Journal of Korean Society of Industrial and Systems Engineering
https://doi.org/10.11627/jksie.2024.47.2.134

Vol. 47, No. 2 : 134-146, June 2024
ISSN : 2005-0461 / EISSN : 2287-7975

Instruction Fine-tuning and LoRA Combined Approach for Optimizing Large Language Models

Sang-Gook Kim† · Kyungran Noh · Hyuk Hahn · Boong Kee Choi

Korea Institute of Science and Technology Information

대규모 언어 모델의 최적화를 위한 지시형 미세 조정과 LoRA 결합 접근법

김상국† · 노경란 · 한 혁 · 최붕기

한국과학기술정보연구원

This study introduces and experimentally validates a novel approach that combines Instruction fine-tuning and Low-Rank Adaptation (LoRA) fine-tuning to optimize the performance of Large Language Models (LLMs). These models have become revolutionary tools in natural language processing, showing remarkable performance across diverse application areas. However, optimizing their performance for specific domains necessitates fine-tuning of the base models (FMs), which is often limited by challenges such as data complexity and resource costs. The proposed approach aims to overcome these limitations by enhancing the performance of LLMs, particularly in the analysis precision and efficiency of national Research and Development (R&D) data. The study provides theoretical foundations and technical implementations of Instruction fine-tuning and LoRA fine-tuning. Through rigorous experimental validation, it is demonstrated that the proposed method significantly improves the precision and efficiency of data analysis, outperforming traditional fine-tuning methods. This enhancement is not only beneficial for national R&D data but also suggests potential applicability in various other data-centric domains, such as medical data analysis, financial forecasting, and educational assessments. The findings highlight the method's broad utility and significant contribution to advancing data analysis techniques in specialized knowledge domains, offering new possibilities for leveraging LLMs in complex and resource-intensive tasks. This research underscores the transformative potential of combining Instruction fine-tuning with LoRA fine-tuning to achieve superior performance in diverse applications, paving the way for more efficient and effective utilization of LLMs in both academic and industrial settings.

Keywords : Large Language Model(LLM), Instruction Fine-tuning, Low-Rank Adaption(LoRA) Fine-tuning, National Research & Development(R&D) Data, Data Analysis Efficiency and Precision

1. 서 론

대규모 언어 모델(Large Language Models, LLMs)은 자연어 처리(NLP) 분야에서 혁신적인 도구로 자리잡고 있으며, 다양한 응용 분야에서 뛰어난 성능을 보여주고 있다. LLM은 방대한 양의 데이터로부터 유의미한 정보를 추출하고, 복잡한 언어적 과제를 해결하는 데 있어 중요한 역할을 하고 있다. 그러나 특정 도메인에서의 성능 최적화를 위해서는 기반 모델(Foundation Models, FMs)의 미세 조정

Received 10 June 2024; Finally Revised 19 June 2024;
Accepted 20 June 2024
† Corresponding Author : sgkim@kisti.re.kr

(fine-tuning)이 필수적이다.

기존의 미세 조정 방법론은 데이터의 복잡성과 다양성을 효과적으로 관리하는 데 한계를 보인다. 예를 들어, 특정 도메인에 특화된 데이터셋을 활용하는 과정에서 발생하는 데이터 부족 문제, 모델의 과적합, 그리고 계산 자원과 비용의 문제 등이 있다. 이에 본 연구에서는 사용자 지시 따르기 미세 조정(Instruction fine-tuning)과 저순위 적응 미세 조정(Low-Rank Adaptation, LoRA fine-tuning)을 결합한 새로운 통합 접근 방식을 제안한다. 이 접근 방식은 대규모 언어 모델의 성능을 최적화하고, 특히 국가 연구개발(R&D) 데이터의 분석 정밀도와 효율성을 향상시키는 것을 목표로 한다.

먼저, 본 연구에서는 LLMs의 실제 활용 사례와 성능 평가 및 개선 사례를 탐색한 시사점을 토대로, 특별 지식 도메인에서의 모델 성능 향상을 위한 해결 방법으로써 중요한 Instruction fine-tuning과 고사양·고비용 하드웨어 요구 자원을 회피하여 FMs 학습과 운영이 가능한 LoRA fine-tuning의 이론적 기초와 기술적 구현 방법을 설명한다.

Instruction fine-tuning은 모델이 특정 작업을 수행할 수 있도록 지시하는 방법을 통해 모델의 성능을 향상시키는 접근 방식이다. 이는 모델이 더 나은 문맥 이해와 정확한 응답 생성을 가능하게 한다. 반면, LoRA fine-tuning은 모델의 파라메터 수를 줄여 효율성을 높이는 방법으로, 고성능 하드웨어 자원 없이도 모델의 성능을 향상시킬 수 있다. 이 두 가지 방법을 결합함으로써, 우리는 LLM의 성능을 최대화하고, 데이터 분석의 효율성을 극대화할 수 있다.

본 연구에서는 국가 R&D 프로젝트 데이터를 대상으로 한 사례 연구를 통해, 제안된 미세 조정 기법들의 데이터 분석 정밀도와 효율성 향상 효과를 실험적으로 검증한다. 특히 해당 사례를 대상으로 한 이유는 특정 지식 도메인에서의 미세 조정 모델의 분석 정밀도를 탐색하기 위함이다. LLM을 사용하여 대규모 국가 R&D 데이터셋을 분석하고, 이를 통해 식별된 주요 연구 주제와 트렌드를 도출한다. 이 과정에서 Instruction fine-tuning을 통해 모델이 더 나은 문맥 이해와 정확한 응답 생성을 가능하게 하고, LoRA fine-tuning을 통해 모델의 효율성을 높인다. 실험 결과, 제안된 접근 방식이 데이터 분석의 정밀도와 효율성을 향상시키는 것으로 나타났다. 더 나아가, 이러한 통합 미세 조정 방법론이 국가 R&D 프로젝트 분석을 넘어 다른 데이터 집중형 도메인에도 적용 가능함을 시사한다. 예를 들어, 의료 데이터 분석, 금융 데이터 예측, 교육 데이터 평가 등 다양한 분야에서 본 방법론의 적용 가능성을 기대할 수 있다.

이 연구의 결과는 특정 분야에서 LLMs의 성능을 최적화하기 위한 새로운 방법론을 제시함으로써, 데이터 분석의 정밀도와 효율성을 크게 향상시킬 수 있음을 보여준다. 이를 통해, LLM을 활용한 데이터 분석의 새로운 가능성을 열고, 특별 지식 도메인 관련 분야에서 방법론적으로 크게 기여할 것으로 기대된다. 이 연구는 향후 다양한 도메인에서 LLMs의 적용을 확장하는 데 있어 중요한 참고 자료가 될 것이며, 정책 결정자와 연구자들에게 유용한 통찰을 제공할 것이다.

본 연구의 구성은 다음과 같다. 제 2장에서는 대규모 언어 모델의 실제 활용 사례와 향후 발전 방향, FM 기반 미세 조정 방법에 관한 연구 동향을 살펴본다. 제 3장에서는 본 연구에서 제안하는 특정 지식 도메인을 위한 생성형 인공지능 모델 개발에 필요한 사용자 지시 따르기 데이터셋 생성 방법, Instruction following fine-tuning 방법, 모델 경량화를 위한 LoRA 기반 파라메터 효율화 미세조정 방법(Paremter Efficient Fine-Tuning, PEFT) 방법을 소개하고 이를 최적화하는 프레임워크를 설명한다. 제 4장에서는 제안 방법의 실제 적용 사례를 보이기 위해 미국의 NASA 연구개발 문헌을 활용하여 미세 조정된 FM의 성능 평가를 수행하기 위해, 기본 FM과 미세 조정된 FM 간의 응답 성능을 비교 분석한 결과를 소개한다. 또한 정성 및 정량적 측면에서의 주요 발견 사실을 제시하고 미세 조정된 FM의 유의성에 대해 설명한다. 마지막으로 제 5장에서는 사용자 지시 따르기 미세 조정의 중요한 영향력과 PEFT 방법에서의 LoRA의 중요한 영향력, 그리고 본 연구의 한계와 향후 추가적인 연구의 필요성을 논의하고 결론을 맺는다.

