Vol. 57, Issue 5, 539–550, 2024

Economic and Environmental Geology

THE KOREAN SOCIETY OF ECONOMIC AND ENVIRONMENTAL GEOLOGY

pISSN : 1225-7281
eISSN : 2288-7962

Special Research Paper on "Data Science and Artificial Intelligence in Economic and Environmental Geology"

Development of a Large-scale Korean Language Model in the Field of Geosciences

Sang-ho Lee*

Mineral Resources Division, Korea Institute of Geosciences and Mineral Resources, Daejeon 34132, Republic of Korea

*Corresponding author : energy@kigam.re.kr

ARTICLE INFORMATION

Manuscript received 30 August 2024
Received in revised form 8 October 2024
Manuscript accepted 10 October 2024
Available online 29 October 2024

DOI : http://dx.doi.org/10.9719/EEG.2024.57.5.539

Research Highlights

- Development of a Korean large language model specialized in geoscience
- The validity of the overall training approach was verified through multiple evaluations
- The result showed significant improvements in scientific and Korean-related abilities

ABSTRACT

With the rapid development and commercialization of large-scale generative language models, concerns regarding the appropriateness of model outputs, expertise, and data security have been emerged. In particular, Korean generative language models specialized in the field of geoscience have not yet been studied due to difficulties in data processing, preprocessing and a lack of development cases. This study conducted the entire process for developing a Korean language model specialized in the field of geoscience and evaluated its applicability in related fields. To achieve this, academic data related to geoscience were collected and preprocessed to create a dataset suitable for the training of the language model. The dataset was applied to the Llama2 model for the training. The trained model was quantitatively evaluated using 19 different evaluation datasets from various fields. The results demonstrated improved functionalities related to scientific question-answering and Korean text interpretation compared to the original model. The language model developed through this study can potentially enhance research productivity in the field of geoscience, offering benefits such as idea generation. The outcomes of this study are expected to stimulate further research and the utilization of generative language models in geoscience in the future.

Keywords : large language model, generative model, natural language processing, artificial intelligence, geoscience

Citation: Lee, S.-h. (2024) Development of a Large-scale Korean Language Model in the Field of Geosciences. Korea Economic and Environmental Geology, v.57, p.539-550, doi:10.9719/EEG.2024.57.5.539.

Journal homepage: http://www.kseeg.org/main.html

자원환경지질, 제57권, 제5호, 539-550, 2024
Econ. Environ. Geol., 57(5), 539-550, 2024
http://dx.doi.org/10.9719/EEG.2024.57.5.539

pISSN 1225-7281
eISSN 2288-7962

연구논문 ("자원환경지질 분야의 데이터과학 및 인공지능 활용" 특별호)

지질과학 분야 한국어 대규모 언어 모델 개발

이상호*
한국지질자원연구원 광물자원연구본부 선임연구원

*책임저자 : energy@kigam.re.kr

요 약

최근 대규모 생성형 언어 모델의 급격한 발달과 상용화가 이루어지면서 모델 출력의 적정성, 전문성 문제 및 데이터 보안 문제가 제기되고 있다. 특히 지질과학 유관 분야에서는 가공된 자료 및 전처리의 어려움과 개발 사례의 부족으로 인해 해당 분야에 특화된 한국어 언어 모델 개발은 아직 진행된 사례가 없다. 이에 따라 본 연구에서는 지질과학 분야에 특화된 한국어 언어 모델 개발을 위한 전반적인 과정을 수행하고 이를 평가함으로써 유관 분야에서의 적용 가능성을 알아보고자 하였다. 이를 위하여 지질과학 유관 분야의 학술 자료를 수집하고 전처리하여 언어 모델의 학습에 적합한 자료를 준비하고, 이를 Llama 2 모델에 적용하여 사전학습 및 미세조정을 수행하였다. 학습된 모델은 19종의 분야별 평가용 데이터셋을 이용하여 정량적으로 평가하였으며, 그 결과 원본 모델 대비 과학 관련 질의응답 및 및 한국어 지문 해석 관련 기능이 향상된 것으로 나타났다. 본 연구를 통해 개발된 언어 모델은 유관 분야에서 아이디어 창출과 같은 연구 생산성 제고에 기여할 수 있으며, 향후 언어 모델을 활용한 연구 및 활용을 활성화할 수 있을 것으로 기대된다.

주요어 : 대규모 언어 모델, 생성형 모델, 자연어 처리, 인공지능, 지질과학

1. 서 론

최근 인공지능 관련 기술의 급격한 발전으로 생성형 언어 모델의 성능이 인간의 능력에 비견될 정도로 높이 평가되고 있다(Brown et al., 2020). 이에 OpenAI를 비롯한 많은 기업에서 언어 모델을 상용화하고 있으며, 이러한 모델들의 학술 분야에 대한 적용도 활발히 이루어지고 있다. 언어 모델의 전반적 성능은 모델을 구성하는 인자(parameter)의 수에 비례하는 것으로 잘 알려져 있으며, 상용 모델의 경우 1천억 개 이상의 많은 인자를 갖는다(Zhao et al., 2023). 모델의 학습 및 운용을 위해서는 그 크기에 비례하는 수준의 하드웨어가 필요하므로 이러한 초대규모 모델의 개발은 단일 조직에 적합하지 않으며, 일반적으로 각 분야에서의 특화 모델은 이보다 작은 수준의 모델을 활용한다.

상용 모델의 높은 성능과 범용성에도 불구하고 각 분야별, 용도별 소규모 모델의 개발이 지속적으로 이루어지는 이유로는 언어 모델 활용에 있어서의 보안 관련 문제를 들 수 있다. ChatGPT를 비롯한 상용 모델은 인터넷을 통한 활용이 필수적이므로, 네트워크를 통한 자료의 유출이나 입력된 자료가 학습 자료로 활용되어 유출의 가능성이 발생하는 등의 문제가 발생할 수 있다. 또한 최근 밝혀진 언어 모델의 기술적 취약점(Nasr et al., 2023)은 모델의 구조적인 문제로 인하여 발생하는 만큼, 보안적 측면이 우선적으로 요구되는 환경에서는 범용적 성능이 비교적 낮더라도 특정 분야나 목적에 특화하여 폐쇄된 환경에서 구동하는 방법이 우선적으로 고려될 수 있다.

특정한 연구 분야에 특화된 언어 모델의 개발을 위해서는 해당 분야에서 활용되는 용어와 지식에 대한 집중적인 학습이 요구되며, 학습의 성공적 진행을 위해서는 데이터의 수집, 가공과 다양한 시험이 필요하다. 그러나 지질과학 분야에서는 영어 및 중국어 자료를 기반으로 한 소수의 연구 사례들만이 존재하며(Lawley et al., 2022, Deng et al., 2024), 특히 한국어로 구성된 학습용 학술 자료를 이용한 연구 사례는 전무하므로 이러한 모델의 기능성이 파악되지 않은 상태이다.

이에 본 연구에서는 지질과학 분야에 특화된 학술 언어 모델을 개발하기 위하여 유관 분야의 학술 논문을 수집, 가공하여 언어 모델에 적용 가능한 데이터셋으로 구성하고, 이를 공개 모델의 학습에 적용하여 사전학습 및 미세조정을 포함한 전체 학습 과정을 수행하였다. 학습 대상 자료는 지질과학 분야와 직, 간접적으로 연관된 넓은 범위를 대상으로 수집 후 정제, 가공함으로써 구성하였다. 이러한 자료를 통해 학습된 언어 모델은 다수의 분야별 시험용 데이터셋을 활용하여 학술적 출력 성능 향

상에 대한 정량적인 평가를 수행하였다.

2. 연구 방법

2.1. 학습 자료 수집 및 전처리

언어 모델의 학습을 위한 데이터는 최대한의 지질과학 분야 학술 자료를 확보하면서도 내용의 지나친 편중을 방지하기 위하여 유관 분야 논문을 대상으로 하되, 다수의 출판사별 학술지 데이터를 학술지별로 일괄 수집 후 관련이 적은 분야의 학술지를 제외하는 방법을 사용하였다. 이를 위하여 XML(eXtensible Markup Language) 형식의 논문 자료를 확보하고, 이로부터 사람이 읽을 수 있는 일반 문자열 형태의 자료를 추출하여 활용하였다. 대부분의 출판사에서는 학술 논문을 마크업 언어 형식으로 제공하며, 이러한 포맷은 일반적인 구독을 위해 활용되는 PDF(Portable Document Format) 형식과 달리 논문을 구성하는 제목, 내용, 수식, 표 및 그림 등이 요소별로 구조화되어 있다. 또한 가독성을 위한 편집이나 줄넘김과 같은 가공이 이루어지지 않은 상태이므로 정확한 정보의 추출이 가능하다.

