ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
https://www.jiisonline.org
J Intell Inform Syst 2024 March: 30(1): 93~120 http://dx.doi.org/10.13088/jiis.2024.30.1.093

도메인 특화 LLM: Mistral 7B를 활용한 금융 업무분야 파인튜닝 및 활용 방법

정천수
삼성SDS AI Automation Team
(csu.jeong@samsung.com)

최근 사전학습된 범용적인 LLM(Large Language Model) 출시가 활발해지고 있지만, 도메인 특화 파인튜닝된 LLM 연구와 생성 방법을 제시하는 것은 부족한 실정이다. 본 연구는 도메인에 특화된 LLM의 파인튜닝과 활용에 대한 방안을 탐구하고 LLM의 최신 동향, 파운데이션 모델 및 LLM의 사전학습, 그리고 도메인 특화 LLM 파인튜닝에 대한 방법을 제시한다. 특히, 금융 분야에서의 언어 모델 활용이 중요하기 때문에 금융 특화 데이터셋의 선정과 전처리 방법, 모델 선정 및 파인튜닝 절차, 그리고 금융 특화 LLM 파인튜닝 시 고려해야 할 사항들에 대해 구체적으로 제시한다. 금융 데이터 특성을 고려한 도메인 특화 어휘의 구축과 보안 및 규정 준수에 대한 고려사항을 다룬다. LLM 파인튜닝의 적용과 활용 연구에서는 SLM인 Mistral 7B을 활용한 실제 보험 금융 도메인 LLM을 생성하는 방법 및 구현 절차와 다양한 금융 분야에 대한 사례를 제시한다. 이를 통해 본 연구는 LLM을 금융 도메인 분야에 적용하는 가능성을 알아보고 한계점과 개선 방향을 제안함으로써 향후 연구 방향성을 제시한다. 따라서 본 연구는 업무 도메인 분야에서 자연어 처리 기술의 적용과 발전에 기여함과 동시에 다양한 산업 분야에서의 LLM 활용 방향성을 제공함으로써 기업 내 금융 서비스 및 다양한 산업군에 LLM을 적극적으로 활용할 수 있도록 하는데 의미와 가치가 있다.

주제어 : 금융 도메인 LLM, 소형 언어모델, 사전학습 언어모델, 파인튜닝 언어모델, PEFT

논문접수일 : 2024년 1월 12일 논문수정일 : 2024년 2월 27일 게재확정일 : 2024년 2월 27일
원고유형 : Regular Track 교신저자 : 정천수

1. 개요

최근 몇 년 동안, 인공지능과 딥러닝 기술의 발전으로 인해 특히, 자연어 처리 분야에서 놀라운 성과를 얻고 있다. 그 중에서도 대규모 언어 모델(Large Language Models, LLM)은 풍부한 문맥과 언어 이해 능력으로 인간 수준의 언어 생성 및 이해에 도달하고 있다. 따라서 여러 산업군에서 LLM도입을 활발히 검토하거나 도입하고 있다. 특히, McKinsey에 따르면 LLM 등 생성형 AI는 은행업 내에서 '마케팅/판매, 고객지원/관리, 프로그래밍, 규제준수' 비즈니스 분야에서 생산성 제고에 기여할 것으로 보고 있다. 생성형 AI는 글로벌 은행산업 내에서 약 2,000억~3,400억 달러의 가치를 창출할 것으로 전망되며, 이는 산업 전체의 매출에서 2.8~4.7%에 상응하는 수치다. 금융회사는 내부적으로 직원의 업무 수행을 지원하고 자동화하며, 자연어 기반 정보를 수집·분석해 전략적 판단을 내리기 위하여 LLM을 활용한다(장갑수, 2023). 이렇게 가장 늦게 신기술을 도입하는 금융권에서도 생성형 AI인 LLM의 활용이 확대되고 있다. 하지만 일반적인 LLM을 가지고 도입을

검토하기 때문에 금융권의 특화된 요구사항을 충족시키는데 한계가 있다. 도메인 특화는 특정 분야에 깊은 지식을 갖춘 LLM을 개발하는 것을 목표로 하는데 해당 분야의 특화된 요구사항을 충족하는 것이 필수적이다. 예를 들어, 의료 분야에서는 정확하고 상세한 의학 용어와 지식이 필요하며, 법률 분야에서는 복잡한 법률 용어와 개념을 정확히 이해하고 처리할 수 있는 능력이 요구된다. 이러한 특화된 요구사항을 충족시키기 위해서는 일반적인 LLM보다는 해당 분야에 특화된 모델이 필요하며 해당 분야의 세밀한 요구사항을 만족시키고, 보다 정확하고 신뢰할 수 있는 결과를 제공하는 데 중요한 역할을 해야 한다.

따라서 본 연구에서는 이러한 도메인 특화된 요구사항을 충족할 수 있는 금융 분야에 적용되는 LLM을 대상으로 파인튜닝 및 활용사례를 탐구하고자 한다. 금융 분야에서는 급변하는 시장 환경, 다양한 금융 이벤트, 그리고 방대한 양의 금융 데이터가 존재한다. 이에 대응하기 위해서는 빠르고 정확한 정보 처리 및 의사결정 능력이 필수적이다. 최근의 연구들은 LLM이 이러한 도전에 효과적으로 대응할 수 있는 가능성을 제시하고 있으며 빅테크 기업에서는 범용적인 LLM 출시가 활발해지고 있다. 그러나 특정 업무에 특화된 LLM을 구축하고 활용하는 과정에서 발생하는 다양한 문제들을 심층적으로 이해하고 개선하는 연구는 아직 미비한 상태이다.

본 연구는 도메인에 특화된 LLM을 적용하는 방법을 구체적으로 제안하고자 한다. 특히, 금융 산업은 타 산업 대비 신뢰, 소비자 보호 규제, 포용 금융 등의 특수성을 보유하고 있다(장갑수, 2023). 이렇게 금융 분야는 기술 혁신에 민감하게 반응하는 동시에 안정성과 정확성을 중요시하는 특성을 가지고 있다. 따라서 금융 분야에 특화된 LLM의 개발과 활용은 금융 비즈니스의 효율성과 경쟁력을 향상시킬 수 있는 중요한 과제이다.

본 연구는 이러한 맥락에서 금융 분야에 특화된 LLM의 핵심 기술과 응용 가능성에 대한 심층적인 이해를 제공하여 금융 기업 및 연구 기관에 실질적인 가치를 제공할 것으로 기대된다. 본 연구에서는 특히 보험 금융 분야에 한정하여 LLM의 파인튜닝 및 활용사례를 다룰 것이다. 그러나 금융 분야는 광범위하며, 다양한 하위 분야가 존재한다. 따라서 이 연구에서 다루지 못하는 세부 분야에 대한 연구는 향후 보완될 필요가 있다. 또한, 언어 모델의 한계를 고려하여 연구 결과를 해석할 것이다. 본 연구의 주된 목적은 금융 분야에 특화된 LLM을 구축하기 위한 파인튜닝의 효과적인 방법론을 탐구하고, 이를 실제 금융 업무에 적용하는 활용사례를 연구하는 것이다. 이를 통해 금융 분야의 전문 지식을 반영한 언어 모델이 향상된 성능을 보이며, 금융 업무의 생산성 향상과 의사결정 과정의 지원에 기여할 것으로 기대한다.

본 논문은 다음과 같이 구성되어 있다. 2장에서는 LLM과 파인튜닝에 대해 살펴보고, 3장에서는 금융 특화 LLM 생성 방안에 대해 상세히 설명한다. 4장에서는 LLM생성 방법 및 구현 코드와 금융 분야에서의 LLM 활용 방안을 다양한 관점에서 살펴보며, 5장에서는 본 연구의 결론과 향후 연구 방향에 대해 논의한다.

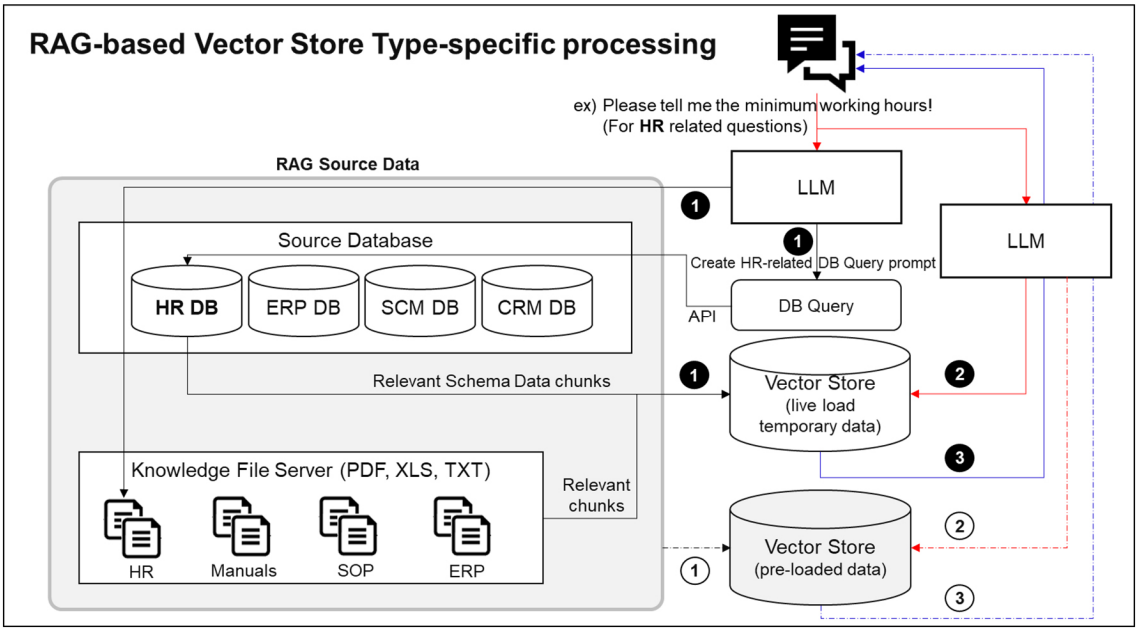
2. 이론적 배경

본 연구를 위해 생성형 AI 및 LLM과 관련된 자료에 대하여 최근에 나온 주요 연구논문 및 저널, 기사, 도서를 조사하였으며 본 장에서는 LLM 개요로 생성형 AI와 파운데이션모델에 대하여 알아본다.

그리고 LLM의 파인튜닝에 대하여 상세하게 알아보고, 본 논문에서 다루는 금융분야에서의 언어모델 적용 영역으로 나누어 설명한다.