2. 선행연구

2.1 대규모 언어 모델(LLMs)의 실제 활용

대규모 언어 모델(LLMs)은 다양한 분야에서 인공지능의 성능을 향상시키며 중요한 역할을 수행하고 있다. 본 절에서는 여러 연구들을 통해 LLMs의 실제 활용 사례와 그 영향을 분석하였다.

2.1.1 사회적 영향 및 윤리적 문제

Sætra[20]는 생성형 인공지능(Generative AI)이 현대 사회에 다양한 형태로 큰 영향을 미치고 있으며, 그 영향력의 범위와 잠재적 위험성에 대한 논의가 필요하다고 주장하였다. 이 연구에서는 Generative AI가 사회에 미치는 영향을 면밀히 검토할 필요가 있다고 지적하였다. 기술 변화가 사회 변화와 불가분의 관계에 있으므로, Generative AI가 사회적, 문화적, 정치적 측면에서 부정적인 영향을 초래할 가능성을 염두에 두어야 한다. Generative AI의 접근성과 활용도가 개인, 집단, 국가 간에 차이가 발생할 수

있으며, 이로 인해 불평등이 심화될 수 있다. 또한, 이 기술은 기존 권력 구조에 변화를 가져올 수 있어 사회적 갈등이 발생할 가능성도 있다고 지적하였다. 이외에도, 생성형 인공지능을 통해 생성된 콘텐츠의 진실성, 투명성, 책임성 등 윤리적 문제도 대두될 수 있음을 언급하였다.

2.1.2 통계적 공정관리(SPC)와 교육 분야 활용

Megahed et al.[15]의 연구에서는 생성형 인공지능 모델이 SPC 실무, 교육, 연구 분야에서 활용되고 있으며, 이에 따른 문제점도 존재한다고 시사하였다. ChatGPT는 SPC 관련 코드 번역, 기본 개념 설명 등의 구조화된 작업에서 효과적이지만, 덜 알려진 용어 설명이나 새로운 코드 생성과 같은 복잡한 작업에서는 한계가 있었다. 생성형 인공지능은 SPC의 기본 개념 설명, 예제 코드 생성 등에서 교육적 활용 가치가 있으며, 연구자들의 생산성을 높이는 데 기여하고 있다. 그러나 모델의 오류 및 편향으로 인해 연구 결과의 정확성이 저하될 수 있어 다른 방법과 병행 사용이 필요하다고 지적하였다.

2.1.3 과학 및 공학 분야에서의 활용

Kamnis[7]은 GPT 모델이 과학 분야에서 데이터 범위, 최신성, 복잡성, 유료 콘텐츠 접근성 등의 이유로 지식이 제한적일 수 있다고 지적하였다. 특히, 표면 공학과 같은 특정 과학 분야에 특화된 GPT 모델을 개발하면 성능을 향상시킬 수 있으며, 전문 용어 이해, 최신 연구 정보 제공, 모호성 및 오류 감소, 맞춤형 작업 수행 등의 이점을 제공할 수 있다고 하였다. 표면 공학 분야 논문을 참고하여 개발된 맞춤형 GPT 모델은 일반 GPT 모델보다 우수한 성능을 보였으며, 이를 통해 특정 기술 지식 분야에서 GPT 모델의 활용 가치를 확인하였다.

2.1.4 혁신 관리와 디지털 프로토타이핑

Bilgram and Laarmann[1]은 생성형 인공지능이 혁신 관리 분야에서 인공지능 활용의 대중화를 유도하고 있다고 주장하며, 이 기술이 탐색, 아이디어 생성, 디지털 프로토타이핑 등 혁신의 초기 단계에서 창의성을 높일 수 있을 것으로 기대하였다. 특히 GPT와 같은 LLMs는 디지털 프로토타이핑 과정을 가속화하고, 반복 작업 속도를 높이며, 비용을 절감할 수 있을 것으로 기대하였다. 생성형 인공지능은 혁신팀과의 협업 과정에서 워크플로우 통합에 기여할 수 있으며, 창의성 향상, 프로세스 가속화, 비용 절감 등에 기여할 수 있다.

2.1.5 의료 및 정신 건강 분야

Mazumdar et al.[14]은 GPT-3의 임베딩과 미세 조정을 활용하여 정신 건강 장애를 감지하고 설명하는 새로운 프레임워크를 제안하였다. GPT-3는 자연어 처리에서 뛰어난 성능을 보여주며, 제안된 GPTFX 프레임워크는 정신 건강 장애를 분류하는 데 약 87%의 정확도를 보였고, 예측된 결과에 대한 설명 생성에서 Rouge-L 점수가 약 0.75로 높은 성능을 보였다. GPT 임베딩과 기계 학습 모델을 통합하여 정신 건강 장애를 분류하는 접근법은 기존의 정신 건강 감지 알고리즘보다 높은 신뢰성과 정확성을 제공할 수 있다.

2.1.6 생의학 분야의 언어 모델

Karkera et al.[8] 연구는 여러 사전 학습된 대규모 언어 모델(GPT-3, BioGPT, BioMedLM, BERT, BioMegatron, PubMedBERT, BioClinicalBERT, BioLinkBERT)을 사용하여 미생물-질병 관계를 추출하는 데 초점을 맞추고 있다. 도메인 특화된 데이터로 미세 조정된 언어 모델들이 뛰어난 성능을 보여 전이 학습(transfer learning)의 효과를 입증하였으며, 특히 생의학 분야에서 NLP와 대규모 언어 모델의 적용 가능성을 시사하였다. 연구에서는 제로샷(zero-shot)과 소수샷(few-shot) 학습 관점에서 초기 평가 결과, 모델들이 도메인 특화된 데이터가 필요함을 보여주었으며, 이는 미세 조정의 필요성을 강조하였다.

2.1.7 기반 모델 (FMs)과 사전 학습 모델 (PLMs)의 응용

Kolides et al.[11]은 BERT, T5, GPT-3, Codex, DALL-E, Whisper, CLIP과 같은 FMs와 PLMs의 기본 원리, 응용 분야, 기회, 사회적 영향을 종합적으로 분석하며, 다학제적 협력의 필요성을 강조하였다. 연구는 FMs가 AI 분야에서 새로운 패러다임을 제시하며 다양한 응용 분야에서 혁신을 이끌 것이라 기대하였다. FMs는 컴퓨터 비전, 단백질 서열 연구, 음성 인식, 코딩 등 다양한 분야에서 높은 성능을 발휘하며 혁신적인 결과를 도출하고 있다. 그러나 이러한 모델의 광범위한 사용으로 인한 사회적 파급효과와 문제점들에 대해서도 지적하였다. 긍정적인 영향이 기대되지만, 오해와 잘못된 정보의 확산, 개인정보 침해, 지적재산권 문제 등 윤리적 문제의 발생 가능성도 존재하므로 적절한 정책과 규제의 필요성이 대두되고 있다.

2.1.8 로봇 조작 작업의 언어 기반 계획

Lin et al.[13]은 로봇이 자연어 지시를 기반으로 복잡한 조작 작업을 수행할 수 있도록 돕기 위해 설계된 Text2Motion에 관한 연구를 진행하였다. Text2Motion 프레임워크는 주어진 자연어 지시를 기반으로 작업 수준과 동작 수준의 계획을 모두 구성한다. 이를 통해 기존의 언어 기반 계획 방법이 직면한 한계를 극복하고자 하였다. Text2Motion의 성능을 평가하기 위해 다양한 문제 집합을 사용하여 실험을 수행한 결과, 82%의 성공률을 기록했으

며 이는 이전의 최첨단 언어 기반 계획 방법이 달성한 13%에 비해 상당히 높은 수치를 나타낸다. Text2Motion이 기하학적 종속성이 있는 다양한 순차적 조작 작업에 대해 유망한 일반화 특성을 보이며, 로봇 조작 작업의 효율성과 정확성을 크게 향상시킬 수 있음을 입증하였다.