논문의 수집은 출판사별로 제공하는 API(Application Programming Interface) 또는 별도의 마크업 문서 제공 페이지를 통해 수집하였으며, 일괄 수집이 허용된 저널을 대상으로만 수행하였다. 이러한 방법으로 국내 출판사 4개(233개 학술지, 129,177개 논문), 국외 출판사 6개(1,347개 학술지, 170,969개 논문)에 대한 자료 수집을 진행하였으며, 이를 다시 선별하여 지질과학 분야와의 연관성을 고려하여 최종적으로 464종의 학술지에 대한 171,815건의 논문을 학습 대상으로 선정하였다.

선별된 자료는 언어 모델로의 적용을 고려하여 지나치게 짧은 문장, 참고문헌 인용, 중복된 문구 등을 일반적인 문자열 처리 기법과 정규표현식을 활용하여 제거하였다. 참고문헌 인용부의 경우 해당 부분을 제거하지 않은 상태로 학습 시험을 수행한 결과 인용 양식만을 따르는 거짓 데이터를 출력하는 환각 현상이 빈번하게 발생하였으며, 언어 모델에서의 실제 참고문헌 인용은 실제 데이터를 검색 후 질문에 포함하여 답변을 수행하는 RAG (retrieval-augmented generation)과 같은 기능을 활용함으로써 정확한 제공이 가능하므로 최종 학습자료에서는 제외하였다.

2.2. 모델 학습

일반적으로 생성형 언어 모델의 학습 과정은 크게 (1) 문장의 구성과 단어의 의미를 학습하기 위한 사전학습(pretraining) 단계와, (2)언어 학습이 완료된 모델에 대화형 입출력 기능을 부여하고 품질을 향상시키는 미세조정(fine-tuning) 단계로 나뉜다. 이 과정들은 순차적으로 진행하여야 하므로 앞서 수집된 학습 자료를 적용하기 위한 대상 모델은 미세조정이 수행되지 않은 것을 활용할 필요가 있다.

생성형 언어 모델에는 다양한 종류가 존재하며, 각 모델은 개발 목적에 따라 구조, 학습된 언어의 종류, 학습량 및 학습 정도가 모두 다르다(Zhao et al., 2023). 예를 들어 프로그래밍에 특화된 모델은 다양한 언어로 구성된 프로그래밍 코드를 학습함으로써 사용자가 요청한 코드를 주석과 함께 작성하는 데에 기능성이 집중되어 있다. 이러한 특화 모델은 특정 분야나 작업에 대한 학습을 통해 해당 분야에 한정된 높은 성능을 발휘할 수 있으며, 이는 분야를 특정하지 않는 고성능의 초거대 모델 개발보다 비용 면에서 매우 효율적이다.

본 연구의 목표는 지질과학 분야에 특화된 언어 모델을 개발하는 것이므로 유관 분야의 자료를 중점적으로 학습할 필요가 있다. 따라서 본 연구에서는 사전학습 단계에 해당 자료를 적용하여 분야에 특화된 성능을 얻고자 하였으며, 이에 따라 Meta에서 공개한 Llama 2 모델 중 사전학습까지만 완료된 모델을 연구의 기반으로 선정하였다. Llama 2는 총 2조 개의 토큰으로 구성된 학습 데이터를 170만 시간(단일 A100 GPU로 환산 시) 동안 학습한 모델로서 성능 측면에서는 70B 규모의 모델이 ChatGPT 3와 유사한 성능을 보이며, 이보다 작은 모델들도 기존의 동급 모델보다 우수한 성능을 나타내는 것으로 보고되었다(Touvron et al., 2023).

모델의 학습 시에는 단순 구동과 달리 그라디언트 계산 등의 추가적인 단계들이 포함되므로 연산용 메모리가 구동 시의 4배 이상 요구된다. 또한 각 학습 과정에서 수행되는 입력 자료의 종류, 구조 및 구체적인 학습 방법은 단계별로 다소 상이하나, 학습 과정에서 대규모 연산이 필수적으로 동반되는 점은 동일하다. 이러한 연산의 대부분은 단순 산술연산으로서, 병렬 산술연산에 특화된 GPU(graphic processing unit)가 범용연산장치인 CPU(central processing unit)에 비해 최대 100배 빠르게 수행할 수 있으므로 통상적으로 GPU가 활용된다(Wang et al., 2019). 본 연구에서는 LLM 개발에 표준적으로 활용되는 NVIDIA A100 80GB GPU 4개를 단일 노드에서 사용하였으며, 해당 장비의 규격에 적합한 13B 규모의 모델을 연구에 활용하였다. 또한 Llama 2는 다국어를 지원하는 모델이나, 출시 초기에는 한국어 관련 성능이 영어보다 현저히 낮아 국내에서는 널리 활용되지 못하였다. 이후 한국어를 추가 학습한 모델이 공개되면서 관련 개발 및 공유가 활발히 이루어졌으며, 이에 본 연구에서는 해당 모델에 약

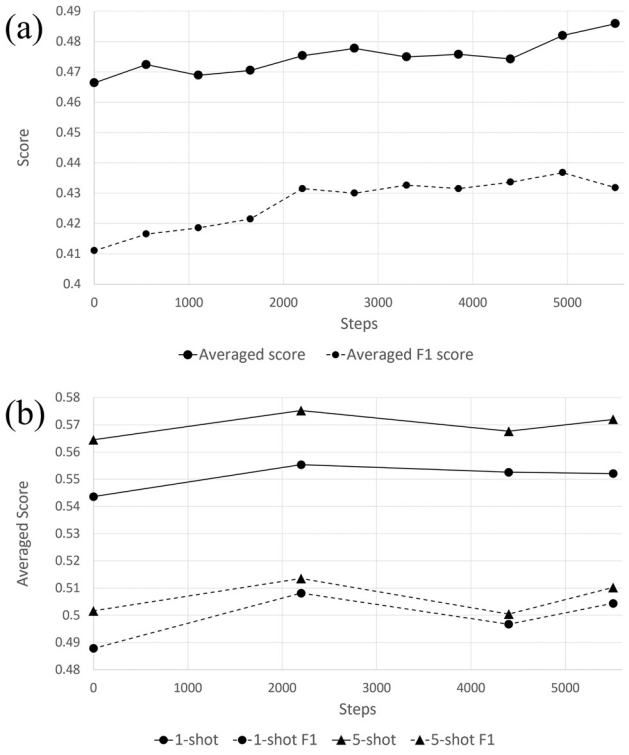

Fig. 1. Evaluation results for each learning stage of the pre-training model using 9 evaluation datasets. (a) 0-shot; (b) 5-shot.

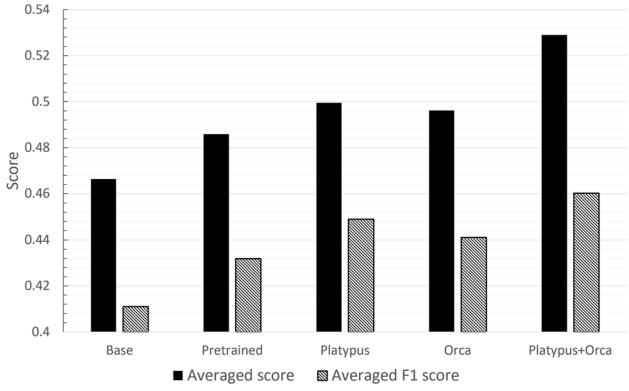

Fig. 2. Comparison of 0-shot scores using 9 evaluation datasets for models trained using each instruction tuning dataset.

60억 개의 국문 및 영문 토큰을 추가로 학습시켜 한국어 성능을 향상시킨 Llama-2-KoEn-13B 모델을 사용하였다(Lee and Choi, 2023).