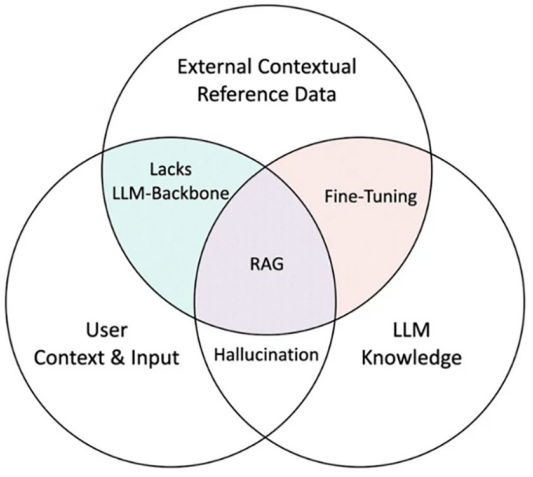
2.1. LLM(Large Language Model)의 개요

최근 몇 년 동안, LLM은 NLP(자연어 처리) 분야에서 차별화된 발전을 이루고 있다. Transformer 구조를 기반으로 한 BERT (Bidirectional Encoder Representations from Transformers), GPT (Generative Pre-trained Transformer) 등의 모델은 대량의 텍스트 데이터를 사전 훈련하여 다양한 자연어 태스크에 적용할 수 있는 강력한 표현을 학습하였다. 이들 모델은 텍스트의 문맥을 파악하고, 다양한 문법 구조와 의미적 관계를 이해하는 데에 우수한 성과를 보여주고 있다. 생성형 AI(Generative AI)는 방대한 양의 학습된 데이터모델을 바탕으로 텍스트, 이미지, 오디오, 비디오와 같은 새로운 콘텐츠를 생성할 수 있는 인공지능의 한 형태이다(Jeong, 2023). 2024년 2월에는 OpenAI에서 Text-to-Video모델인 SORA를 발표해 비디오를 쉽게 생성할 수 있는 모델을 출시하였다(OpenAI, 2024). 또한 NLP 분야중 챗봇은 NLU(Natural Language Understanding) 기술의 발전으로 Context 모델과 Transformer 언어모델 활용으로 복잡한 대화 처리가 가능하다(정천수, 2023a). 또한 챗봇에 RPA 및 OCR 등 타 솔루션과 연계하여 챗봇을 업무에 직접적으로 활용하여 효율성을 높이고 있다(정천수, 정지환, 2020). 이렇게 LLM과 생성형 AI는 AI의 딥러닝 안에 포지셔닝하고 있어 딥러닝 기반으로 LLM을 활용하여 생성형 AI 서비스를 할 수 있게 된다(Mayank, 2023; 정천수, 2023d). 2022년 11월에는 OpenAI에서 GPT-3에 인간 전문가 집단이 피드백(RLHF, Reinforcement Learning from Human Feedback)시키는 학습과정을 거치면서 대화형으로 발전시킨 GPT-3.5모델을 적용하여 사람과 유사한 자연어를 생성하도록 학습된 AI 챗봇인 ChatGPT를 공개하여 출시 2개월만에 월간 이용자가 1억명을 넘기며 많은 관심을 받았다(정천수, 2023b). ChatGPT같은 오픈된 도메인에 대한 챗봇 대화에서는 ChatGPT와 LLM모델과의 Prompt로 인터페이스를 하게 되는데, 이때 인터페이스를 하기 위한 어플리케이션 사이에 사용하는 매개변수를 적용할 때 주의해야 할 정보는 개인정보에 관한 것은 회피해야 한다. 개발자가 동의 없이 사용자로부터 데이터를 수집하고 사용하는 것을 방지하는 법이 이미 있지만, 실제 생활에서 사용자는 개발자가 데이터를 얼마나 많이 가져오고, 해당 데이터가 어디에 있는지 알기 어렵기 때문이다(Jeong and Jeong, 2022). 또한 ChatGPT에서 대화 시에는 질문이나 요청인 Prompt를 얼마나 자세하게 전달하느냐에 따라 완성도 높은 답변을 얻을 수 있기 때문에 LLM으로부터 프롬프트 입력 값들의 조합을 찾는 작업을 탐구하는 프롬프트 엔지니어링도 중요한 요소로 작용한다(정천수, 2023c). 특히 금융분야에서는 고객 응대 챗봇을 통한 최신 정보 제공의 중요성에 대두되며, LLM의 정보 제한성과 환각(Hallucination) 문제는 이러한 모델들의 도전과제로 지적되고 있다. 이를 해결하기 위한 접근방식으로는 새로운 데이터로의 파인튜닝과 프롬프트 콘텍스트에 직접 정보를 삽입하는 방안이 있으나, 파인튜닝의 경우에 학습을 위한 인프라 준비 등 상당한 비용이 발생하며, 모든 정보를 프롬프트에 넣어주는 것도 현실적으로 어렵기 때문에 이에 대안으로 RAG모델이 제안되었으며, <그림 1>과 같이 정보를 벡터 데이터베이스에 저장하고, 필요한 정보를 검색하여 LLM에 전달하는 방식으로 구현되기도 한다(정천수, 2023d).



〈그림 1〉 RAG기반 Vector Store 구성 유형 및 처리절차

따라서 기업에서는 <그림 2>와 같이 LLM에 파인튜닝된 모델과 함께 RAG를 활용하여 환각을 최대한 줄 일수 있는 RAG를 조합하여 사용하는 것이 필요하다(Greyling, 2023).



〈그림 2〉 LLM, Fine-Tuning and RAG Relation Diagram

이렇듯 LLM 활용에는 여러 도전 과제가 존재한다. 이는 비용, 속도, 개인 정보 보호, 언어적 한계, 그리고 도메인 지식 부족 등의 측면에서 나타난다. 이러한 문제를 해결하기 위해 LLM의 경량화를 고려하게 된다. 경량화된 모델은 비용 효율적이며, 빠른 처리 속도를 제공하고, 오프라인 환경에서도 운영이 가능하며, 특정 언어나 도메인에 대해 특화된 경량 모델의 개발을 통해 언어적 정확성과 도메인 지식 부족 문제를 해결할 수 있다. 이러한 기술 개발은 비즈니스 환경에서 LLM의 효과적인 도입과 활용을 위해 필수적이다.

2.1.1. LLM 동향

2023년부터 초거대AI 기술이 빠르게 발전하였다. <표 1>은 2023년에 출시된 LLM과 SLM(Small Language Model) 출시 현황을 보여주고 있으며

3월에 오픈AI는 GPT-4를 발표했으며, 구글은 5월에 PaLM2를 발표했는데, 이전 모델에 비해 파라미터 수는 줄였지만 약 5배 더 많은 토큰(텍스트 데이터)을 학습하여 실제 성능을 더 높였다. 또한 삼성전자는 11월에 삼성 가우스(Samsung Gauss)를 첫 공개하였으며 향후 출시될 갤럭시 S24등 제품들에 삼성 가우스를 단계적으로 탑재할 계획이라고 밝혔다. 카카오 등 유수의 기업들도 자체 생성형 AI를 구축 중이거나 검토하고 있으며, 모델 크기보다는 학습량에 초점을 둔 메타의 라마(LLaMA) 와 팰컨(Falcon) 등 오픈소스 LLM에 대한 관심이 높아지고 있다(정천수, 2023d).

또한, LLM의 경량 모델인 소형언어모델의 경쟁이 치열해질 것으로 예상되고 있으며 업계는 매개변수가 300~340억개를 사용한 모델을 모델명 뒤에 30B, 34B를 붙여서 LLM으로 불러왔으며,

〈표 1〉 최근 LLM 및 SLM 출시 현황

Model Size	Company	Foundation Model	Parameters	Source	Release Date	Service
LLM	OpenAI	GPT-4 Turbo	1.75T	Closed	2023.11	ChatGPT, GPTs / MS Bing AI, MS Copilot, MS 365 Copilot
	Google	Gemini	미공개	Closed	2023.12	Gemini
	Naver	HyperClovaX	미공개	Closed	2023.08	폴라리스오피스 AI, 루이스 등
	LG	EXAONE2.0	300B	자체활용	2023.07	AI아티스트 틸다 등
	NC Soft	VARCO	미공개	Closed	2023.08	VARCO Art/Text/Human/Studio
	SKT	A. Enterprise	미공개	자체활용	2023.08	문서요약, 문서생성, Q&A기능
	KT	MI:DEUM2	200B	자체활용	2023.10	지니TV, AICC, AI 통화비서
	SAMSUNG	Samsung Gauss	미공개	자체활용	2023.11	Gauss Language/Code/Image
	Huawei	PanGu 3.0	100B	Open	2023.07	Pangu-Weather, etc.
	Baidu	Ernie 3.5	130B(추정)	자체활용	2023.06	Ernie Bot 3.5
	SKT	A. Enterprise	미공개	자체활용	2023.08	문서요약, 문서생성, Q&A기능
	KT	MI:DEUM2	200B	자체활용	2023.10	지니TV, AICC, AI 통화비서
SLM	META	LLaMA2	7B, 13B, 70B	Open	2023.07	-
	META	Code LLaMA	7B, 13B, 34B	Open	2023.07	-
	Google	Gemini Nano-1, 2	1.8B, 3.25B	Closed	2023.12	온디바이스용
	Google	Gemma	2B, 7B	Open	2024.02	-
	Stanford Univ.	Alpaca	7B	Open	2023.03	LLaMA 7B Fine-tuning Model
	Nomic	GPT4All v2	3B~13B	Open	2023.04	LLaMA 7B Fine-tuning Model
	Alibaba	Tongyi Qianwen	7B	Open	2023.04	DingTalk, Tmall Genie
	TII	Falcon	7B, 18B	Open	2023.09	-
	Mistral AI	Mistral 7B	7B	Open	2023.09	-
	MS	Phi-2	2.7B	Closed	2023.12	온디바이스용
	Upstage	Solar Mini	10.7B	Closed	2023.12	-

이 이하 매개변수를 사용한 모델들을 SLM이라 부르고 있다. SLM의 대표주자에는 메타의 인기 모델 라마2(Llama-2 7B), 알리바바의 큐원(Qwen 7B), 그리고 최근 급부상중인 Mistral(Mistral 7B) 모델이 있고, Mistral AI가 2023년 12월 추가로 출시한 Mistral 7B 모델기반 “Mixtral 8x7B”는 Transformer내 Feed Forward 블록을 8배 크기로 확장한 모델이다. 과거에는 오픈 소스 진영이나 스타트업에서 인기를 끌던 SLM에 이제는 빅테크까지 뛰어들며 치열한 경쟁을 예고하고 있다. 메타의 인기 모델인 ‘라마2 7B(매개변수 70억개)’를 비롯해 13B, 22B 모델이 주를 이루고 있다. 특히 최근에는 온디바이스 AI가 대세로 떠오르며 7B보다 작은 모델이 속속 등장하는 추세이다. 2023년 12월에 구글은 차세대 모델 제미나이(Gemini)를 공개하며 온디바이스용 나노-1(1.8B), 나노-2(3.25B)를 내놓았으며 마이크로소프트(MS)는 2.7B 매개변수의 ‘파이-2’를 출시하며 온디바이스 AI용이라고 강조했다(인공지능신문, 2023.12.16). 또한 2024년 2월에는 구글이 젬마(Gemma) 2B와 7B를 오픈소스로 공개하여 상업용으로도 이용이 가능한 모델을 발표 하였다(Google DeepMind, 2024.02.21). 이와 같이 2023년부터 많은 빅테크 기업들이 <표 1>과 같이 LLM 및 SLM을 경쟁처럼 만들어내고 학계에서 기술을 발전시켜 왔다. 앞으로는 기술의 발전과 더불어 산업/서비스에 LLM을 얼마나 고객이 만족하는 방향으로 실제로 잘 적용하는 지가 중요한 요소가 될 것이다. 이중에 최근에 많이 공개되고 있는 주요 SLM의 평가표인 <표 2>에 따르면 Mistral 7B는 모든 측정 항목 벤치마크에서 Llama2 13B보다 뛰어난 성능을 발휘하는 것으로 나타나기도 했다(Jiang et al., 2023). 이것은 파라미터가 많다고 좋은 성능을 발휘한다고 단정지을 수 없다는 것을 나타낸다.