2.1.9 작업 시퀀스 생성을 위한 ProgPrompt 프레임워크

Singh et al.[19]는 로봇 작업 계획에서 필요한 방대한 도메인 지식을 정의하는 노력을 줄이기 위해 LLMs를 활용하는 방법을 탐색하였다. 이 연구는 자연어 지시를 바탕으로 가능한 다음 작업을 평가하고, 직접적으로 작업 시퀀스를 생성할 수 있는 프로그램 구조인 ProgPrompt 프레임워크를 제안하였다. 이를 통해 로봇이 실제 환경에서 수행해야 할 작업 계획을 효율적으로 수립할 수 있도록 돕는다. ProgPrompt는 다양한 작업 시나리오에서 높은 성공률과 정확도를 나타내었으며, 특히 인간-로봇 상호작용에서 효율적인 협업을 가능하게 하는 잠재력을 지니고 있다.

2.2 LLMs의 향후 발전 방향

LLMs는 다양한 응용 분야에서 그 성능을 평가받고 있으며, 여러 연구를 통해 개선되고 있다. 본 절에서는 LLMS의 성능 평가 및 개선과 관련하여 향후 발전 방향을 탐색하였다.

2.2.1 도메인 특화된 대규모 언어 모델의 개발

Kamnis[7]의 연구는 특정 과학 분야에 특화된 대규모 언어 모델의 개발이 필요함을 강조하였다. 예를 들어, 표면 공학과 같은 특정 과학 분야에 특화된 GPT 모델을 개발하면 성능을 향상시킬 수 있으며, 이를 통해 전문 용어 이해, 최신 연구 정보 제공, 모호성 및 오류 감소, 맞춤형 작업 수행 등의 이점을 제공할 수 있다. 이러한 도메인 특화 모델은 일반 모델보다 우수한 성능을 보일 것으로 기대된다.

2.2.2 다양한 분야에서의 GPT 모델 활용

Yin et al.[21] 연구에서는 대규모 언어 모델이 주로 의료, 금융, 법률 분야에 집중되었지만, 전력 에너지와 같은 중요 도메인에 대한 연구가 상대적으로 부족하다고 지적하였다. 이 연구에서는 LLaMA 아키텍처를 기반으로 중국의 전력 분야 도메인 지식을 활용하여 PowerPulse 모델을 개발하였고, 이 모델이 텍스트 생성, 요약 추출, 주제 분류 등의 작업에서 뛰어난 성능을 나타냈음을 강조하였다. 특정 도메인에서의 정확성과 신뢰성을 높이기 위해 관련 사전 학습 데이터와 전력 에너지 도메인에 맞춘 교육 데이터셋을 활용하여 모델을 미세 조정하였다.

2.2.3 AI 모델의 윤리적 문제 해결

Kolides et al.[11]의 연구는 대규모 언어 모델의 윤리적 문제를 해결하기 위한 다학제적 협력의 필요성을 강조하였다. AI 모델의 광범위한 사용으로 인한 사회적 파급효과와 문제점들을 해결하기 위해 적절한 정책과 규제가 필요하다. 긍정적인 영향이 기대되지만, 오해와 잘못된 정보의 확산, 개인정보 침해, 지적재산권 문제 등 윤리적 문제의 발생 가능성도 존재하므로 이에 대한 대비가 필요하다.

이와 같이 대규모 언어 모델의 발전과 함께 다양한 분야에서의 활용 가능성이 확대되고 있으며, 이를 통해 새로운 혁신이 기대된다. 그러나 동시에 윤리적 문제와 사회적 영향을 고려한 신중한 접근이 필요하다. 이러한 측면에서 다학제적 협력과 적절한 정책 마련이 요구되고 있다.

2.3 FM 기반 미세 조정

기반 모델(FMs)은 일반적으로 라벨링 되지 않은 대규모 데이터를 자기 지도 방식으로 학습한 거대 인공지능 모델을 말한다. 광범위한 데이터를 대상으로 대규모 사전 학습을 수행한 모델로, 질문에 답하기, 그림 그리기, 글쓰기, 번역하기 등 다양한 서비스를 제공할 수 있다. 특허 생성형 인공지능 부문에서는 PLMs로 불리기도 하며, 다양한 용도의 임무에 맞추어 미세 조정 또는 맥락 내 학습(in-context learning) 후에 바로 사용이 가능하다. 이 같은 FMs는 학습에 막대한 자원이 소요되며, 주로 인터넷에 공개된 정보를 활용하여 학습되었기 때문에, 특정 지식 도메인에 따라 여전히 학습해야 요소가 많다는 것을 의미한다.

openAI사에서 출시한 대표적인 FM인 ChatGPT 이후에, Meta, 스탠포드 대학, UC Berkeley 등에 의해 다양한 LLM 파생 모델들이 소개되었으며, 다양한 국가의 언어를 반영할 수 있는 LLM들이 지속적으로 소개되고 있다. 이때 모델의 크기를 7B, 13B, 33B, 65B, 70B 등 다양한 규모로 학습된 모델들을 배포하고 있으며, 이것은 완전한 순위(full-rank)와 모든 파라메터(all the parameters)를 갱신하는 완전한 미세 조정(Full fine-tuning) 방법이 아닌 파라메터 효율화 미세조정 방법(PEFT)을 주로 활용하고 있다.

언어 처리에 대해 특화된 FMs는 목적성이 없기때문에 특화한 지식 및 답변 세트에 맞춰 미세 조정하거나, 실제 데이터 등은 외부 검색엔진 및 데이터베이스를 참조하도록 중간에 코드를 삽입하고, 이외에도 저순위 적응(LoRA) 모델을 생성하여 이미 구축된 기반 모델에 적용이 가능하다. 결과적으로 특정 지식 도메인에 특화된 서비스형 LLMs를 개발하기 위해서는 FMs와 미세 조정 방법을 기본적으로 적용해야 할 필요성이 존재한다.

Hu et al.[6] 연구에서는 구글이나 네이버와 같은 거대 자본이 부족한 일반 기업과 연구소에서도 보다 경량화된

형태로 미세 조정을 할 수 있는 방법인 LoRA를 소개하였다. LoRA의 기본적인 개념은 기존 PLM의 가중치는 고정시키고 몇 개의 밀집 또는 완전 연결 층(dense or fully connected layer)만을 학습시켜 다운스트림 태스크의 연산량을 획기적으로 줄일 수 있는 방법을 개발하였다. 이 연구에서는 두 가지 문제에 대한 고민을 통해 해결 방법을 탐색하였으며, 첫 번째 문제는 "모든 매개변수를 찾아 조정해야 만 하는가" 였다. 두 번째 문제는 "미세 조정하는 가중치 행렬의 경우에 행렬 순위 측면에서 갱신이 얼마나 표현되어야 하는가?" 하는 문제였다.

결과적으로 이 연구에서는 GPT-3의 checkpoints 크기를 1TB(1,750억 개 매개변수)에서 25MB(470만 개 매개변수) 수준으로 획기적인 감소를 실현하였으며, 기존 매개변수에 갱신 내용을 추가하여 FM과 문자 그대로 동일한 방식으로 추론 수행할 수 있도록 하여, 추가 추론 지연시간을 도입하지 않고 모델을 빠르게 전환할 수 있도록 하였다. 이외에도 RAM에 LoRA 모듈을 캐싱할 수 있도록 하였으며, 하나의 배치(batch)에 상이한 LoRA 모듈들을 병렬로 학습이 가능하도록 하였다. 마지막으로 LoRA 모듈의 계층구조를 활용하여 다양한 데이터셋 크기를 수용하기 위해서 상위 구조 뿌리(root) 근처에서 더 크고, 하위 구조 잎(leaf) 근처에서 더 작을 수 있도록 하여, 모델 전환은 트리 순회가 가능하도록 하였으며, FM을 두 번 이상 로딩할 필요가 없도록 하였다.

3. 연구 방법론

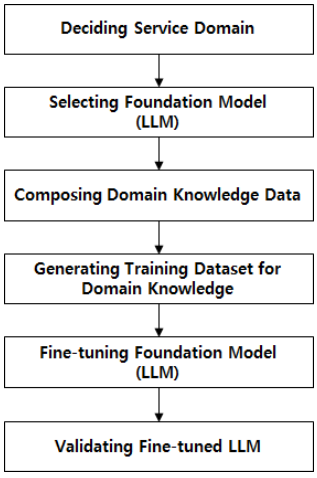
본 연구에서는 국가 연구개발 사업과 관련된 지식 도메인 중 미국 NASA 연구개발 사업 정보를 토대로, 생성형 인공지능 서비스 모델을 개발하기 위해서 해당 도메인의 학습 데이터를 생성하고, 한국어 지원이 가능한 FM을 적용하여 LoRA 미세 조정 방법을 활용하여 최적화하는 새로운 서비스 모델 개발 프레워임워크를 제시하고자 하였다.