각 학습 과정에서는 단계별 학습이 적합하게 완료되었는지를 신속하게 시험하기 위하여 Table 4에 제시된 평가용 데이터셋 중 9종을 선별하여 정량 평가를 수행하였다(Fig. 1, 2, 3). 각 평가에서의 점수는 정답률(전체 문제

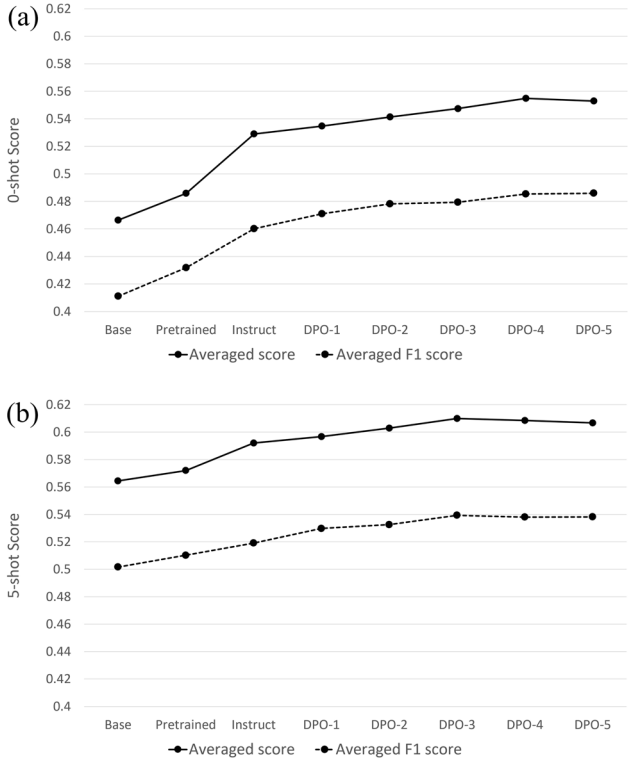

Fig. 3. Comparison of epoch-by-epoch evaluation results using 9 evaluation datasets for alignment tuning models. (a) 0-shot; (b) 5-shot.

중 정답을 출력한 비율)을 의미한다. F1 score는 정답 클래스의 분포가 불균형한 경우에서의 모델 성능을 정량적으로 평가하기 위한 것으로서 다음과 같이 계산된다.

$$F1\,score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

여기서 TP(True Positive)는 참을 참으로 분류한 개수, FP(False Positive)는 거짓을 참으로 분류한 개수, FN(False Negative)은 참을 거짓으로 분류한 개수를 나타내며, 사지선다형 문제와 같이 선택지가 2개를 초과하는 경우에는 동일한 방법으로 각각의 클래스에 대한 값을 구한 뒤 평균한다.

학습 과정 적합성 평가용 데이터셋은 과학 관련 성능 및 지능을 정량화하기 위하여 지능화, 전문화, 일반 지능, 과학적 사실 유추 등을 포함하여 구성하였으며, 각 데이터셋에 대한 세부 사항은 Table 4에 구체적으로 기재하였다.

2.2.1. 사전학습

사전학습은 주어진 문맥에서 다음에 출현할 단어(토큰)의 분포를 학습하는 비지도 학습 과정이다. 해당 학습 과정에서는 대량의 문장 데이터를 입력하여 언어의 구성 요소, 문장의 구조 및 의미 등을 학습하게 되며, 이러한 과정이 완료되면 언어 모델은 주어진 문장 이후 각 단어가 출현할 확률을 연산할 수 있게 된다.

본 학습에서는 앞선 가공 학술 자료 전체를 이용하였으며, 학습률은 자료의 활용을 극대화하기 위하여 상수로 설정하였다(Table 1). 사전학습 과정은 총 7.6일이 소요되었으며, 학습 과정의 평가를 위하여 매 10%의 학습이 진행될 때마다 모델을 저장하였다. 또한 사전학습과 이후의 미세조정 과정에서는 학습 효율을 최적화하기 위해 DeepSpeed 라이브러리를 활용하였다(Rasley et al., 2020). DeepSpeed는 PyTorch 라이브러리를 이용한 학습의 수행 시 학습 모델의 성능을 저하시키지 않으면서 소요되는 메모리를 최적화하여 학습 속도를 증가시키고 동일한 하드웨어에서 더 큰 모델을 학습할 수 있도록 한다. DeepSpeed의 핵심 알고리즘인 ZeRO(Zero Redundancy Optimizer)는 ZeRO-1, 2, 3의 세 단계로 구성되어 있으며, 본 연구에서는 메모리의 효율성과 속도 증가에 대한 균형적 측면에서 일반적으로 사용되는 Stage 2를 활용하였다.

생성형 언어 모델을 활용하여 질문이나 평가를 수행하는 경우, 해당 질문과 유사한 문답의 예시를 함께 제시하면 출력물의 정확도와 품질이 크게 향상된다는 점이 널리 알려져 있다(Brown et al., 2020). 예시가 전혀 주어지지 않은 상태(0-shot)에서도 언어모델이 준수한 성능을 보인다면 모델의 학습이 성공적으로 이루어진 것으로 평가할 수 있다.

사전학습에서는 원본 모델, 학습 중 저장한 모델 및 완료된 모델에 대한 평가를 수행하였다. Fig. 1a는 원본 모델부터 사전학습이 완료된 모델까지의 0-shot 평가 결과에 대한 평균 점수(정확도)를 나타낸다. 학습 도중에는 지속적으로 다른 자료가 입력되므로 정확도에서 다소 등락이 있으나 전반적으로는 성능의 향상이 나타났다. 이는 언어 모델이 학술 자료를 추가적으로 학습하더라도 모델의 일반적 성능이 하락하지 않으며, 여전히 다양한 작업을 처리할 수 있음을 의미한다. 예시가 제시된 상태에서의 모델 평가는 각각 1개, 5개의 예시를 제시하는 1-shot, 5-shot으로 수행하였으며, 해당 시험은 모델의 정상 작동을 평가하기 위함이므로 학습 중간 단계에서는 2개 모델에 대해서만 수행하였다. 평가 결과 일반적 사례와 일치하는 전체적 점수의 향상이 나타났으며(Fig. 1b), 예시를 제시하였을 때 정확도의 향상이 나타나는 점은 모델의 언어적 기능이 정상적으로 작동함을 의미한다. 이를 통해 생성형 언어 모델에 대한 지질과학 분야로의 특화를 위하여 통상적으로 수행하는 전체 인자 학습을 기반 모델에 추가로 수행하더라도 성능상의 문제가 나타나지 않음을 확인하였다.

Table 1. Hyperparameters and time required for the pre-training

Hyperparameter and variable	Value
Learning rate	1e-5
Learning rate scheduler type	constant with warmup
Warmup steps	100
Torch data type	bfloat16
Per-device batch size	4
Gradient accumulation steps	8
Training steps	5,506
Training runtime(hr)	182.4

2.2.2. 미세조정(Fine-tuning)

사전학습이 완료된 생성형 언어 모델은 학습 기법의 특성상 주어진 문맥에 적합한 후속 문장들을 생성하는 방식(text completion)으로만 작동 가능하다. 이러한 모델의 출력 결과는 언어적으로 유창함에도 불구하고 대화형 작업이 불가능하므로 특정 요청을 처리하거나 일관된 대화를 수행하는 데에 한계가 있으며, 출력 결과의 방향성을 예측하기 어려운 문제가 있다. 또한 사전학습에 사용되는 데이터는 그 규모가 방대하고 자료의 출처 및 주제의 범위가 매우 넓기 때문에 모든 내용의 적합성이나 품질을 철저히 검수하기 어렵다. 이로 인하여 공격적이거나 불확실한 내용, 비윤리적인 답변과 같이 부적절한 출력이 발생할 위험이 존재하므로, 모델이 사용자 요청에 적합한 답변을 생성할 수 있도록 미세조정 학습을 수행하여야 한다.

