 다가오는 연구의 중요한 도전 과제로 여겨진다.

즉 편향성 (LLM은 학습 데이터에 포함된 편향을 반영할 수 있음), 안전성 (LLM을 악의적인 목적으로 사용할 수 있음), 설명 가능성 (LLM의 예측 결과를 설명하기 어려움), 최신성 (최신정보를 반영하기 어려움)의 문제점을 여전히 LLM의 한계점으로 보유하고 있으며 이러한 문제는 장기적으로 해결하기 위해 연구되어야할 것이다.

본 논문은 초거대 언어모델(LLM)에 대한 전반적인 동향을 다루고자 작성되었다. 첫째로 초기의 언어모델부터 현재의 초거대 언어모델까지의 연구 및 발전 과정을 소개한다. 둘째로, 한국어 초거대 언어모델의 특징 및 최근 동향을 조명한다. 셋째로, 최신 초거대 언어모델 연구 동향을 심층적으로 살펴본다. 넷째로, 초거대 언어모델의 성능 평가 방식과 그 변화에 대해 논의한다. 마지막으로, 초거대 언어모델 연구와 활용에 있어 중요하게 여겨지는 윤리적 원칙과 관련된 최근의 동향을 소개 한다. 본 논문을 통해 초거대 언어모델에 관한 전반적인 동향과 중요한 주제들에 대한 체계적인 이해를 제공하고, 이 분야의 연구자 및 관련 전문가들에게 유용한 통찰과 지침을 제시하고자 한다.

2. 언어모델부터 초거대언어모델까지

자연언어란 “인간의 언어”를 의미하며, 자연언어처리는 자연언어를 컴퓨터가 처리하는 것을 의미한다. 자연언어처리를 위해서는 인간의 언어표현 체계를 컴퓨터가 이해할 수 있는 형태로 변환해주는 것이 필요하다. 이러한 역할을 하는 것이 바로 언어모델이다. 이번 섹션에서는 전통적인 언어모델 연구에 대해 먼저 살펴보고, 의미기반 언어모델 연구, 문맥기반 언어모델 연구, 초거대 언어모델 연구들에 대해 차례로 살펴본다.

전통적인 언어모델 연구 전통적인 언어모델은 인간이 사용하는 단어를 컴퓨터가 이해할 수 있는 숫자 체계로 변환하는 데에 초점을 맞춰 발전했다. 이를 위해 전통적인 언어모델은 단어 집합 (vocabulary)을 생성하고, 단어 집합을 이용하여 자연언어를 컴퓨터가 이해할 수 있는 형태로 변환했다. 단어 집합을 이용하여 자연언어를 표현하는 전통적인 방법 중 대표적으로 사용되는 방법은 원-핫 인코딩 (one-hot encoding)이다. 원-핫 인코딩은 표현하고자 하는 단어의 색인(index)에만 1을 표시 하고, 다른 단어의 색인에는 0을 표시하여 단어를 벡터로 나타내는 방법이다. 따라서, 원-핫 인코딩을 사용하면 모든 단어를 단어 집합의 크기를 가지는 벡터로 표현할 수 있다. 이러한 벡터를 희소 벡터 (sparse vector)라고 부른다. 예를 들어, 단어 집합에 ‘강아지’, ‘고양이’라는 2개의 단어만 존재한다고 했을 때, 원-핫 인코딩 방식으로 이를 표현하면 ‘강아지’는 [1, 0], ‘고양이’는 [0, 1]로 표현된다.

그러나, 원-핫 인코딩은 단어 간의 의미적인 연관성을 고려할 수 없다는 치명적인 한계를 가진다. 예를 들어, ‘강아지’와 ‘고양이’는 포유류 동물이며 사람들에게 사랑받는 애완동물이라는 공통점을 가지고 있지만, 원-핫 인코딩은 단어를 의미와는 관계없이 희소 벡터 형태로 표현하다보니 이들 간의 의미적인 연관성을 전혀 고려하지 못한다.

의미기반 언어모델 연구 단어 간의 의미적인 연관성을 고려하여 언어를 표현하기 위해, 의미기반 언어모델은 단어의 의미가 반영되도록 단어를 밀집 벡터 (dense vector) 공간에 표현하는 데에 초점을 맞춰 발전했다. 가장 대표적인 의미기반 언어모델은 Word2Vec [11]이다. 이는 주변 단어들로부터 중심 단어를 예측하거나 중심 단어로 주변 단어들을 예측하도록 학습함으로써, 유사한 의미의 단어들을 밀집 벡터 공간 상 가까운 거리에 분포하도록 학습시킨다. 이러한 패러다임을 기반으로 GloVe [12]와 FastText [13]와 같은 다양한 연구들이 이루어졌다.

이는 전통적인 언어모델의 한계를 극복했지만, 문맥 정보를 이해하지 못한다는 한계를 가지고 있다. 예를 들어, ‘사과를 먹고 싶다.’와 ‘내가 사과할게.’에서 전자에서의 사과는 과일을 의미하지만, 후자에서의 사과는 잘못을 인정하고 용서를 빈다는 의미를 지닌다. 즉, 같은 문자임에도 사용되는 문맥에 따라 다른 의미를 지닐 수 있다. 하지만, 의미기반 언어모델은 같은 문자라면 같은 밀집 벡터로 표현하기 때문에 문맥 정보를 반영하지 못한다는 한계를 지닌다.

문맥기반 언어모델 연구 문맥 정보를 반영하여 언어를 표현하기 위해, 텍스트 내의 정보를 이용하는 RNN (Recurrent Neural Network)이 등장했다. 그러나, RNN은 입력 텍스트의 길이가 길어질수록 앞쪽에 위

치하는 정보들을 기억하지 못하는 장기 의존성 문제가 존재한다. 이러한 문제를 극복하기 위해, LSTM (Long Short-Term Memory) [14]과 GRU (Gated Recurrent Unit) [15]가 등장했다. 하지만, 이들은 모두 텍스트에 존재하는 단방향 문맥 정보만 활용한다는 한계를 지닌다.

양방향 문맥 정보를 활용하기 위해, ELMo [16]는 주어진 텍스트에 존재하는 순방향 문맥 정보와 역방향 문맥 정보를 함께 활용하는 양방향 학습을 제안했다. 이를 위해, ELMo는 순방 향 LSTM과 역방향 LSTM를 동시에 활용한다. 하지만, 이는 LSTM을 기반으로 하기 때문에, LSTM이 지니는 다음과 같은 한계를 그대로 가진다: 1) 하나의 벡터에 텍스트의 모든 정보를 담기 때문에 정보 손실이 발생하고, 2) 입력 텍스트의 길이가 길어지면 기울기 소실 (gradient vanishing)이 발생한다.

이러한 한계를 해결하기 위해 나온 것이 바로 Attention Mechanism [2]과 이를 활용한 Transformer Architecture [3] 이다. Attention Mechanism은 하나의 벡터에 텍스트의 모든 정보를 담는 RNN, LSTM, GRU와 다르게, 텍스트 내 단어들의 벡터들을 필요에 따라 적절히 활용하는 메커니즘이다. 현재 언어모델의 근간이 되는 Transformer가 바로 이러한 Attention Mechanism을 기반으로 한다. Transformer는 크게 인코더와 디코더로 구성되는데, 인코더는 주어진 텍스트를 이해하는 역할을 하고 디코더는 이해한 텍스트를 기반으로 언어를 생성해내는 역할을 수행한다. 이러한 Transformer의 인코더를 기반으로 발전한 대표적인 모델이 Google[2]의 BERT (Bidirectional Encoder Representations from Transformers) [17] 이고, 디코더를 기반으로 발전한 대표적인 모델이 OpenAI[3]의 GPT (Generative Pretrained Transformer) [18]이다.

BERT는 입력 텍스트의 약 15%에 해당하는 임의의 토큰을 마스킹하고 마스킹된 토큰이 무엇인지 예측하는 MLM (Masked Language Modeling) 방식으로 학습된다. 한편, GPT는 이전 텍스트를 기반으로 다음에 나올 토큰이 무엇인지 예측하는 NTP (Next Token Prediction) 방식으로 학습된다. 이들은 별도의 레이블링 작업 없이 텍스트 데이터만 있으면 학습을 할 수 있다는 강점을 가진다. 이러한 강점을 바탕으로, 이후 문맥기반 언어모델들은 대용량의 텍스트 데이터로 사전학습 (Pretraining) 하고, 이후 특정 태스크로 미세조정(Fine-tuning)하는 Pretrain-Finetune 패러다임을 중심으로 발전한다.

초거대 언어모델 연구 문맥기반 언어모델 이후, 다양한 연구들에서 모델 및 학습 데이터의 크기와 모델의 성능은 긍정적인 상관 관계를 보인다는 'scaling law' [4, 19, 20]가 밝혀지면서, 초거대 언어모델 (Large Language Model, LLM)이 등장하기 시작했다. LLM은 기존 언어모델에서와 다르게, 모델의 가중치 업데이트 없이도 새로운 태스크를 수행할 수 있는 In-context learning (Zero-shot learning [21]과 Few-shot learning [22]) 능력을 가진다. 이처럼 작은 크기의 모델에서는 발현되지 않던 LLM의 능력을 창발 능력 (Emergent ability) [23]이라고 부른다.

이러한 LLM의 창발 능력을 잘 이끌어내기 위한 연구 분야가 바로 프롬프트 엔지니어링이다. 프롬프트 엔지니어링이란 LLM이 모델 가중치 업데이트 없이 특정 태스크를 더욱 잘 해결하게 하기 위해, 입력으로 주는 프롬프트를 어떻게 설계할 것인지에 대한 연구 분야이다. 가장 대표적인 프롬프트 엔지니어링 연구는 Chain-of-Thought (CoT) [24]가 있다. 이는 해결하고자 하는 태스크의 예시를 일련의 중간 추론 단계와 함께 넣어줌으로써, 복잡한 문제를 여러 단계로 나누어 해결하는 CoT 프롬프트를 제안했다. 또한, 이러한 프롬프트 엔지니어링까지도 LLM으로 대체하고자 하는 연구도 활발히 진행되고 있다 [25].

