

화학 도메인에 특화된 거대 언어 모델 구축을 위한 지시어 튜닝

이정민¹ · 김현지¹ · 이성수² · 김연수² · 이규황² · 김성범^{1*}

¹고려대학교 산업경영공학과 / ²LG 화학

Instruction Tuning for Developing Large Language Models Specialized in Chemical Domain

Jungmin Lee¹ · Hyeonji Kim¹ · Sungsu Lee² · Yunsoo Kim² · Kyuhwang Lee² · Seoung Bum Kim¹

¹Department of Industrial and Management Engineering, Korea University

²LG Chemical

In recent years, notable progress in natural language processing (NLP) has been attributed to the advent of large language models (LLM), exemplified by innovations such as chat generative pre-trained transformer (ChatGPT) developed by OpenAI. Complementary LLMs such as Llama and Vicuna, are also being rapidly developed. Despite these advancements, LLMs still have limitations because of their training on generic text sources like Wikipedia. While some LLMs are trained on data spanning multiple domains, they often exhibit suboptimal performance in specialized domains like chemistry or finance. This study aims to address this limitation by developing an LLM specialized for the chemical domain through instruction tuning. We use a diverse dataset in both Korean and English, encompassing multiple tasks such as summarization, keyword extraction, and question answering for instruction tuning. The baseline model for our study is Llama2. Evaluation using both Korean and English datasets demonstrates that our specialized Llama2 outperforms its untuned counterpart in both languages. This demonstrates the effectiveness of our approach in enhancing the performance of models across languages and domains, particularly in chemistry. Additionally, we propose efficient instruction tuning strategies through various experiments.

Keywords: Chemical, Large Language Models, Llama2, Instruction Tuning

1. 서 론

지속적인 언어 모델들의 등장으로 자연어 처리(natural language processing, NLP) 분야는 최근 몇 년 동안 큰 발전을 이루어 왔다. Transformer(Vaswani *et al.*, 2017)의 등장과 함께, 이를 활용한 bidirectional encoder representations from transformers (BERT)(Devlin *et al.*, 2018)와 같이 많은 양의 텍스트 데이터를 활용해 일반적인 수준의 언어 이해가 가능하도록 사전 학습된 언어 모델(pre-trained language models, PLM)이 등장하기 시작했다. 그러나 PLM은 새로운 작업을 수행할 때, 그 작업에 대한 충분한 양의 데이터로 미세 조정을 추가로 수행해주어야 한다

는 한계점을 지니고 있다. 거대 언어 모델(large language models, LLM)은 기존보다 방대한 양의 텍스트 데이터와 함께 수천 개 이상의 파라미터를 활용하여 이러한 문제점을 극복하였다. LLM은 사전 학습된 단일 모델이 여러 작업에 사용될 수 있으며 특정 작업을 위한 미세 조정을 진행할 시, 데이터를 많이 필요로 하지 않는다. 또한, 학습 데이터와 모델의 파라미터를 증가시킴으로써 성능을 지속적으로 향상시킬 수 있어 generative pre-trained transformers(GPT-1)(Radford *et al.*, 2018)을 확장시킨 GPT-2, GPT-3, GPT-3.5, GPT-4, 그리고 GPT-4o(Radford *et al.*, 2019; Brown *et al.*, 2020; Ye *et al.*, 2023; Achiam *et al.*, 2023; Shahriar *et al.*, 2024)가 연속적으로 개발되었다. 이에 더

* 연락저자 : 김성범 교수, 02841, 서울특별시 성북구 안암로 145 고려대학교 산업경영공학부, Tel: 02-3290-3397, Fax: 02-929-5888,

E-mail : sbkim1@korea.ac.kr

2024년 7월 29일 접수; 2024년 9월 13일 수정본 접수; 2024년 11월 5일 게재 확정.

해, GPT 모델에 reinforcement learning with human feedback (RLHF)을 적용시킨 InstructGPT(Ouyang *et al.*, 2022)와 현재도 많은 사용자들을 보유한 chat generative pre-trained transformers(ChatGPT)(OpenAI, 2023)가 개발되었다. 그러나 이러한 LLM은 모델 학습에 큰 컴퓨팅 비용을 필요로 한다는 한계점을 가지고 있어 이를 극복하기 위한 Switch Transformer(Fedus *et al.*, 2022), Chinchilla(Hoffmann *et al.*, 2022)와 같은 모델들이 개발되었다. 메타(Meta)에서는 오픈 소스로 누구나 쉽게 사용할 수 있도록 Llama(Touvron *et al.*, 2023)를 개발하여 사용자들에게 큰 편리함을 제공해주었으며 다양한 응용 연구의 가능성을 열었다. Llama는 학습 컴퓨팅 비용보다 추론 컴퓨팅 비용에 초점을 두었고 Chinchilla에서 제안한 것 보다 더 적은 수의 파라미터와 다량의 고품질 데이터를 통해 보다 우수한 성능을 보였다. 그 후, 기존 Llama보다 더 많은 데이터셋으로 학습된 Llama2(Touvron *et al.*, 2023)가 개발되었다.

이렇듯 LLM은 지속적으로 발전해왔지만, 사용자가 원하는 형태의 답변 생성이 어려운 경우가 발생했고, 이 한계점을 극복하고자 LLM에 프롬프트 엔지니어링(prompt engineering)(Liu *et al.*, 2023)과 지시어 튜닝(instruction-tuning)(Wei *et al.*, 2021)이 도입되었다. 프롬프트 엔지니어링이란 사전 학습된 LLM에 텍스트 형태의 프롬프트를 입력하여 필요한 배경 정보나 추가적인 세부 사항을 LLM에 제공하는 방식을 의미한다. 이는 프롬프트의 작은 변화만으로도 성능이 크게 달라질 수 있어 프롬프트를 미세하게 조정하는 과정이 필요하며, 작업마다 적합한 프롬프트를 새로 설정해주어야 한다. 지시어 튜닝이란 “(지시어, 입력 문장, 출력 문장)”을 쌍으로 하는 데이터셋을 통해 모델을 지도 학습 형식으로 학습시키는 과정을 말한다. 지시어 튜닝은 여러 작업에 대해 지시어를 통해 LLM을 학습시키기 때문에 프롬프트 엔지니어링과 비교했을 때, 새로운 작업에 대해 보다 일반화된 성능을 도출할 수 있다. 이러한 지시어 튜닝을 통해 LLM은 few-shot 환경에서 뿐만 아니라 zero-shot 환경에서도 우수한 성능을 달성했다. 그러나 다량의 지시어 튜닝 데이터셋을 확보하는 데에는 큰 비용이 들기 때문에, LLM을 통해 지시어 튜닝 데이터셋을 확보하는 연구가 진행되었고, 이를 통해 LLM은 한 단계 더 발전하게 되었다.

LLM은 대량의 파라미터를 가지고 있기 때문에 이를 튜닝할 때는 방대한 계산 자원이 필수적이다. 이를 위해, LLM의 튜닝을 효율적으로 수행하기 위해 low-rank adaptation(Lora) 방법론이 제안되었다(Hu *et al.*, 2021). 해당 방법론은 기존의 사전 학습된 모델을 전체 재학습하는 대신, 일부 파라미터를 저차원 공간으로 분해하여 적은 수의 추가 파라미터만 학습하는 방식이다. 이를 통해, 학습 과정에서 계산 비용을 줄이면서 기존 모델의 사전 학습 지식을 유지할 수 있다.