비즈니스에 비용 효율적이고 성능 좋은 LLM 도입하기위해서는 <표 3>의 LLM 활용 분야를 고려하여 충분한 검토를 통해 사용하는 것이 중요하다.

〈표 2〉 SLM의 비교(Mistral7B와 Llama)

Model	Modality	MMLU	HellaSwag	WinoG	PIQA	Arc-e	Arc-c	NQ	TriviaQA	HumanEval	MBPP	MATH	GSM8K
LLaMA 2 7B	Pretrained	44.4%	77.1%	69.5%	77.9%	68.7%	43.2%	24.7%	63.8%	11.6%	26.1%	3.9%	16.0%
LLaMA 2 13B	Pretrained	55.6%	**80.7%**	72.9%	80.8%	75.2%	48.8%	**29.0%**	**69.6%**	18.9%	35.4%	6.0%	34.3%
Code-Llama 7B	Finetuned	36.9%	62.9%	62.3%	72.8%	59.4%	34.5%	11.0%	34.9%	**31.1%**	**52.5%**	5.2%	20.8%
Mistral 7B	Pretrained	**60.1%**	**81.3%**	**75.3%**	**83.0%**	**80.0%**	**55.5%**	**28.8%**	**69.9%**	**30.5%**	47.5%	**13.1%**	**52.2%**

〈표 3〉 LLM 활용 유형

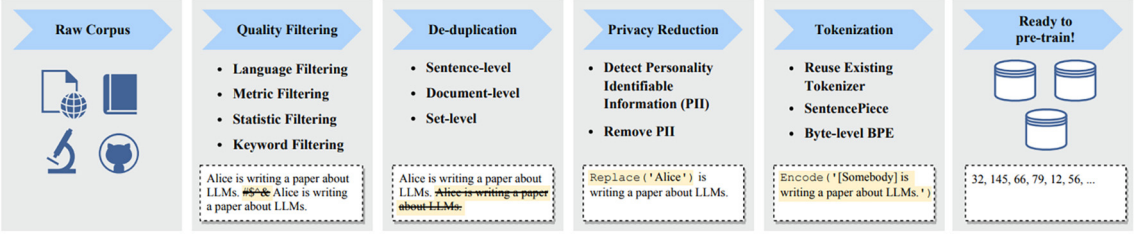
LLM 구분	활용 분야	Model Example
Multimodal LLM	이미지와 텍스트를 같이 분석하여, 이미지 설명을 생성, 텍스트 질문에 이미지를 활용해 답변을 생성할 수 있는 분야	GPT-4V, KOSMOS-2, Gemini
Domain Specific LLM	제조, 금융, 의료, 법 등 특정 분야에 대한 깊은 지식을 갖춘 업무 분야	Med-PaLM2, BloombergGPT
On-Device LLM	디바이스에 탑재되어 리소스가 제한된 환경에서도 고급 LLM을 배포하고 활용할 수 있는 분야	Gemini Nano, Phi-2, Samsung Gauss

2.1.2. 파운데이션 모델(Foundation Model) 및 LLM 사전학습(Pre-training)

생성형 AI 모델은 어떤 출력을 생성하는가에 따라서 언어모델, 이미지 모델, 동영상 모델 등을 사용한다. 하지만 현재는 이미지와 텍스트를 동시에 학습하는 멀티모달(Multi-modal) 모델들이 하루가 다르게 성능과 기능이 업그레이드고 있고 있으며 기초모델로 자리잡아가고 있다(정천수, 2023d). 파운데이션 모델(Foundation Model)의 데이터는 텍스트, 이미지, 음성, 정형데이터, 3D 시그널 등 구분하지 않고 학습에 이용되며 인간의 창의력과 추론력을 포함한 일을 수행하며 이러한 기초모델은 방대한 양의 데이터를 비지도 학습(Unsupervised learning)을 통해 모델을 학습시킨 후 배포되어 사용자가 원하는 목적에 맞게 다운스트림 작업에 대해 파인튜닝이나 문맥 내 학습(In-context learning)등과 같은 과정을 거쳐 완성되는 것이 파운데이션 모델이라고 볼 수 있다(Bommasani, et. al., 2021).

파운데이션 모델 중에서 언어모델로서의 LLM은 일반 Text 데이터인 Wikipidia 등 거대한 일반적인 지식들을 수집하여 자기지도학습(Self-Supervised Learning)이나 반자기지도학습(Semi-Supervised Learning)을 사용하여 레이블링되지 않은 상당한 양의 텍스트로 사전 학습된 언어 모델(Pre-trained Language Model, PLM)이다. 사전학습하기 위한 첫 번째 작업은 LLM이 학습될 리소스인 학습 데이터 세트를 수집하는 것이다. 데이터는 책, 웹사이트, 기사, 공개 데이터세트 등 다양한 소스에서 가져올 수 있다. 유능한 LLM을 개발하기 위해 사전 학습된 자료로 텍스트 데이터 세트를 사용한다. 사전 학습된 코퍼스의 소스는 크게 일반 데이터와 전문 데이터의 두 가지 유형으로 분류할 수 있으며 웹 페이지, 서적 및 대화 텍스트와 같은 일반 데이터는 크고 다양하며 접근 가능한 특성으로 인해 대부분의 LLM에서 활용되며 LLM의 언어 모델링 및 일반화 능력을 향상시킬 수 있다. 또한 다국어 데이터, 과학 데이터 및 코드와 같은 보다 전문화된 데이터 세트로 확장하여 LLM에 특정 작업 해결 기능을 부여하는 연구도 발표되고 있다(Chowdhery et al., 2023; Nijkamp et al., 2022).

<그림 3>은 일반적인 LLM의 데이터 수집 및 사전학습 절차를 보여주고 있다(Zhao et al., 2023). 모델 학습은 지도 학습(Supervised learning)을 사용하여 전 처리된 텍스트 데이터에 대해 학습된다. 또한 모델과 데이터의 크기가 크기 때문에 모델을 학습하려면 엄청난 계산 능력이 필요하며 학습 시간을 줄이기 위해 모델 병렬화라는 기술이 사용된다. 이렇게 대규모 언어 모델을 처음부터 학습하려면 상당한 투자가 필요하기 때문에 보다 경제적인 대안으로 기존 언어 모델을 특정 사용 사례에 맞게 파인튜닝을 하게 된다(정천수, 2023d).



〈그림 3〉 LLM 사전 학습 절차

2.1.3. 도메인 특화 LLM 사전학습

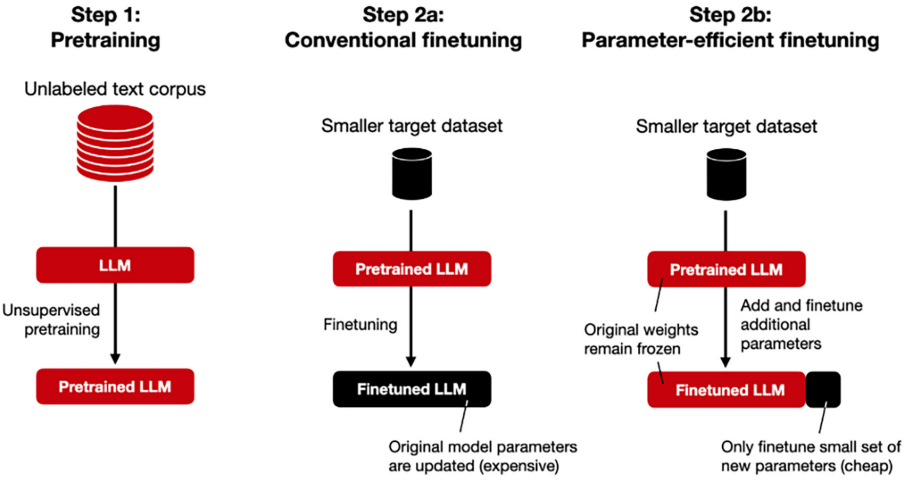
특정 도메인에 특화된 LLM을 생성하고자 할 때는 먼저 생성하고자 하는 도메인을 선택하고 관련 데이터를 수집하여 일반적인 지식만 알고 있는 일반적인 LLM에 특정 도메인 데이터 정보를 주입한 후(Adaptive Pre-training), 특정 도메인에 대한 대화 모델 학습을 통해 해당 도메인에 대한 대화가 가능한 모델을 생성한다.

도메인 선택 시 몇 가지 기준을 살펴보면, 이미 학습된 도메인 일 가능성이 낮은 특이한 도메인이나 최근 정보가 있는 도메인, 최소한 해당 도메인을 설명하는 1만자 이상의 텍스트 데이터가 있는 도메인, 공식적인 위키피디아(Wikipedia), 나무위키(namu.wiki) 정보는 대부분 이미 학습되었을 확률이 높기 때문에 학습 여부를 확인한 후 선택한다.

2.2. LLM 파인튜닝

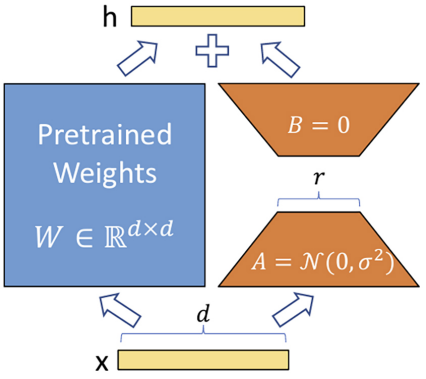
LLM은 대규모 데이터 모음에 대해 훈련되고 일반적인 지식을 가지고 있으나 파인튜닝 없이는 특정 작업에서 제대로 수행되지 않을 수 있기 때문에 <그림 4>와 같이 파인튜닝 단계를 거쳐 모델의 성능을 향상시킨다. BERT (450M), RoBERTa (1.3G) 처럼 크지 않은 모델을 사용할 때는 Full Fine-tuning을 진행한다. 하지만 LLaMA가 나오면서 이 모델들을 Full Fine-tuning을 하기에는 컴퓨팅 소스가 매우 크기 때문에 LoRA와 같이 기존의 Pre-trained Layer의 가중치는 고정하고, 새로운 레이어의 가중치만을 학습을 시키는데도, 실제 성능의 차이가 많지 않은 것으로 검증됨에 따라 최근에는 모든 매개변수를 튜닝하는 것이 아닌 사전 훈련된 LLM에 소수의 새로운 매개변수를 추가하고, 추가된 매개변수만 파인튜닝하여 적은 비용으로 더 나은 성능을 발휘하도록 하는 PEFT(Parameter-Efficient Fine-Tuning)방법을 주로 사용한다(Raschka, 2023).