한편, LLM의 조종성 (Steerability)을 높이기 위해, Instruction Tuning [26]과 Reinforcement Learning from Human Feedback (RLHF) [27]과 같은 학습 기법이 등장했다. Instruction Tuning은 다양한 태스크를 (지시, 입력, 출력) 형태의 데이터로 구성하여, 해당 데이터를 통해 LLM을 미세조정하는 학습 기법이다. RLHF는 LLM이 생성할 수 있는 다양한 답변들 중 사용자가 선호할만한 답변을 출력하도록 LLM을 학습하는 기법이다. 하지만, 사용자 선호도를 학습하기 위해 강화학습을 사용하는 RLHF는 학습 과정이 복잡하다는 한계가 있어서, 이를 완화하기 위한 연구도 활발히 진행되고 있다 [28].

3. 한국어 초거대 언어모델 동향

GPT [21, 22, 29], PALM [30, 20]과 같은 대규모 LLM 뿐만 아니라, Falcon [31, 32], Llama [33, 34], Claude [35], Qwen [36]과 같은 비교적 작은 크기의 오픈소스 LLM이 전 세계적으로 공개되고 활발히 연구되고 있다. 하지만, 이러한 LLM들은 일반적으로 한

2) https://www.google.com/
3) https://openai.com/

국어를 비효율적으로 토큰화하고, 학습한 한국어 토큰 수가 매우 부족하다는 한계를 가진다. 실제로, GPT-3 [22] 의 경우 학습된 한국어 토큰의 비율은 0.01697% 밖에 되지 않으며, 오픈소스 LLM인 Llama 2 [34]의 경우도 0.06% 밖에 되지 않는다. 이에 따라, 한국어 사용자를 위한 한국어 LLM의 필요성이 대두되고 있다.

이러한 필요성에 따라, 최근 많은 국내 기업에서 한국어 LLM을 자체적으로 학습하기 시작했다. Naver Clova[4]의 HyperClova [37]를 시작으로, Kakao Brain[5]의 KoGPT, KT Enterprise[6]의 믿음, LG AI Research[7]의 Exaone, NCSOFT[8]의 VARCO, SALTLUX[9]의 Luxia, 코난테크놀로지[10]의 코난 LLM 등 다양한 한국어 LLM이 공개되고 있다. 이들의 공통점은 자체적으로 보유한 한국어 데이터와 공개되어 있는 한국어 데이터, 크롤링 데이터를 적극적으로 활용하여, 한국어 토큰 비율을 높여서 학습하고 있다는 것이다. 더불어 업스테이지의 경우 Llama2를 파인튜닝하여 Solar-0-70b 모델을 개발하였고, 글로벌 LLM 플랫폼 중 하나인 Poe.com에 서비스하고 있다[11]. 해당 모델은 한국어와 영어 모두 지원하고 있다.

한편, 오픈소스 한국어 LLM도 존재한다. 가장 대표적인 모델이 EleutherAI[12]의 Polyglot-Ko [38]이다. Polyglot-Ko는 1.3B, 3.8B, 5.8B, 12.8B의 다양한 크기로 공개되어, 한국어 LLM 연구 발전에 큰 도움이 되고 있다. 실제로, Polyglot-Ko를 미세조정한 KoAlpaca, KORani, KULLM, NA-LLM과 같은 한국어 LLM이 활발히 연구되고 있다. 이처럼 한국어 LLM의 빠른 발전을 위해서는, 위와 같은 오픈소스 한국어 LLM이 더 많이 공개되는 것이 필요하다.

이러한 한국어 LLM 학습을 위해 사용된 한국어 공개 데이터셋은 AI-Hub[13], 모두의 말뭉치[14], 위키백과[15], 청와대 국민청원[16] 등이 있다. 이러한 한국어 공개 데이터셋은 한국어 LLM 학습에 큰 도움이 되지만, 여전히 학습할 고품질 한국어 데이터가 부족하여 한국어 LLM은 아직까지 사용자가 만족하기에는 불충분한 성능을 보이고 있다.

4) https://clova.ai/
5) https://www.kakaobrain.com/
6) https://enterprise.kt.com
7) https://www.lgresearch.ai/
8) https://kr.ncsoft.com/
9) https://www.saltlux.com/
10) https://www.konantech.com/
11) https://poe.com/Solar-0-70b
12) https://www.eleuther.ai/
13) https://www.aihub.or.kr/
14) https://corpus.korean.go.kr/
15) https://ko.wikipedia.org/
16) https://www1.president.go.kr/petitions/

4. 최신 초거대 언어모델 연구 동향

해당 섹션에서는 LLM의 최신 연구 동향에 대해 조금 더 자세히 살펴보고자 한다. 보다 구체적으로, 학계와 산업계에서 발표된 관련 최신 연구들을 1) 사전학습, 2) 미세 조정, 3) 활용 및 증강 (Utilization & Augmentation) 의 세 가지 관점에서 살펴보고자 한다. 이러한 분류는 대규모 데이터로부터 자가 지도 학습 (self-supervised learning) 또는 준 지도 학습 (semi-supervised learning)을 통해 다양한 하위 태스크에 접목할 수 있는 기반 모델 (foundation model) [39]로서의 LLM을 개발하는 **사전 학습** 과정과, 하위 태스크를 보다 잘 풀기 위한 목적으로, LLM을 도메인 또는 태스크 별로 적응시키거나 (domain or task-specific adaptation) 또는 사람이 기대하는 바와 일치시키는 (human alignment) **미세조정** 단계를 포함하고 있다. 마지막으로, LLM 능력을 **활용**하는 다양한 활용 전략과 더불어, LLM의 내재적인 한계로 지적받는 환각 현상을 해결하고, 복잡한 기호 및 산술 추론 등의 태스크를 해결하기 위해 외부 도구를 활용하는 전략인 **증강** 관점의 사례들을 소개한다.

4.1 사전학습

사전학습 단계는 언어 생성 및 문맥 이해 능력 등을 모델에 학습시킴으로써 LLM의 근간을 형성하는 과정이다. 이 단계에서는 대량의 코퍼스와 컴퓨팅 자원을 활용함으로써, LLM으로 하여금 세상에 대한 기본 지식 (world knowledge)을 습득할 수 있도록 한다. 결과적으로 LLM은 전통적인 언어모델에 비해 뛰어난 문맥 이해력 및 상식, 기호, 논리 추론 능력을 보유하게 되고 [22], 기초적인 수준의 범용적인 자연어 태스크 솔버 (general-purpose natural language task solver)로서 기능할 수 있다 [40]. 해당 섹션에서는 LLM의 사전학습 관련 내용 중 데이터 활용 현황 및 전처리에 대해서 논의할 것이다.

4.1.1 데이터 활용 현황

LLM의 사전학습을 위한 데이터 활용 현황을 살펴보면, 웹페이지, 책, 대화 데이터, 학술 데이터, 코드 등 다양한 종류의 이질적인 코퍼스를 혼합하여 활용하는 추세이다. 이러한 추세는 LLM의 성능에 사전학습용 코퍼스의 품질뿐만 아니라 그것의 다양성이 중

요한 역할을 한다는 연구 결과들 [41, 42]에 의해 뒷받침된다. 예를 들어, 2019년에 발표된 T5 [43]는 웹 페이지만을 사전학습에 활용하였으나, 이후에 공개된 GPT-3 [22]는 웹페이지를 비롯한, 책 및 뉴스 데이터를 함께 활용하였다. 더불어, Llama-1 65B 모델 [33]에서는 사전학습 데이터 중 웹페이지가 차지하는 비중이 87%에 달하지만, 남은 13%의 데이터는 대화 데이터, 책 및 뉴스, 학술 데이터, 코드 데이터가 골고루 차지하고 있다.

이러한 사전학습 데이터의 다양성을 강조하는 추세에도 불구하고, LLM의 성능 향상을 위한 다양한 코퍼스의 최적 혼합 비율과 필요한 데이터 양에 관한 연구는 아직 초기 단계에 머물러 있다. 이와 관련하여 주목할 만한 연구로는 [41]이 있다. 해당 연구에서는 사전 학습 코퍼스를 시간대, 필터링 기법, 도메인 혼합 비율 조합에 따라 28개로 구분하고, 이를 대상으로 1.5B 파라미터를 갖는 Transformer decoder-only 모델을 학습 하였다. 이들은 사전학습 데이터와 평가 데이터 사이의 시간적 차이 (temporal shift) 때문에 발생하는 성능 저하는 미세조정 만으로는 극복하기 어려움을 발견했으며, 데이터 품질 필터링 및 독성 필터링의 중요성을 정량적으로 증명하였다. 또한, 사전 학습시 이질적인 도메인 코퍼스를 활용하는 것이 전체적으로 도움이 된다는 것을 재확인했다. 또 다른 사례로 GPT-4 [29] 에서는 사전학습에 많은 자원과 시간이 소요되는 문제를 완화하기 위해 predictable scaling 기법을 소개했다. 이를 활용하면 LLM 사전학습 중에 적은 양의 컴퓨팅으로 최종 성능을 정확히 예측할 수 있는 것으로 알려져 있다.

코퍼스의 다양성을 강조하는 방향과는 별개로, 하위 태스크에 특화된 LLM을 위한 사전학습에서는 관련된 코퍼스의 비중을 증가시키는 전략도 활용되고 있다. Google에서 발표한 대화 어플리케이션을 위한 언어 모델인 LaMDA [44]는 전체 사전학습 데이터 비중의 약 절반 (50%) 가량을 대화 데이터로 할당하였으며, 교육 및 콘텐츠 추천 영역에서 해당 모델의 효용성을 입증하였다. 다국어 특화 LLM인 BLOOM [45] 및 PaLM [30]은 타겟 언어인 영어 이외의 다국어 텍스트를 사전학습에 함께 활용함으로써, 다국어 기반의 번역, 요약, QA 태스크에서 뛰어난 성능을 달성하였다. 과학 도메인 특화 LLM 인 Galactica [46]는 사전학습 데이터의 약 86%를 과학 데이터로 사용하였고, 코드 생성에 특화된 LLM인 AlphaCode [47]는 사전학습 데이터를 전부 코드 데이터로 사용하기도 했다.[17)]