3.1 연구방법론 설계

미우주항공국(NASA) 연구개발 사업 도메인에 특화된 생성형 인공지능 모델을 개발하기 위하여 6단계의 절차를 구성하였다. 첫 번째 단계는 서비스 도메인을 결정하는 단계로, 과학기술 문헌 정보를 선택하는 단계이다. 두 번째는 FM을 선택하는 단계이며 향후 서비스 모델에서 제공가능한 언어를 고려하여 해당 언어 서비스가 원할한 FM을 선택하는 것이 중요하다. 세 번째는 해당 지식 도메인의 데이터를 구성하는 단계로, 원시 텍스트 데이터를 확보하는 단계이다. 네 번째는 도메인 지식을 적용하기 위한 학습 데이터 생성 단계로, 사용자 지시 따르기 학습 데이터를 생성하기 위한 핵심적인 단계이다. 다섯 번째 단계는 앞서 선택한 FM을 해당 도메인에 특화한 사용자 지시 따르기 학습 데이터를 적용하여 FM을 학습시키는 단계이며, 본 연구에서는 파라메터 효율화 방법으로 LoRA 방법을 적용하는 단계이다. 마지막으로, 특정 도메인에 특화하여 미세 조정한 FM의 모델 신뢰성을 평가하는 단계이며, 이때 기존 openAI사의 GPT-4.0을 활용하여 질의에 대한 생성 답변과 해당 도메인 원시자료로부터 확보된 실제 답변 간의 유사성과 차이점을 비교 평가하는 단계이다.

3.2 데이터 구성과 사전 처리

미국의 SBIR(Small Business Innovation Research)과 STTR(Small Business Technology Transfer)은 중소기업프로그램으로 통칭되며, 미국의 시드 펀드(America's Seed Fund)라고도 불리운다. SBIR은 미국 내 중소기업이 상용화 가능성이 있는 Federal Research/Research & Development에 참여하도록 장려하는 프로그램이며, STTR은 중소기업과 비영리 연구기관 간의 파트너십으로 중소기업이 1단계 및 2단계에서 연구기관과 공식적으로 협력할 것을 요구하며, 기초 과학의 성과와 그에 따른 혁신의 상용화 사이의 격차를 해소하기 위한 목적이 있다.



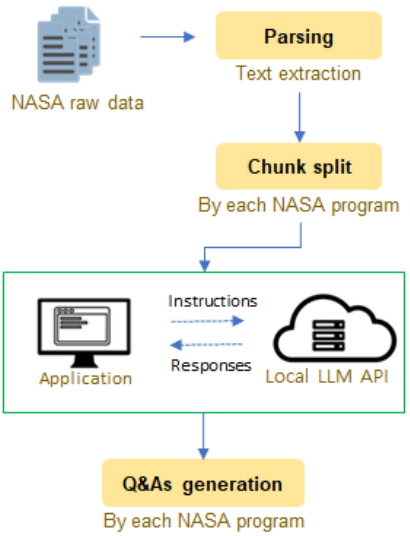
<Figure 1> Research Framework

따라서 본 연구를 위해서 상기 2개의 프로그램을 통해 수행되어 왔던 연구개발 프로젝트 중에서, NASA에서 수행한 프로젝트를 대상으로 온라인 검색을 통해 53,255건의 데이터를 구성하였으며, 데이터 무결성(integrity)과 정합성(consistency)을 위해 USA Spending, NASA Federal Reporter, NSSC(NASA Shared Services Center)의 온라인 검색자료와 비교 검토를 수행하였다. 결과적으로 FM 학

습에 필요한 대상 데이터는 Program의Title, Abstract, NASA Application fields, Keyword, Program Type, Phase status, Firm Name을 대상으로 하였다.

3.3 FM의 미세 조정 절차

특화된 지식 도메인의 서비스 생성 모델 개발을 위해서, 지도 학습 기반의 미세 조정 방법을 선택하면서 사용자 지시 따르기 데이터셋 구성(instruction following dataset)을 위한 방법을 새롭게 제시하였다. 기본적으로 특화된 지식 도메인에서 사용자 지시 따르기 데이터 생성에는 많은 자원과 높은 비용, 그리고 시간이 소요된다. 모델의 정확도를 개선하기 위해서는 필수적인 요소이지만 현실적으로 많은 제약이 따르는 과정으로, 이 부분에 대한 현실적 해결 방안이 필요하다.



<Figure 2> Instruction Generation for NASA Knowledge Domain

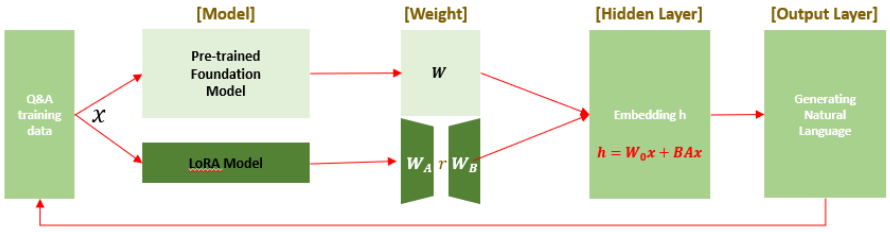
따라서 본 연구에서는 비용이 소요되지 않은 기존 FM에서 제공하는 API와 특화된 지식 도메인의 원시자료를 토대로 사용자 지시 따르기 데이터셋 생성을 위한 방법을 개발하였다. 이때 기존 FM은 HuggingFace Github에서 제공하고 있는 "davidkim205_komt-llama2-13b-v1" 모델을 사용하였고, 기존 FM에게 제공할 사용자 지시(instruction)는 "You are an API that converts text provided by a user role a single question and answer in perfect JSON format"와 같이 정의하고, 정의된 instruction과 함께 API에게 제공할 질의 input은 원시데이터로부터 순차적으로 자동 생성된 각각의 질의와 답변을 활용한다.

본 연구에서 제안하여 생성한 사용자 지시 따르기 데이터셋은 총 21,543건이 생성되었으며, NASA 연구개발 사업 문헌 정보당 평균 2.5건의 사용자 지시 따르기 데이터가 생성된 것을 의미한다.

사용자 지시 따르기 데이터셋 생성에 사용된 하드웨어 사양은 Dell Precision 7920 서버 기종에 Intel 제온 골드 6242R 3.1GHz CPU 2개, 256GB DDR4 3200MHz 램메모리, Nividia RTX A6000 D6 48G GPU 3개를 운영하여 총 5일이 소요되었다.

3.4 LoRA 기반 PEFT 절차

LoRA 기반 PEFT는 3단계인 초기 파라메터 상태 설정(Pre-trained weight) 단계, 파라메터 가중치 갱신(Update Weight) 단계, 파라메터 갱신 가중치 적용(Adapted Weight) 단계를 통해 진행된다. <Figure 3>에 초기 상태 설정 단계에서는 미리 학습된 가중치 $W \in R^{d \times d}$를 사용하며 입력 x는 d차원 벡터이다. 가중치 W와 입력 x를 곱한 후에 결과를 다음 층으로 전달한다. 이 과정에서 일부 임의의 매개변수 A와 B가 사용되며, B는 0으로 설정되고, A는 평균 0, 분산 σ^2을 가지는 정규분포에서 표본 추출된다. 두 번째 단계에서는 학습 과정에서 가중치 갱신이 된다. 기존의 가중치 W와 갱신된 가중치 ΔW가 결합된다. 이때 입력 x는 기존과 동일하게 d차원 벡터이며, 갱신된 가중치 ΔW는 동일한 차원 $d \times d$를 갖게되며, 입력과 결합하여 새로운 출력을 생성하게 된다. 마지막 단계에서는 업데이트된 가중치가 적용되어 새로운 가중치 $W \in R^{d \times d}$로 변경된다. 이 단계에서 적응된 가중치는 새로운 입력 x와 결합되어 최종 출력을 생성한다. 결과적으로, 갱신된 가중치와 입력의 조합을 통해 모델의 성능이 향상된다.