대표적인 것인 LoRA는 LLM의 일부 Weight Matrix들에 대해서만 추가적인 학습을 허용하며 각각의Transformer Layer 내에서 일부의 Weight Matrix에 대해 파인튜닝 대상을 결정하며 해당



〈그림 4〉 LLM의 Fine-tuning

Weight Matrix의 형태를 그대로 복제하여 새롭게 파인튜닝 수행에서 학습을 허용하고 Weight Matrix은 Low-Rank Matrix Decomposition을 수행하여 두개의 Matrix로 구성되고 <그림 5>와 같이 A 매트릭스는 Random Gaussian 초기화가 적용되며, B 매트릭스는 Zero로 초기화되어 학습이 시작된다(Hu, et. al., 2021).

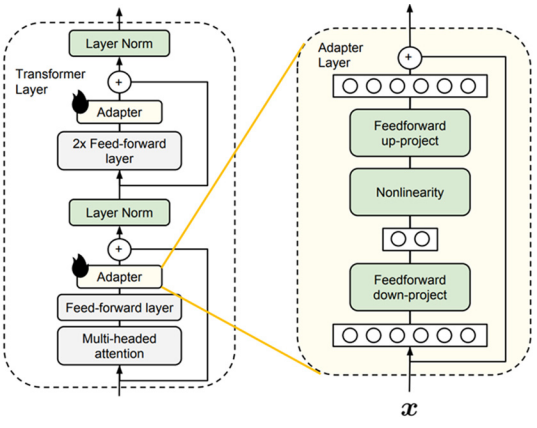


〈그림 5〉 LoRA Re-Parameterization

특히, 사전 학습 데이터가 LLM의 성능에 크게 영향을 미치기 때문에 소규모 언어 모델과 비교하여 LLM은 모델 사전 학습을 위한 고품질 데이터에 대한 수요가 더욱 필요하며, 모델 용량은 사전 학습을 위한 말뭉치(Corpus) 데이터 수집과 사전학습 처리 방법에 크게 의존하게 된다(정천수, 2023d). 이렇듯 LLM은 NLP(자연어 처리) 및 NLG (자연어 생성) 작업에서 딥 러닝을 활용하는 기반 모델로서, 언어의 복잡성과 연결성을 학습할 수 있도록 돕기 위해 LLM은 방대한 양의 데이터에 대해 사전 학습되며 파인튜닝, In-context learning, Zero/one/few-shot learning 같은 기술을 사용한다 (Dilmegani, 2023). 사전학습 후에는 모델 성능을 측정하기 위한 학습 데이터 세트로 사용되지 않은 테스트 데이터 세트에서 모델을 평가하고 평가 결과에 따라 모델의 성능을 향상시키기 위해 하이퍼파라미터를 조정하거나, 아키텍처를 변경하거나, 추가 데이터에 대한 교육을 통해 일부 파인튜닝을 한다(정천수, 2023d). 이렇게 파인튜닝 된 언어 모델(Fine-tuned Language Model, FLM)은 다양한 도메인 특화된 소규모 언어 모델(SLM)로 활용된다.

2.2.1. 최신 파인튜닝(Fine Tuning)

최근의 PEFT 파인튜닝 방법으로는 Prompt modification, Adapter methods, Parameterization으로 분류하고 있다. Prompt modification은 Hard prompt tuning과 Soft prompt tuning, Prefix-tuning이 있다. Adapter methods는 LLaMA-Adapter 등이 있으며 Adapter를 통한 PEFT로 전체 모델의 파인튜닝 부작용을 최소화하기 위해 모듈화한 파라미터를 <그림 6>과 같이 추가 삽입하여 학습한다. 하나의 Adapter 모듈은 Bottleneck Layer를 중심으로 Down-projection과 Up-projection의 선형변환을 수행하고 파인튜닝 과정에서 Pre-trained LLM은 Frozen된다.



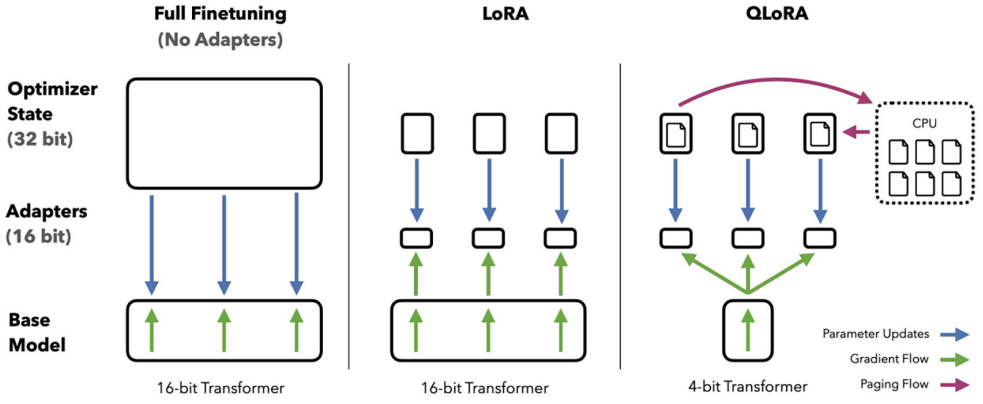
〈그림 6〉 Adapter methods

최근에는 LoRA보다 향상된 QLoRA도 발표되었으며, QLoRA는 개인용 PC 수준에서도 LLM 모델을 테스트할 수 있도록 QLoRA로 LLM을 PEFT 할 수 있는 방법을 제공하고 있다. LoRA가 Base 모델의 네트워크는 그대로 둔 채, 추가 데이터 만을 학습하여 본래의 Network에 연결하는 것이라면, <그림 7>과 같이 QLoRA는 여기에 더해 16bit Network Node를 4bit로 양자화하고, 부족한 메모리로 큰 모델을 조정할 수 있도록 바이너리를 나누어 2차 저장 장치에 스와핑(Swapping)하는 페이징(Paging)을 추가한 것으로 16비트를 4비트로 바꾸었으나 정보량 손실은 크지 않다 (Dettmers, et. al., 2024)

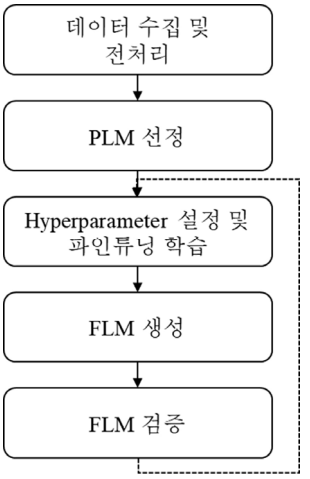
또한 Instruction Tuning은 최근에 많이 활용하는 파인튜닝으로 Task를 자연어로 옮겨 놓는 것으로 Vanilla-LLM은 단순히 다음 텍스트를 완성하지만, Instruction Tuning LLM은 사용자의 명령에 따라 다음 텍스트를 완성한다. 모델이 Task Instruction을 이해하고 수행하도록 하는 방식으로 학습하여 신규 Task에 대해서도 주어진 Instruction을 따라 추론하게 하는데 Alpaca LLM은 LLaMA 7B를 기본모델로 두고 Instruction tuning한 모델이다.

3. 연구 방법

본 장에서는 금융 도메인에 특화된 LLM의 생성을 위한 체계적인 접근 방법을 제시한다. 파인튜닝 진행 방법은 <그림 8>에서 제시하는 절차로 진행하며 데이터 수집과 전처리 단계에서는 금융 특화 데이터셋을 선정하고, 효과적인 전처리 방법을 도입한다. 모델 선정과 파인튜닝 절차에서는 적절한 사전 훈련된 LLM인 PLM을 선정하고, 하이퍼파라미터를 조정하여 튜닝한다. 금융 분야의 특성을 고려한 파인튜닝 고려사항에서는 금융 데이터 특성, 도메인 특화 어휘, 파인튜닝 알고리즘에 대한 고려사항을 다룬다. 이후 금융 분야를 위한 LLM 구성에서는 처음부터 학습하는 방법과 이미 존재하는 모델을 튜닝하는 방법에 대한 구성 방안을 소개한다. 마지막으로, 평가 기준과 지표에서는 정량적 성능 지표와 정성적 성능 지표를 활용한 모델의 평가기준을 가이드한다. 이를 통해 금융 도메인에 특화된 LLM의 생성 방안을 전반적으로 제시하고자 한다.



〈그림 7〉 LoRA보다 향상된 QLoRA



〈그림 8〉 파인튜닝 절차

3.1. 데이터 수집 및 전처리

3.1.1. 금융 특화 데이터셋 선정

우선, 금융 분야에 특화된 데이터셋을 선정해야 한다. 이를 위해 금융 리서치 보고서, 금융 뉴스, 금융 시장 거래 내역 등 다양한 소스에서 수집된 데이터를 활용할 것이다.

금융 데이터셋은 금융 관련 텍스트, 코드, 이미지, 오디오 등 다양한 형태로 존재할 수 있으며 금융 특화 데이터셋을 선정할 때는 다음과 같은 요소들을 고려해야 한다.

금융 특화 데이터셋을 선정하는 방법은 크게 두 가지로 나눌 수 있다.

첫 번째 방법은 기존에 공개된 데이터셋을 활용하는 방법이다. 금융 분야에서 공개된 데이터셋으로는 <표 4>와 같은 것들이 있다.

두 번째 방법은 자체적으로 데이터셋을 구축하는 방법이다. 자체적으로 데이터셋을 구축하는 경우, 활용하고자 하는 목적에 맞는 데이터를 수집하고 정제할 수 있다는 장점이 있다. 그러나 데이터 수집 및 정제에 많은 시간과 노력이 소요될 수 있으며 자체적으로 데이터셋을 구축할 때는 다음과 같은 사항들을 고려해야 한다.

1) 데이터 수집 방법: 데이터는 인터넷, 데이터베이스, 센서 등 다양한 방법으로 수집할 수 있다. 활용하고자 하는 목적에 맞는 데이터 수집 방법을 선택해야 한다.

2) 데이터 정제 방법: 데이터는 수집 과정에서 오류나 편향이 있을 수 있다. 데이터 정제를 통해 오류나 편향을 제거해야 한다.