미세조정은 입력된 요청에 부합하는 대화형 답변을 출력하도록 학습하는 instruction tuning과, 출력의 방향성 및 적합성을 조정하여 전체적인 품질을 향상시키는 alignment tuning으로 나뉜다. 최근에는 적은 학습 데이터로도 모델의 성능을 향상시킬 수 있는 다양한 기법들이 개발되어 널리 적용되고 있다(Sanh et al., 2021).

미세조정 단계에서는 모델의 전체 또는 일부 파라미터를 직접 튜닝하는 것과 LoRA (Low-Rank Adaptation)를 이용한 학습 방식 간의 성능 차이가 거의 없음이 Llama 2 모델을 기반으로 한 시험을 통해 알려져 있다(Niederfahrenhorst et al., 2023). LoRA는 학습 시 모델의 인자를 직접 수정하지 않고, 모델의 각 부분에 대한 출력을 변조하는 별도의 어댑터 모델을 생성하여 일종의 간접 학습을 수행하는 기법으로서 원래 모델보다 훨씬 작은 모

델을 학습하므로 특정 목적을 위한 학습에 있어서 매우 높은 효율성을 갖는다(Hu et al., 2021). 따라서 본 연구에서도 LoRA를 활용하여 미세조정을 수행하였다.

Instruction tuning은 단순 생성형 언어 모델을 대화식 출력이 가능하도록 학습하는 주요 과정으로서, 모델이 사용자의 요청사항에 부합하는 답변을 출력하도록 한다. 학습 자료는 일반적으로 질문과 입력 인자, 답변이 하나의 대화 예시로서 구성되며, 다양한 분야에 걸친 단순 질의, 논리적 추론, 번역, 분류 등의 작업 사례가 포함된다. 추가적으로 언어 모델이 특정 작업을 수행할 수 없을 때의 대응 방안을 포함한 출력 예시도 자료에 포함되어 답변이 불가능한 질문에 대해 잘못된 답변을 출력하지 않도록 유도한다.

이처럼 미세조정 과정에서는 특정 형태, 양식, 태도, 구조로 요청된 질의에 적합한 답변을 생성하는 학습을 수행해야 하므로, 언어 모델의 성능을 최적화할 수 있도록 설계된 특정 데이터셋이 필요하다. 사전학습 자료와 유사하게 미세조정 시 사용하는 데이터셋 역시 언어별로 다르게 구성되어야 하며, 한국어 미세조정 데이터셋의 경우 주로 영문으로 개발된 데이터셋을 번역하거나 재구축하여 사용되고 있다. 본 연구에서는 기존 사례에서 높은 성능의 증가가 보고되었으며 이론적 기반이 확립되어 있는 Open-platypus와 OpenOrca 데이터셋을 사용하였다(Lee et al., 2023, Mukherjee et al., 2023). Open-Platypus는 다수의 데이터셋에서 선별된 데이터를 종합하여 재구성한 것으로, 이를 활용해 Llama 2를 미세조정한 모델인 Platypus는 HuggingFace의 Open LLM Leaderboard에서 1위를 기록한 바 있다. OpenOrca는 본 연구에서 활용한 13B급 모델의 튜닝을 위해 설계된 데이터셋으로서 더 큰 모델의 추론 과정을 모사하는 기법을 학습할 수 있도록 구성되어 있으며, 해당 데이터셋을 학습한 Orca 모델은 ChatGPT 3 수준의 성능을 보이는 것으로 보고되었다.

이상의 학습 자료들을 기반으로 수행한 학습 인자 및 소요 시간은 Table 2에 제시된 바와 같다. LoRA의 적용 대상은 가장 효과적으로 알려져 대부분의 미세조정에서 활용되는 query projection(q_proj), value projection(v_proj) 모듈로 지정하였다. 이러한 인자들을 고정한 상태에서 앞선 2종의 미세조정 데이터셋을 각각 적용하는 학습과 두 데이터를 결합하여 적용하는 학습을 모두 진행하였으며, 그 결과는 Fig. 2와 같다. 어떠한 미세조정 데이터셋을 적용하더라도 원본 모델 및 학술 논문으로 사전학습된 모델에 비해 높은 성능이 일관적으로 나타났으며, 특히 두 데이터셋을 병용한 경우에서 가장 높은 성능을 나타내었으므로 해당 모델을 이용하여 다음 학습을 수행하였다.

Alignment tuning은 모델의 출력 품질 향상을 주요 목표로 하는 학습 과정으로서, 기존에는 주로 인간의 피드백을 기반으로 한 RLHF(Reinforcement Learning from Human Feedback, Ouyang et al., 2022)가 널리 사용되었다. 이 방법은 인간이 선정한 출력 예시의 순위를 학습에 활용하여 보상 모델이 해당 순위를 토대로 점수를 유추하는 방식으로 동작하나, 학습률과 같은 하이퍼파라미터에 민감하며 연산이 복잡하다는 단점이 있다.

Table 2. Hyperparameters and time required for the fine-tuning

Hyperparameter and variable	Instruction tuning	Alignment tuning
Learning rate	3e-5	
Learning rate scheduler type	cosine	
Torch data type	bfloat16	
Per-device batch size	4	
Gradient accumulation steps	8	
Target modules	q_proj, v_proj	
Training epochs	5	
Training steps per epoch	363	484
Training runtime per epoch(hr)	5.91	21.4

최근에는 보상 모델의 명시적 설계 없이도 높은 성능을 발휘하는 DPO(Direct Preference Optimization, Rafailov et al., 2023)가 제안되었다. DPO는 RLHF와 유사한 성능을 보이면서도 구현이 비교적 간단하고 연산이 효율적이라는 장점이 있다. 이로 인해 최근 공개되고 있는 대부분의 고성능 LLM에서는 미세조정에서 DPO를 활용하고 있으며, 본 연구에서도 해당 기법을 활용하여 최종 모델의 성능을 극대화하고자 하였다. 이에 따라 Table 2와 같은 인자를 사용하여 5 epoch에 걸쳐 학습을 진행하였으며, 각 epoch의 종료 시점마다 모델을 저장하여 이후 평가에 활용할 수 있도록 하였다.

Alignment tuning을 수행한 모델의 정량 평가 결과 Fig. 3과 같이 학습의 진행에 따라 점진적인 평가 지표의 상승이 나타났다. 그러나 각 평가에서 epoch 4 전후의 지표가 다소 하락하는 현상이 나타났으며, 따라서 모든 평가용 데이터셋을 활용한 모델의 정량 기능 측정시에는 이상의 각 모델별 평가를 학습 종료 모델뿐 아니라 이전의 모델까지 모두 진행하였다.

3. 모델 평가

3.1. 정성적 평가

본 연구에서는 사전 학습 단계에서 학술 자료를 활용하여 모델에 관련 지식을 입력하도록 하였으며, 이에 따라 해당 학습 이후부터 지질과학 관련 학술적 답변이 가

능할 것으로 예상할 수 있다. 이후의 단계에서는 LoRA를 이용한 미세 조정을 통해 모델이 대화형 작업을 수행할 수 있도록 학습하였으나, 해당 과정에서는 학술 자료가 활용되지 않았으므로 사전 학습에서의 지식을 여전히 유지하고 있는지에 대한 시험이 필요하였다. 따라서 학술 자료가 학습되지 않은 원본 모델, 사전학습 완료 모델 및 미세조정 완료 모델에 각각 동일한 입력(프롬프트)을 수행하여 출력 결과를 비교하는 정성적 평가를 수행하였다.

생성형 언어 모델의 출력은 주어진 입력 문장의 다음에 출현 가능한 각 토큰의 확률이다. 이 확률을 이용하여 다음에 출력할 토큰을 선택하는 과정, 즉 샘플링은 모델의 구동 범위에서 벗어나므로 모델간의 비교를 위해서는 샘플링에 요구되는 인자들 또한 동일하게 구성하여야 한다. 일반적으로 샘플링 과정은 다양한 기법과 인자들을 사용하여 행해지나, 본 평가에서는 필수적인 인자로서 가장 널리 사용되는 top_p와 temperature만을 활용하였다. top_p는 누적 확률이 p 이하가 되도록 샘플링 대상 토큰의 후보를 제한하는 인자이며, temperature는 모델의 확률값 출력 시 각 값의 차이를 조절하는 역할을 수행한다. 두 인자 모두 0에 가까울수록 높은 확률을 갖는 토큰을 위주로 출력하게 되며, 1인 경우 해당 인자들을 사용하지 않는 것과 같다. 본 평가에서는 세 모델 모두에 대하여 top_p와 temperature를 0.5로 고정하여 출력을 수행하였다.