지시어 튜닝을 통해 LLM이 zero-shot 상황에서 우수한 성능을 보여주고 있지만, 화학과 같은 특수한 도메인에 대해서는 최적의 성능을 보여주지 못하고 있다. 화학 도메인 데이터는 일반적인 문장과 달리 특징적인 형태를 가지고 있다. 예를 들

어, 화학 물질 명은 알파벳과 특수 기호, 숫자 등이 혼합되어 일반 LLM이 특정 개체명으로 인식하는 것이 어렵다. 이렇게 화학 도메인에서만 존재하는 독특한 개체와 개념들로 인해 화학 도메인과 일반 도메인 데이터의 단어 분포는 크게 다르다. 이와 더불어 최신 기술의 개발로 인해 새로운 개체와 개념들이 지속적으로 등장하고 있어 이를 효과적으로 처리하기 위한 방법론이 요구된다. 따라서 화학 도메인에 특화된 LLM을 구축하기 위해 화학 도메인의 데이터를 활용한 추가적인 학습이 필수적인 상황이다.

본 연구에서는 화학 도메인에 특화된 LLM을 구축하기 위해 이에 대한 다양한 한국어 및 영어 데이터셋을 통해 지시어 튜닝을 수행한다. 우선적으로, 화학 도메인에 대한 데이터가 부족한 상황에서, GPT-3와 GPT-4 모델을 통해 요약 및 추출, 요약 및 번역 그리고 텍스트 분류 등의 작업에 대해 지시어 튜닝 데이터셋을 생성한다. 생성된 데이터셋과 더불어 허깅페이스(Huggingface)에 공개된 오픈 데이터셋을 같이 사용하여 기존 Llama2 모델에 추가 지시어 튜닝을 수행한다. 본 연구에서는 지시어 튜닝을 통해 화학 도메인에서 LLM의 성능 향상을 보여주고 다양한 실험을 통해 효과적인 지시어 튜닝 활용 방안을 제시한다. 또한 다양한 특수 도메인에서 지시어 튜닝을 통한 LLM의 성능 향상에 대한 가능성을 보여준다. 본 연구의 주요 기여점은 다음과 같다.

- 화학 도메인의 한국어 및 영어 지시어 튜닝 데이터셋을 직접 생성 및 구축하여 사용한 첫 번째 사례이다. 특수한 도메인의 지시어 튜닝 데이터셋이 충분하지 않은 상황에서 LLM APIs를 이용하여 추가적인 데이터셋을 생성한다.
- 화학 도메인에서 한국어 및 영어 데이터셋을 활용한 다양한 지시어 튜닝을 통해 두 가지 언어에 대한 데이터셋에서 LLM의 성능 향상을 실험적으로 보여주었다. 이를 통해 다양한 지시어 튜닝의 효과를 확인했을 뿐만 아니라 이러한 지시어 튜닝을 활용할 수 있는 효과적인 방안을 제시한다. 최종적으로, 다른 특수한 도메인에서도 지시어 튜닝을 수행할 때 LLM의 성능 향상 가능성을 보여주었다.

본 논문은 다음과 같이 구성된다. 제2장에서는 제안 방법론과 관련한 연구에 대해 소개하고, 제3장에서는 제안하는 화학 도메인 특화 LLM에 대해 설명한다. 제4장에서는 사용한 데이터셋과 실험 방법 및 결과를 설명하고, 마지막 제5장에서는 본 연구의 결론과 기대 효과를 다루도록 한다.

2. 배경

2.1 거대 언어 모델(Large Language Models)

LLM은 수천 개 이상의 파라미터를 가지고, 방대한 데이터로 광범위하게 학습된 대규모 언어 모델을 말한다. 최근

ChatGPT와 같은 우수한 성능의 모델들이 급속히 발전하면서 큰 관심을 받고 있다. LLM은 방대한 양의 텍스트 데이터로 다양한 언어 현상과 지식을 습득하여 자연어에 대한 높은 이해도를 가진다. 이러한 이해도를 바탕으로, 질의응답, 텍스트 요약 등 다양한 작업에서 뛰어난 능력을 보여주고 있다. ChatGPT는 대화형 질문을 이해할 뿐만 아니라 대화 내용을 기억하고 맥락에 맞는 답변을 생성해주어 후속 LLM 연구에 큰 영향을 주었다. 이후 등장한 오픈 소스 LLM은 누구나 접근하고 활용할 수 있어 이를 활용한 연구들이 활발히 진행되고 있다. Meta에서 공개한 Llama는 OpenAI의 GPT-3나 DeepMind의 Chinchilla 보다 적은 수의 파라미터를 가지고 있지만, 다량의 고품질 데이터셋을 사용한 학습을 통해 효율성을 높여 훨씬 적은 리소스로도 뛰어난 성능을 보여준다. 이후 공개된 Llama2는 Llama보다 약 40% 이상의 학습 데이터셋을 사용하여, 기존 오픈 소스 LLM 중 가장 우수한 성능을 보였고 상업적으로도 이용이 가능해 다양한 연구에서 활발히 활용되고 있다 (Touvron *et al.*, 2023; Liu *et al.*, 2023).

2.2 지시어 튜닝 (Instruction Tuning)

LLM 발전의 핵심 요소 중 하나는 지시어 튜닝이다. 지시어 튜닝은 LLM의 zero-shot 성능을 향상시키는 전략으로, 각 작업에 대한 문제와 답을 명시적으로 학습함으로써 언어 모델에 의미 있는 정보를 제공하며, 사용자의 선호에 맞는 답변을 생성할 수 있도록 한다. 학습에 사용되는 지시어가 다양할수록 모델을 더 강건하게 학습시킬 수 있다. 137B 크기의 파라미터를 가지는 사전 학습 모델에 60개의 다양한 NLP 데이터셋을 통한 지시어 튜닝으로 LLM의 성능을 향상시킬 수 있었다. 그러나 이러한 지시어 튜닝 데이터셋을 수집하는 데는 많은 비용이 들며, 충분한 다양성을 확보하기 어렵다. 이러한 문제를 해결하기 위해 LLM을 사용하여 지시어 튜닝 데이터셋을 반복적으로 생성하는 Self-instruct(Wang *et al.*, 2022) 프레임워크가 제안되었다. 또한, ChatGPT와 GPT-4의 우수한 성능은 지시어 튜닝을 통해 오픈 소스 LLM을 개선하는 데 큰 기회를 제공한다. 예를 들어, Stanford Alpaca(Taori *et al.*, 2023)는 GPT-3.5로 생성한 52,000개의 지시어 데이터를, Vicuna(Chiang *et al.*, 2023)는 shared user-ChatGPT에서 공유된 약 700,000개의 지시어 데이터를, 그리고 GPT-4-LLM(Peng *et al.*, 2023)는 GPT-4로 생성한 52,000개의 지시어 데이터를 사용하여 오픈 소스 LLM인 Llama를 개선시켰다. 이와 같은 확장 가능한 지시어 튜닝 기법을 통해 기초 모델을 의학이나 법률과 같은 다양한 도메인에 쉽게 활용하는 연구들이 활발히 진행되고 있다.