4.1.2 전처리

사전학습 시 수집한 데이터를 그대로 사용하는 것은 데이터의 크기와 노이즈, 중복, 독성 데이터 등의 존재로 인해 여러 문제를 야기할 수 있다. 따라서 사전학습 용도로 데이터를 전처리하는 것이 필수적이다. 이러한 전처리 과정은 크게 품질 필터링, 중복 제거, 개인정보 제거, 토큰화의 순서로 이루어진다 [49]. **품질 필터링** (quality filtering) 단계에서는 수집된 데이터로부터 저품질의 데이터를 걸러낸다. 해당 단계에서는 고품질의 텍스트 데이터로부터 학습된 분류기를 통해 저품질 데이터를 걸러내거나 [22, 30], 정교하게 디자인된 규칙에 기반한 휴리스틱스 [45, 50]을 사용하는 것이 일반적이다. 사전학습 데이터에 중복되는 데이터가 존재할 경우 LLM의 성능을 저해하는 것으로 알려져 있다 [51]. 이를 방지하기 위해서, **중복 제거** (de-duplication) 단계에서는 반복되는 단어를 갖는 저품질 문장이나, 단어 및 N-그램 기반 겹침 비율에 기반하여 유사한 내용을 갖는 중복 문서들을 필터링한다 [33, 50, 45, 52]. 또한 information leakage를 방지하기 위해, 학습 데이터와 평가 데이터 사이의 중복 데이터도 제거되어야 한다 [30]. 다음으로 **개인정보 제거** (privacy reduction) 단계가 수행된다. 대부분의 LLM 사전학습 데이터는 웹 텍스트를 포함하므로 이메일 주소나 전화번호 같은 민감 정보를 포함할 수 있다. 실제로 몇몇 연구들 [53, 54]에서는 정교한 프롬프팅을 통해서 LLM으로부터 개인 식별 정보 (Personally Identifiable Information, PII) 또는 Github Copilot secret API keys와 같은 민감 정보를 추출할 수 있음을 보인 바 있다. 이러한 이유로, LLM을 윤리적으로 사용하고 개인정보 침해 위험을 제거하기 위해 사전학습 데이터에서 민감 정보를 제거하는 것이 필수적이다. 마지막 전처리 단계로, 원본 텍스트를 토큰이라 불리는 작은 단위의 시퀀스로 분리하는 **토큰화** (tokenization) 작업이 수행된다. 이 작업은 LLM이 등장하기 이전의 전통적인 NLP 태스크에서도 중요한 연구 분야였으며, LLM이 도래한 이후에도 컴퓨팅 비용, 언어 의존성, 정보 손실 등을 고려하여 토큰화를 개선하기 위한 연구가 계속되고 있다. LLM과 관련된 토크나이저의 중요성에 관한 논의는 [55]를 참고하길 바란다.

17) 코드 데이터의 중요성과 관련하여, 최근 연구 [48]는 CoT와 같은 LLM의 복잡한 추론 능력의 출현이, 텍스트와 차별화되는 코드 데이터의 독특한 특성에 기인하는 것으로 추정하고 있다.

4.1.3 기타 고려사항

위에서 언급한 데이터 관련 논의 외에도, LLM 사전학습에는 모델의 아키텍처, 모델의 상세 설정, 목적함수, 학습 환경 세팅 및 학습 테크닉 등 여러 고려사항이 있다. 해당 논의를 모두 다루는 것은 이 논문의 범위를 벗어나므로, 관심 있는 독자는 다음의 서베이 논문 [49]을 참고하길 바란다.

4.2 미세조정

사전학습이 완료된 LLM은 다양한 하위 태스크를 해결하기 위한 기본적인 준비가 마련된 상태라 할 수 있다. 그럼에도 불구하고, 최근 연구 동향에 따르면 LLM을 특정 목적에 맞게 미세조정 하는 경우가 증가하고 있다. 이러한 미세조정의 대표적인 전략으로는 Instruction Tuning과 Alignment Tuning이 주목받고 있다. 전자는 기존에 본 적 없는 태스크에 대한 일반화 능력 (unseen task generalization ability)을 향상시키는 방법론이며, 후자는 LLM의 출력을 인간의 가치와 기준에 부합하도록 조정하는 접근법이다. 마지막으로, LLM의 계산 집약적 특성으로 인한 한계를 개선하기 위한 방법론인 자원 효율적인 (resource-efficient) 미세조정 방법에 대해서도 간략히 언급하고자 한다.

4.2.1 Instruction Tuning

Instruction Tuning 이란 사전학습된 LLM을 대상으로 자연어로 이루어진 포맷팅된 지시사항 (formatted instructions)과 그에 대응하는 출력 (output) 쌍으로 이루어진 데이터를 기반으로 미세조정하는 추가 훈련 과정을 의미한다 [26]. 이것은 기존의 지도학습 패러다임과 유사하나, Instruction Tuning은 비교적 적은 수의 예제만으로도 뛰어난 성능 및 새로운 태스크에 대한 일반화가 가능한 것으로 알려져, 더욱 효율적인 학습 패러다임이라 할 수 있다 [56].

Instruction Tuning을 위해서는 태스크의 의미를 LLM이 이해할 수 있도록 관련 지시사항과 이에 대응하는 출력으로 이루어진 자연어 포맷의 데이터를 구성해야 한다. 해당 과정은 기존에 존재하는 NLP 분야의 특정한 태스크 (e.g., 번역, 요약, QA 등) 관련 데이터를 형식화하는 것뿐만 아니라, 일상적인 대화 데이터 [27] 또는 모델로부터 합성된 데이터 [57]를 표준화하는 방법을 포함한다. 사전학습에서 데이터셋의 품질이 중요한 것과 마찬가지로, 데이터셋의 형식 및 품질이 Instruction Tuning의 성공 여부에 중요한 역할을 하는 것으로 알려져 있으며, 지시사항의 다양성과 품질이 예제의 개수보다 더 중요한 것으로 알려져 있다 [58]. 특히, 태스크 당 지시사항의 개수가 너무 많을 경우 오히려 오버피팅이 발생하고, 모델 성능을 저해 하는 것으로 밝혀져서 Instruction Tuning을 위한 데이터 수집 및 생성에 주의가 필요하다 [59, 60, 61].

일반적으로 Instruction Tuning을 통해 LLM은 태스크의 성능 향상을 도모할 수 있다. 최근의 연구들은 [59, 62] 77M에서 540B에 이르는 다양한 규모의 언어모델에서 Instruction Tuning을 통한 성능 향상 효과가 나타났다고 보고하였다. 더불어 태스크의 일반화 측면에서 볼 때, Instruction Tuning은 언어 모델이 자연어 형태의 지시사항을 이해할 수 있도록 돕는다. 이 과정에서 LLM이 인간의 지시에 따라 (instruction following) 특정 태스크를 수행할 수 있는 형태의 창발 능력을 획득하게 된다 [23]. 결과적으로, Instruction-tuned LLM은 기존 태스크뿐만 아니라 처음 보는 태스크에도 적응하고 일반화하는 능력을 갖게 된다 [59]. 또한, 도메인 특화된 데이터셋을 이용한 Instruction Tuning을 통해, 일반 LLM을 특정 분야의 전문가로 학습시킬 수 있다. 이러한 시도는 LLM이 범용적인 태스크 솔버 (general-purpose task solver) 측면을 넘어서서, 의학 [63, 64], 법률 [65], 금융 [66] 및 전자상거래 [67]와 같은 특화 도메인에 대해 전문화된 태스크 솔버 (domain-specialized task solver)로 활용될 수 있음을 시사한다. 보다 자세한 Instruction Tuning에 대한 논의는 다음의 서베이 논문들을 참고하길 바란다 [49, 56, 68].

4.2.2 Alignment Tuning

LLM의 사전학습 과정을 살펴보면, 주로 MLM 또는 NTP 형태의 목적함수를 가지고 학습되기 때문에, 주변 또는 이전 컨텍스트를 기반으로 단어를 예측함으로써 학습이 이루어진다. 따라서 사전학습 과정에서는 인간의 선호가 반영된다고 보기 어렵다. 이로 인해 LLM은 종종 유해한 또는 잘못된 정보의 제공이나 편향된 표현을 생성하기도 한다. Alignment Tuning은 이러한 LLM의 의도치 않은 행동을 방지하기 위한 방법론이다. 대표적인 전략으로, LLM을 인간의 기대치에 맞게 조정하는 'human alignment' [27, 69, 70] 방식이 있다. 그러나 이 방법은 '도움이 되는지 (helpfulness)', '정직한지 (honesty)', '무해한지 (harmlessness)'와 같은 사전학습 및 Instruction Tuning의 목적함수와는 전혀 다른, 주관적인 형태의 alignment criteria를 고려해야 한다. 또한 alignment criteria를 올바르게 측정하기 위해서는 고품질의 human feedback 수집이 필수적이라 상대적으로 많은 비용이 소모된다.

이러한 alignment criteria는 대부분 인간의 인식을 기반으로 하므로 LLM에 직접 최적화 목표로서 차용하기에는 어려움이 따른다. 이에, LLM을 인간의 가치와 일치시키기 위한 방법으로 인간의 피드백을 기반으로 한 강화 학습 (RLHF) [27, 69]이 제안되었다. RLHF는 수집된 인간의 피드백 데이터를 활용하여 LLM을 미세조정하는 방법으로 상술한 alignment criteria를 개선하는 데 유용하다. RLHF는 강화 학습 알고리즘을 사용하여 인간의 피드백을 바탕으로 보상 모델을 학습하면서 LLM을 적응시킨다. InstructGPT [27] 또는 ChatGPT와 같은 성공 사례에서 알 수 있듯이, 인간을 학습 루프에 포함시키는 이러한 방법은 LLM을 well-aligned 형태로 개선하는 데 중요한 역할을 한다. 결과적으로, 개선된 LLM은 편향이 적고, 더욱 안전한 내용을 생성하게 된다. Alignment Tuning이 LLM의 사용성 개선을 위해 중요함에도 불구하고, 주관적인 alignment criteria의 특성 상 의도치 않은 부작용이 발생하기도 한다. 실제로, alignment 과정이 LLM의 기본 능력을 일정부분 감소시킬 수도 있음이 밝혀졌으며, 이러한 현상을 alignment tax라고 부른다 [70].

4.2.3 Resource-Efficient Fine-Tuning

다음으로 LLM의 계산 집약적 특성으로 인한 한계를 개선하기 위한 방법론인 자원 효율적인 (Resource-Efficient) 미세조정 방법에 대해서도 간략히 언급할 것이다. LLM들은 수많은 모델 파라미터를 가지고 있기 때문에, 각 미세조정 시에 모든 파라미터를 튜닝하는 것은 비용 관점에서 비효율적이다. 따라서, 가능한 한 좋은 성능을 유지하면서, 학습가능한 파라미터의 수를 줄이는 Parameter-Efficient Fine-Tuning (PEFT) 방법에 대해 살펴볼 것이다.