Table 3은 동일한 입력에 대한 세 모델의 출력 결과 예시를 나타낸다. 학습 이전 모델(Base)과 사전학습된 모델(Pretrained)은 대화형 입출력이 불가하여 입력된 문장 이후를 완성하는 형태로 내용이 출력되며, 미세조정이 완료된 모델(DPO-4)은 입력된 문장을 하나의 완결된 질문으로 간주하고 이에 대한 답변의 형태로 출력이 수행된다. 따라서 각 모델에 동일한 입력을 수행하기 위하여 모든 입력 문장은 기호가 제외된 질문의 형태로 구성하였다.

출력 예시에서는 전체적으로 사전학습과 미세조정이 수행된 모델에서 입력 문장에 더 적합한 연관 단어 및 내용이 나타나는 경향이 관찰되며, 특히 일반적으로는 사용되지 않는 단어나 개념에 관한 질문일수록 그러한 점이 두드러지므로 학습이 유효하게 수행된 것으로 판단할 수 있다. 질문에서 언어모델에 학습되지 않은 내용을 요구할 경우 모델은 실질적인 내용이 없거나 잘못된 답변(환각)을 출력하게 되며, Table 3에 나타난 학습 이전 모델의 답변 모두가 이에 해당한다. 대표적으로 2번 예시의 경우 기본 모델에서 학습되었던 내용 중 질문의 제시어인 THMC(Thermal-Hydraulic-Mechanical-Chemical)와 가장 유사한 트리할로메탄(Trihalomethane, THM)이 잘못된 답변으로서 표출된 것으로 보이나, 지질과학 분야에 대한 중점적 학습 이후 답변의 내용이 정상적으로 나타난다. 그러나 사전학습은 주어진 내용을 모두 기억하는 과정이 아니므로 이미 학습된 내용에 대해 질문하더라도 항상 적합하거나 올바른 내용으로 표출되지는 않으며(2c, 5b, 7b, 7c, 8c), 특히 과학 분야에서 흔히 활용되는 단어를 통해 지엽적인 개념을 묻는 경우(7, 8) 출력되는 단어의 방향성이 질문의 의도에서 벗어나거나 미세조정 단계에서 관련 내용을 쉽게 손실하는 것으로 보인다.

이러한 점을 해결하기 위해서는 사전학습 데이터의 양을 크게 늘리거나 미세조정 단계에서 활용되는 데이터를 연관 내용으로 구축하여야 할 필요가 있으나, 한정된 분야에서의 학습 데이터량 증가는 일반적으로 활용되는 단어에 대한 망각이나 학습 자료에 포함된 주요 연관 문구 및 사례를 환각으로 표출하는 문제(5b)를 발생시킬 수 있다. 따라서 지질과학 분야로의 전문화를 수행하면서도 이러한 문제를 방지하기 위해서는 지질과학 분야에 적합한 학습용 데이터의 개발이 추가적으로 필요할 것으로 보인다. 또한 최근에는 답변에 필요한 내용을 검색 후 이를 참조하여 정확한 답변을 수행하는 검색증강생성(Retrieval-Augmented Generation, RAG) 기법이 상용 서비스를 중심으로 활발히 활용되고 있으므로(Lewis et al., 2020) 이러한 문제를 해결하기 위한 효율적인 방법으로서 고려될 수 있다.

3.2. 정량 평가용 데이터 구성

모델 성능의 정량적 평가를 위해 사용되는 기법과 평가 기준에는 많은 종류가 있으며, 최근에는 단순히 모델이 특정 사실을 알고 있는지를 평가하는 것을 넘어 특정 상황이 주어졌을 때 그 사실을 인지하여 올바른 대답을 하는지를 평가하는 일종의 지능 복합 평가의 형태로 이루어지고 있다.

과학 및 기술 분야에서도 다양한 데이터셋이 제시되어 있으며, 이러한 평가들은 주로 과학적 사실이나 개념에 대한 직접적인 질문을 포함하고 있다. 그러나 지질학, 자원공학, 지구과학 등 특정 분야에 특화된 표준 평가 자료는 존재하지 않으며, 과학 관련 다분야 데이터셋인 ARC나 SciQ 등에서 일부 지질과학 관련 기초 문항이 포함된 정도에 그치고 있다. 일부 선행 연구(Deng et al., 2023)에서는 대학 1학년 수준의 지질학 시험 내용을 포함하는 데이터셋이 사용되었으며, 본 연구에서도 해당 데이터셋을 한국어로 번역하여 활용하였다.

모델 개발 과정에서는 중간 평가를 위하여 9개 데이터셋을 제한적으로 활용하였으나, 본 평가에서는 다양한 유형별, 기준별 정량적 성능 지표를 측정하기 위해 한국어 및 영어로 된 19개의 데이터셋을 사용하였다(Table 4). 데