<Figure 3> The Procedure of LoRA

앞서 생성된 사용자 지시 따르기 데이터셋과 상기 LoRA 기법을 적용하여 파라메터 효율화를 수행하였으며, 이때 적용한 Rank 수는 32, Alpha 값은 64, Batch size 128, Micro batch size 4, Cutoff length 256, Epoch 3, Learning rate 3e-4, LR Scheduler는 Linear, LoRA Dropout값은 0.05를 적용하였다.

이때 앞서 설명한 동일 하드웨어 사양에서 학습을 위해 총 6시간이 소요되었으며, 최종 Epoch 수는 1.7에서 학습이 종료되어 최적화가 완료되었다.

4. 주요 실험 결과

4.1 생성형 인공지능 모델 평가 방법

본 연구에서는 FM과 NASA program의 사용자 지시 따르기 데이터셋으로 부터 미세 조정된 모델이 해당 분야에서의 기술내용 질의에 대한 답변을 원활하게 잘 생성하는지 평가하기 위해 다음과 같은 실험을 진행하였다. 먼저 10개의 질의를 랜덤으로 추출하고, 이를 토대로 기본 FM과 미세 조정된 FM이 생성하는 답변을 상호비교 평가하였다. 이때 적용한 평가 방식은 두 모델에 의한 유사점과 차별점에 대한 비교평가를 수행하고, NASA Program의 원시 자료와 최선으로 매칭되는 답변을 생성한 후에, 이를 근거로 기본 FM과 미세 조정된 FM의 생성 답변들에 대해서 1) 매칭 모델 선택, 2) 매칭 점수 부여, 3) 매칭 이유에 대한 평가를 GPT-4에게 질의하여 진행하였다. 이때 사용한 매칭 점수는 5점 척도(5: 아주 강한 매칭, 4: 적절한 매칭, 3: 보통 매칭, 2: 약간 약한 매칭, 1: 아주 약한 매칭)를 사용하였다.

4.2 미세 조정된 FM의 성능 평가

4.2.1 질의-답변 임의 추출

앞서 생성된 사용자 지시 따르기 데이터셋으로 부터 10개 질의를 임의 추출한 후에, 해당 질의에 최선으로 매칭되는 답변을 생성하여 정답지를 구성하였다. 그 결과는 아래 <Table 1>과 같이 구성되었다.

<Table 1> Randomly Sampled 10 Questions and the Best Matched Answers Based on NASA Program Raw Data

No.	Question	Best matched answer
1	What is the abstract of the project related to HighStrainComposite Deployable Radiators for CubeSats?	The abstract of the project related to HighStrainComposite Deployable Radiators for CubeSats is In response to NASA's need for compact, lightweight, and efficient low-cost deployable radiators for CubeSats, Roccorproposes to develop a high-strain laminate composite deployable radiator capable of dissipating over 50W
2	What is the chemical detection capability of the LIDAR system?	The chemical detection capability of the LIDAR system can detect one molecule thick layers adsorbed on surfaces, allowing detection of leaks in pressurized or vacuum conditions with a range of 30m.
3	How will Mainstream regenerate the collection electrodes in the carbon removal system?	Collection electrodes can be regenerated in place through gas reaction or removed and scraped for safe storage and reuse or disposal, retaining all carbon particles.
4	What are the potential NASA applications of the Universal Sample Preparation Module?	The potential NASA applications of the Universal Sample Preparation Module include microbial detection in advanced life support systems, invitro testing for infectious diseases during space flight, and gene expression analysis for radiation exposure monitoring.
5	What is the potential use of ITR bonding for NASA's deep space missions?	ITR bonding can facilitate the construction of habitable architectures and reduce buckling, leaking, and thermal effects in space. It can also enrich solar electric propulsion project requirements by offering readjustablesolar array configurations.
6	What is the purpose of the Automated Multiple Object Optical Tracking and Recognition System?	The purpose of the Automated Multiple Object Optical Tracking and Recognition System is to develop an optical tracking system capable of recognizing and tracking up to 50 different objects within an approximately 2-degree x 3-degree field of view. The system output will be used to determine the tracking performance of the system.
7	What are the potential NASA applications of the HeatAssistedCutter HAC and accompanying Sierra Lobo IceCarvingExcavator for Regolith SLICER?	The potential NASA applications of the HeatAssistedCutter HAC and accompanying Sierra Lobo IceCarvingExcavator for Regolith SLICER include harvesting water ice from permanently shadowed regions on the Moon as a source of water for human consumption or as a rocket propellant, and excavation of ice on Mars.
8	What are the potential non-NASA applications of the receiver technology developed by Virginia Diodes Inc.?	The potential non-NASA applications of the receiver technology developed by Virginia Diodes Inc. include the realization of highly compact full waveguidebandreceivers with excellent sensitivity and no mechanical tuners. These integrated components are easy to manufacture and extremely reliable, making commercial applications of terahertz technology possible. The initial market will consist mostly of scientists and engineers developing terahertz components and systems and manufacturers of research and test equipment. As costs are reduced, high data rate point-to-point communication systems, medical diagnostic tools, and methods are envisioned.
9	What is the objective of comparing the system mass and energy efficiency of the novel plasma technique for oxygen extraction from Mars' atmosphere?	The objective is to compare the system mass and energy efficiency of the novel plasma technique with those for existing systems.
10	How does the dual plenopticdesign of the PlenopticAttitude Monitoring System work?	One plenopticimager analyzes and corrects for turbulence effects introduced by airflow around the model, while the second imager extracts the attitude information from the model itself

4.2.2 미세 조정된 FM에 의한 답변 생성

앞서 제시한 방법론적인 절차에 따라, 기존 FM인 “davidkim205_komt-llama2-13b-v1”와 사용자 지시 따르기 데이터셋에 의한 미세 조정 및 LoRA peft를 통해 최종적으로 미세 조정된 FM을 토대로 <Table 2> 형식의 프롬프트를 상기 10개 질의를 제공하여 답변을 생성하도록 하였다.

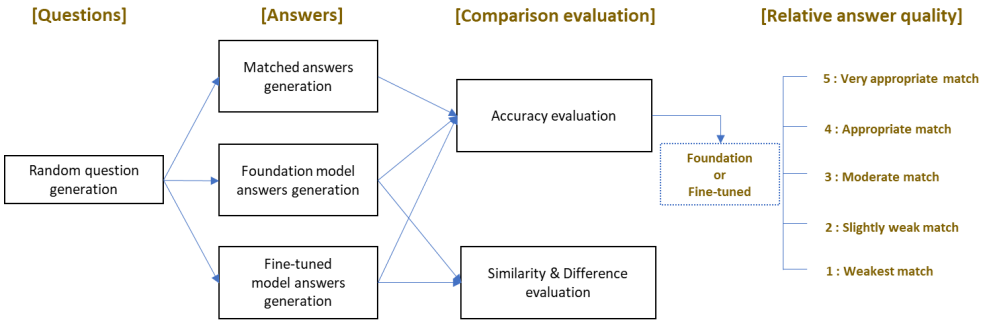
<Table 2> A Prompt Provided to the Fine-tuned Model

```
Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

### Instruction:
You are a national R&D planning expert and you will answer questions related to R&D planning inquiries.

### Input:
What is the abstract of the project related to HighStrainComposite Deployable Radiators for CubeSats?

### Response:
```

상기 절차에 따라 미세 조정된 FM에서 10개의 질에 따라 답변들이 생성되었으며, 생성 답변 중 9개의 답변(7번 질의를 제외한)이 기본 FM 대비 2배 이상의 월등히 많은 내용을 생성하고 있는 것으로 나타났다.