Adaptor Tuning [71, 72]은 Transformer 구조에 adaptor라 부르는 작은 신경망 모듈을 추가한다. 이 과정에서 원래의 언어 모델의 파라미터는 고정된 상태로, adaptor 모듈의 파라미터만 특정 태스크 목적을 달성하기 위해 최적화된다. Prefix Tuning [73]은 학습가능한 연속 벡터로 구성된 일련의 prefix 시퀀스를 각 Transformer 레이어에 추가한다. 이러한 prefix vector들은 태스크 별로 할당되며, 일종의 가상 토큰 임베딩으로 볼 수 있다. 마찬가지로 prefix 파라미터만 학습되기 때문에, 파라미터 효율적인 방식의 최적화가 가능하다.

Transformer 모델 계층에 학습가능한 벡터를 추가하는 Pre-fix Tuning과는 대조적으로, Prompt Tuning [74, 75]은 학습 가능한 프롬프트 벡터를 입력 계층에 추가하는 형태로 이루어진다. 입력 텍스트에 프롬프트 토큰을 덧붙이고, 학습 과정에서 프롬프트 임베딩만 최적화되기 때문에 효율적인 태스크 특화 미세조정이 가능하다. Low-Rank Adaptation (LoRA) [76]은 이름에서 알 수 있듯이 PEFT에 low-rank approximation을 차용한다. 모델의 파라미터 W_0를 업데이트한다고 가정하자. 이 과정은 $W_0 \leftarrow W_0 + \triangle W$로 서술할 수 있다. 이때, 원래의 파라미터 행렬 $W_0 \in R^{d\times k}$는 고정한 뒤, 업데이트 행렬 $\triangle W$ 를 low-rank 행렬 분해를 통해 근사함으로써 업데이트 식을 다음과 같이 표현할 수 있다: $W_0 + \triangle W \simeq W_0 + BA$, 이때, $B \in R^{d\times r}$, $A \in R^{r\times k}$, 그리고 rank $r \ll \min(d, k)$이다. 결과적으로, LoRA는 메모리 및 스토리지 비용을 크게 절약할 수 있으며 태스크 별로 효과적인 모델 적응을 가능케 한다. 지금까지 PEFT 방법을 간략하게 살펴보았다. 이에 대한 보다 심도있는 논의는 다음 논문들을 참고하길 바란다 [77, 78].

또다른 자원효율적인 미세조정 방법으로는 Memory-Efficient Fine-Tuning 방법이 있다. LLM은 많은 모델 파라미터로 인해 추론 시에 대용량의 메모리를 필요로 하며, 이는 LLM의 활용 관점에서 매우 큰 장애물이다. 이를 해결하기 위해서, 양자화 (quantization) [79]와 같은 모델 압축 (model compression) 접근법을 통해 LLM의 메모리 사용량을 줄이는 방법들이 활발하게 연구되고 있다 [80, 81].

4.3 활용 및 증강

4.3.1 Utilization of LLMs

해당 섹션에서는 LLM을 활용하는 방법들에 살펴볼 것이다. LLM을 활용하는 가장 대표적인 방법 중 하나는 태스크를 해결하기 위한 적절한 프롬프팅 전략을 수립하는 것이고, 대표적인 프롬프팅 방법으로는 in-context learning (ICL)이 있다. ICL은 시연 (demonstration) 형태의 몇 가지 예시만으로 언어 모델이 태스크를 학습하게 하는 방식이다. 이는 잘 훈련된 언어 모델이 시연에 기반하여 태스크의 잠재적인 특성을 파악할 수 있음을 전제로 한다. ICL을 위한 프롬프트는 자연어 텍스트 형태의 태스크 설명, 시연을 위한 몇 가지 예시 및 테스트 쿼리로 구성된다. 최신 연구 [82]에 따르면, ICL은 다음과 같은 다양한 이점을 보유하고 있다. 첫째, 자연어 형태로 제공되는 시연은 LLM과의 명확하고 이해하기 쉬운 소통 방식을 제공한다 [22]. 둘째, ICL은 유사성에서 학습하는

인간의 의사결정 과정과 비슷한 측면이 있다. 마지막으로, ICL은 전통적인 지도학습 방식에 비해 training-free learning 구조를 가지고 있으므로, 새로운 태스크 적응에 필요한 계산 비용을 크게 줄일 수 있으며, 확장 가능한 (scalable) 특성을 지닌다.

이렇듯 유용한 ICL은 어떤 원리로 작동하는 것일까? 연구자들은 LLM의 대표적인 활용 패러다임으로 자리 잡은 ICL의 작동 원리를 규명하기 위해 다양한 측면에서 가설을 제안하였다. Chan et al. (2022) [83]은 ICL 능력이 학습 데이터의 분포 특성으로부터 기인한다고 주장하였다. Xie et al. (2022) [84]은 ICL을 암시적 베이지안 추론으로 설명하면서, 사전학습 분포가 은닉 마코프 모델의 혼합 (mixture of hidden Markov models) 형태를 따를 때 ICL 능력이 나타난다는 것을 증명하기 위해 합성 데이터 세트를 구성하였다. Garg et al. (2022) [85]은 알맞은 시연 예제가 주어질 경우, Transformer가 본 적 없는 선형 함수를 학습할 수 있는 효과적인 학습 알고리즘을 인코딩할 수 있음을 증명하였다. 그들은 또한 ICL에 인코딩된 학습 알고리즘이 최소 제곱 추정기의 오류와 비슷한 수준의 오류를 달성할 수 있음을 발견하였다. 또다른 연구들은 ICL과 경사 하강법 (gradient descent) 사이의 관계를 발견하려고 시도했으며, 특히 최근의 연구 [86]는 Transformer 기반의 in-context learner가 표준 미세조정 알고리즘을 암시적으로 구현할 수 있음을 발견했다. Dai et al. (2023) [87]은 Transformer attention과 경사 하강법 사이의 dual form을 밝혀냈고, 이에 따라 ICL을 암시적 미세조정 (implicit fine-tuning)으로 이해할 것을 제안하였다. 또한, GPT 기반 ICL과 실제 태스크에 대한 명시적인 미세조정을 비교한 결과, 여러 관점에서 ICL이 미세 조정과 유사하게 동작함을 발견하였다. Olsson et al. (2022) [88]은 Transformer 내에서 이전 패턴을 복사하여 다음 토큰을 완성하는 유도 헤드 (induction head)들이 존재함을 밝혔고, 이러한 기능이 ICL을 구현할 수 있음을 제시했다.

ICL 관점에서, 추론 능력을 보다 강화하기 위한 연구로 CoT [24]가 소개되었다. CoT는 입력과 출력 사이의 중간 추론 단계 (intermediate reasoning steps)을 시연 형태로 추가함으로써 이루어진다. CoT 프롬프팅은 입력-출력 매핑을 여러 중간 단계로 분해함으로써, 산술 추론 [89], 상식 추론 [90] 및 기호 추론 [24] 등의 복잡한 추론 태스크에서 LLM의 성능을 향상 시킬 수 있다. 최근에는 다양한 추론 경로 (multiple reasoning paths)를 생성하고 도출된 답변들간의 합의점을 찾는 형태로 기존의 CoT를 강화하는 연구들이 제안되기도 했다 [57, 91]. 이 외에도 재귀적인 프롬프팅 [92]을 통해서 compositional generalization [93] 능력이 요구되는 복잡한 태스크를 해결한 사례도 존재한다.

4.3.2 Augmented LLMs

LLM은 missing token prediction 목적 함수를 최적화하는 형태로 학습되기 때문에, 사실이 아니지만 구조적으로 그럴듯하게 보이는 컨텐츠를 생성하는 환각 등의 내재적인 한계를 지니고 있다. 또한, 자연어 코퍼스를 활용하여 학습되기 때문에, 주요 NLP 태스크가 아닌 산술 추론 (e.g., 1234+4321=?) 등에 약점을 보이기도 한다. 모델 크기 관점에서는, LLM의 창발 능력을 발휘하기 위해서는 대용량의 지식 등을 기억해야 하고 결과적으로 많은 수의 파라미터를 요구하게 된다. 기존 연구 [23] 에 따르면, 파라미터 수가 적은 언어 모델은 LLM에서 나타나는 in-context learning, instruction following, step-by-step reasoning과 같은 창발 능력이 발현되지 않는다고 한다. 즉, 좋은 성능의 LLM은 필연적으로 많은 수의 파라미터를 요구하게 되는 것이다. 이러한 내재적인 한계를 해결하고 적은 수의 파라미터로도 목적을 달성할 수 있도록, LLM을 추론 (reasoning) 및 도구 사용 (use tools) 관점에서 강화한 모델을 Augmented LLMs이라 부른다 [94]. 이 중에서 추론에 관한 내용은 프롬프팅을 통해서 고도화된 추론 능력을 LLM에게 부여하는 것으로, 앞서 4.3.1에서 논의한 ICL 및 CoT와 연관이 깊다. 따라서 해당 섹션에서는 도구 사용에 대해서 주로 논의할 것이다. 이외의 Augmented LLMs에 대한 심도 있는 논의는 다음 논문을 참조하길 바란다 [94].

Retrieval Augmented Generation LLM의 파라미터는 일종의 내부 메모리 모듈의 역할을 수행하는데, 특정한 태스크의 해결을 위해서는 context 내에 명시되지 않은 정보를 내재적으로 갖춰야 하는 경우가 많고, 그 결과로 파라미터의 수가 증가하게 된다. 그러나, 만약 LLM이 외부의 지식 또는 정보에 효과적으로 접근하며 그 정보를 활용할 수 있다면, 모든 지식을 내부 메모리에 저장하는 대신, 필요한 정보를 외부에서 추출하여 사용하는 방식으로 파라미터 수를 줄일 수 있을 것이다. 이러한 관점에서 볼 때, 검색 엔진과 같은 도구를 외부 메모리 모듈로 활용하는 LLM은 특정 쿼리와 관련된 정보를 빠르게 색인하고 추출하여 사실 기반의 답변 제공 및 최신 정보를 반영이 가능하며, 불필요한 지식의 저장을 최소화함으로써, 모델의 파라미터 수를 획기적으로 줄일 수 있다. 이러한 방법론을

Retrieval Augmented Generation (RAG) [95, 96, 97, 98]이라 한다.