Table 3. Examples of responses to the same prompt entered in 3 different models

No.	Prompt / Response		
	a. Base	b. Pretrained	c. DPO-4
1	부유선별법이란		
	부유물질을 제거하기 위한 방법 중 하나로, 부유물질을 물에 뜨게 하여 제거하는 방법입니다.	부유선별기(flotation cell)에서 기포(bubble)를 이용하여 광물입자를 기포에 부착시켜 광물입자를 분리하는 방법이다.	“부유선별”은 광업에서 광석과 같은 유용한 물질을 분리하고 제거하려는 불필요한 물질을 제거하기 위해 사용되는 기법입니다. 부유선별은 물에 가라앉는 물질(침강성)과 물에 뜨는 물질(부유성)을 분리하는 데 사용됩니다.
2	THMC 복합모델이란		
	THM(트리할로메탄)의 생성 및 제거 메커니즘을 고려한 모델로서, THM의 생성 및 제거 메커니즘을 고려하여 THM의 농도를 예측하는 모델입니다.	지반의 역학적 거동과 수리학적 거동, 그리고 열적 거동과 화학적 거동 등 다양한 거동을 동시에 고려하는 모델을 의미한다.	THMC는 열역학(T), 유체역학(H), 재료 역학(M), 화학(C)의 머리글자를 조합하여 만든 용어입니다. THMC 복합 모델은 다양한 화학적, 기계적, 열적, 유체역학적 과정을 통합하여…
3	SAR 영상의 활용 사례로는		
	1991년 걸프전 당시 이라크의 핵시설 파괴를 위해 SAR를 이용한 것이 가장 대표적인 사례이다. SAR는 지상에서 관측할 수 없는 지하 시설물이나, 지형지물의 변화에 대한 정보를 획득할 수 있기 때문에…	2010년 아이슬란드 화산 분화 시에 ESA의 Sentinel-1 SAR 영상과 TerraSAR-X SAR 영상을 이용하여 화산 분화 시의 화산재 분포와 화산재 확산 범위를 분석한 사례가 있다. 또한,…	SAR(Synthetic Aperture Radar) 이미지는 레이더 신호를 사용하여 지형을 스캔하여 이미지를 생성하기 때문에 악천후에 영향을 받지 않습니다. SAR 이미지는 다양한 애플리케이션에 매우 유용하며, 다음은 몇 가지 사례입니다:1. 재난 관리…
4	석유 저류층에서 수압파쇄를 사용하는 이유는		
	석유 저류층에서 석유와 물이 혼합되어 있는 경우에 물을 제거하기 위함이다. 수압파쇄는 석유 저류층에서 석유와 물을 분리시키는 방법이다.	저류층 내의 공극률과 투수율을 증가시켜 생산성을 향상시키기 위함이다. 따라서 저류층 내의 공극률과 투수율을 증가시키는 것은 수압파쇄의 성공을 결정하는 중요한 요소이다.	수압파쇄는 석유 및 가스 저류층에서 석유를 생산하기 위해 사용되는 기술입니다. 수압파쇄는 저류층 내의 암석이나 지층을 깨뜨리기 위해 고압의 물을 주입하여 생산성을 개선하는 방법입니다.
5	다분광 영상을 이용하여 지표피복을 분류하려면		
	먼저 영상에서 피복의 종류에 따라 다른 특성을 갖는 피복의 분류를 위한 기준(threshold)을 설정해야 한다. 본 논문에서는 다분광 영상에서 피복의 종류에 따라 다른 특성을 갖는 피복의 분류를 위한 기준을 설정하기 위해…	영상의 분광정보가 지표피복의 분광반사특성을 반영해야 한다. 따라서 본 연구에서는 2000년 10월 28일 Landsat ETM+ 영상과 2000년 10월 29일 SPOT-5 영상의 분광반사특성을 비교하였다. 두 영상의 분광반사특성을 비교하기 위해…	다분광 영상을 사용하여 지표 피복을 분류하려면 다음 단계를 따르세요: 1. 영상 수집: 다분광 영상은 위성, 항공기 또는 지상 센서에서 수집할 수 있습니다. 영상은 다양한 분광 밴드(각각 고유한 파장 범위)를 포함해야 합니다.
6	절토사면에서 지하수에 의한 간극수압이 증가하면		
	토양의 압축강도가 감소하게 된다. 이러한 압축강도의 감소는 토양의 압축변형에 의해 발생되는 변형에너지의 감소로 인하여 발생한다.	사면의 전단강도가 감소하게 된다. 따라서 사면의 안정해석시 지하수위를 고려하는 것이 필요하다. 본 연구에서는 지하수위를 고려한…	절토사면에서 지하수에 의한 간극수압이 증가하면, 절토사면의 안정성에 영향을 미칠 수 있는 다양한 요인이 있습니다. 간극수압이 증가하면 절토사면의 전단 강도가 감소하여 사면이 더 쉽게 미끄러질 수 있습니다.
7	한반도 백악기 화성암류의 형성은		
	백악기 초에 형성된 선캄브리아기 변성암류의 열수변질작용에 의해 형성된 변성암류의 관입에 의해 형성된 것으로 해석된다.	동아시아 대륙 아래로 섭입하는 고태평양판(Izanagi plate)의 영향으로 인해 발생한 것으로 알려져 있다. 이 섭입작용은 한반도 남동부 경상분지 내 백악기 화산암류의 분출과 함께 한반도 남서부 해남지역에서부터…	한국에서는 발견되는 백악기 화성암류의 형성은 주로 지질학적 시간과 지질학적 과정의 결과입니다. 한반도 백악기 화성암류의 형성은 주로 지질학적 시간에 따라 일어났습니다. 백악기는 약 1억 4,500만 년 전부터 6,600만 년 전까지 지속된 지질학적 시대입니다.
8	매질의 투수성이 낮은 대수층에서 수행하기에 적합한 수리시험은		
	시추공에 일정한 간격으로 시료채취공을 설치하고, 시료채취공에서 시료를 채취하여 시료의 수분함량을 측정함으로써, 시추공의 투수계수를 구하는 방법이다.	양수시험이다. 양수시험은 대수층에서 지하수 흐름을 측정하는 가장 일반적인 방법이며, 양수정의 양수량과 수위강하량의 관계를 이용하여 대수층의 수리상수를 결정한다.	대수층에서 투수성이 낮은 경우에는 저유속 투수성 시험(LTT)이 가장 적합한 수리 시험입니다. LTT는 저유속 조건에서 투수성 측정 장치를 사용하여 대수층에서 투수성을 측정하는 수리 시험입니다.

Table 4. Datasets used for the evaluation of the language model

Dataset	Abbreviation	Description
ARC Challange	ARC-C	AI2 Reasoning Challange, Basic level scientific reasoning test
ARC Easy	ARC-E	
BoolQ	BoolQ	A test to determine the consistency of content between short passages
BoolQ (Korean)	BoolQ-K	
COPA	COPA	Choice of Plausible Alternatives, A test that infers which of two causes given for a fact is more reasonable
COPA (Korean)*	COPA-K	
AP Exam (Geoscience)	AP-Geo	Questions related to geology in Advanced Placement examinations
HT Astronomy*	HT-Astron	Hendrycks Test, a science-related dataset among multiple choice tests organized by various fields
HT College Biology*	HT-Bio	
HT College Chemistry*	HT-Chem	
HT College Physics*	HT-Phys	
HT Conceptual Physics*	HT-CPhys	
HT High School Geography*	HT-Geo	
HellaSwag (Korean)	HellaSwag-K	A test that predicts the next situation using common sense after a sentence describing a certain state
SentiNeg (Korean)*	SentiNeg-K	A test that classifies whether a sentence contains a positive or negative meaning
OpenBookQA	OBQA	Multiple choice questions on basic scientific facts
PIQA	PIQA	Physical Interaction: Question Answering, multiple-choice questions on common-sense level physics problems
SciQ*	SciQ	13,679 multiple choice questions on physics, chemistry, biology, etc.
WinoGrande	WG	A two-choice test of the appropriate word to fill in the blank in the sentence

이터셋들은 과학적 사실에 근거한 인과 추론(ARC), 과학 관련 언어 이해(OpenBookQA), 과학 지식 평가(AP Exam, Hendrycks Test, SciQ), 일상적 논리 추론(COPA, HellaSwag, PIQA), 언어 및 문맥 이해(BoolQ, SentiNeg, WinoGrande) 등으로 분류할 수 있으며 언어 모델의 성능 평가에 널리 사용되는 COPA와 BoolQ의 경우 한국어, 영어 데이터셋을 모두 포함하였다.

19개의 평가 지표 중 8개(AP-Geo, HT, SciQ)는 별도의 지문 없이 과학적 사실이나 개념에 대한 모델의 지식 및 인지 능력을 직접적으로 평가하는 데이터셋으로서, 본 연구에서 학습한 지질학 및 자원 분야의 학술 자료와 달리 주로 기초 과학에 해당하는 데이터를 포함한다. 4개의 데이터셋(BoolQ-K, COPA-K, AP Exam, HellaSwag-K)은 한국어로 구성되었으며, 이러한 평가들은 한국어로 구성된 지질과학 학술 자료의 학습 이후 모델의 한국어 관련 기능성을 별도로 평가할 필요가 있다고 판단하여 포함하였다.

본 평가는 일반적 성능 지표로서 중시되는 0-shot 시험만으로 구성하였으며, 평가 방식에 따라 F1 score가 산출되지 않는 경우가 있으므로 해당 점수의 산출은 생략하였다. 평가 대상 모델은 원본 모델(Base), 사전학습이 완료된 모델(Pretrained), 미세조정 중 instruction tuning이 완료된 모델(Instruct), DPO 학습의 epoch별 출력 모델(DPO-1, 2, 3, 4, 5)로 총 8개이다. 각 모델에 대한 평가에는 약 1시간 45분이 소요되었다.

3.3. 모델 정량 평가 결과 및 토의

각 평가 지표에 대한 정량적 평가 결과는 Table 5 및 Fig. 4와 같다. 전체적인 지표는 모델 개발 도중의 평가 결과와 유사한 양상을 보였으며, DPO epoch 5에서 이전 모델보다 평균 점수가 하락하는 경향도 동일하게 나타났다. DPO-4와 DPO-5의 평가에서 학습 이후 정확도가 상

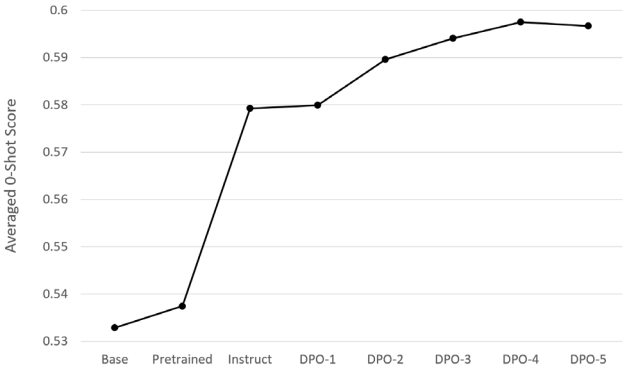

Fig. 4. Average score of final evaluation results using 19 evaluation datasets.