〈표 4〉 금융특화 공개 데이터셋

데이터셋	내용	데이터 특성
금융 뉴스	금융 뉴스 기사를 수집하여 만든 데이터셋 : 금융 시장의 동향 및 흐름을 파악하는 데 활용	• 데이터의 종류 : 텍스트 • 데이터의 양 : 10만 건 이상 • 데이터의 품질 : 금융 뉴스에 대한 감성 라벨이 부여된 데이터
금융 문서	용어집, 금융 보고서, 계약서, 규정 등 금융 관련 문서를 수집하여 만든 데이터셋 : 금융 상품 및 서비스에 대한 이해를 높이는 데 활용	• 데이터의 종류 : 텍스트 • 데이터의 양 : 100만 건 이상 • 데이터의 품질 : 금융 상품에 대한 설명 및 고객 정보가 포함된 데이터
금융 코드	금융 관련 소프트웨어 코드를 수집하여 만든 데이터셋 : 금융 시스템을 이해하고 자동화하는 데 활용	• 데이터의 종류 : 텍스트 또는 코드 • 데이터의 양 : 10라인 이상 • 데이터의 품질 : 금융 관련 로직 구현 및 알고리즘 코드 데이터

이러한 데이터셋을 활용하면 금융 뉴스의 감성을 보다 정확하게 분석할 수 있으며 고객의 특성에 맞는 금융 상품을 보다 효과적으로 추천할 수 있다.

3.1.2. 데이터 전처리 방법

수집한 금융 데이터는 불규칙하고 복잡한 형태를 가지고 있을 것이다. 따라서 데이터를 모델 학습에 적합한 형태로 전처리하는 작업이 필요하다. 이 단계에서는 텍스트 정규화, 토큰화, 불용어 처리 및 형태소 분석 등의 기법을 활용하여 데이터를 정제하고 모델 학습에 적합한 형태로 가공한다.

금융 분야에서는 <표 5>와 같은 데이터 전처리 방법들이 주로 사용된다.

이러한 데이터 전처리 과정을 거치면, 금융 상품의 특성을 보다 정확하게 이해할 수 있다.

3.2. 모델 선정 및 LLM 파인튜닝 절차

이번 섹션에서는 금융 도메인에 특화된 LLM을 만들기 위한 모델 선정과 파인튜닝의 핵심 단계에 대해 다룬다. 먼저, 적절한 사전 훈련된 LLM 모델을 선정한 후, 해당 모델을 금융 도메인에 특화 시키기 위한 파인튜닝 절차를 수행한다. 이를 통해 금융 분야의 특수성에 민감한 언어 모델을 구축할 수 있다.

3.2.1. 사전 훈련된 LLM 모델 선정

금융 분야에 특화된 LLM을 개발하기 위해서는 사전 훈련된 언어 모델을 선택해야 한다. 모델 선택 기준으로 금융 도메인에 특화된 용어 및 문맥을 이해할 수 있는 언어 이해 능력 이 있는 모델을 선택한다. 성능 및 활용 측면에서 모델의 성능과 금융 분야 응용에 적합성을 고려하며 사전 훈련 데이터셋 부분에서 모델의 선정은 해당 모델이 훈련된 데이터셋의 특성과 금융 분야의 데이터 특성이 일치하는지 확인하는 것이 중요하다.

사전 훈련된 LLM 모델을 선정할 때는 다음과 같은 요소들을 고려해야 한다.

1) 모델의 크기: 모델의 크기는 모델의 성능에 영향을 미친다. 일반적으로 모델의 크기가 클수록 성능이 향상된다.

〈표 5〉 금융분야 전처리 방법

전처리 구분	방법	설 명
텍스트 데이터 전처리	문자 정규화	특수문자, 이모티콘, 공백 등을 제거하는 과정
	단어 토큰화	문장을 단어로 나누는 과정
	불용어 제거	의미 없는 단어를 제거하는 과정
	어휘 사전 구축	단어의 의미와 특성을 사전에 저장하는 과정
코드 데이터 전처리	코드 정리	코드의 형식을 통일하고, 오류를 수정하는 과정
	코드 구조 분석	코드의 구조를 파악하고, 의미 있는 단위로 분리하는 과정
	코드 의미 분석	코드의 의미를 파악하고, 추출하는 과정
이미지 데이터 전처리	이미지 정규화	이미지의 크기, 밝기, 대비 등을 조정하는 과정
	이미지 특징 추출	이미지의 특징을 추출하는 과정
	이미지 분류	이미지를 분류하는 과정

2) 모델의 용도: 모델의 용도에 따라 적합한 모델이 다를 수 있다. 예를 들어, 금융 뉴스의 감성 분석을 수행하기 위해서는 자연어 처리에 특화된 모델이 적합하다.

3) 모델의 가용성: 모델이 공개되어 있지 않다면, 모델을 직접 학습해야 한다. 모델을 직접 학습하는 경우, 많은 시간과 자원이 소요될 수 있다.

금융 분야에서 사용할 수 있는 사전 훈련된 LLM 모델로는 <표 6>과 같은 것들이 있다.

사전 학습 모델로는 GPT-4, BERT, 또는 최근에 등장한 금융 특화 모델인 BloombergGPT, FinBERT와 모델 가용성 측면을 고려한 오픈된 SLM인 LLaM2와 Mistral-7B 같은 모델들이 후보로 고려될 수 있다. 선택한 모델은 금융 도메인에서의 특성을 충분히 학습한 것이 중요하고 최종적으로 사전 훈련된 LLM 모델을 선정할 때는 활용하고자 하는 목적에 따라 적합한 모델을 선택해야 한다.

3.2.2. 하이퍼파라미터 조정

하이퍼파라미터 조정은 모델의 성능을 극대화하기 위해 중요한 단계이다. 금융 특화 LLM을 구축하는 과정에서 주요한 하이퍼파라미터들을 적절히 조정함으로써 모델의 학습과 일반화 능력을 향상시킬 수 있다.

1) 학습률 (Learning Rate): 학습률은 모델이 가중치를 얼마나 크게 업데이트할지를 결정하는 요소이다. 초기 학습률을 적절히 설정하여 수렴 속도를 조절하고, 발산이나 수렴 부족 등의 문제를 예방한다(예시: 초기 학습률을 0.0001로 설정).

2) 에폭 수 (Number of Epochs): 에폭 수는 전체 데이터셋이 모델에 한 번 전달되는 횟수를 나타낸다. 너무 적은 에폭은 모델이 충분한 학습을 하지 못하게 하고, 너무 많은 에폭은 과적합의 가능성을 증가시킬 수 있다(예시: 에폭 수를 10 또는 20으로 설정).

〈표 6〉 금융 분야 특화용으로 활용 가능한 LLM

모 델	설 명
GPT-4	OpenAI에서 개발한 대규모 언어 모델로, 1,750억 개 이상의 매개변수를 가지고 있고 자연어 처리, 기계 번역, 코드 생성 등 다양한 분야에서 사용
BERT	Google AI에서 개발한 대규모 언어 모델로, 1억 개의 매개변수를 가지고 있고 자연어 처리, 질문 응답, 감성 분석 등 다양한 분야에서 사용
LLaMA2	Meta에서 개발한 대규모 언어 모델로, 70억~650억개의 매개변수를 가지고 있고 연구 및 상용화를 위해 오픈 소스로 공개하여 다양한 분야에서 사용
Falcon	UAE TII에서 개발한 대규모 언어 모델로, 70억~1,800억개의 매개변수를 가지고 있고 연구 및 상용화를 위해 오픈 소스로 공개하여 다양한 분야에서 사용
Mistral-7B	프랑스의 스타트업체인 미스트랄 AI가 매개변수 73억개의 기업용 소규모 LLM(sLLM)으로 출시하여 상용화를 위해 오픈 소스로 공개하여 다양한 분야에서 사용
BloombergGPT	Bloomberg에서 개발한 금융 특화 대규모 언어 모델로, 100억 개의 매개변수를 가지고 있고 금융 뉴스의 감성 분석, 금융 상품의 추천 등 금융 분야에서 사용
FinBERT	FinBERT는 BERT의 변형 모델로, 금융 데이터로 사전 학습되었고 금융 분야에서 자연어 처리 작업을 수행하는 데 사용

3) 배치 크기 (Batch Size): 배치 크기는 한 번에 모델이 처리하는 데이터의 양을 결정한다. 작은 배치 크기는 더 많은 업데이트를 허용하지만 메모리 사용량이 증가하므로 적절한 크기를 찾는 것이 중요하다(예시: 배치 크기를 32 또는 64로 설정).

4) 드롭아웃 비율 (Dropout Rate): 드롭아웃은 훈련 중에 일부 뉴런을 랜덤하게 제외하여 모델이 특정 패턴에 과도하게 의존하는 것을 방지한다. 적절한 드롭아웃 비율을 선택하여 모델의 일반화 능력을 향상시킨다(예시: 드롭아웃 비율을 0.2로 설정).

5) 기타 모델 관련 하이퍼파라미터: 모델에 따라 다양한 특화된 하이퍼파라미터가 존재한다. 예를 들어, Transformer 기반 모델에서는 헤드의 수나 레이어의 개수를 조정하여 모델의 복잡성과 성능을 조절한다(예시: Attention Mechanism의 헤드 수, Transformer 레이어의 개수 등).

하이퍼파라미터를 조정할 때는 실험을 통해 최적의 조합을 찾아가는 것이 일반적이며, 그에 따른 성능 변화를 지속적으로 평가하여 최적의 학습 설정을 찾아야 한다.

3.2.3. 파인튜닝 실행 환경 설정

파인튜닝 실행 환경 설정은 모델의 학습을 최적화하고 높은 성능을 달성하기 위해 필요한 단계이다. 모델을 파인튜닝하기 위해서는 적절한 실행 환경을 설정해야 한다. GPU (Graphics Processing Unit) 가속화를 포함한 고성능 컴퓨팅 자원을 사용하여 학습 속도를 향상시키고, 모델의 성능을 실시간으로 모니터링하기 위한 환경을 구축하게 된다. 이렇게 다양한 요소를 고려하여 실행 환경을 설정하면 모델의 효율성과 성능이 향상된다.

1) 하드웨어 선택: 대용량 데이터셋 및 복잡한 모델의 파인튜닝에서는 고성능의 하드웨어가 필요하다. GPU 또는 TPU (Tensor Processing Unit)를 사용하여 모델 학습 속도를 향상시키고, 대량의 데이터를 효율적으로 처리한다.

2) 분산 학습 (Distributed Training): 대규모 모델의 학습 시간을 단축하기 위해 분산 학습을 적용한다. 데이터 및 모델을 여러 장치에 나누어 동시에 학습하므로 학습 속도가 향상된다. RAY는 분산컴퓨팅을 위한 AI라이브러리를 제공하고 허깅페이스의 DeepSpeed는 초거대 모델 학습에 특화된 딥러닝 학습 최적화 라이브러리를 제공하여 1,700억개 파라미터를 가진 모델 학습도 가능하고 Accelerate를 통해 코드를 병렬화하고 DeepSpeed를 이용하여 속도 및 메모리 효율을 개선하는 방식으로 활용이 가능하다. 또한 MS의 ZeRO는 분산학습시 메모리 사용률을 최대화하는데 사용할 수 있다.