그러나 LLM은 무수히 많은 양의 파라미터를 가지고 있기 때문에, 이를 추가적으로 튜닝하기 위해서는 방대한 계산 자원 뿐만 아니라 긴 시간이 필수적이다. 이를 해결하기 위해, LLM에 추가적인 학습을 수행 시, 학습을 효율적이고 자원 절약적으로 수행하기 위해 Lora 방식을 적용하기 시작하였다(Qin *et al.*,

2024). Lora는 기존 사전 학습된 모델을 학습할 때, 모델을 전체적으로 업데이트하지 않고 일부 파라미터를 저차원 공간으로 분해하여 특정 도메인에 맞춘 적은 수의 추가 파라미터만 업데이트하는 방식으로 적용된다. 사전 학습된 LLM의 가중치 행렬 $W \in R^{d \times s}$ 를 저차원 행렬 $A \in R^{d \times r}$, $B \in R^{r \times s}$ 로 근사하여 사용하며, r 은 d 와 s 에 비해 매우 작은 차원을 의미한다. 가중치 행렬 W 을 저차원으로 분해하는 과정은 식 (1)과 같다.

$$W_0 + \Delta W = W^0 + AB \quad (1)$$

이 때, W_0 는 파라미터 업데이트 시, 고정되는 가중치 행렬을 의미한다. 이렇게 저차원의 행렬 A 와 B 만 업데이트하여, 학습 과정에서 계산 비용을 줄이면서 기존 모델의 사전 학습 지식을 유지할 수 있다.

2.3 특수 도메인에서의 LLM

LLM을 특정 도메인에 특화시키기 위한 연구들이 활발히 진행되고 있다. 전자 상거래 도메인을 위한 지시어 데이터셋인 EcomInstruct가 구축되었고, 이를 사용해 BLOOMZ(Le *et al.*, 2023)을 학습시켜 EcomGPT가 개발되었다(Li *et al.*, 2023). 또한, 금융 도메인을 위한 오픈 소스 LLM인 FinGPT(Yang *et al.*, 2023)와, 재료 과학 도메인을 위한 Llama2 기반의 오픈 소스 LLM 챗봇이 개발되었다(Yang *et al.*, 2024). 의학 분야에서는 환자와 의사 데이터셋을 활용해 Llama-7B를 미세 조정된 ChatDoctor(Yunxiang *et al.*, 2023)와 방사선 보고서 데이터셋을 사용해 Alpaca-7B를 추가 학습한 Radiology-GPT(Liu *et al.*, 2023)가 개발되었다. 화학 도메인에 맞게 LLM을 학습시키기 위해서는 화학 분자 구조와 텍스트로 이루어진 데이터셋을 통해 GPT-3를 세 가지 작업으로 미세 조정된 연구가 수행되었다(Jablonka *et al.*, 2024). 이 연구에서 수행한 세 가지 미세 조정 작업은 분자 구조와 텍스트를 통한 분류 및 회귀 작업, 그리고 텍스트를 분자 구조로 변환하는 작업들로 구성된다. 본 연구에서는 텍스트 데이터만을 이용하였으며, 화학 도메인에 대해 보다 다양한 작업 뿐만 아니라, 한국어와 영어, 두 가지의 언어에 대해 지시어 튜닝을 수행하여 두 가지 언어 모두에서의 화학 도메인에 특화된 LLM을 구축한다.

3. 방법론

제안 방법론은 화학 도메인에 특화된 LLM을 구축하기 위한 지시어 튜닝 방법론으로, 지시어 튜닝 데이터셋 구축 단계와 지시어 튜닝 단계로 구성된다. 제안 방법론의 전체적인 구조는 <Figure 1>과 같다. Step 1인 지시어 튜닝 데이터셋 구축 단계에서는 EcomInstruct(Li *et al.*, 2023)와 같이 LLM APIs와 기존 화학 도메인의 다양한 텍스트 데이터를 활용하여 지시어

튜닝 데이터셋을 구축했으며, LLM으로부터 생성되는 답변을 pseudo-label로 정의했다. 이 때 사용한 LLM은 화학 도메인에 관한 정보보다는 텍스트의 맥락을 이해하는 능력이 강하게 요구되며, LLM은 다양한 텍스트로부터 맥락을 잘 이해할 수 있도록 사전 학습되었기 때문에 지시어 튜닝 데이터셋을 생성할 때 충분히 신뢰할 수 있다. 이를 활용하여 지시어 튜닝 데이터셋을 생성 시, 화학 도메인 내 일반화 성능과 적용 가능성을 향상시키기 위해, 다양한 작업에 대해 데이터셋을 생성했다. Step 2인 지시어 튜닝 단계에서는 충분히 확보된 튜닝 데이터셋을 기반으로 Llama2 모델을 지시어 튜닝하여 최종 모델을 구축했다.

3.1 지시어 튜닝 데이터셋 생성

화학과 같이 특수한 도메인에 대한 데이터셋은 일반적인 데

이터셋에 비해 부족하기 때문에, 본 연구에서는 <Figure 1>의 Step 1과 같이 화학 분야 텍스트 데이터를 추가적으로 생성하였다. 추가적인 지시어 튜닝 데이터셋은 기존 화학 분야 데이터셋을 기반으로 GPT-3.5, GPT-4를 사용해 생성하였다. 생성된 데이터셋은 “(쿼리, 지시어, 답변)” 쌍으로 구성되며, 요약 및 추출, 요약 및 번역, 정보 추출, 텍스트 분류 작업의 지시어 튜닝에 사용했다.

요약 및 추출 작업에 사용한 데이터셋은 화학 분야 논문 초록을 바탕으로 구성된 기존 Royal Society of Chemistry Summary(RSCSUM)(Chen *et al.*, 2020) 데이터셋을 변형하여 사용했다. 이 데이터셋은 “(요약문, “제목”, “초록)” 형태로 구성되어 있으며, 평가 데이터를 챗(chat) 스타일로 변환하여 사용했다. 논문 초록을 GPT 모델에 입력하여, 이에 대한 요약문과 핵심 단어를 출력하도록 한다. 이 때, 논문 초록과 함께 ‘해당 초록의 요약문 및 언급된 핵심 화학 물질을 작성하라’