Table 5. Final quantitative evaluation results for each model

	Raw	Pretrained	Instruct	DPO-1	DPO-2	DPO-3	DPO-4	DPO-5
ARC-C	0.4130	0.4172	0.4445	0.4531	0.4599	0.4642	0.4616	0.4642
ARC-E	0.6886	0.7012	0.7247	0.7319	0.7311	0.7315	0.7298	0.7298
BoolQ	0.7789	0.7495	0.7679	0.7927	0.7927	0.7969	0.8064	0.8034
COPA	0.8800	0.8700	0.8900	0.8900	0.8900	0.8800	0.8800	0.8800
AP-Geo-K	0.2889	0.2853	0.3118	0.3168	0.3118	0.3111	0.3111	0.3125
HT-A	0.4605	0.4803	0.4671	0.4474	0.4671	0.4803	0.4737	0.4737
HT-CB	0.2708	0.2986	0.3542	0.3889	0.3819	0.3958	0.3889	0.4028
HT-CC	0.2500	0.2500	0.3400	0.3300	0.3400	0.3500	0.3900	0.3800
HT-ColP	0.2255	0.2451	0.2451	0.2941	0.2941	0.3235	0.3529	0.3333
HT-ConP	0.3787	0.3234	0.4000	0.3787	0.3872	0.3660	0.3660	0.3702
HT-HG	0.3687	0.3636	0.5303	0.5303	0.5354	0.5354	0.5404	0.5404
BoolQ-K	0.5677	0.5556	0.7934	0.7080	0.8405	0.8618	0.8618	0.8597
COPA-K	0.7970	0.7770	0.7860	0.7900	0.7930	0.7920	0.7920	0.7910
HellaSwag-K	0.5000	0.4780	0.5040	0.5000	0.4940	0.4980	0.4960	0.4960
SNeg-K	0.5340	0.7128	0.7053	0.7103	0.7305	0.7380	0.7456	0.7431
OBQA	0.3260	0.3020	0.3220	0.3180	0.3200	0.3240	0.3220	0.3240
PIQA	0.7802	0.7720	0.7824	0.7851	0.7818	0.7840	0.7840	0.7824
SciQ	0.9120	0.9220	0.9330	0.9430	0.9430	0.9460	0.9440	0.9420
WG	0.7040	0.7080	0.7040	0.7103	0.7088	0.7088	0.7064	0.7088
Averaged Score	**0.5329**	**0.5375**	**0.5792**	**0.5799**	**0.5896**	**0.5941**	**0.5975**	**0.5967**

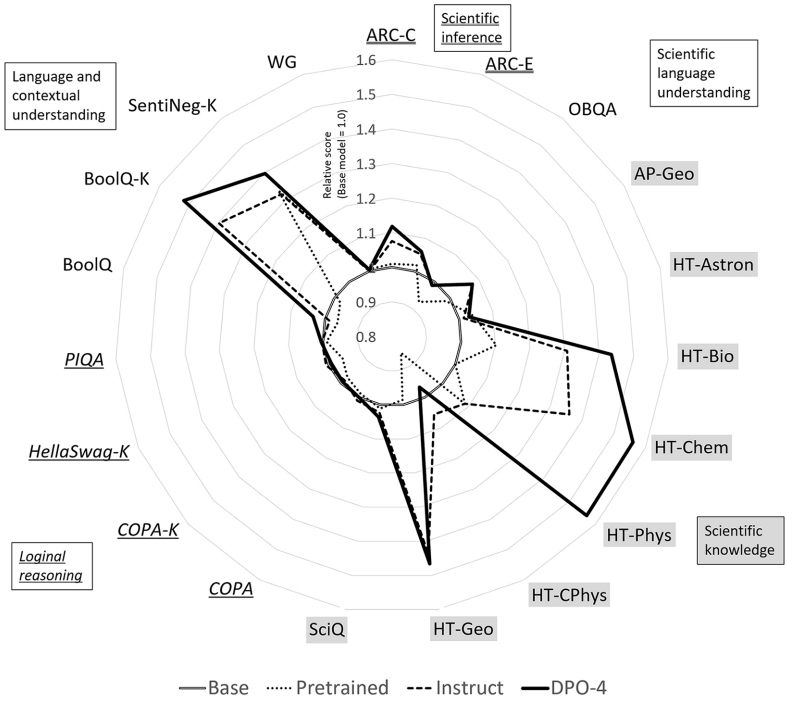

Fig. 5. Relative comparison of the final evaluation results of each stage model, with the original base model's score set to 1.

승한 지표는 6개, 하락한 지표는 8개로 나타났으므로 과적합에 따른 오류를 피하기 위해 DPO-4를 최종 모델로 선정하였다.

Fig. 5는 원본 모델의 각 평가 지표별 점수를 1.0으로

하여 각 학습 단계별 모델의 분야별 성능을 상대적으로 비교한 것이다. DPO-4 모델과 원본 모델을 비교한 결과 19개의 평가 지표 중 14건에서 성능 향상이 있었으며, 4건에서 미세한 성능 하락이 나타났다. 특히 수학적 풀이가 없는 물리학 질문으로 구성된 HT-CPhys의 경우 원본 모델과 비교해 1.3%p의 차이를 보였는데, 이는 사전학습 과정에서 큰 하락을 보인 점이 주요 원인으로 보인다. 언어 이해와 논리 추론 관련 평가에서는 미세한 하락이 나타났으나, 가장 낮은 성능이 원본 모델의 98.7% 수준이므로 한정된 분야의 자료의 학습에 의한 부정적 효과는 매우 낮은 것으로 파악된다.

단순 사실에 대한 질의를 수행하는 8개 지표 중에서는 HT-CPhys를 제외한 모든 지표에서 원본 모델보다 높은 점수를 보였다(Fig. 6). 이러한 단순 사실에 대한 평가의 특성상 대량의 자료를 직접 학습하는 사전학습 단계에서 주된 차이를 보이고 이후에는 큰 차이가 없을 것으로 기대되었으나, 실제로는 대부분의 경우에서 미세조정 단계의 성능 향상이 더 크게 나타났다. 이는 모델의 평가가 사람의 평가와 유사하게 문제의 제시와 답변이 언어적으로 수행되므로 미세조정 학습 이후 이해 능력이 발달하면서 문제를 이해하고 풀이하는 과정에 요구되는 성능의 향상에 기인하여 정답률이 높아진 것으로 유추할 수 있다.

한국어로 구성된 4종의 시험에서는 원본 모델 대비 유사하거나 일부 상승된 성능이 관찰되었으며, 기초과학 분야에서는 대체적으로 높은 성능 향상이 나타났다. 가장 큰 상대적 성능 향상이 나타난 평가는 BoolQ-K(지문간 동일성 해석 평가, 51.8%), SNeg-K(문장 긍정-부정 평가, 39.6%), HT-Bio(대학 생물학, 43.6%), HT-Chem(대학 화학, 56.0%), HT-Phys(대학 물리, 56.0%), HT-Geo(고교 지

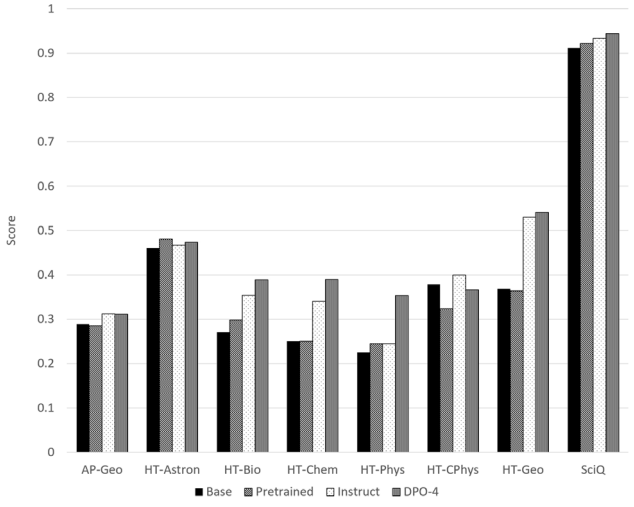

Fig. 6. Comparison of evaluation results of science-related datasets by model.

리학, 46.6%) 등이며, 동일한 평가에 해당하나 언어가 다른 COPA의 경우 BoolQ와 달리 원본 모델과 거의 차이가 없는 결과를 나타내었다. 이는 한국어 문헌에 집중된 사전학습과 미세조정을 수행하더라도 향상 가능한 성능이 세부 분야별로 매우 상이함을 나타낸다.