3) 가속화 기술 활용: 하드웨어에서 제공하는 가속화 기술을 활용하여 모델 학습 속도를 높인다. CUDA와 cuDNN은 NVIDIA GPU에서 딥러닝 모델을 가속화하는데 도움이 되는 라이브러리이다.

4) 배치 정규화 (Batch Normalization): 배치 정규화는 모델의 안정성을 높이고 학습 속도를 가속화하는 데 도움이 된다. 특히 딥러닝 모델에서 레이어 간의 분포를 안정화시켜 성능을 향상시킨다(예시: 모델 구조에 배치 정규화 레이어 추가).

5) 데이터 증강 (Data Augmentation): 텍스트

데이터의 다양성을 높이기 위해 데이터 증강을 활용한다. 문장 내 단어 순서 변경, 동의어 삽입 등의 기법을 사용하여 모델이 다양한 문맥을 학습하도록 도움을 준다.

6) 실험 로깅 및 모니터링: TensorBoard나 원하는 로깅 도구를 사용하여 학습 중 성능을 모니터링하는데 학습 중에 중요한 지표들을 로깅하여 모델의 성능을 실시간으로 확인하고, 학습 과정에서 발생한 문제를 신속하게 파악할 수 있다.

7) 하이퍼파라미터 튜닝: 하이퍼파라미터 최적화 도구를 활용하여 자동으로 최적의 하이퍼파라미터 조합 찾기를 하게 되는데 Grid Search 또는 Random Search와 같은 튜닝 기법을 사용하여 하이퍼파라미터를 조정하면 모델의 성능을 극대화할 수 있다.

이렇게 모델 파인튜닝 실행 환경 설정은 실험의 효율성 및 모델의 수렴을 보장하기 위해 신중한 계획과 감독이 필요하다. 설정된 환경에서 수행되는 실험 결과를 기반으로 계속해서 최적화를 진행해야 한다.

3.3. 금융 특화 LLM 파인튜닝시 고려사항

금융 분야의 LLM 파인튜닝은 다음과 같은 단계를 포함하여 진행한다.

3.3.1. 금융 데이터 특성을 고려한 도메인 특화 어휘 구축

금융 데이터는 독특한 특성을 지니고 있으며, 주식 가격의 변동, 금융 뉴스의 감성 등 다양한 측면이 모델에 영향을 미친다. 따라서, 이러한 특성을 고려하여 모델을 파인튜닝하는 것이 중요하다. 금융 데이터는 다음과 같은 특성을 가지고 있다.

1) 금융 전문 용어의 사용: 금융 데이터에는 다양한 전문 용어가 사용된다. LLM이 이러한 전문 용어를 이해하고 인식할 수 있도록 해야 한다. 예를 들어, “주가 하락”, “환율 상승”, “금리 인상” 등이 있다. 또한 금융 상품에는 “적립식 펀드”, “변액 보험”, “대출” 등 다양하게 있다. 그리고 금융 상품에는 다양한 특성이 있다. 예를 들어, “적립식 펀드는 장기 투자에 적합하다.”, “변액 보험은 저축과 보험의 기능을 모두 가지고 있다.”, “대출은 부채를 의미한다.” 등이 있다. LLM이 이러한 전문 용어를 이해하고 인식할 수 있도록, 사전 또는 어휘 사전을 통해 전문 용어 및 금융 상품에 대한 지식을 학습해야 한다.

2) 숫자의 사용: 금융 데이터에는 다양한 숫자가 사용된다. LLM이 이러한 숫자를 정확하게 처리할 수 있도록 해야 한다. 예를 들어, “코스피 지수 2,300”, “달러 환율 1,300원”, “기준금리 2.0%” 등이 있다. LLM이 이러한 숫자를 정확하게 처리할 수 있도록, 숫자 처리 관련 기능을 활용해야 한다.

3) 규칙의 복잡성: 금융 데이터는 복잡한 규칙을 가지고 있다. LLM이 이러한 규칙을 이해하고 적용할 수 있도록 해야 한다. 예를 들어, 과거 주식 시장 데이터를 기반으로 학습된 모델이 있을 때, 갑작스러운 뉴스 사건으로 인해 시장이 급락하면 모델의 예측 정확도가 크게 떨어질 수 있기 때문에 다양한 시장 상황을 반영하는 데이터를 사용하여 모델을 학습해야 한다.

3.3.2. 파인튜닝 알고리즘 적용

일반적인 텍스트 데이터뿐만 아니라 금융 데이터의 특성을 고려한 파인튜닝 알고리즘을 선택하고 적용한다. 금융 데이터는 시계열 데이터, 특정 도메인의 패턴 등을 포함하고 있기 때문에, LSTM, RNN, Attention Mechanism 등과 같은 알고리즘을 적용하여 모델이 이러한 특성을 잘 학습하도록 한다.

LSTM, RNN은 시계열 데이터를 처리하는 데 적합한 알고리즘으로 과거의 정보를 기억하여 현재의 정보를 처리하는 데 활용할 수 있다. 따라서, LSTM이나 RNN을 적용하면 모델이 과거의 시장 상황을 고려하여 현재의 시장 상황을 예측하는 데 도움이 될 수 있다.

Attention Mechanism은 특정 부분에 집중하여 정보를 처리하는 알고리즘으로 금융 데이터의 특정 패턴을 학습하는 데 활용할 수 있다. 금융 데이터에는 다양한 패턴이 존재하기 때문에, Attention Mechanism을 적용하면 모델이 이러한 패턴을 보다 정확하게 학습하는 데 도움이 될 수 있다. 예를 들어, 주가 예측에서 특정 이벤트나 뉴스가 주가에 미치는 영향을 강조하여 모델이 해당 정보에 더 집중하도록 도와줄 수 있다.

특화된 금융 알고리즘의 예로는 금융 리스크 관리에 사용되는 통계적 방법이나 옵션 가치 평가에 사용되는 Black-Scholes 모델 등이 있다. 이러한 알고리즘을 활용하여 모델이 금융 도메인의 특정 작업에 더 효과적으로 적응할 수 있다.

3.3.3. 보안 및 규정 준수

금융 데이터는 민감한 정보를 포함할 수 있으므로 데이터 수집 및 사용 시 관련된 보안 및 규정을 준수하는 것이 중요하다. 이 단계에서는 데이터 보안 및 개인정보 보호를 고려하여 데이터셋을 선정한다. 금융 분야의 LLM 파인튜닝을 수행할 때는 다음과 같은 사항을 고려해야 한다.

1) 데이터 보안: 금융 뉴스 데이터는 금융 시장의 동향을 파악할 수 있는 중요한 정보이므로, 데이터의 암호화, 접근 제어, 백업 등 데이터 보안을 위한 조치를 취해야 한다.
2) 개인정보 보호: 금융 뉴스 데이터에는 개인정보가 포함될 수 있으므로, 개인정보의 수집, 사용, 처리에 대한 규정을 준수해야 한다. 예를 들어, 개인정보를 수집할 때는 개인의 동의를 받고, 개인정보를 사용하는 목적을 명확히 해야 한다.
3) 규정 준수: 금융 뉴스 데이터를 수집하거나 사용할 때는 금융 관련 규정을 준수해야 한다. 예를 들어, 전자금융거래법, 외국환거래법 등을 준수하여 외환 거래 관련 정보를 수집하거나 사용할 수 있다. 이렇게 금융 관련 규정을 준수해야 한다

이와 같은 과정을 통해 금융 분야의 LLM 파인튜닝을 수행하면, 모델은 특정 금융 작업에 높은 정확성과 유용성을 보일 것이다. 이를 통해 금융 전문가들은 효과적으로 예측, 리서치, 보고서 작성 등 다양한 작업에 LLM의 활용을 보다 안전하게 할 수 있게 된다.

3.4. 금융 분야를 위한 LLM 구성

금융 분야에서의 자연어 처리에 특화된 모델로는 주로 BloombergGPT와 FinGPT가 주목받고 있으며 이러한 모델들은 일반적인 언어 모델의 파인튜닝을 통해 금융 용어와 도메인 특성을 학습

하여, 금융 분야의 특정 작업에 높은 성능을 보이고 있다.

3.4.1. 금융 분야를 위한 사전학습 LLM (LLM pre-trained for finance from Scratch)

처음부터 금융 도메인에서 학습된 모델로는 2023년 2월에 출시된 Fin-T5와 3월에 등장한 BloombergGPT가 있다. Fin-T5는 770M-T5 모델을 기반으로 하며, 80B Finance tokens 규모의 데이터셋으로 학습되었다. BloombergGPT는 <표 7>과 같이 금융 전문 언어를 학습한 GPT 기반 모델로, Bloomberg의 금융 데이터와 뉴스를 활용하여 사전 훈련되었고, 50B-BLOOM 모델을 기본으로 363B Finance tokens과 345B public tokens 규모의 데이터로 파인튜닝 되었다. 대부분 NER (Named Entity Recognition) 및 감정 분석(Sentiment Analysis) 같은 토큰 및 문장 분류 Task로 구성되어 있으며(Wu, et. al., 2023) 이 모델은 금융 시장 동향 예측, 뉴스 감성 분석 등의 작업에 뛰어난 성능을 보여주고 있다.

또한, 금융 뉴스 데이터로 사전학습한 모델 중 FinBERT는 BERT 아키텍처를 사용하여 금융 문맥에서의 단어, 구문, 의미를 추가로 학습하였다. 이 모델은 주로 금융 리서치나 트레이딩과 관련된 문제에 적용되며, 도메인 특화된 정보를 추출하는 데 성공적으로 활용되고 있다.

3.4.2. 금융 특화 파인튜닝 LLM (LLM fine-tuned for finance)

금융 분야에 특화된 파인튜닝이 적용된 모델로는 2023년 7월에 소개된 FinGPT와 Fin-LLaMA가 있다. FinGPT는 OpenAI의 GPT 아키텍처를 금융 분야에 특화시킨 모델로, ChatGLM-6B를 기본모델로 하여 50,000개의 샘플 데이터를 경량의 LoRA기술을 사용하여 파인튜닝하여 금융 텍스트의 특수성을 학습하였다. FinGPT는 금융 리서치, 예측 분석, 자동 트레이딩과 같은 응용 분야에서 뛰어난 성과를 보여주고 있으며 Fin-LLaMA는 LLaMA-33B를 기본모델로 적용하여 16,900개의 데이터를 Instruction 파인튜닝한 모델로, 높은 성능을 나타내고 있다. 이렇게 일반적으로 LLaMA2, Falcon, BLOOM같은 오픈소스 LLM을 기본모델로 활용하여 금융 특화된 정보를 추가하여 사전학습하거나 파인튜닝을 하는 것을 볼 수 있다. 또한 모델마다 각각의 특징과 장단점을 지니고 있는데 FinBERT는 특정 작업에 대한 성능이 우수하지만 데이터 양에 따라 성능이 크게 좌우될 수 있고 BloombergGPT는 실제 금융 데이터와 강력한 파인튜닝 메커니즘을 활용하여 실전 적용에 강점을 지닌다. FinGPT는 GPT 아키텍처를 기반으로 하되, 금융 도메인에 특화된 데이터셋과 파인튜닝을 통해 높은 성능을 달성한다. 이러한 금융 분야에서의 LLM 모델들은 향후 더욱

〈표 7〉 Evaluation Benchmarks of BloombergGPT.