Other Tools LaMDA [44]는 대화 애플리케이션에 특화된 LLM으로, 검색 모듈, 계산기 및 번역기 등의 외부 도구 호출 기능을 가지고 있다. WebGPT [99]는 웹 브라우저와의 상호작용을 통해 검색 쿼리에 사실 기반의 답변과 함께 출처 정보를 제공 한다. PAL [100]은 Python 인터프리터를 통한 복잡한 기호 추론 기능을 제공하며, 여러 관련 벤치마크에서 뛰어난 성능을 보여주었다. 다양한 종류의 API (e.g., 계산기, 달력, 검색, QA, 번역 등 단순한 API에서부터 Torch/TensorFlow/HuggingFace Hub에 이르는 복잡한 API까지) 호출 기능을 갖춘 연구들 [101, 102, 103, 104, 105] 역시 존재한다. Microsoft는 최근 발표한 position paper [104]에서 LLM과 같은 기반 모델을 뇌 (brain)와 같은 중앙 통제 시스템으로 사용하여 다양한 API를 연동하는 방식으로 자사 AI 제품의 청사진을 제시하였다. 이외의 참고할만한 사례로는, 오픈소스 프로젝트인 LangChain[18]과 상용 제품인 ChatGPT Plugins[19] 등이 있으며, 이들은 LLM을 기반으로 사용자가 원하는 외부 API와의 연동을 쉽게 할 수 있도록 설계되었다. 국내의 비슷한 사례로는 CLOVA X[20]가 있으나, 아직까지는 내부 및 제휴사 API만 연동된 것으로 보인다. 이외에도, 단순 도구 사용을 넘어서서 LLM을 활용한 가상 에이전트 [106] 및 물리적 로봇 [107] 제어에 활용한 사례도 존재한다.

5. 초거대 언어모델 평가 동향

예전부터 NLP 분야에서는 인간 수준의 성능을 달성하기 위하여 다양한 벤치마크 데이터셋이 개발되었다. 기존 벤치마크 데이터셋은 대부분 고도의 언어 이해 능력 [108, 109]과 일반 상식기반의 추론 능력을 측정하는데 초점을 맞추고 있다 [110, 111, 112, 113, 114]. LLM의 등장 이후, 대부분의 벤치마크 데이터셋의 유효성은 크게 낮아지고 변별력이 줄어들고 있는 상황이다. 한국에서도 KoBEST [115], KLUE [116]와 같은 벤치 마크 데이터셋이 물론 존재하지만, LLM을 평가하는데 적합한 형태로 보기 어렵다. 즉 LLM이 얼마나 정확한 지식을 내재하고 있으며, 이를 얼마나 적절하게 발현할 수 있는지에 대한 새로운 평가 척도의 개발이 절실하다. 이러한 연구와 평가 방법론의 변화는 언어 모델의 복잡성과 다양성이 증가함에 따라 NLP분야에 있어서 신중하게 고려되어야 할 중요한 주제이다.

5.1 OpenLLM Leaderboard

최근, HuggingFace는 OpenLLM Leaderboard를 공개하면서, 복수의 벤치마크 데이터셋을 통해 LLM의 성능을 체계적으로 평가하고 있다. LLM이 여전히 정복하지 못한 추론능력, 환각현상, 상식능력 등을 종합적으로 검증할 수 있는 리더보드이다. 이러한 평가 방식은 GLUE [108]와 SuperGLUE [109]를 중심으로 하던 언어 모델 평가의 패러다임을 전환시키고 있다. HuggingFace의 OpenLLM Leaderboard에서는 다음과 같은 4가지 종류의 평가 방법을 제시한다. **ARC** (**A**I2 **R**easoning **C**hallenge) [117]는 초등학교 수준의 과학 문제를 바탕으로 모델의 추론 능력을 평가한다. ARC는 2,590개의 challenge set과 5,197개의 easy set으로 구성되고 있으며, challenge set은 단어 중첩과 정보검색 알고리즘을 활용하여 재구성함으로써 모델이 오답을 선택하거나 어렵도록 유도한다. 이를 통해 모델의 추론 능력을 다각도로 평가할 수 있다.

HellaSWAG [118]는 일반 상식을 기반으로 한 추론 능력을 평가한다. 사람에게는 약 95%의 정답율을 지니는 쉬운 평가이지만, 적대적 필터링으로 인해서 모델에게 난해할 수 있는 선택지를 포함하였다. 이를 통해 모델의 일반상식 능력을 평가할 수 있다.

MMLU [119]는 언어모델이 광역 도메인의 지식에 대해서 사전 훈련 과정에서 얼마나 이를 습득하고 발현하는지 평가한다. 인문학, 사회학, 과학 등 57개의 도메인에 대해서 초등학교 수준부터 전문가의 영역까지의 문제 해결을 포함한다. 총 15,908개의 질의응답 쌍을 지니고 있으며, 각 도메인마다 최소 100개 이상의 예시를 포함하고 있다. 이를 통해 모델의 언어종합 이해능력도를 측정할 수 있다.

TruthfulQA [120]는 언어모델이 얼마나 높은 정보력을 바탕으로 신뢰성 있는 정보를 생산하는지 평가한다. 온라인에서 수집한 텍스트 정보는 허위 정보를 포함할 가능성이 높으며, 모델 사이즈가 커지는 것은 오히려 허위 정보를 모방할 가능성이 높아진다는 것을 가정한다. 38개의 도메인에 817개의 질의 쌍을 구성하고, zero-shot 세팅을 기본으로 모델의 성능을 측정 한다. 성능에 대한 평가는 진실성과 정보전달 측면으로 나누어 진행하며, 사람 평가자의 점수와 해당 점수로 학습을 진행한 모델을 활용한다. 이를 통해 모델의 환각현상에 얼마나 강건한지 평가할 수 있다. 국내

18) https://www.langchain.com
19) https://openai.com/blog/chatgpt-plugins
20) https://clova-x.naver.com

표 1 AI 윤리 원칙

	Humanity Human-centred	Responsibility Accountability	Privacy Security	Safety Reliability	Transparency Explainability
OECD	O	O	O	O	O
UNESCO	O		O	O	O
European Commission	O	O	O	O	O
미국 국가정보장실	O		O		O
호주 산업과학자원부	O	O	O	O	O
대한민국 과학기술정보통신부	O	O	O	O	O
사우디 데이터인공지능청	O	O	O	O	O
Google	O	O		O	
Microsoft		O	O	O	O
IBM	O		O		O
Adobe		O			O
OpenAI	O			O	
LG AI Research	O	O		O	O
Kakao	O		O		O
NAVER	O		O	O	O
SAMSUNG	O		O		O
SK Telecom	O		O	O	O
Upstage	O	O	O	O	O

의 경우도 많은 모델들이 OpenLLM Leaderboard에 참가하고 있으며, 특히 업스테이지가 두드러진 성과를 보였다. 업스테이지는 해당 리더보드에서 두 번이나 세계 1위의 자리를 차지한 뛰어난 성과를 보였다. 이로 인해 다양한 국내 기업들이 이 리더보드에서의 경쟁에 참여하게 되었으며, 국내 LLM 연구 분야 활성화에 일조하였다.

5.2 Open Ko-LLM Leaderboard

한국어에서도 Open LLM 리더보드가 운영되고 있다. Open Ko-LLM Leaderboard[21]라는 이름으로 NIA와 업스테이지에서 공동 주관을 하고 있으며, KT Cloud의 인프라 지원으로 운영되고 있다. Ko-HellaSwag, Ko-MMLU, Ko-Arc, Ko-Truthful QA, Ko-CommonGen V2의 총 5가지 태스크로 운영되고 있다. 기존 영어 OpenLLM Leaderboard에서 운영하고 있는 4개의 태스크를 한국어화 시킨 데이터에, 고려대학교 자연언어처리 연구실에서 구축한 Ko-CommonGen V2 밴치마크 데이터셋을 추가하여, 평가 지표로 활용하고 있는 리더보드이다.

해당 리더보드는 오픈 후 2주만에 100개가 넘는 모델들이 참여할 뿐만 아니라, 한국의 대표적인 Open LLM인 Polyglot-Ko [38], KULLM[22], KoAlpaca[23]와 더불어 42MARU[24], ETRI[25], Maum.AI[26] 등 다양한 기업들이 참여하고 있다. 오픈 초기 모델들은 평균 점수가 대부분 30점대 초반이었으나 2주만에 대부분 45점을 돌파하여 50%의 큰 향상폭을 보여주고 있다. 즉 다양한 모델들의 활발한 참여와 치열한 경쟁이 펼쳐지고 있다. 해당 리더보드를 통해 한국어 LLM 평가 생태계에 큰 기여를 하고 있으며, 현재 평가 허브 역할을 감당하고 있다.

6. 초거대 언어모델 윤리 원칙 동향

LLM을 포함한 인공지능 모델에 대한 적절한 개발과 올바른 활용을 위한 윤리 원칙이 필수적이다. 각 국제기구, 정부, 기업에서는 인공지능 윤리 원칙을 마련하여 인공지능을 개발하고 활용하는 주체들이 이를 준수하도록 방향을 제시하고 있다.

과학기술정보통신부가 2020년 12월 23일에 마련한 **인공지능 (AI) 윤리기준**은 최고 가치인 인간성(Humanity)을 위한 3대 기본원칙과 **10대 핵심요건**을 제시하고 있다. **3대 기본원칙**에는 인간성을 구현하기 위해 인공지능의 개발 및 활용 과정에서 1) 인간의 존엄성 원칙, 2) 사회의 공공선 원칙, 3) 기술의 합목적성 원칙을 지켜야 한다는 내용을 담고 있다. **10대 핵심요건**에는 3대 기본원칙을 실천하고 이행할 수 있도록 인공지능 개발부터 활용 전 과정에서 1) 인권 보장, 2) 프라이

21) https://huggingface.co/spaces/upstage/open-ko-llm-leaderboard
22) https://github.com/nlpai-lab/KULLM
23) https://github.com/Beomi/KoAlpaca
24) https://www.42maru.ai/kr/
25) https://www.etri.re.kr/intro.html
26) https://maum.ai/

버시 보호, 3) 다양성 존중, 4) 침해금지, 5) 공공성, 6) 연대성, 7) 데이터 관리, 8) 책임성, 9) 안정성, 10) 투명성 등의 요건이 충족되어야 한다는 내용이 포함되어 있다.

다양한 윤리 원칙에 명시된 내용들은 크게 6가지로 인간성, 책임성, 보안성, 안전성, 투명성, 다양성으로 구분된다.

인간성 (Humanity & Human-centered)은 인공지능의 개발과 활용은 인간과 사회에 유익한 가치를 제공하며 인간의 권리와 자유를 침해하지 않는다는 내용이다.