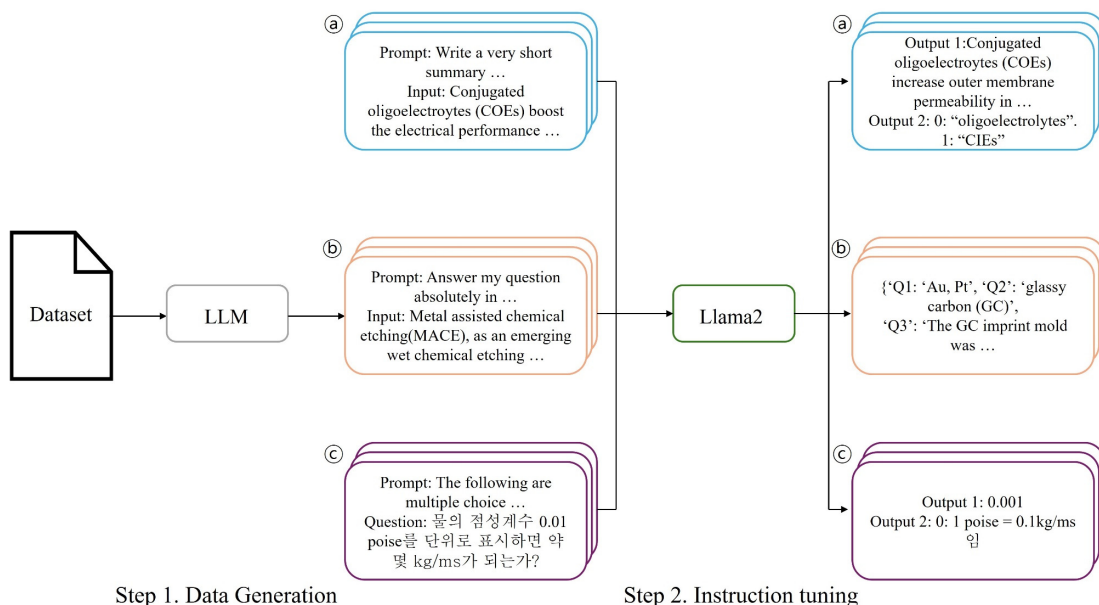


Figure 1. Overall framework of our instruction tuning. In Step 1, instruction tuning datasets are generated through LLM APIs. In Step 2, instruction tuning is performed on Llama2 for various tasks. (a) Example of a summarization and extraction task, (b) Example of an information extraction task, and (c) Example of a text classification task

Table 1. Example of a Summarization and Extraction Task

Input
Abstract: “Conjugated oligoelectrolytes (COEs) boost the electrical performance of a wide range of bioelectrochemical systems, yet their mechanism of action remains incompletely understood. One possible ...”
Prompt: You are an expert in Chemistry and reviewer for the Royal Society of Chemistry. Write a very short summary of this text and a list of chemical compounds mentioned in the text.”
Output
Summary: “Conjugated oligoelectrolytes (COE) increase outer membrane permeability in Escherichia coli, improve transport of small molecules through the cell envelope and thus accelerate whole-cell catalysis.”
Chemnames: {0: “oligoelectrolytes”, 1: “COEs”}

는 지시어를 모델에 입력한다. 이러한 입력 데이터와 지시어를 기반으로 GPT 모델은 요약문과 핵심 화학 물질을 출력하여, 최종적으로 (‘Abstract’: 논문 초록, ‘Prompt’: 지시어, ‘Output’: {요약문, 언급된 화학 물질}) 형태의 데이터셋을 얻게 된다. 해당 작업에 대한 데이터셋 예시는 <Table 1>에 정리하였다.

요약 및 추출 작업 외 작업들의 데이터셋은 LG 화학(주) 내부의 화학 도메인 문서 데이터를 기반으로 생성하였다. 요약 및 번역 작업은 화학 도메인 문서와 이에 대한 지시어를 GPT 모델에 입력한다. 전체 지시어는 해당 문서에 대한 지시문들과 “{‘Q1’: your answer, ‘Q2’: your answer, ‘Q3’: your answer}” 형태로 구성되며, 각 지시문에 대해 답변하도록 되어 있다. 여기서 ‘Q1’은 해당 문서 전체를 한국어로 요약하라는 지시어이고, ‘Q2’와 ‘Q3’은 각각 문서의 목적과 결과를 한국어로 요약하라는 지시어이다. 이는 LLM이 화학 분야 텍스트 내용을 이해할 뿐만 아니라, 영어로 작성된 텍스트를 한국어로 답하는 형태의 지시어 튜닝을 통해 영어와 한국어 모두에서 화학 도메인을 잘 이해하도록 한다. 해당 작업에 대한 데이터셋 예시는 <Table 2>에 정리하였다.

정보 추출 작업 1의 지시어는 요약 및 추출 작업과 마찬가지로 “{‘Q1’: your answer, ‘Q2’: your answer, ‘Q3’: your answer}”

형태를 유지하나, 요약과 관련된 지시어 대신 입력 문서에 관한 질문들로 구성된다. 정보 추출 작업 2의 데이터셋에서는 화학 도메인 텍스트에 관한 추가적인 정보가 작업 1의 지시어에 추가되어, LLM이 보다 풍부한 정보를 바탕으로 질문에 대한 답할 수 있도록 한다. 또한, 지시어가 추출하고자 한 용어를 <Table 4>의 시스템 지시어와 같이 표현하며, 추출하고자 하는 정보가 없을 때에는 없다고 답변을 하도록 한다. 이러한 정보 추출 작업을 통해 다양한 질문에 적절히 답변함으로써, LLM이 화학 도메인에 대해 폭넓게 이해하게끔 한다. 각 정보 추출 작업에 대한 데이터셋 예시는 <Table 3>과 <Table 4>에 정리하였다.

텍스트 분류 작업은 화학 도메인의 한국어 객관식 시험 문제 데이터셋을 사용하였다. 각 화학 시험 문제는 객관식 보기를 포함한 질문으로 구성되며, 모델에는 해당 질문에 대한 답과 해설을 {‘Answer’: your answer, ‘Explanation’: your answer} 형태로 도출하도록 지시어가 입력된다. 정답 뿐만 아니라 그에 대한 해설도 함께 작성하게 함으로써, LLM이 화학 도메인 내용을 보다 잘 이해하도록 하였다. 이는 본 연구에서 수행한 요약 및 번역 작업과 마찬가지로 LLM이 한국어에 대해서도 잘 이해하도록 돕는다. 해당 작업에 대한 데이터셋 예시는 <Table 5>에 정리하였다.

Table 2. Example of a Summarization and Translation Task

Input
Text: “Metal assisted chemical etching (MACE), as an emerging wet chemical etching method for fabricating semiconductor micro / nano-structures (MNS), has attracted wide attentions because of the advantages of ...”
Prompt: “From now on, answer my question absolutely in accordance with the JSON format for the preceding article. ‘Key’ in JSON format is ‘Q1’, ‘Q2’, ‘Q3’ and corresponding ‘VALUE’ for each question record your answer. {‘Q1’: your answer, ‘Q2’: your answer, ‘Q3’: your answer}. Q1) Summarize this document in three sentences in Korean. Q2) Write in Korean within 20 words about the purpose of writing this document. Q3) Write in Korean within 20 words about the result of writing this document.”
Output
Answer: “{‘Q1’: ‘금속 지원화학적 에칭(MACE) 기술을 개선하기 위한 전기화학 나노임프린트 방법을 제안함.’, ‘Q2’: ‘금속 촉매와 금속 이온 오염을 방지하면서 반도체 마이크로/나노 구조물을 제작하는 것이 목적임.’, ‘Q3’: ‘GC 나노임프린트 금형을 사용하여 고성능 반도체 마이크로/나노 구조물을 제작하는 방법을 제안함.’}”