Suit	Tasks	What does it measure?
Public Financial Tasks	5	Public datasets in the financial domain
Bloomberg Financial Tasks	12	NER and sentiment analysis tasks
Big-bench Hard (Suzgun et al., 2022)	23	Reasoning and general NLP tasks
Knowledge Assessments	5	Testing closed-book information recall
Reading Comprehension	5	Testing open-book tasks
Linguistic Tasks	9	Not directly user-facing NLP tasks

높은 정밀도와 효율성을 갖추어 금융 전문가들에게 실질적인 가치를 제공할 것이다.

3.5. 평가 기준 및 지표

평가 기준과 지표는 금융 특화 LLM의 성능을 정량적 및 정성적으로 평가하기 위한 핵심 요소이다. 모델의 정량적 성능을 평가하기 위해 다양한 지표를 활용하는데 정확도, 정밀도, 재현율, F1 스코어 등을 정량적으로 측정하여 모델이 특정 금융 작업에 얼마나 효과적으로 수행되었는지를 평가한다. 또한 전문가의 주관적인 평가를 수용하여 모델의 성능을 정성적으로 평가하게 되는데, 이때 모델이 금융 도메인에 대한 지식을 얼마나 잘 학습하였는지, 얼마나 효과적으로 응용되었는지 등을 종합적으로 고려한다. <표 8>은 금융 도메인에 특화된 언어 모델의 평가를 위한 기준과 지표로 다양한 측면에서 모델을 평가하는데 활용될 수 있는 평가 기준과 지표이다(Jeong, 2024).

이러한 정량적 및 정성적인 성능 지표를 종합적으로 고려하여 모델의 금융 도메인 적합성을 평가한다. 평가 결과를 통해 모델을 지속적으로 개선하고 최적화하는 데 활용하며 평가 기준과 지표는 모델의 강점과 약점을 식별하고, 최종적으로 금융 도메인에서의 유용성을 평가하는 데 도움을 줄 것이다.

4. LLM 파인튜닝 적용 및 활용 연구

본 장에서는 3장에서 소개한 LLM 파인튜닝 방법을 기반으로 구축하기 위한 적용 방법 및 구현 코드 제시 그리고 금융 특화 LLM의 활용 분야에 대하여 논의한다. 금융 분야에서 LLM의 활용 방안은 다양한 작업과 응용을 포함하고 있으며, 이를 통해 보험 금융 분야에서의 LLM의 효과적인 적용 가능성을 확인하고, 금융 도메인에서의 의사결정과 업무 효율성을 향상시킬 수 있는 분야를 제안한다.

4.1. LLM 파인튜닝 적용

본 절에서는 <그림 8>의 절차에 대한 파인튜닝 절차를 기본으로 적용하여 주요 구현된 코드를 제시하고자 한다. 적용 언어는 Python을 활용하여 구현하였으며, 모델의 weights와 biases와 같이 파라미터를 추적할 수 있는 대시보드를 제공하는 MLOps로 Wandb를 사용하였다. 학습을 위한 개발 인프라는 별도 설치 없이 GPU와 Python을 바로 적용이 가능한 Google Colab을 사용하였다. 또한 SLM은 <표 2>에서 보여준 적은 파라미터로 상대적으로 높은 성능을 보여준 Mistral 7B모델을 선택하여 사용하였다.

4.1.1. 데이터 수집 및 전처리

금융 분야에 특화된 파인튜닝을 위하여 자체 데이터셋을 구축하거나 오픈된 데이터셋을 활용할 수 있다. <그림 9>는 손해보험금융 FAQ 및 용어 데이터를 QA set형태로 전처리하여 CSV파일로 준비한 데이터이다. QA쌍으로 된 전처리 데이터인 ‘csujeong/Non_life_insurance’ 금융 데이터셋은 <그림 10>과 같이 파인튜닝 할 데이터셋으로 로드 하여 준비하였다.

〈표 8〉 금융 도메인 특화 LLM 평가기준 및 지표

평가구분	지표	내용
정량적	금융 예측 정확도 (Financial Prediction Accuracy)	주어진 금융 데이터에 대한 모델의 예측 정확도(금융 분야에서는 주가 예측이나 금융 이벤트 발생 예측과 같은 작업에서 모델의 정확도를 평가하는 것이 중요하다)
	문장 생성 정확도 (Sentence Generation Accuracy)	생성된 문장이 정답과 일치하는 정확도(금융 특화 LLM의 주요 목적 중 하나는 정확하고 의미 있는 금융 문장을 생성하는 것이므로, 생성된 문장의 정확도를 측정한다)
	감성 분석 정확도 (Sentiment Analysis Accuracy)	금융 뉴스나 리서치에 대한 감성 분석 결과의 정확도(금융 텍스트의 감성을 정확하게 분석하는 데 모델이 얼마나 효과적인지 측정한다)
	정밀도 (Precision)	긍정 클래스로 예측한 샘플 중 실제 긍정인 샘플의 비율로 계산(긍정으로 예측한 것 중에서 실제로 긍정인 비율을 나타내어 모델이 긍정으로 분류한 샘플 중 얼마나 정확한지를 측정한다)
	재현율 (Recall)	실제 긍정 클래스에 속한 샘플 중 모델이 긍정으로 예측한 비율로 계산(실제 긍정인 샘플 중에서 모델이 얼마나 많은 것을 감지할 수 있는지를 측정하여 놓치는 샘플이 적은지 확인한다)
	F1 스코어 (F1 Score)	정밀도와 재현율의 조화평균으로 계산(정밀도와 재현율 사이의 균형을 측정하는 지표로, 둘 중 어느 하나에 치우치지 않고 모델의 성능을 평가한다)
정성적	도메인 특화 용어 활용 (Domain-specific Terminology Usage)	모델이 금융 분야에서 특수 용어를 올바르게 활용하는 능력(금융 용어의 올바른 사용은 모델이 도메인 특성을 얼마나 잘 학습했는지를 나타낸다)
	도메인 특화 평가 지표 (Domain-specific evaluation metrics)	금융 분야에서 중요한 특성에 초점을 맞춘 평가 지표 도입(금융 데이터와 작업에 특화된 지표를 도입하여 모델이 도메인 특화 작업에 얼마나 적합한지를 평가한다)
	문맥 이해능력 (Context Understanding)	모델이 금융 도메인에서 특정 문맥을 얼마나 잘 이해하는지 평가(주어진 금융 문장에서 모델이 도메인 특화된 어휘 및 문맥을 얼마나 잘 파악하는지를 평가한다)
	특이한 상황 대응 능력 (Adaptability to Unusual Scenarios)	모델이 금융 분야에서 예상치 못한 상황에 대해 얼마나 적절하게 대응하는지 평가(금융 시장은 동적이며 예측할 수 없는 상황이 발생할 수 있으므로, 모델의 적응력이 중요하다)
	생성된 텍스트의 일관성 (Text Coherence)	모델이 생성한 텍스트가 일관성 있는지 평가(일관성 있는 텍스트는 금융 리포트나 예측 분석에서 중요하며, 모델의 품질을 평가하는 중요한 기준 중 하나이다)
	텍스트 생성 평가 (Text generation evaluation)	BLEU 스코어, ROUGE 스코어 등을 활용한 텍스트 생성 평가(모델이 금융 리포트, 뉴스 요약 등을 생성할 때 문법적으로 적절하고 의미 있는 결과물을 생성하는 능력을 평가한다)
	전문가의 주관적인 평가 (Subjective evaluation by experts)	금융 전문가들이 모델의 결과를 평가하고 피드백 제공(전문가들의 경험과 지식을 활용하여 모델이 특정 금융 작업에서 얼마나 실용적이고 효과적인지를 주관적으로 평가한다)

```
df = data['train'].to_pandas()
df.head(6)
```

1 to 6 of 6 entries Filter

index	QA_text	Category
0	##Question: 골프보험 알려줘? ##Answer: 골프장에서의 사고를 대상으로, 자동차 보험과는 다르게 골프에 관련된 위험을 포괄적으로 담보하는 보험으로서, 보상하는 손해는 골프의 연습 또는 경기, 지도 중에 타인의 신체에 장해를 입히거나 타인의 재물을 손괴함으로써 생긴 제3자에 대한 손해배상책임, 자기신체손해, 골프용품의 도난 파손 등의 용품손해, 홀인원/앨바트로스를 달성한 경우에 관행상 불가피하게 지출해야 하는 축하회 비용 같은 홀인원 비용등 입니다.	장기보험-물건,사물
1	##Question: 고지의무 알려줘 ##Answer: 보험계약의 체결에 있어서 보험계약자 또는 피보험자가 보험자에 대하여 중요한 사실을 고하지 않거나 중요한 사항에 대하여 부실(不實)한 고지를 하여서는 안 된다는 의무입니다.	보험용어
2	##Question: 손해보험과 생명보험의 차이 알려줘 ##Answer: 손해보험은 재산상의 손해를 보험의 목적으로 하며 실손 방식으로 보상함. 생명보험은 사람의 생존과 사망을 보험의 목적으로 하며 정액 방식으로 보상함.제3 보험은 상해, 질병, 간병과 관련된 상품을 말하며, 생 손보사에서 모두 판매 가능함.	보험용어
3	##Question: 골프클럽하우스의 가입업종 알려줘? ##Answer: 골프클럽하우스의 가입업종은 골프클럽하우스를 적용합니다. 골프클럽하우스 건물 및 골프장 구내에 있는 모든 시설에 대하여 적용합니다.	장기보험-물건,사물
4	##Question: 단체보험 가입자의 개인실손 전환 무심사 조건중 10대 중대질병 알려줘? ##Answer: 단체실손 가입자의 개인실손 전환 무심사 조건 중 10대 중대질병 미발생자가 의미하는 바는 다음과 같습니다. ①암 ②백혈병 ③고혈압 ④협심증 ⑤심근경색 ⑥심장판막증 ⑦간경화증 ⑧뇌졸중증(뇌출혈, 뇌경색) ⑨당뇨병 ⑩에이즈(AIDS)및 HIV로균이 미발생한 자	장기보험-대인,사람
5	##Question: 운전면허 미보유 보험가입 어떻게 하는지 알려줘? ##Answer: 피보험자가 등록증상의 소유자이면, 피보험자의 면허여부는 상관 없이 보험가입 가능합니다.	자동차보험

Show 25 per page

〈그림 9〉 금융 데이터셋 준비

```
data = load_dataset("csujeong/Non_life_insurance")

data
```

Downloading data: 100% 186k/186k [00:01<00:00, 130kB/s]

Generating train split: 544/0 [00:00<00:00, 13906.96 examples/s]

```
DatasetDict({
    train: Dataset({
        features: ['QA_text', 'Category'],
        num_rows: 544
    })
})
```

〈그림 10〉 파인튜닝 할 데이터셋 로드

4.1.2. PLM 선정

사전 학습된 언어 모델인 PLM선정을 본 연구에서는 모델의 가용성 측면을 고려하여 연구 및 상용화를 위해 오픈 소스로 공개하여 다양한 분야에서 사용될 수 있는 Mistral 7B모델을 선정하여 진행하였다.