이와 같은 평가 결과를 종합하면, 주로 과학 분야에서의 성능이 크게 향상되었고, 한국어 관련 지문 해석 능력 또한 상승한 것으로 평가할 수 있다. 일부 지표에서의 하락이 있었지만 그 정도가 크지 않으므로, 본 연구의 접근법과 평가 과정은 적절하게 수행되었으며, 향후 연구에서도 동일한 접근법을 사용할 수 있을 것으로 판단된다.

다만, 본 연구에서 사용된 대부분의 평가용 데이터셋은 지질과학 분야의 학술 논문과 큰 관련성이 없으며, 관련된 평가 지표들도 다소 기초적인 수준에 그치고 있어 본 연구에서 개발한 모델의 성능을 정량적으로 측정하기에는 한계가 있다. 향후 유관 분야에 특화된 언어 모델 개발을 지속하기 위해서는 지질과학 분야의 편향되지 않은 정량 평가용 데이터셋의 개발이 필요할 것으로 보인다.

4. 결 론

본 연구에서는 지질과학 분야에 특화된 생성형 거대 언어모델의 개발을 위하여 지질학 및 자원공학 등의 유관 분야에 대한 학술 자료의 수집, 가공과 이를 이용한 언어 모델의 개발 및 평가에 이르는 전 과정을 수행하였으며, 이를 통해 개발된 언어 모델은 한국어 해석 및 과학 관련 성능에서 원본 모델보다 현저히 높은 성능을 갖는 것으로 평가되어 유관 분야에서의 충분한 활용 가능성을 나타내었다.

본 연구에서 개발된 언어 모델의 개별적인 활용을 위해서는 약 48GB의 GPU 메모리가 필요하며, 이는 일반적인 업무에서 사용되는 GPU 메모리 용량인 24GB의 약 2배에 해당하므로 추후 보다 효율적인 모델 배포 및 활용을 위해서는 양자화(quantization)를 통한 모델의 규모 축소 및 추가적인 평가가 요구된다.

또한 본 연구의 결과를 바탕으로 확장 개발을 진행하기 위해서는 기술적인 부분 이외에도 사전학습, 미세조정 및 정량 평가용 데이터셋에 대한 질적, 양적 강화가 필요할 것으로 사료된다. 본 연구에서는 취약점 등의 유출 문제를 고려하면서도 최대한 많은 수의 자료를 확보하기 위하여 오픈액세스 논문만을 활용하였으나, 학습 자료의 분야적, 학술적 심화도를 다양화하면 언어 모델의 더욱 넓은 활용성을 부여할 수 있다. 본 연구의 결과는 지질과학 분야에서의 대규모 언어 모델 구축과 활용에 대한 기초를 제공함으로써 향후 보다 활발한 연구가 이루어질 수 있을 것으로 기대된다.

Acknowledgements

본 연구는 한국지질자원연구원 자체연구사업인 “지질자원분야 대규모 언어 모델 시범개발(23-7512)” 과제의 일환으로 수행되었습니다.

References

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger G., Henighan T., Child R., Ramesh A., Ziegler D., Wu J., Winter C., Hesse C., Chen M., Sigler E., Litwin M., Gray S., Chess B., Clark J., Berner C., McCandlish S., Radford A., Sutskever I. and Amodei, D. (2020) Language models are few-shot learners. Advances in Neural Information Processing Systems, v.33, p.1877-1901. https://doi.org/10.48550/arXiv.2005.14165

Deng, C., Zhang, T., He, Z., Xu, Y., Chen, Q., Shi, Y., Fu, L., Zhang, W., Wang, X., Zhou, C., Lin, Z. and He, J. (2024, March) K2: Learning A Foundation Language Model for Geoscience Knowledge Understanding and Utilization. In Proceedings of the 17th ACM International Conference on Web Search and Data Mining, Association for Computing Machinery, p.161-170. https://doi.org/10.1145/3616855.3635772

Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L. and Chen, W. (2021) Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685. doi: 10.48550/arXiv.2106.09685

Lawley, C.J., Raimondo, S., Chen, T., Brin, L., Zakharov, A., Kur, D., Hui, J., Newton, G., Burgoyne, S.L. and Marquis, G. (2022) Geoscience language models and their intrinsic evaluation. Applied Computing and Geosciences, v.14, 100084. doi: 10.1016/j.acags.2022.100084

Lee. J. and Choi, T. (2023) Llama-2-KoEn-13B. doi: 10.57967/hf/1280

Lee, A.N., Hunter, C.J. and Ruiz, N. (2023) Platypus: Quick, cheap, and powerful refinement of llms. arXiv preprint arXiv:2308.07317. doi: 10.48550/arXiv.2308.07317

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S. and Kiela, D. (2020) Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in Neural Information Processing Systems, v33, p.9459-9474. https://doi.org/10.48550/arXiv.2005.11401

Mukherjee, S., Mitra, A., Jawahar, G., Agarwal, S., Palangi, H. and Awadallah, A. (2023) Orca: Progressive learning from complex explanation traces of gpt-4. arXiv preprint arXiv:2306.02707. doi: 10.48550/arXiv.2306.02707

Nasr, M., Carlini, N., Hayase, J., Jagielski, M., Cooper, F.A., Ippolito, D., Choquette-Choo, C.A., Wallace, E., Tramèr, F. and Lee, K. (2023) Scalable extraction of training data from (production) language models. arXiv preprint arXiv:2311.17035. doi: 10.48550/arXiv.2311.17035

Niederfahrenhorst, A., Hakhamaneshi, K. and Ahmad, R. (2023, September 6) Fine-Tuning LLMs: LoRA or Full-Parameter? An in-depth Analysis with Llama 2. Anyscale. https://www.anyscale.com/blog/fine-tuning-llms-lora-or-full-parameter-an-in-depth-analysis-with-llama-2

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C.L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J. and Lowe, R. (2022) Training language models to follow instructions with human feedback. In Proceedings of the Advances in Neural Information Processing Systems 35 (NeurIPS 2022), Curran Associates, Inc., v.35, p.27730-27744. https://doi.org/10.48550/arXiv.2203.02155

Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C.D. and Finn, C. (2023) Direct preference optimization: Your language model is secretly a reward model. arXiv preprint arXiv:2305.18290. doi: 10.48550/arXiv.2305.18290

Rasley, J., Rajbhandari, S., Ruwase, O. and He, Y. (2020, August) Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Association for Computing Machinery, p.3505-3506. https://doi.org/10.1145/3394486.3406703

Sanh, V., Webson, A., Raffel, C., Bach, S. H., Sutawika, L., Alyafeai, Z., Chaffin, A., Stiegler, A., Le Scao, T., Raja, A., Dey, M., Bari, M. S., Xu, C., Thakker, U., Sharma, S.S., Szczechla, E., Kim, T., Chhablani, G., Nayak, N., Datta, D., Chang, J., Jiang, M.T., Wang, H., Manica, M., Shen, S., Yong, Z.X., Pandey, H., Bawden, R., Wang, T., Neeraj, T., Rozen, J., Sharma, A., Santilli, A., Fevry, T., Fries, J.A., Teehan, R., Bers, T., Biderman, S., Gao, L., Wolf, T. and Rush, A.M. (2021) Multitask prompted training enables zero-shot task generalization. arXiv preprint arXiv:2110.08207. doi: 10.48550/arXiv.2110.08207

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C.C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P.S., Lachaux, M., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E.M., Subramanian, R., Tan, X.E., Tang, B., Taylor, R., Williams, A., Kuan, J.X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S. and Scialom, T. (2023) Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288. doi: 10.48550/arXiv.2307.09288

Wang, Y.E., Wei, G.Y. and Brooks, D. (2019) Benchmarking TPU, GPU, and CPU platforms for deep learning. arXiv preprint arXiv:1907.10701. doi: 10.48550/arXiv.1907.10701

Zhao, W.X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., Liu, P., Nie, J., Wen, J. (2023) A survey of large language models. arXiv preprint arXiv:2303.18223. doi: 10.48550/arXiv.2303.18223