책임성 (Responsibility & Accountability)은 인공지능을 개발하고 활용하는 주체들의 역할과 책임을 명확히 설정하여 발생할 수 있는 피해를 최소화 한다는 내용이다.

보안성 (Privacy & Security)은 인공지능 개발 및 활용하는 과정에서 사용자의 개인정보와 프라이버시를 보호하기 위해 정보 보안을 고려하여 설계한다는 내용이다.

안전성 (Safety & Reliability)은 인공지능의 개발과 활용 과정에서 발생할 수 있는 잠재적 위험에 대응하고 안전하게 작동할 수 있도록 한다는 내용이다.

투명성 (Transparency & Explainability)은 인공지능의 작동 방식 또는 데이터 활용 방안에 대해 투명하게 공개하여 사용자들의 이해를 높이고 신뢰할 수 있도록 한다는 내용이다.

다양성 (Fairness & Diversity)은 인공지능을 개발하고 활용하는 과정에서 성별·연령·국적·인종·지역·종교 등에 대한 차별을 최소화하여 다양한 가치를 존중한다는 내용이다.

이외에도 각 기업마다 인공지능 윤리원칙을 제시하고 있으며 이에 대한 정보는 표 1과 같다.

7. 결 론

본 논문은, 초거대언어모델 (LLM)의 발전과 활용에 대한 근본적인 이해를 제공하려 하였다. LLM의 등장은 자연언어처리 (NLP)의 다양한 분야에서 혁신적인 변화를 가져왔으며, 이로 인해 번역, 요약, 질의응답, 형태소분석 등 다양한 태스크들이 하나의 모델로 수행될 수 있게 되었다. 데이터의 양적 확대, 컴퓨팅 기술의 진보, 그리고 알고리즘 및 기술의 발전은 LLM의 발전을 이끌었다. 이러한 기술적, 연구적 발전은 교육, 의료, 금융, 제조 등 다양한 산업 분야에서 LLM의 활용 가능성을 넓혔다. 그러나, LLM의 활용과 발전에는 여러 가지 도전 과제와 문제점이 존재한다. 편향성, 안전성, 설명 가능성 및 최신성 문제는 LLM의 한계점으로 지속적으로 고려되어야 하며, 이러한 문제점들을 해결하는 것은 다가오는 연구에서의 중요한 도전 과제로 남아 있다. 이러한 문제와 도전을 극복함으로써, LLM은 향후 NLP 연구와 산업계에서 더욱 중요한 역할을 차지할 것으로 예상된다. 이 분야의 연구가 계속 진행됨에 따라, 더욱 정교하고 다양한 어플리케이션의 등장이 기대되며, 이를 통해 사회적, 산업적 가치가 더욱 향상될 것이다. LLM의 연구와 활용은 계속되는 윤리적 고민과 함께 발전해 나가야 하며, 향후 연구는 이러한 모델의 가능성과 한계를 더욱 탐색하고 활용 하는 방향으로 전개될 것으로 보인다. 본 논문이 LLM에 대한 깊이 있는 이해를 제공하고, 이 분야의 연구자 및 전문가들에게 유익한 인사이트와 지침을 제공할 수 있기를 희망한다.

참고문헌

[1] J. Zhang, H. Feng, B. Liu, and D. Zhao, “Survey of technology in network security situation awareness,” *Sensors*, Vol. 23, No. 5, p. 2608, 2023.

[2] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” *arXiv preprint arXiv:1409.0473*, 2014.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L- . Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in neural infor- mation processing systems*, Vol. 30, 2017.

[4] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, “Scaling laws for neural language models,” *arXiv preprint arXiv:2001.08361*, 2020.

[5] I. L. Alberts, L. Mercolli, T. Pyka, G. Prenosil, K. Shi, A. Rominger, and A. Afshar-Oromieh, “Large lan- guage models (llm) and chatgpt: what will the impact on nuclear medicine be?” *European journal of nuclear medicine and molecular imaging*, Vol. 50, No. 6, pp. 1549 – 1552, 2023.

[6] M. Fraiwan and N. Khasawneh, “A review of chat- gpt applications in education, marketing, software engineering, and healthcare: Benefits, drawbacks, and research directions,” *arXiv preprint arXiv:2305.00237*, 2023.

[7] M. Sallam, N. Salim, M. Barakat, and A. Al-Tammemi, “Chatgpt applications in medical, dental, pharmacy, and public health education: A descriptive study high-

lighting the advantages and limitations," *Narra J*, Vol. 3, No. 1, pp. e103 – e103, 2023.

[8] A. Bahrini, M. Khamoshifar, H. Abbasimehr, R. J. Riggs, M. Esmaeili, R. M. Majdabadkohne, and M. Pasehvar, "Chatgpt: Applications, opportunities, and threats," *2023 Systems and Information Engineer- ing Design Symposium (SIEDS)*, pp. 274 – 279, 2023.

[9] O. Shaikh, H. Zhang, W. Held, M. Bernstein, and D. Yang, "On second thought, let's not think step by step! bias and toxicity in zero-shot reasoning," *arXiv preprint arXiv:2212.08061*, 2022.

[10] B. Hu, Q. Sheng, J. Cao, Y. Shi, Y. Li, D. Wang, and P. Qi, "Bad actor, good advisor: Exploring the role of large language models in fake news detection," *arXiv preprint arXiv:2309.12247*, 2023.

[11] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Ef- ficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[12] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532 – 1543, 2014.

[13] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the association for computational linguistics*, Vol. 5, pp. 135 – 146, 2017.

[14] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, Vol. 9, No. 8, pp. 1735 – 1780, 1997.

[15] K. Cho, B. Van Merri¨enboer, C. Gulcehre, D. Bah- danau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder- decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[16] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 2227 – 2237, Jun. 2018. [Online]. Available: https://aclanthology.org/N18-1202

[17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transform- ers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[18] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever *et al.*, "Improving language understanding by genera- tive pre-training," 2018.

[19] D. Hernandez, J. Kaplan, T. Henighan, and S. Mc- Candlish, "Scaling laws for transfer," *arXiv preprint arXiv:2102.01293*, 2021.

[20] R. Anil, A. M. Dai, O. Firat, M. Johnson, D. Lep- ikhin, A. Passos, S. Shakeri, E. Taropa, P. Bailey, Z. Chen *et al.*, "Palm 2 technical report," *arXiv preprint arXiv:2305.10403*, 2023.

[21] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, Vol. 1, No. 8, p. 9, 2019.

[22] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Ka- plan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sas- try, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, Vol. 33, pp. 1877 – 1901, 2020.

[23] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler *et al.*, "Emergent abilities of large language models," *arXiv preprint arXiv:2206.07682*, 2022.

[24] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in Neural Information Processing Systems*, Vol. 35, pp. 24 824 – 24 837, 2022.

[25] C. Yang, X. Wang, Y. Lu, H. Liu, Q. V. Le, D. Zhou, and X. Chen, "Large language models as optimizers," *arXiv preprint arXiv:2309.03409*, 2023.

[26] J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le, "Finetuned language models are zero-shot learners," *arXiv preprint arXiv:2109.01652*, 2021.

[27] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wain- wright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray *et al.*, "Training language models to follow in-structions with human feedback," *Advances in Neural Information Processing Systems*, Vol. 35, pp. 27 730 – 27 744, 2022.

[28] Z. Yuan, H. Yuan, C. Tan, W. Wang, S. Huang, and F. Huang, "Rrhf: Rank responses to align language models with human feedback without tears," *arXiv preprint arXiv:2304.05302*, 2023.

[29] OpenAI, "Gpt-4 technical report," 2023.

[30] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton,

S. Gehrmann *et al.*, “Palm: Scaling language modeling with pathways,” *arXiv preprint arXiv:2204.02311*, 2022.

[31] E. Almazrouei, H. Alobeidli, A. Alshamsi, A. Cap- pelli, R. Cojocaru, M. Debbah, E. Goffinet, D. Hes- low, J. Launay, Q. Malartic *et al.*, “Falcon-40b: an open large language model with state-of-the-art perfor- mance,” Technical report, Technology Innovation In- stitute, Tech. Rep., 2023.

[32] G. Penedo, Q. Malartic, D. Hesslow, R. Cojocaru, A. Cappelli, H. Alobeidli, B. Pannier, E. Almazrouei, and J. Launay, “The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only,” *arXiv preprint arXiv:2306.01116*, 2023.

[33] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozi`ere, N. Goyal, E. Hambro, F. Azhar *et al.*, “Llama: Open and efficient foundation language models,” *arXiv preprint arXiv:2302.13971*, 2023.

[34] H. Touvron, L. Martin, K. Stone, P. Albert, A. Alma- hairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, “Llama 2: Open foundation and fine-tuned chat models,” *arXiv preprint arXiv:2307.09288*, 2023.

[35] Anthropic, “Model card and evaluations for claude models,” 2023.

[36] A. Group, “Qwen technical report,” 2023.

[37] B. Kim, H. Kim, S.-W. Lee, G. Lee, D. Kwak, D. H. Jeon, S. Park, S. Kim, S. Kim, D. Seo *et al.*, “What changes can large-scale language models bring? intensive study on hyperclova: Billions-scale korean generative pretrained transformers,” *arXiv preprint arXiv:2109.04650*, 2021.

[38] H. Ko, K. Yang, M. Ryu, T. Choi, S. Yang, S. Park *et al.*, “A technical report for polyglot-ko: Open-source large-scale korean language models,” *arXiv preprint arXiv:2306.02254*, 2023.

[39] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill *et al.*, “On the opportuni- ties and risks of foundation models,” *arXiv preprint arXiv:2108.07258*, 2021.

[40] C. Qin, A. Zhang, Z. Zhang, J. Chen, M. Yasunaga, and D. Yang, “Is chatgpt a general-purpose natu- ral language processing task solver?” *arXiv preprint arXiv:2302.06476*, 2023.

[41] S. Longpre, G. Yauney, E. Reif, K. Lee, A. Roberts, B. Zoph, D. Zhou, J. Wei, K. Robinson, D. Mimno *et al.*, “A pretrainer’s guide to training data: Measur- ing the effects of data age, domain coverage, quality, & toxicity,” *arXiv preprint arXiv:2305.13169*, 2023.

[42] A. Lee, B. Miranda, and S. Koyejo, “Beyond scale: the diversity coefficient as a data quality metric demonstrates llms are pre-trained on formally diverse data,” *arXiv preprint arXiv:2306.13840*, 2023.