Table 3. Example of an Information Extraction Task 1

Input
Text: “Metal-assisted chemical etching (MACE), as an emerging wet chemical etching method for fabricating semiconductor micro / nanostructures (MNS), has attracted wide attention because of the advantages of ...”
Prompt: “From now on, answer my question absolutely in accordance with the JSON format for the preceding article. ‘Key’ in JSON format is ‘Q1’, ‘Q2’, ‘Q3’ and the corresponding ‘VALUE’ for each question records your answer. {‘Q1’: your answer, ‘Q2’: your answer, ‘Q3’: your answer, ‘Q4’: your answer}. Q1) Tell me what catalyst material is used in this literature, if there is no mention then answer ‘N/A’. Q2) Tell me which support material is used in this literature to improve the activity of the catalyst or to stably immobilize it, if there is no mention then answer ‘N/A’. Q3) If a special processing method has been applied to support material, tell me that method. Q4) Tell me for what purpose the treatment was introduced into Support material.”
Output
Answer: “{‘Q1’: ‘Au, Pt’, ‘Q2’: ‘glassy carbon (GC)’, ‘Q3’: ‘The GC imprint mold was connected to a Pt sheet in the cathodic chamber’, ‘Q4’: ‘To overcome the poor electrocatalytic activity of GC.’}”

Table 4. Example of an Information Extraction Task 2

Input
Text: "A high-performance anion exchange membrane (AEM) is critical for the development of alkaline fuel cell. In this work, AEMs with an interpenetrating polymer network (IPN) are synthesized. ..."
Prompt: "The water uptake is expressed with various expressions such as 'WU', 'water absorption', 'water affinity'. Do not extract the properties of the following stop list exception rate, relatively humidity (RH), acid recovery, power density, desalination rate, coulombic efficiencies, energy efficiencies, mechanical strength. Extract all the water uptake value including its unit and replace the ANSWER2 with them. If the document has no water uptake, write '-' without attaching any other explanation for the value. Extract all the material name matched with the water uptake above and replace the ANSWER1 with it. If the document has no material name, write '-'."
System: "You are my engineering assistant, an expert in anion exchange membrane research. Answer the questions about the above document accurately by strictly matching the dictionary FORMAT below. {'AEM_MAT': {'ANSWER1': {'WU': ANSWER2}}}"
Output
Answer: "{ 'AEM_MAT': { '-': { 'WU': '-' } } }"

Table 5. Example of a Text Classification Task

Input
Question: "물의 점성계수() 0.01 poise를 SI 단위로 표시하면 약 몇 kg/ms가 되는가? 보기: ①0.001 ②0.01 ③0.1 ④1 ⑤10"
Prompt: 문제 질문에 해당하는 답을 질문의 "보기"에서 찾고 정답과 해설을 {'Answer': your answer, 'Explanation': your answer} 형태로 출력하시오.
Output
Answer: "{ 'Answer': ② 0.01, 'Explanation': 1 poise = 0.1 kg/ms }"

3.2 지시어 튜닝

다양한 작업에 대한 지시어 튜닝 데이터셋을 구축한 후, 다량의 일반 언어로 사전 학습된 Llama2에 대해 <Figure 1>의 Step 2와 같이 지시어 튜닝을 수행하였다. 본 연구에서 수행한 지시어 튜닝 작업에는 정보 추출, 요약 및 추출, 요약 및 번역, 키워드 추출, 질의응답, 텍스트 분류 작업이 포함된다. 각 작업의 지시어 튜닝은 독립적으로 이루어지며, 모델은 입력된 텍스트를 바탕으로 지시어에 맞는 답변을 잘 도출하도록 학습된다. 특히, 요약 및 번역 작업에서는 영어로 된 화학 도메인 텍스트를 한국어로 요약함으로써, 화학 도메인에 한국어 이해도를 향상시킬 수 있었다. 텍스트 분류 작업 역시 한국어 데이터셋을 사용하므로 유사한 효과를 기대할 수 있다. 이처럼 다양한 작업 및 언어로 Llama2를 지시어 튜닝을 함으로써, Llama2가 영어와 한국어 모두에서 화학 도메인 텍스트에 대한 전반적인 이해를 증진시키고, 화학 도메인과 관련된 새로운 작업에서도 사용자가 원하는 답변을 도출할 수 있도록 한다.

본 연구에서는 사전 학습된 Llama2를 지시어 튜닝할 때, Lora 기술을 사용하였다. Lora는 LLM을 미세 조정할 때 사전 학습된 파라미터 중 대부분은 업데이트 하지 않고, 저장원으로 축소된 신경망 층의 파라미터만 업데이트 하는 방식이다. 축소된 신경망 층의 파라미터 크기는 사전 학습된 LLM의 파라미터 크기보다 작기 때문에, LLM의 파라미터 전체를 새로 업데이트 하는 방식보다 효율적으로 LLM을 미세 조정할 수

있다. 또한, 미세 조정 시 사전 학습된 파라미터를 업데이트 하지 않음으로써, 사전 학습된 지식의 손실을 방지할 수 있다. 본 연구에서는 화학 도메인 데이터셋을 통한 지시어 튜닝 시, Lora 방법론을 통해 기존의 사전 학습된 일반적인 언어에 대한 지식을 유지하면서 화학 도메인에 대한 지식을 효율적으로 추가할 수 있도록 한다. 본 연구에서 사용한 Lora 방법론은 식 (1)과 같이 적용한다. 추가적으로, Llama2에 대한 지시어 튜닝 및 Lora 방법론 적용 시, 별도의 dropout은 적용하지 않으며 출력 토큰을 autoregressive한 방식으로 도출할 때, 최대 logit 값을 가지는 토큰으로 도출하게 된다.

4. 실험

4.1 데이터셋

(1) 지시어 튜닝 데이터셋

본 연구에서는 앞서 언급했듯이, GPT-3.5와 GPT-4를 활용하여 화학 도메인 데이터셋을 다양한 지시어 튜닝 작업들에 맞게 증강 및 구축하였다. 크게 세 가지 유형의 데이터셋을 사용했으며, 본 논문에서는 각각을 LG Chem, Public, 그리고 Korean 데이터셋으로 지칭한다. LG Chem 데이터셋은 3.1장에서 설명한 요약 및 추출, 요약 및 번역, 그리고 정보 추출 작업에 대해 LLM APIs로 생성한 데이터셋으로 구성되어 있다.

Public 데이터셋은 오픈 소스로 공개된 화학 도메인 텍스트 데이터이며, 키워드 추출, 질의응답 작업에 대한 데이터셋으로 구성되어 있다. 마지막으로, Korean 데이터셋은 3.1장에서 설명한 텍스트 분류 작업에 대한 데이터셋이다. LG Chem과 Public 데이터셋은 영어로, Korean 데이터셋은 한국어로 이루어져 있으며, 각 데이터셋에 대한 요약은 <Table 6>에 정리하였다.

Table 6. Description of Instruction Tuning Datasets

Dataset	Task	The number of dataset samples
LG Chem	Summarization and Extraction	30,776
	Summarization and Translation	1,345
	Information Extraction 1	4,035
	Information Extraction 2	3,510
Public	Chemrxiv: Keyword Extraction	6,269
	Adnerson_chemistry: Question Answering	150
	HydraLM_chemistry: Question Answering with long answers	255
Korean	Text Classification	12,280

1) LG Chem 데이터셋

LG Chem 데이터셋은 LG 화학 내부의 친환경 그린 수소 관련 연구 분야 데이터셋을 활용하여 구축되었다. 이 데이터셋은 proton exchange membrane(PEM) 촉매 연구 분야와 관련하여 개체명 인식, 문서 분류, 문서 요약, 국문 번역 데이터셋 및 anion exchange membrane(AEM) 분리막 연구 분야와 관련된 물성 추출 데이터셋으로 구성되어 있다. AEM 분리막 연구 분야 데이터셋은 ion conductivity, ion exchange capacity(IEC), 그리고 water uptake(WU) 등의 물성들과 관련되어 있다. 이러한 데이터셋은 LLM APIs를 통해 다양한 지시어 튜닝 형태에 맞게 증강 및 구축되었으며, 정보 추출, 요약 및 추출, 그리고 요약 및 번역 작업의 지시어 튜닝에 사용된다.