4.2.3 각 질의에 대한 유사점과 차이점 분석

생성형 인공지능 모델의 성능 평가에는 기본적으로 다양성이 존재한다. MMLU, Vicuna, OA 벤치마크 평가 등 다양한 검증 방식이 존재한다. 하지만 이 같은 평가 방식의 결과가 BigBench, RAFT, HELM 등 다른 벤치마크 평가에서의 일반화 가능성으로 연결되지는 않는다. 이 같은 사실은 기존 벤치마크 테스트의 타당성에 의문을 제기하게 된다. 벤치마크가 그 이름이나 설명이 제안하는 것을 정말로 테스트 하는지에 대해서는 항상 의문을 갖게 되는 게 사실이다. 모델이 벤치마크를 해결하기 위해 이용하는 ‘지름길’ 발견에 따른 타당성 문제는 생성형 인공지능 모



<Figure 4> Evaluation Procedure between Basic FM and fine-tuned FM

<Table 3> Analysis of Similarities and Differences between FM and the Fine-tuned Model

Category	Base Model	Fine-tuned Model
Objectives and Roles	Contribution Explanation: Describes how the technology can contribute to research or industry.	Contribution Explanation: Describes how the technology can contribute to research or industry.
	Both models describe how the technology can contribute to research or industry. Technology Description: Explains the functionality and potential of the technology.	
Importance of Research	Emphasis on Applicability: Highlights the broad research applicability of the technology.	Emphasis on Specific Applications: Explains how the technology can be applied in specific fields.
	Both models emphasize the applicability of the technology and suggest potential uses in various fields. NASA Mission Contribution: Both models mention the technology's potential contribution to NASA missions.	
Information Provision	Technical Information: Broadly explains the overall functionality and theoretical potential of the technology.	Detailed Information: Discusses the development process, commercialization potential, and changes in specific industries in detail.
	Both models explain the technical details and applications of the technology. Analysis and Results: Each model aims to demonstrate the value of the technology through analysis and results.	
Research and Development Direction	Technology Potential: Broadly covers the potential applications in various industries.	Specific Applications: Emphasizes the potential applications of the technology in specific industries. Provides specific examples, such as the use of LIDAR for chemical detection.
Commercial and Industrial Impact	Broad Application: Discusses how the technology can be utilized in various industries.	Specific Benefits: More specifically discusses the concrete benefits and commercial potential of the technology in certain industries.
Contribution to NASA Missions	Broad Contribution: Explains how the technology can contribute to NASA missions broadly.	Specific Mission Contribution: Explains the specific potential contributions of the technology to particular NASA missions.

델을 평가하는데 있어서 기본적으로 경험하게 되는 쟁점이다. 따라서 이와 같은 문제를 해결하기 위한 수단으로 정성적 분석이 필요하며, 정량적 분석만으로는 포착하지 못하는 특정 패턴과 예시를 설명하는 것이 필요하다. 따라서 본 연구에서는 미세 조정 모델의 성능을 평가하기 위한 방법으로 GPT-4를 이용한 정성적 평가를 수행하였다.

기본 FM과 미세 조정된 FM에 의해 생성된 답변들에 대해서 분석된 결과, 연구개발 목표와 역할, 연구의 중요성, 기술적 정보 및 분석 결과 제공 측면에서는 두 모델이 유사한 결과를 생성하는 것으로 나타났다. 하지만 내용의 깊이와 범위, 연구개발 방향, 상업적 및 산업적 영향, NASA 임무에 대한 기여도 측면에서는 두 모델이 상이한 결과를 생성하고 있는 것으로 나타났다. 특히 미세 조정된 모델이 연구개발 방향 측면에서 구체적인 응용 사례와 기술 적용 가능성을 강조하였고, 상업적 및 산업적 영향 측면에서 특정 산업에서 기술이 가져올 수 있는 구체적 이점과 상업적 잠재력을 제공하는 것으로 나타났다. 또한 NASA 임무에 대한 기여도 측면에서는 특정 NASA 임무에 대한 기술의 구체적인 기여 가능성을 제공하였다 (<Table 3> 내용 참조).

4.2.4 미세 조정된 FM의 답변 생성 정확도 평가

<Table 4> Evaluation Prompt Instruction

Below are the matched answers deemed relatively accurate for each question, along with the answers generated by the foundation model and the fine-tuned model. For each question, please: 1.Select the model (Foundation or Fine-tune) that generated the answer closest to the matched answer, or indicate "Nothing" if neither model's answer matches. 2.Rate the matching on a 5-point scale (1: weakest match, 2: somewhat weak match, 3: average match, 4: appropriate match, 5: very appropriate match). 3.Briefly explain the reason for your rating.

<Figure 4> 절차에 따라 두 모델 간의 답변 생성 정확도 평가를 수행하였으며, 두 모델에 의한 생성 답변들에 대해서 GPT-4를 활용하여 평가를 수행하였다. GPT 프롬프트 형식은 <Table 4>와 같이 구성하였고, 각 질문과 생성 답변들에 대한 정보를 추가로 제공하여 평가 결과를 생성하도록 하였으며 그 결과는 <Table 7>과 같이 나타났다.

두 모델에 대한 GPT-4 평가 결과, 기본 FM의 평균값이 2.2점, 표준오차 0.3590, 표준편차 1.1353 인 것에 비해 미세 조정된 FM의 평균값은 2.6점, 표준오차 0.3055, 표준편차 0.9661로 분석되었다.

이외에도 두 모델에 대한 평가 점수를 대상으로 t-검정과 F-검정을 수행하였다(<Table 5>와 <Table 6> 참조 가능). t-통계량이 양측 검정의 t-citical 값(±2.262157163)보다 작기 때문에, 이를 통해 두 모델 평균 간의 차이가 통계적으로 유의하지 않음을 확인할 수 있다. 추가적으로 F-검정 통계량이 한쪽 꼬리 검정에서의 F-critical 값인 3.178893104 보다 작아, 이는 두 모델의 분산이 통계적으로 유의한 차이가 있다고 볼 수 없음을 확인하였다. 결과적으로, 두 모델 간의 평균값에 차이가 존재하지 않고, 두 모델 간의 분산 역시 서로 다르지 않아 비슷한 일관성을 갖고 있다고 할 수 있다. 하지만, 이와 같은 분석 결과는 통계적 표본 수가 많지 않아 일반화하기에 한계가 있다.

<Table 5> t-test Results between Two Models

Categories	FM	Fine-tuned model
Mean	2.2	2.6
Variance	1.288888889	0.933333333
Observations	10	10
Pearson Correlation	0.08104409	-
Hypothesized Mean Difference	0	
df	9	
t Stat	-0.884651737	
P(T<=t) one-tail	0.19968068	
t Critical one-tail	1.833112933	
P(T<=t) two-tail	0.399361359	
t Critical two-tail	2.262157163	

<Table 6> F-test Resutls between Two Models

Categories	FM	Fine-tuned model
Mean	2.2	2.6
Variance	1.288888889	0.933333333
Observations	10	10
df	9	9
F	1.380952381	-
P(F<=f) one-tail	0.319200279	
F Critical one-tail	3.178893104	

4.3 주요 발견

이 연구에서는 기본 FM과 미세 조정된 FM 간의 모델 응답의 공통적인 특징과 세부 사항 및 응용분야에서의 차별성, 상업적 및 산업적 연관성에서의 차별성, 기술적인 진부와 혁신 강조, 기술의 확장성 및 유연성, 정책 결정과 전략적 기획의 영향에 관한 응답에서 분명한 차별성을 확인하였다.