[43] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, “Exploring the limits of transfer learning with a unified text-to- text transformer,” *The Journal of Machine Learning Research*, Vol. 21, No. 1, pp. 5485–5551, 2020.

[44] R. Thoppilan, D. De Freitas, J. Hall, N. Shazeer, Kulshreshtha, H.-T. Cheng, A. Jin, T. Bos, L. Baker, Y. Du *et al.*, “Lamda: Language models for dialog applications,” *arXiv preprint arXiv:2201.08239*, 2022.

[45] T. L. Scao, A. Fan, C. Akiki, E. Pavlick, S. Ili’c, D. Hesslow, R. Castagn´e, A. S. Luccioni, F. Yvon, M. Gall´e *et al.*, “Bloom: A 176b-parameter open- access multilingual language model,” *arXiv preprint arXiv:2211.05100*, 2022.

[46] R. Taylor, M. Kardas, G. Cucurull, T. Scialom, A. Hartshorn, E. Saravia, A. Poulton, V. Kerkez, and R. Stojnic, “Galactica: A large language model for sci- ence,” *arXiv preprint arXiv:2211.09085*, 2022.

[47] Y. Li, D. Choi, J. Chung, N. Kushman, J. Schrit- twieser, R. Leblond, T. Eccles, J. Keeling, F. Gimeno, A. Dal Lago *et al.*, “Competition-level code genera-tion with alphacode,” *Science*, Vol. 378, No. 6624, pp. 1092–1097, 2022.

[48] H. Fu, Yao; Peng and T. Khot, “How does gpt obtain its ability? tracing emergent abilities of language mod- els to their sources,” *Yao Fu’s Notion*, Dec 2022.

[49] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong *et al.*, “A survey of large language models,” *arXiv preprint arXiv:2303.18223*, 2023.

[50] J. W. Rae, S. Borgeaud, T. Cai, K. Millican, J. Hoff- mann, F. Song, J. Aslanides, S. Henderson, R. Ring, S. Young *et al.*, “Scaling language models: Meth-ods, analysis & insights from training gopher,” *arXiv preprint arXiv:2112.11446*, 2021.

[51] D. Hernandez, T. Brown, T. Conerly, N. DasSarma, D. Drain, S. El-Showk, N. Elhage, Z. Hatfield-Dodds, T. Henighan, T. Hume *et al.*, “Scaling laws and interpretability of learning from repeated data,” *arXiv preprint arXiv:2205.10487*, 2022.

[52] K. Lee, D. Ippolito, A. Nystrom, C. Zhang, D. Eck, C. Callison-Burch, and N. Carlini, "Deduplicating training data makes language models better," *arXiv preprint arXiv:2107.06499*, 2021.

[53] N. Lukas, A. Salem, R. Sim, S. Tople, L. Wutschitz, and S. Zanella-B´eguelin, "Analyzing leakage of per-sonally identifiable information in language models," *arXiv preprint arXiv:2302.00539*, 2023.

[54] L. Niu, S. Mirza, Z. Maradni, and C. P¨opper, "*{*CodexLeaks*}*: Privacy leaks from code generation language models in *{*GitHub*}* copilot," *32nd USENIX Security Symposium (USENIX Security 23)*, pp. 2133 – 2150, 2023.

[55] J. Kaddour, J. Harris, M. Mozes, H. Bradley, R. Raileanu, and R. McHardy, "Challenges and ap- plications of large language models," *arXiv preprint arXiv:2307.10169*, 2023.

[56] R. Lou, K. Zhang, and W. Yin, "Is prompt all you need? no. a comprehensive and broader view of instruc- tion learning," *arXiv preprint arXiv:2303.10475*, 2023.

[57] Y. Wang, Y. Kordi, S. Mishra, A. Liu, N. A. Smith, D. Khashabi, and H. Hajishirzi, "Self-instruct: Align- ing language model with self generated instructions," *arXiv preprint arXiv:2212.10560*, 2022.

[58] C. Zhou, P. Liu, P. Xu, S. Iyer, J. Sun, Y. Mao, X. Ma, A. Efrat, P. Yu, L. Yu *et al.*, "Lima: Less is more for alignment," *arXiv preprint arXiv:2305.11206*, 2023.

[59] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, E. Li, X. Wang, M. Dehghani, S. Brahma *et al.*, "Scaling instruction-finetuned language mod- els," *arXiv preprint arXiv:2210.11416*, 2022.

[60] Y. Wang, S. Mishra, P. Alipoormolabashi, Y. Ko- rdi, A. Mirzaei, A. Arunkumar, A. Ashok, A. S. Dhanasekaran, A. Naik, D. Stap *et al.*, "Super- naturalinstructions: Generalization via declarative in- structions on 1600+ nlp tasks," *arXiv preprint arXiv:2204.07705*, 2022.

[61] H. Chen, Y. Zhang, Q. Zhang, H. Yang, X. Hu, X. Ma, Y. Yanggong, and J. Zhao, "Maybe only 0.5% data is needed: A preliminary exploration of low training data instruction tuning," *arXiv preprint arXiv:2305.09246*, 2023.

[62] S. Longpre, L. Hou, T. Vu, A. Webson, H. W. Chung, Y. Tay, D. Zhou, Q. V. Le, B. Zoph, J. Wei *et al.*, "The flan collection: Designing data and meth- ods for effective instruction tuning," *arXiv preprint arXiv:2301.13688*, 2023.

[63] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl *et al.*, "Large language models encode clinical knowledge," *arXiv preprint arXiv:2212.13138*, 2022.

[64] K. Singhal, T. Tu, J. Gottweis, R. Sayres, E. Wul- czyn, L. Hou, K. Clark, S. Pfohl, H. Cole-Lewis, D. Neal *et al.*, "Towards expert-level medical question answering with large language models," *arXiv preprint arXiv:2305.09617*, 2023.

[65] C. Xiao, X. Hu, Z. Liu, C. Tu, and M. Sun, "Law- former: A pre-trained language model for chinese legal long documents," *AI Open*, Vol. 2, pp. 79 – 84, 2021.

[66] S. Wu, O. Irsoy, S. Lu, V. Dabravolski, M. Dredze, S. Gehrmann, P. Kambadur, D. Rosenberg, and G. Mann, "Bloomberggpt: A large language model for finance," *arXiv preprint arXiv:2303.17564*, 2023.

[67] Y. Li, S. Ma, X. Wang, S. Huang, C. Jiang, H.-T. Zheng, P. Xie, F. Huang, and Y. Jiang, "Ecomgpt: Instruction-tuning large language model with chain-of-task tasks for e-commerce," *arXiv preprint arXiv:2308.06966*, 2023.

[68] S. Zhang, L. Dong, X. Li, S. Zhang, X. Sun, S. Wang, J. Li, R. Hu, T. Zhang, F. Wu *et al.*, "Instruction tuning for large language models: A survey," *arXiv preprint arXiv:2308.10792*, 2023.

[69] D. M. Ziegler, N. Stiennon, J. Wu, T. B. Brown, A. Radford, D. Amodei, P. Christiano, and G. Irv- ing, "Fine-tuning language models from human pref- erences," *arXiv preprint arXiv:1909.08593*, 2019.

[70] A. Askell, Y. Bai, A. Chen, D. Drain, D. Ganguli, T. Henighan, A. Jones, N. Joseph, B. Mann, N. Das- Sarma *et al.*, "A general language assistant as a labora- tory for alignment," *arXiv preprint arXiv:2112.00861*, 2021.

[71] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for nlp," *International Conference on Machine Learning*, pp. 2790 – 2799, 2019.

[72] Z. Hu, Y. Lan, L. Wang, W. Xu, E.-P. Lim, R. K.-W. Lee, L. Bing, and S. Poria, "Llm-adapters: An adapter family for parameter-efficient fine-tuning of large lan- guage models," *arXiv preprint arXiv:2304.01933*, 2023.

[73] X. L. Li and P. Liang, "Prefix-tuning: Optimizing continuous prompts for generation," *arXiv preprint arXiv:2101.00190*, 2021.

[74] B. Lester, R. Al-Rfou, and N. Constant, "The power of scale for parameter-efficient prompt tuning," *arXiv*

preprint arXiv:2104.08691, 2021.

[75] X. Liu, Y. Zheng, Z. Du, M. Ding, Y. Qian, Z. Yang, and J. Tang, "Gpt understands, too," *AI Open*, 2023.

[76] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[77] J. He, C. Zhou, X. Ma, T. Berg-Kirkpatrick, and G. Neubig, "Towards a unified view of parameter-efficient transfer learning," *arXiv preprint arXiv:2110.04366*, 2021.

[78] N. Ding, Y. Qin, G. Yang, F. Wei, Z. Yang, Y. Su, S. Hu, Y. Chen, C.-M. Chan, W. Chen *et al.*, "Parameter-efficient fine-tuning of large-scale pre- trained language models," *Nature Machine Intelli- gence*, Vol. 5, No. 3, pp. 220 – 235, 2023.

[79] A. Gholami, S. Kim, Z. Dong, Z. Yao, M. W. Mahoney, and K. Keutzer, "A survey of quantization methods for efficient neural network inference," *Low-Power Computer Vision*, pp. 291 – 326, 2022.

[80] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "Qlora: Efficient finetuning of quantized llms," *arXiv preprint arXiv:2305.14314*, 2023.

[81] Z. Yao, X. Wu, C. Li, S. Youn, and Y. He, "Zeroquant- v2: Exploring post-training quantization in llms from comprehensive study to low rank compensation," *arXiv preprint arXiv:2303.08302*, 2023.

[82] Q. Dong, L. Li, D. Dai, C. Zheng, Z. Wu, B. Chang, X. Sun, J. Xu, and Z. Sui, "A survey for in-context learning," *arXiv preprint arXiv:2301.00234*, 2022.

[83] S. Chan, A. Santoro, A. Lampinen, J. Wang, A. Singh, P. Richemond, J. McClelland, and F. Hill, "Data distributional properties drive emergent in-context learn- ing in transformers," *Advances in Neural Information Processing Systems*, Vol. 35, pp. 18 878 – 18 891, 2022.

[84] S. M. Xie, A. Raghunathan, P. Liang, and T. Ma, "An explanation of in-context learning as implicit bayesian inference," *arXiv preprint arXiv:2111.02080*, 2021.

[85] S. Garg, D. Tsipras, P. S. Liang, and G. Valiant, "What can transformers learn in-context? a case study of sim- ple function classes," *Advances in Neural Information Processing Systems*, Vol. 35, pp. 30 583 – 30 598, 2022.