2) Public 데이터셋

Public 데이터셋은 Huggingface에 공개된 세 가지 데이터셋을 사용하였다. Chemrxiv(https://huggingface.co/datasets/marianna13/chemrxiv_reprocessed) 데이터셋은 Chemrxiv 플랫폼에서 수집된 논문 데이터를 전처리한 것으로, 키워드 추출 작업에 사용되었다. Adnerson_chemistry(<https://huggingface.co/datasets/adnersonbcdefg/chemistry>) 데이터셋은 화학적 반응 및 분자 구조 예측을 위한 텍스트 데이터를 포함하고 있으며, HydraLM_chemistry(https://huggingface.co/datasets/HydraLM/chemistry_dataset_alpaca) 데이터셋은 화학 분야의 질의응답 작업을 위한

instruction following 데이터셋이다. 이 두 가지 데이터셋은 질의응답 작업을 위해 사용하였다. 특히, HydraLM_chemistry 데이터셋은 Adnerson_chemistry 데이터셋보다 긴 텍스트로 이루어져 있어 난이도가 높은 데이터셋으로 평가된다. 본 연구에서는 이 세 가지 Public 데이터셋을 LLM APIs를 활용한 추가적인 작업 없이 그대로 사용하였다.

3) Korean 데이터셋

Korean 데이터셋은 국내 소재 관련 과학 기술 분야의 자격증 30가지 시험 기출 문제를 사용하였다. 본 연구에서는, 이들 시험 중 일부를 선별하여 사용하였으며, 데이터셋은 난이도에 따라 상(professional, 해당 분야 최고 수준), 중(college, 대학생 전공 심화 수준), 하(high school, 일반적 과학 기술 지식 수준)로 분류하였다. 이후, LLM APIs를 이용하여 선별된 데이터를 “(문제 및 답변 보기, 정답 보기, 해설)” 쌍으로 구성된 데이터셋을 구축하였다.

(2) 평가 데이터셋

평가 데이터셋으로는 오픈 데이터인 MMLU 데이터셋(Hendrycks *et al.*, 2020)과 지시어 튜닝에 사용한 Korean 데이터셋 중 일부를 해설 없이 문제와 답으로 구성하여 사용하였다. MMLU 데이터셋은 화학 도메인과 관련된 고등학교 및 대학교의 객관식 오지선다 시험 문제 데이터이며 영어로 이루어져 있다. Korean 데이터셋은 지시어 튜닝에서 사용하지 않은 Korean 데이터셋을 해설 텍스트 없이 사용하였다. 두 가지 데이터셋 모두 텍스트 분류 문제에 대한 데이터셋이다.

4.2 평가 지표

본 연구에서는 LLM의 오지선다 분류 작업을 평가하기 위한 평가 지표로 F1 score를 사용하였다. F1 score는 정밀도와 재현율의 조화 평균을 나타내어 두 지표를 모두 반영할 수 있다. 정밀도, 재현율, 그리고 F1 score에 대한 수식은 식 (2) ~ 식 (4)으로 표현된다. 이 때, TP(true positive)는 실제로 positive인 데이터를 모델이 올바르게 분류한 개수, TN(true negative)는 실제로 negative인 데이터를 모델이 올바르게 분류한 개수, FP(false positive)는 실제로 negative인 데이터를 모델이 positive로 오분류한 개수, 마지막으로 FN(false negative)는 실제로 positive인 데이터를 모델이 negative로 오분류한 개수를 의미한다.

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

$$F1score = 2 \times \frac{Recall \times Precision}{Recall + Precision} \quad (4)$$

4.3 실험 결과

본 연구에서는 다양한 지시어 튜닝의 효과를 검증하기 위한 LLM 모델로 Llama2 모델을 사용하였으며, 모델 파라미터 크기에 따라 Llama2 - 7b와 Llama2 -13b으로 구분하여 실험하였다. 또한 각 평가 데이터셋의 소분류 데이터셋에 대한 평균 성능을 “Average”로 표기하였다. 평가 데이터셋 별로 모델의 파라미터 크기에 따른 비교 실험은 <Table 7>과 <Table 8>에 정리하였다. 우선, 두 가지 모델 모두 추가적인 지시어 튜닝이 수행되지 않은 상태에서 Llama2를 사용하여 성능을 비교하였다. <Table 7>과 <Table 8>에서 확인할 수 있듯이, 두 가지 데이터셋의 모든 소분류 데이터셋에서 Llama2 - 13B 모델이 보다 우수한 성능을 보였다. 일반적으로, 모델의 파라미터 크기가 클수록 더 많은 정보를 포함하고 있어 성능이 향상된다. 본 실험을 통해 화학 도메인에서도 이러한 현상이 확인되었다. 이 실험 결과를 바탕으로, 이어지는 실험들은 모두 Llama2 - 13B 모델을 사용하여 진행하였다.

Table 7. Evaluation results of F1 score with “no tuning” model for each sub-category dataset in MMLU dataset. Bold indicates the best performance between results among model sizes

Model	High School	College	Average
Llama2 - 7B	0.3596	0.2700	0.3148
Llama2 - 13B	0.4286	0.3900	0.4093

Table 8. Evaluation results of F1 score with “no tuning” model for each sub-category dataset in Korean dataset. Bold indicates the best performance among model sizes

Model	High School	College	Professional	Average
Llama2- 7B	0.2732	0.2672	0.2665	0.2741
Llama2- 13B	0.3319	0.3161	0.3110	0.3277

다음으로, 본 연구에서는 지시어 튜닝에 따른 성능 비교 실험으로 진행하였으며, 평가 데이터셋 별 실험 결과는 <Table 9>와 <Table 10>에 정리하였다. 기존 사전 학습된 Llama2 모델에 일반적인 언어들을 사용하여 대화 형식으로 지시어 튜닝하여 공개된 Llama2-chat 모델을 베이스 라인으로 사용하였으며 화학 도메인에 대해서는 추가적인 지시어 튜닝을 수행하지 않았기 때문에 “no tuning”이라고 표기하였다. 영어로 이루어진 LG Chem과 Public 데이터셋을 활용하여 지시어 튜닝을 수행한 모델은 “English”로, Korean 데이터셋을 활용한 모델은 “Korean”, 전체 데이터셋을 활용한 모델은 “Total”로 표기하였다. 지시어 튜닝 데이터셋으로 Korean 데이터셋을 사용한 경우, 모든 경우에서 성능이 향상되었으며, English 데이터셋을 사용한 경우에도 대부분의 경우에서 성능이 향상되었다. English 데이터셋으로 지시어 튜닝을 할 때, LG Chem 데이터

셋의 경우 LLM APIs를 통해 증강된 데이터셋이 함께 활용되기 때문에 약간의 노이즈가 포함되어 전체 데이터셋을 활용할 때보다 성능이 저하될 수 있지만, 베이스라인 모델보다는 대부분 성능 향상이 이루어지기 때문에 증강된 데이터셋에 노이즈만 존재하지는 않는다는 것을 확인할 수 있었다. 이를 통해 한국어와 영어 두 가지 언어에서 모두 적용 가능하도록 하는 화학 도메인 특화 지시어 튜닝의 효과를 확인할 수 있었다. 또한, 평가 데이터셋으로 MMLU 데이터셋을 사용할 때, Korean 데이터셋으로 지시어 튜닝을 한 경우가 더 우수한 성능을 도출했음을 보았을 때, 지시어 튜닝 데이터셋에서 언어적인 측면보다는 데이터의 품질이 더 중요하다는 것을 확인할 수 있었다.