<Table 7> Evaluation Results between FM and the Fine-tuned Model by GPT-4

Question No.	FM points	Fine-tuned FM points	Why
1	4	1	• The response from the base model closely aligns with the matched response in focusing on the primary goal of developing deployable radiators for CubeSats, thereby enhancing CubeSat performance and sustainability. However, the fine-tuned model's response deviates from the topic by adding too many technical details and complexities.
2	1	2	• While the response from the base model provides a general description of the basic functions of LIDAR systems, it hardly mentions specific chemical sensing capabilities mentioned in the matched response. • The response from the fine-tuned model, while not detailing specific chemical detection capabilities such as “single-molecule layer detection” mentioned in the matched response, suggests the ability to detect various chemical substances and particles, thus relating to the general theme of the matched response.
3	2	2	• The base model's response offers a general explanation, speculating on various methods for regenerating electrodes, which does not directly align with the specific content of the matched response. • The response from the fine-tuned model presents specific methods for regenerating electrodes, albeit different from those mentioned in the matched response. Nonetheless, it shows some alignment regarding the theme of “regeneration.”
4	1	2	• The response from the base model does not delve into specific content related to certain application areas mentioned in the matched response, focusing instead on extraterrestrial sample analysis, weakening its connection to the matched response. • The response from the fine-tuned model mentions the potential integration of USPM into atmospheric sampling systems, suggesting the possibility of sample collection and processing in environments similar to those mentioned in the matched response. However, it still does not perfectly align with the matched response.
5	3	4	• The response from the base model discusses how ITR coupling can ensure the stability and functionality of spacecraft components, partially applicable to tasks related to constructing habitable structures and reducing buckling as mentioned in the matched response. However, it does not comprehensively cover all elements of the matched response. • In contrast, the response from the fine-tuned model provides specific explanations of how ITR coupling can support deep-space missions through lightweighting and performance enhancement, including the adjustability of solar array configurations related to solar electric propulsion mentioned in the matched response, which aligns more closely with the technical requirements mentioned.
6	2	4	• The response from the base model broadly discusses the wide range of applications for AMOORS but does not delve into specific functionalities such as recognizing specific fields of view and objects, which were crucially addressed in the matched response. • The fine-tuned model's response clearly explains the system's purpose of detecting objects like space debris in real-time and tracking multiple objects simultaneously, aligning better with the functionalities demanded in the matched response.
7	2	3	• The base model's response elaborates on the potential uses of HAC and SLICER but does not specifically address applications such as harvesting water ice in the permanently shadowed regions of the Moon, as mentioned in the matched response. • The response from the fine-tuned model mentions the use of equipment for digging up and processing lunar surface deposits, explicitly stating its direct applicability to future NASA missions, thus having a higher relevance to the matched response.
8	2	3	• The base model's response broadly covers various potential applications of receiver technology but lacks specific information related to the commercialization of terahertz technology, which was prominently discussed in the matched response. • The fine-tuned model's response specifically mentions military and commercial space applications of receiver technology, emphasizing the development of high-performance microwave components and the commercialization potential of terahertz technology, albeit lacking direct mentions of terahertz technology.
9	4	3	• The base model's response evaluates the mass and energy efficiency of new technologies for extracting oxygen from Mars, emphasizing the importance of comparing them with existing methods, aligning closely with the matched response in stressing the comparison with existing systems. • The response from the fine-tuned model specifies the purpose of comparison but is less clear on the specific existing systems being compared, leading to less perfect alignment with the matched response.
10	1	2	• The base model's response provides a general explanation of how the dual plenoptic system works but lacks specific descriptions of functionalities such as turbulence effect analysis and correction and posture information extraction, as described in the matched response. • The fine-tuned model's response explains the process by which the system uses two beams to enable precise posture determination, emphasizing its applicability in vibrational environments. While partially aligning with the process of extracting posture information from the model mentioned in the matched response, the overall description does not perfectly match the matched response.

기반 모델과 미세 조정 모델 모두 각 기술의 핵심 목적과 기능을 설명하는 데 있어 일관된 정보를 제공하고, 기술의 기본 개념과 운영 원리에 대한 깊은 이해를 반영하고 있다. 하지만 세부 사항 및 응용 분야에 관한 응답에서 미세 조정 모델은 기반 모델 보다 훨씬 더 구체적인 기술적 세부 사항과 다양한 응용 분야를 제시하였으며, 미세 조정 모델이 특정 산업 또는 잠재 고객의 요구에 더 잘 맞춰져 있음을 시사하였다. 상업적 및 산업적 연관성 측면에서 미세 조정 모델은 기술의 상업적 적용 가능성과 특정 산업에 대한 적합성을 강조하였고, NASA 및 기타 정부 기관 뿐만 아니라, 군사, 우주, 의료, 자동차 등 다양한 분야에서의 상업적 응용에 대한 더 광범위한 통찰력을 제공하였다. 기술적인 진보와 혁신을 강조하는 측면에서 미세 조정 모델은 기존의 기술적 접근법을 향상시키고 새로운 가능성을 탐구하는 데 더 큰 중점을 두었으며, 개발 과정에서의 진보가 기술의 효과성과 시장 적합성을 어떻게 향상시킬 수 있는지에 대한 중요한 시사점을 제공하였다. 기술의 확장성과 유연성 측면에서 미세 조정 모델은 기술이 다양한 환경과 조건에서 어떻게 작동할 수 있는지에 대한 시나리오를 제공함으로써 기술의 확장성과 유연성을 강조하였으며, 특히 새로운 시장 진입 전략이나 다양한 응용 프로그램 개발에 관한 중요성을 강조하였다. 마지막으로 정책 결정과 전략적 기획의 영향 측면에서 미세 조정 모델은 기술의 구체적인 적용 사례를 통해 정책 결정자와 기술 전략가들에게 유용한 통찰력을 제공하였으며, 기술의 정책 및 전략적 적용 가능성을 평가하는 데 도움이 되는 정보를 제공하였다.

또한 10개 질의에 대해서 NASA 프로그램 원시 자료에 기반한 매칭 답변과의 두 모델의 생성 답변 간의 적합성 평가를 수행한 결과, 통계적 유의성을 발견하지는 못하였으며 표본 수의 한계로 일반화의 어려움이 존재한다. 하지만, 10건의 표본 내에서는 미세 조정 모델이 기반 모델보다 원시 자료와의 일치성 측면에서 더 높은 성능을 보이고 있음을 시사하였으며, 특히, 미세 조정 모델은 원시 자료에 대한 이해를 바탕으로 더 정교한 정보를 제공하며, 주어진 데이터에 대해 더욱 정확하게 반응하는 것으로 정성적 분석을 통해 확인되었다. 또한, 10건의 표본에서 표준편차가 미세 조정 모델에서 기반 모델에 비해 작게 나타난 것은 미세 조정 모델이 결과의 일관성 면에서도 더 좋은 경향이 존재하고 있음을 확인하였다. 이는 미세 조정 모델이 안정적으로 더 정확한 정보를 생성하며, 개별 질문에 대한 답변의 변동성이 적을 가능성이 일정 수준 존재하고 있음을 의미한다.

5. 결 론

본 연구에서는 대규모 언어 모델(LLMs)의 성능을 최적화하기 위해 사용자 지시 따르기 미세 조정과 저순위 적응 미세 조정을 결합한 새로운 접근 방식을 제안하고 실험적으로 검증하였다. 이를 통해 LLMs의 성능을 개선하고, 특히 국가 연구개발(R&D) 데이터의 분석 정밀도와 효율성을 향상시키는 효과를 확인하였다.

기술적 혁신과 진보에 대한 강조는 미세 조정 모델이 특화되고 최적화된 답변을 생성함으로써 산업 요구와 시장 동향에 대응하는 능력이 향상되었음을 시사하고 있다. 모델이 제공하는 상세한 적용 사례는 기술의 정책적, 전략적 수립에 영향을 미칠 수 있는 통찰력을 제공하여, 기술의 사회적 수용성과 경제적 가치를 높이는 방향으로 정책과 전략을 유도하는 데 중요한 역할을 수행할 가능성을 확인하였다. 미세 조정 모델이 특정 지식 기반의 데이터에 비해 더 정확하고 일관된 응답을 생성할 수 있는 중요한 시사점 제공하는 것을 확인하였고, 향후 과학기술문헌 분야 모델 개발과 응용에 있어서 미세 조정의 중요성을 강조하는 근거로 활용 가능할 것으로 기대되었다.

미세 조정된 모델은 기본 모델에 비해 더 정확하고 일관된 답변을 생성하는 것으로 나타났다. 이는 데이터 분석의 정밀도와 응답의 정확성을 높이는 데 중요한 역할을 한다. 특히 NASA 연구개발 문헌을 활용한 사례 연구에서, 미세 조정된 모델이 더 높은 일치성과 정교한 정보를 제공하는 것으로 나타났다. 이는 특정 도메인에 특화된 모델이 원시 자료와의 적합성을 높일 수 있음을 시사한다.

미세 조정 모델은 기존 기술적 접근법을 향상시키고 새로운 가능성을 탐구하는 데 더 큰 중점을 두었으며, 이는 기술의 효과성과 시장 적합성을 향상시키는 데 중요한 시사점을 제공한다.