[86] E. Akyu¨rek, D. Schuurmans, J. Andreas, T. Ma, and D. Zhou, "What learning algorithm is in-context learn- ing? investigations with linear models," *arXiv preprint arXiv:2211.15661*, 2022.

[87] D. Dai, Y. Sun, L. Dong, Y. Hao, S. Ma, Z. Sui, and F. Wei, "Why can gpt learn in-context? language models implicitly perform gradient descent as meta- optimizers," *ICLR 2023 Workshop on Mathematical and Empirical Understanding of Foundation Models*, 2023.

[88] C. Olsson, N. Elhage, N. Nanda, N. Joseph, N. DasSarma, T. Henighan, B. Mann, A. Askell, Y. Bai, A. Chen *et al.*, "In-context learning and induction heads," *arXiv preprint arXiv:2209.11895*, 2022.

[89] S.-Y. Miao, C.-C. Liang, and K.-Y. Su, "A diverse cor- pus for evaluating and developing english math word problem solvers," *arXiv preprint arXiv:2106.15772*, 2021.

[90] A. Talmor, J. Herzig, N. Lourie, and J. Berant, "Commonsenseqa: A question answering challenge targeting commonsense knowledge," *arXiv preprint arXiv:1811.00937*, 2018.

[91] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, and D. Zhou, "Rationale-augmented ensembles in language models," *arXiv preprint arXiv:2207.00747*, 2022.

[92] D. Zhou, N. Sch¨arli, L. Hou, J. Wei, N. Scales, X. Wang, D. Schuurmans, C. Cui, O. Bousquet, Q. Le *et al.*, "Least-to-most prompting enables complex rea- soning in large language models," *arXiv preprint arXiv:2205.10625*, 2022.

[93] B. Lake and M. Baroni, "Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks," *International conference on machine learning*, pp. 2873 – 2882, 2018.

[94] G. Mialon, R. Dess'ı, M. Lomeli, C. Nalmpantis, R. Pasunuru, R. Raileanu, B. Rozi`ere, T. Schick, J. Dwivedi-Yu, A. Celikyilmaz *et al.*, "Augmented language models: a survey," *arXiv preprint arXiv:2302.07842*, 2023.

[95] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Ku¨ttler, M. Lewis, W.-t. Yih, T. Rockt¨aschel *et al.*, "Retrieval-augmented gen- eration for knowledge-intensive nlp tasks," *Advances in Neural Information Processing Systems*, Vol. 33, pp. 9459 – 9474, 2020.

[96] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M. Chang, "Retrieval augmented language model pre-training," *International conference on machine learning*, pp. 3929 – 3938, 2020.

[97] S. Borgeaud, A. Mensch, J. Hoffmann, T. Cai, E. Rutherford, K. Millican, G. B. Van Den Driessche, J.-B. Lespiau, B. Damoc, A. Clark *et al.*, "Improv- ing language models by retrieving from trillions of to- kens,"

International conference on machine learning, pp. 2206 – 2240, 2022.

[98] G. Izacard, P. Lewis, M. Lomeli, L. Hosseini, F. Petroni, T. Schick, J. Dwivedi-Yu, A. Joulin, S. Riedel, and E. Grave, "Few-shot learning with re- trieval augmented language models," *arXiv preprint arXiv:2208.03299*, 2022.

[99] R. Nakano, J. Hilton, S. Balaji, J. Wu, L. Ouyang, C. Kim, C. Hesse, S. Jain, V. Kosaraju, W. Saun- ders *et al.*, "Webgpt: Browser-assisted question- answering with human feedback," *arXiv preprint arXiv:2112.09332*, 2021.

[100] L. Gao, A. Madaan, S. Zhou, U. Alon, P. Liu, Y. Yang, J. Callan, and G. Neubig, "Pal: Program-aided lan- guage models," *International Conference on Machine Learning*, pp. 10 764 – 10 799, 2023.

[101] T. Schick, J. Dwivedi-Yu, R. Dess`ı, R. Raileanu, M. Lomeli, L. Zettlemoyer, N. Cancedda, and T. Scialom, "Toolformer: Language models can teach themselves to use tools," *arXiv preprint arXiv:2302.04761*, 2023.

[102] S. G. Patil, T. Zhang, X. Wang, and J. E. Gonzalez, "Gorilla: Large language model connected with mas- sive apis," *arXiv preprint arXiv:2305.15334*, 2023.

[103] S. Hao, T. Liu, Z. Wang, and Z. Hu, "Toolkengpt: Augmenting frozen language models with massive tools via tool embeddings," *arXiv preprint arXiv:2305.11554*, 2023.

[104] Y. Liang, C. Wu, T. Song, W. Wu, Y. Xia, Y. Liu, Y. Ou, S. Lu, L. Ji, S. Mao *et al.*, "Taskma- trix. ai: Completing tasks by connecting founda- tion models with millions of apis," *arXiv preprint arXiv:2303.16434*, 2023.

[105] Y. Qin, S. Liang, Y. Ye, K. Zhu, L. Yan, Y. Lu, Y. Lin, X. Cong, X. Tang, B. Qian *et al.*, "Toolllm: Facilitat- ing large language models to master 16000+ real-world apis," *arXiv preprint arXiv:2307.16789*, 2023.

[106] S. Li, X. Puig, C. Paxton, Y. Du, C. Wang, L. Fan, T. Chen, D.-A. Huang, E. Akyu¨rek, A. Anandku- mar *et al.*, "Pre-trained language models for interac- tive decision-making," *Advances in Neural Information Processing Systems*, Vol. 35, pp. 31 199 – 31 212, 2022.

[107] A. Zeng, M. Attarian, B. Ichter, K. Choromanski, A. Wong, S. Welker, F. Tombari, A. Purohit, M. Ryoo, V. Sindhwani *et al.*, "Socratic models: Composing zero-shot multimodal reasoning with language," *arXiv preprint arXiv:2204.00598*, 2022.

[108] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "Glue: A multi-task benchmark and analysis platform for natural language understanding," *International Conference on Learning Representations*, 2018.

[109] A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman, "Su- perglue: A stickier benchmark for general-purpose lan- guage understanding systems," *Advances in neural information processing systems*, Vol. 32, 2019.

[110] R. Zellers, Y. Bisk, R. Schwartz, and Y. Choi, "Swag: A large-scale adversarial dataset for grounded commonsense inference," *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 93 – 104, 2018.

[111] L. Huang, R. Le Bras, C. Bhagavatula, and Y. Choi, "Cosmos qa: Machine reading comprehension with contextual commonsense reasoning," *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2391 – 2401, 2019.

[112] A. Talmor, J. Herzig, N. Lourie, and J. Berant, "Commonsenseqa: A question answering challenge target- ing commonsense knowledge," *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Pa- pers)*, pp. 4149 – 4158, 2019.

[113] Y. Bisk, R. Zellers, J. Gao, Y. Choi *et al.*, "Piqa: Reasoning about physical commonsense in natural lan- guage," *Proceedings of the AAAI conference on artifi- cial intelligence*, Vol. 34, No. 05, pp. 7432 – 7439, 2020.

[114] N. Mostafazadeh, A. Kalyanpur, L. Moon, D. Buchanan, L. Berkowitz, O. Biran, and J. Chu- Carroll, "Glucose: Generalized and contextualized story explanations," *Proceedings of the 2020 Con- ference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4569 – 4586, 2020.

[115] M. Jang, D. Kim, D. S. Kwon, and E. Davis, "Kobest: Korean balanced evaluation of significant tasks," *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 3697 – 3708, 2022.

[116] S. Park, J. Moon, S. Kim, W. I. Cho, J. Han, J. Park, C. Song, J. Kim, Y. Song, T. Oh *et al.*, "Klue: Korean language understanding evaluation," *arXiv preprint arXiv:2105.09680*, 2021.

[117] P. Clark, I. Cowhey, O. Etzioni, T. Khot, A. Sabhar- wal, C. Schoenick, and O. Tafjord, “Think you have solved question answering? try arc, the ai2 reasoning challenge,” *arXiv preprint arXiv:1803.05457*, 2018.

[118] R. Zellers, A. Holtzman, Y. Bisk, A. Farhadi, and Y. Choi, “Hellaswag: Can a machine really finish your sentence?” *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4791 – 4800, 2019.

[119] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt, “Measuring massive multitask language understanding,” *Interna- tional Conference on Learning Representations*, 2020.

[120] S. Lin, J. Hilton, and O. Evans, “Truthfulqa: Measur- ing how models mimic human falsehoods,” *Proceed- ings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3214 – 3252, 2022.

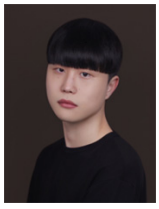
약 력

박 찬 준

2019 부산외국어대학교 언어처리창의융합과 졸업 (학사)
2023 고려대학교 컴퓨터학과 졸업 (박사)
2018~2019 SYSTRAN Research Engineer
2022~현재 Upstage Technical Leader
관심분야 : 자연언어처리, 초거대언어모델, 기계번역, 데이터중심 인공지능
Email : chanjun.park@upstage.ai

이 원 성

2012 연세대학교 정보산업공학과 졸업 (학사)
2018 KAIST 산업및시스템공학과 졸업 (박사)
2018~2021 SK Telecom Data Scientist
2021~현재 Upstage Technical Leader
관심분야 : 추천시스템, 초거대언어모델, 개인화 AI
Email : wonsung.lee@upstage.ai



김 윤 기

2020 한양대학교 산업공학과 졸업 (학사)
2023 한양대학교 컴퓨터소프트웨어학과 졸업 (석사)
2023~현재 Upstage AI Research Engineer
관심분야 : 추천시스템, 초거대언어모델, 초개인화 AI
Email : eddie@upstage.ai

김 지 후

2019 경희대학교 산업경영공학과 졸업 (학사)
2021 한양대학교 컴퓨터소프트웨어학과 졸업 (석사)
2021~현재 Upstage AI Research Engineer
관심분야 : 추천시스템, 초거대언어모델, 초개인화 AI
Email : jerry@upstage.ai

이 활 석

2011 KAIST 전기 및 전자공학 졸업 (박사)
2011~2016 한화테크윈 선행기술 연구원 비전기술 그룹 연구원
2016~2017 NCSOFT AI Center AI Lab Vision TF 연구원
2017~2020 네이버 Clova Visual AI 책임리더
2020~현재 Upstage CTO
관심분야 : 초거대언어모델, OCR
Email : hwalsuk.lee@upstage.ai