Table 9. Evaluation results of F1 score for each sub-category dataset in MMLU dataset. Bold indicates the best performance among results from the instruction tuning dataset

Instruction tuning dataset	High School	College	Average
No tuning	0.4286	0.3900	0.4093
English	0.4286	0.4000	0.4143
Korean	0.4433	0.4200	0.4317
Total	0.4286	0.4200	0.4243

Table 10. Evaluation results of F1 score for each sub-category dataset in Korean dataset. Bold indicates the best performance among results from the instruction tuning dataset

Instruction tuning dataset	High School	College	Professional	Average
No tuning	0.3319	0.3161	0.3110	0.3277
English	0.3480	0.3436	0.3134	0.3435
Korean	0.3587	0.3680	0.3622	0.3645
Total	0.3648	0.3658	0.3436	0.3655

<Table 9>와 <Table 10>에서 사용한 베이스라인 모델 (Llama2-chat)은 기존 사전 학습된 모델에 일반적인 도메인의 텍스트를 사용하여 대화 형식으로 지시어 튜닝한 모델이다. 기존 사전 학습된 모델과 비교했을 때, Llama2-chat 모델에 추가적인 도메인 특화 지시어 튜닝을 한 효과를 확인하기 위한 실험 결과는 <Table 11>과 <Table 12>에 정리하였고, 도메인 특화 지시어 튜닝에 사용한 데이터셋 별로 두 모델 간의 성능을 비교하였다. 기존 사전 학습된 Llama2는 “Original”로, 사전에 대화 형식으로 지시어 튜닝한 Llama2-chat은 “Chat”으로 표기하였다. 두 가지 평가 데이터셋의 결과에서 대부분, 기존 Llama2 모델의 성능이 Llama2-chat 모델의 성능보다 우수함을

확인할 수 있었다. 이는 Llama2-chat이 일반적인 도메인의 텍스트로 지시어 튜닝이 이미 수행되었기 때문에, 이러한 정보가 특정 도메인의 텍스트로 다시 지시어 튜닝을 수행할 때 효과적이지 않다는 것을 본 실험을 통해 알 수 있었다. 따라서, 도메인 특화 LLM을 구축할 때에는 일반적인 언어로 지시어 튜닝이 수행되지 않은 기존 LLM에 특정 도메인의 텍스트로만 지시어 튜닝을 하는 것이 효과적임을 확인할 수 있었다.

Table 11. Evaluation results comparing the F1 score of the original Llama2 and Chat-Llama2 for each sub-category dataset in MMLU dataset. Bold indicates the best performance between the two types of Llama2 models

Instruction tuning dataset	High School	College	Average
Original	0.4680	0.4200	0.4440
Chat	0.4286	0.3900	0.4093
Original + English	0.4286	0.4000	0.4143
Chat + English	0.4138	0.3800	0.3969
Original + Korean	0.4433	0.4200	0.4317
Chat + Korean	0.4532	0.3700	0.4116
Original + Total	0.4286	0.4200	0.4243
Chat + Total	0.4187	0.3900	0.4044

Table 12. Evaluation results comparing the F1 score of the original Llama2 and Chat-Llama2 in for each sub-category dataset Korean dataset. Bold indicates the best performance between the two types of Llama2 models

Instruction tuning dataset	High School	College	Professional	Average
Original	0.3453	0.3466	0.3288	0.3456
Chat	0.3319	0.3161	0.3110	0.3227
Original + English	0.3480	0.3436	0.3134	0.3435
Chat + English	0.3375	0.3225	0.3204	0.3327
Original + Korean	0.3587	0.3680	0.3622	0.3645
Chat + Korean	0.3610	0.3681	0.3524	0.3655
Original + Total	0.3648	0.3658	0.3436	0.3655
Chat + Total	0.3616	0.3715	0.3540	0.3678

5. 결 론

본 연구는 화학 도메인에서 LLM의 성능을 향상시키기 위해 LLM APIs를 통해 직접 생성한 화학 데이터셋을 기반으로 지시어 튜닝을 수행하는 방법론을 제안하였다. 기존의 사전 학습된 LLM은 화학과 같은 특수 도메인에서 데이터 부족과 도메인 특수성으로 인해 성능이 저하되는 문제가 존재했다. 이를 해결하기 위해 본 연구에서는 LLM APIs와 기존 화학 분야

의 다양한 텍스트 데이터셋을 활용하여 지시어 튜닝 데이터셋을 직접 생성하고, 이를 바탕으로 정보 추출, 요약, 번역, 키워드 추출, 질문 응답 등 다양한 자연어 처리 작업에 지시어 튜닝을 적용하였다. 이 과정을 통해 화학 도메인에 특화된 학습을 수행함으로써 모델 성능을 효과적으로 향상시켰다. 실험 결과, 영어와 한국어 모두에서 성능이 향상되었으며, 화학 분야에 적합한 거대 언어 모델의 지시어 튜닝이 효과적임을 입증하였다.

본 연구는 화학 도메인에서 LLM을 활용해 생성된 데이터셋을 기반으로 지시어 튜닝을 수행함으로써 화학 분야에서의 LLM 응용 가능성을 크게 확장 시켰다. 예를 들어, 제약 산업에서는 자동화된 논문 분석, 새로운 화합물 및 약물 후보군의 평가를 빠르고 효율적으로 처리할 수 있으며, 교육 분야에서도 화학 개념을 보다 쉽게 전달하거나 학생들의 질의응답을 자동화할 수 있는 가능성을 열었다. 또한, 본 연구에서 제시한 LLM 기반 데이터 생성 및 지시어 튜닝 방법은 다른 특수 도메인에도 적용될 수 있는 가능성을 보여준다. 법률 문서 분석, 의료 데이터 요약, 금융 데이터 처리 등 다양한 분야에서 복잡한 작업을 자동화하거나 보조하는 데 활용될 수 있으며, 각 도메인에 특화된 LLM을 개발하여 전문 지식이 필요한 작업에서도 효율적인 성능을 기대할 수 있다.

그러나 본 연구에는 몇 가지 한계점이 있다. 첫째, 실험이 화학 도메인에 국한되어 있어, 다른 특수 분야에서도 해당 방법이 효과적인지 검증이 필요하다. 둘째, 구축한 모델을 추론할 때, 텍스트 분류 작업 이외의 다양한 작업에서도 성능 검증이 필요하다. 추후에는 질의응답, 개체명 인식 등의 다양한 작업에서의 성능 검증을 통해 보다 신뢰할 수 있는 화학 도메인 특화 LLM을 설명할 수 있을 것이다.