LoRA 기반 PEFT 방법을 통해 모델의 파라메터 수를 줄여 고성능 하드웨어 자원 없이도 효율적으로 모델을 운영할 수 있음을 확인하였다. 이는 고비용의 하드웨어 요구를 회피하고도 높은 성능을 유지할 수 있음을 의미한다.

본 연구는 몇 가지 한계점을 가지고 있다. 첫째, 실험 데이터의 범위와 규모가 제한적이어서 결과의 일반화에 한계가 있다. 둘째, 미세 조정 과정에서 사용된 데이터의 질과 양에 따라 결과가 달라질 수 있다. 따라서 향후 연구에서는 더 다양한 도메인과 더 큰 규모의 데이터를 활용한 실험이 필요하다.

미래 연구에서는 다음과 같은 방향으로 진행될 수 있다:

다양한 도메인 적용: 본 연구에서 제안한 방법론을 다른 도메인에 적용하여 그 효과를 검증하는 연구가 필요하다.

효율성 향상: LoRA 기반 PEFT 방법의 효율성을 더욱 향상시키기 위한 연구가 필요하다.

정책 및 전략적 기획: 미세 조정된 모델의 응답이 정책 결정과 전략적 기획에 미치는 영향을 평가하는 연구가 필요하다.

본 연구는 대규모 언어 모델의 성능을 최적화하기 위한

새로운 통합 미세 조정 방법론을 제안하고, 이를 통해 데이터 분석의 정밀도와 효율성을 향상시킬 수 있음을 실험적으로 확인하였다. 이 연구는 향후 다양한 도메인에서 LLMs의 적용을 확장하는 데 중요한 참고 자료가 될 것이며, 정책 결정자와 연구자들에게 유용한 통찰을 제공할 것이다. 이를 통해 LLM을 활용한 데이터 분석의 새로운 가능성을 열고, 특별 지식 도메인 관련 분야에서 방법론적으로 크게 기여할 것으로 기대된다.

Acknowledgement

This research was supported by Korea Institute of Science and Technology Information(KISTI) ((KISTI) K-24-L3-M2-C4-02).

References

[1] Bilgram, V., Laarmann, F., Accelerating Innovation With Generative AI: AI-Augmented Digital Prototyping and Innovation Methods, *IEEE Engineering Management Review*, 2023, Vol. 51, No. 2, pp. 18-25.

[2] Crothers, E., Japkowicz, N., Viktor, H., Machine-Generated Text: A Comprehensive Survey of Threat Models and Detection Methods, *IEEE Access*, 2023, Vol. 11, pp. 70977-71002.

[3] Cámara, J., Troya, J., Burgueño, L., Vallecillo, A., On the assessment of generative AI in modeling tasks: an experience report with ChatGPT and UML, *Software and Systems Modeling*, 2023, Vol. 22, No. 3, pp. 781-793.

[4] Hassija, V., Chakrabarti, A., Singh, A., Chamola, V., Sikdar, B., Unleashing the Potential of Conversational AI: Amplifying Chat-GPT's Capabilities and Tackling Technical Hurdles, *IEEE Access*, 2023, Vol.11, pp. 143657-143682.

[5] Hommel, B., Expanding the methodological toolbox: Machine-based item desirability ratings as an alternative to human-based ratings, Personality and Individual Differences, 2023, Vol. 213, 112307.

[6] Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., LoRA: Low-Rank Adaptation of Large Language Models, *ICLR 2022 Conference Poster*, 2022, pp. 1-26.

[7] Kamnis, S., Generative pre-trained transformers(GPT) for surface engineering, *Surface and Coatings Technology*, 2023, Vol. 466, 129680.

[8] Karkera, N., Acharya, S., Palaniappan, S., Leveraging pre-trained language models for mining microbiome-disease relationships, *BMC Bioinformatics*, 2023, Vol. 24, No. 1, Article 290.

[9] Kheddar, H., Himeur, Y., Al Maadeed, S., Amira, A., Bensaali, F., Deep transfer learning for automatic speech recognition: Towards better generalization, *Knowledge-Based Systems*, 2023, Vol.277, pp. 1-34.

[10] Kim, J., Yoon, S., Choi, T., Sull, S., Unsupervised Video Anomaly Detection Based on Similarity with Predefined Text Descriptions, *Sensors*, 2023, Vol.23, No. 14, 6256.

[11] Kolides, A., Nawaz, A., Rathor, A., Beeman, D., Hashmi, M., Fatima, S., Berdik, D., Al Ayyoub, M., Jararweh, Y., Artificial intelligence foundation and pre-trained models: Fundamentals, applications, opportunities, and social impacts, *Simulation Modelling Practice and Theory*, 2023, Vol.126, 102754.

[12] Lankford, S., Afli, H., Way, A., adaptMLLM: Fine-Tuning Multilingual Language Models on Low-Resource Languages with Integrated LLM Playgrounds, *Information(Switzerland)*, 2023, Vol. 14, No. 12, pp. 1-24.

[13] Lin, K., Agia, C., Migimatsu, T., Pavone, M., Bohg, J., Text2Motion: from natural language instructions to feasible plans, *Autonomous Robots*, 2023, Vol.47, No. 8, pp. 1345-1365.

[14] Mazumdar, H., Chakraborty, C., Sathvik, M., Mukhopadhyay, S., Panigrahi, P., GPTFX: A Novel GPT-3 Based Framework for Mental Health Detection and Explanations, *IEEE Journal of Biomedical and Health Informatics*, 2023, PMID:37903039.

[15] Megahed, F., Chen, Y., Ferris, J., Knoth, S., Jones Farmer, L., How generative AI models such as ChatGPT can be(mis)used in SPC practice, education, and research? An exploratory study, *Quality Engineering*, 2023, pp. 278-315.

[16] Nicula, B., Dascalu, M., Arner, T., Balyan, R., McNamara, D., Automated Assessment of Comprehension Strategies from Self-Explanations Using LLMs, *Information (Switzerland)*, 2023, Vol. 14, No. 10, 567.

[17] Pan, W., Jiang, P., Li, Y., Wang, Z., Huang, J., Research on automatic pilot repetition generation method based on deep reinforcement learning, *Frontiers in Neurorobotics*, 2023, Vol. 17.

[18] Porsdam Mann, S., Earp, B., Møller, N., Vynn, S., Savulescu, J., AUTOGEN: A Personalized Large Language Model for Academic Enhancement—Ethics and Proof of Principle, *American Journal of Bioethics*,

2023, Vol. 23, No. 10, pp. 28-41.

[19] Singh, I., Blukis, V., Mousavian, A., Goyal, A., Xu, D., Tremblay, J., Fox, D., Thomason, J., Garg, A., ProgPrompt: Program generation for situated robot task planning using large language models, *Autonomous Robots*, 2023, Vol. 47, No. 8, pp. 999-1012.

[20] Sætra, H., Generative AI: Here to stay, but for good?, *Technology in Society*, 2023, Vol.75, 102372.

[21] Yin, C., Du, K., Nong, Q., Zhang, H., Yang, L., Yan, B., Huang, X., Wang, X., Zhang, X., *PowerPulse: Power energy chat model with LLaMA model fine-tuned on Chinese and power sector domain knowledge*, Expert Systems, 2023.

[22] Zhao, C., Yuan, S., Jiang, C., Cai, J., Yu, H., Wang, M., Chen, Q., ERRA: An Embodied Representation and Reasoning Architecture for Long-Horizon Language-Conditioned Manipulation Tasks, *IEEE Robotics and Automation Letters*, 2023, Vol. 8, No. 6, pp. 3230-3237.

[23] Zhu, Q., Zhang, X., Luo, J., Biologically Inspired Design Concept Generation Using Generative Pre-Trained Transformers, *Journal of Mechanical Designs*, 2023, Vol.145, No. 4, pp. 1-23.

ORCID

Sang-Gook Kim | https://orcid.org/0000-0001-7018-9716
Kyungran Noh | https://orcid.org/0000-0001-9525-1387
Hyuk Hahn | https://orcid.org/0000-0003-1176-2767
Boong Kee Choi | https://orcid.org/0000-0001-6034-1091