결론적으로, 본 연구는 LLM을 활용해 직접 생성한 데이터셋을 통해 화학 도메인에서 지시어 튜닝을 성공적으로 적용하였으며, 이를 통해 특수한 도메인에서도 LLM의 성능을 효과적으로 향상시킬 수 있음을 입증하였다. 향후 이 방법을 다른 분야로 확장하여 전문 작업의 효율성을 더욱 높일 수 있을 것으로 기대된다.

참고문헌

- Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... and McGrew, B. (2023), Gpt-4 technical report, arXiv preprint arXiv:2303.08774.
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... and Amodei, D. (2020), Language models are few-shot learners, *Advances in Neural Information Processing Systems*, **33**, 1877-1901.
- Chen, Y., Polajnar, T., Batchelor, C., and Teufel, S. (2020, November), A corpus of very short scientific summaries. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, 153-164.

- Chiang, W. L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., ... and Xing, E. P. (2023), Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, See <https://vicuna.lmsys.org> (accessed 14 April 2023), 2(3), 6.
- Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. (2018), Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805.
- Fedus, W., Zoph, B., and Shazeer, N. (2022), Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity, *The Journal of Machine Learning Research*, 23(1), 5232-5270.
- Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. (2020), Measuring massive multitask language understanding, arXiv preprint arXiv:2009.03300.
- Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., ... and Sifre, L. (2022), Training compute-optimal large language models, arXiv preprint arXiv:2203.15556.
- Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... and Chen, W. (2021), Lora: Low-rank adaptation of large language models, arXiv preprint arXiv:2106.09685.
- Jablonka, K. M., Schwaller, P., Ortega-Guerrero, A., and Smit, B. (2024), Leveraging large language models for predictive chemistry, *Nature Machine Intelligence*, 6(2), 161-169.
- Le Scao, T., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., ... and Al-Shaibani, M. S. (2023), Bloom: A 176b-parameter open-access multilingual language model.
- Li, Y., Ma, S., Wang, X., Huang, S., Jiang, C., Zheng, H. T., ... and Jiang, Y. (2024, March), Ecomgpt: Instruction-tuning large language models with chain-of-task tasks for e-commerce, In *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17), 18582-18590.
- Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., and Neubig, G. (2023), Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing, *ACM Computing Surveys*, 55(9), 1-35.
- Liu, Z., Li, Y., Shu, P., Zhong, A., Yang, L., Ju, C., ... and Li, X. (2023), Radiology-llama2: Best-in-class large language model for radiology, arXiv preprint arXiv:2309.06419.
- Liu, Z., Zhong, A., Li, Y., Yang, L., Ju, C., Wu, Z., ... and Liu, T. (2023), Radiology-GPT: A Large Language Model for Radiology, arXiv preprint arXiv:2306.08666.
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... and Lowe, R. (2022), Training language models to follow instructions with human feedback, *Advances in Neural Information Processing Systems*, 35, 27730-27744.
- Peng, B., Li, C., He, P., Galley, M., and Gao, J. (2023), Instruction tuning with gpt-4, arXiv preprint arXiv:2304.03277.
- Qin, H., Ma, X., Zheng, X., Li, X., Zhang, Y., Liu, S., ... and Magno, M. (2024), Accurate lora-finetuning quantization of llms via information retention, arXiv preprint arXiv:2402.05445.
- Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018), Improving language understanding by generative pre-training.
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019), Language models are unsupervised multitask learners, *OpenAI blog*, 1(8), 9.
- Shahriar, S., Lund, B., Mannuru, N. R., Arshad, M. A., Hayawi, K., Bevara, R. V. K., ... and Batool, L. (2024), Putting GPT-4o to the Sword: A Comprehensive Evaluation of Language, Vision, Speech, and Multimodal Proficiency, arXiv preprint arXiv:2407.09519.
- Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., ... and Hashimoto, T. B. (2023), Alpaca: A strong, replicable instruction-following model, *Stanford Center for Research on Foundation Models*, 3(6), 7, <https://crfm.stanford.edu/2023/03/13/alpaca.html>.
- Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., ... and Lample, G. (2023), Llama: Open and efficient foundation language models, arXiv preprint arXiv:2302.13971.
- Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... and Scialom, T. (2023), Llama 2: Open foundation and fine-tuned chat models, arXiv preprint arXiv:2307.09288.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... and Polosukhin, I. (2017), Attention is all you need, *Advances in Neural Information Processing Systems*, 30.
- Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N. A., Khashabi, D., and Hajishirzi, H. (2022), Self-instruct: Aligning language model with self generated instructions, arXiv preprint arXiv:2212.10560.
- Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., ... and Le, Q. V. (2021), Finetuned language models are zero-shot learners, arXiv preprint arXiv:2109.01652.
- Yang, H., Liu, X. Y., and Wang, C. D. (2023), FinGPT: Open-Source Financial Large Language Models, arXiv preprint arXiv:2306.06031.
- Yang, X., Wilson, S. D., and Petzold, L. (2024), Quokka: An Open-source Large Language Model ChatBot for Material Science, arXiv preprint arXiv:2401.01089.
- Ye, J., Chen, X., Xu, N., Zu, C., Shao, Z., Liu, S., ... and Huang, X. (2023), A comprehensive capability analysis of gpt-3 and gpt-3.5 series models, arXiv preprint arXiv:2303.10420.
- Yunxiang, L., Zihan, L., Kai, Z., Ruilong, D., and You, Z. (2023), Chatdoctor: A medical chat model fine-tuned on llama model using medical domain knowledge, arXiv preprint arXiv:2303.14070, 2(5), 6.
- Zhou, J., Ke, P., Qiu, X., Huang, M., and Zhang, J. (2023), ChatGPT: potential, prospects, and limitations, *Frontiers of Information Technology & Electronic Engineering*, 1-6.

저자소개

이정민 : 한양대학교 산업공학과에서 2022년 학사 학위를 취득하고, 고려대학교 산업경영공학과에서 석박통합과정에 재학 중이다. 연구 분야는 Applications of Uncertainty Quantification, Large Language Models이다.

김현지 : 고려대학교 통계학과에서 2022년 학사학위를 취득하고, 고려대학교 산업경영공학과에서 2024년 석사학위를 취득하였다. 현재는 삼성전자 SAIT System research center에서 연구원으로 재직 중이다. 연구 분야는 시스템 전력 및 성능 최적화이다.

이성수 : LG화학 재직 중이다.

이규황 : LG화학 재직 중이다.

김연수 : LG화학 재직 중이다.

김성범 : 고려대학교 산업경영공학부 교수로 2009년부터 재직하고 있으며, 인공지능공학연구소 소장, 기업산학협력센터 센터장, 한국데이터마이닝학회 회장을 역임했다. 미국 University of Texas at Arlington 산업공학과에서 교수를 역임하였으며, 한양대

학교 산업공학과에서 학사학위를 미국 Georgia Institute of Technology에서 산업시스템공학 석사 및 박사학위를 취득하였다. 인공지능, 머신러닝, 최적화 방법론을 개발하고 이를 다양한 공학, 자연과학, 사회과학 분야에 응용하는 연구를 수행하고 있다.