<그림 11> 코드는 Mistral-7B 모델을 로드하고, 이를 BitsAndBytes 설정에서 nf4로 양자화한 모델에 LoRA를 모델에 삽입하는 QLoRA를 구현하여 메모리 효율적으로 처리하도록 구성하는 과정을 나타내고 있다.

```
model_name = "mistralai/Mistral-7B-v0.1"  # Mistral-7B model

bnb_config = BitsAndBytesConfig(
    load_in_4bit=True,              # 4비트 정밀도로 모델 로드
    bnb_4bit_quant_type="nf4",      # pre-trained model은 4비트 NF 형식으로 양자화 되어야 함
    bnb_4bit_use_double_quant=True, # QLoRA 에서 언급한 이중 양자화 사용
    bnb_4bit_compute_dtype=torch.bfloat16, # Pre-trained model은 BF16 형식으로 로드되어야 함
)

model = AutoModelForCausalLM.from_pretrained(
    model_name,
    quantization_config=bnb_config, # bitsandbytes config 사용
    device_map="auto",  # "auto" -> HF Accelerate가 모델의 각 레이어에 배치할 GPU를 결정
    trust_remote_code=True, # Mistral-7B 모델을 사용을 위한 True값 설정
)
```

config.json: 100% 571/571 [00:00<00:00, 41.4kB/s]

model.safetensors.index.json: 100% 25.1k/25.1k [00:00<00:00, 1.75MB/s]

Downloading shards: 100% 2/2 [02:05<00:00, 59.04s/it]

model-00001-of-00002.safetensors: 100% 9.94G/9.94G [01:20<00:00, 181MB/s]

〈그림 11〉 PLM 선정

4.1.3. 하이퍼파라미터 설정 및 파인튜닝 학습

<그림 11>에서 양자화를 위한 준비가 되었으면 하이퍼파라미터 Lora Alpha 및 Rank 값 등 QLoRA관련하여 LoraConfig를 설정하고 PEFT 모델 가져오기 위해 <그림 12>와 같이 적용한다.

```
#레이어에 어댑터 추가
model = prepare_model_for_kbit_training(model)

lora_alpha = 32 # 가중치 matrices를 위한 scaling factor
lora_dropout = 0.05 # LoRA 레이어의 드롭아웃
lora_rank = 32 # Low-Rank matrices dimension

peft_config = LoraConfig(
    lora_alpha=lora_alpha,
    lora_dropout=lora_dropout,
    r=lora_rank,
    bias="none",  # 편향 대신 가중치 매개변수만 훈련하는 경우 'none'값 설정
    task_type="CAUSAL_LM",
    target_modules=["q_proj", "k_proj", "v_proj", "o_proj","gate_proj"]
)
peft_model = get_peft_model(model, peft_config)
```

```
output_dir = "/content/gdrive/MyDrive/LLM/Mistral-7B-Finetuning-Insurance"
per_device_train_batch_size = 2 # 메모리 부족 오류가 발생하면 배치 크기를 2배로 줄
gradient_accumulation_steps = 2  # 배치 크기가 줄어들면 기울기 누적 단계가 2배 증가
optim = "paged_adamw_32bit" # 더 나은 메모리 관리를 위해 페이징을 활성화
save_strategy="steps" # 학습 중에 채택할 체크포인트 save strategy
save_steps = 10 # 두 개의 체크포인트가 저장되기 전의 업데이트 단계 수
logging_steps = 10  # 두 로그 사이의 업데이트 단계 수
learning_rate = 2e-4  # AdamW 최적화 프로그램의 학습률
max_grad_norm = 0.3 # 최대 그라데이션 표준(gradient clipping)
max_steps = 60        # 60단계 동안 학습
warmup_ratio = 0.03 # 0에서 learning_rate까지 선형 준비에 사용되는 단계 수
lr_scheduler_type = "cosine"  # 학습률 스케줄러
```

〈그림 12〉 하이퍼파라미터 설정

이 코드는 모델의 파인튜닝을 수행하는 동안의 학습 설정을 정의한 주요 학습 관련 하이퍼파라미터와 저장, 로깅, 그리고 최적화 관련 설정이 포함되어 있다. 이러한 설정들은 모델 파인튜닝을

위한 학습 설정을 제어하며, 하이퍼파라미터 및 학습 전략에 관련된 다양한 요소를 조정할 수 있다. 또한 <그림 13>은 Hugging Face의 TRL(Transformer Reinforcement Learning) 라이브러리가 사용자 친화적인 API를 제공하여 최소한의 코딩으로 데이터셋에서 지도학습 파인튜닝(Supervised Fine-Tuning, SFT) 모델을 생성하고 훈련할 수 있도록 SFTTrainer에 필요한 구성 요소들을 제공하고, 이는 모델, 데이터셋, Lora 설정, 토크나이저, 그리고 훈련 매개변수 등을 포함한다.

```
trainer = SFTTrainer(
    model=peft_model,
    train_dataset=data['train'],
    peft_config=peft_config,
    dataset_text_field="QA_text",
    max_seq_length=1024,
    tokenizer=tokenizer,
    args=training_arguments,
)

for name, module in trainer.model.named_modules():
    if "norm" in name:
        module = module.to(torch.float32)
```

Map: 100% 541/541 [00:00<00:00, 2302.02 examples/s]

〈그림 13〉 SFT parameters

마지막은 <그림 14>와 같이 모델 학습의 순서로 이루어 지는데 PLM에 추가적인 데이터셋을 학습하는 파인튜닝 실행 코드 및 결과이다.

```
peft_model.config.use_cache = False
trainer.train()
```

```
wandb: Logging into wandb.ai. (Learn how to deploy a W&B server locally: https://wandb.me/wandb-server)
wandb: You can find your API key in your browser here: https://wandb.ai/authorize
wandb: Paste an API key from your profile and hit enter, or press ctrl+c to quit: · · · · · · · · · ·
wandb: Appending key for api.wandb.ai to your netrc file: /root/.netrc
Tracking run with wandb version 0.16.1
Run data is saved locally in /content/gdrive/MyDrive/LLM/Mistral-7B-Finetuning-Insurance/wandb/run-20240107_083215-enryt6zo
Syncing run sparkling-silence-7 to Weights & Biases (docs)
View project at https://wandb.ai/llm_git/huggingface
View run at https://wandb.ai/llm_git/huggingface/runs/enryt6zo
You're using a LlamaTokenizerFast tokenizer. Please note that with a fast tokenizer, using the `__call__` method is faste
[60/60 11:06, Epoch 0/1]
```

Step	Training Loss
10	1.741400
20	1.526900
30	1.308000
40	1.428400
50	1.412400
60	1.293700

```
TrainOutput(global_step=60, training_loss=1.4518062591552734, metrics={'train_runtime': 756.7394,
'train_samples_per_second': 0.317, 'train_steps_per_second': 0.079, 'total_flos': 1588809031680000.0, 'train_loss':
1.4518062591552734, 'epoch': 0.44})
```

〈그림 14〉 파인튜닝 학습 실행

이 코드는 PEFT 모델의 캐시 사용 여부를 비활성화하고, 그 후에 트레이너를 사용하여 모델을 훈련시키는 과정을 나타내는 것으로 <그림 12> 하이퍼파라미터 설정에서 'max_steps = 60'으로 설정하여 학습이 진행되는 총 스텝 횟수가 60이 완료된 것을 볼 수 있으며, 파인튜닝된 'Mistral-7B-Fine-tuning-Insurance' FLM이 로컬 드라이브에 저장된 것을 확인할 수 있다. 이렇게 생성된 FLM은 관련업무에 맞게 자유롭게 로컬에서 활용 수 있다.

4.1.4. FLM 테스트

생성된 FLM을 테스트하기 위해 <그림 15>와 같이 PEFT FLM을 로드 하는 과정을 나타내고 있다. FLM 모델에 <그림 16>과 같이 질문하여 정상적으로 모델이 작동하는 것을 확인할 수 있다.

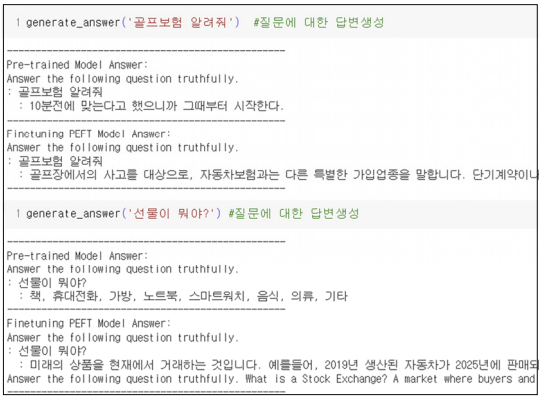
```
# Loading PEFT model
PEFT_MODEL = "csujeong/Mistral-7B-Finetuning-Insurance"

config = PeftConfig.from_pretrained(PEFT_MODEL)
peft_base_model = AutoModelForCausalLM.from_pretrained(
    config.base_model_name_or_path,
    return_dict=True,
    quantization_config=bnb_config,
    device_map="auto",
    trust_remote_code=True,
)

peft_model = PeftModel.from_pretrained(peft_base_model, PEFT_MODEL)

peft_tokenizer = AutoTokenizer.from_pretrained(config.base_model_name_or_path)
peft_tokenizer.pad_token = peft_tokenizer.eos_token
```

adapter_config.json: 100% 646/646 [00:00<00:00, 28.5kB/s]
Loading checkpoint shards: 100% 2/2 [01:08<00:00, 31.81s/it]
adapter_model.safetensors: 100% 185M/185M [00:08<00:00, 21.8MB/s]

〈그림 15〉 PEFT Model 로드

사전 학습된 PLM과 PEFT 튜닝된 PLM에 '골프보험 알려줘'라는 질문을 했을 때 파인튜닝되기 전 PLM 모델에서는 보험 내용이 아닌 전혀 다른 답변을 생성한 것 비해 FLM에서는 보험에 특화된 데이터셋으로 PEFT 파인튜닝된 모델에서는 보험에 관련된 내용으로 학습된 답변(예시: "골프장에서의 사고를 대상으로,")을 생성하였으며, 또

다른 질문인 “선물이 뭐야?” 라고 질문했을 때도 PLM에서는 일반적인 책, 휴대전화 등과 같은 제품 선물에 대한 답변을 했으나 FLM PEFT 모델에서는 금융 용어에 적합한 “미래의 상품을 현재에서 거래하는 것입니다.”라고 생성해주는 것을 확인할 수 있었다.



〈그림 16〉 PEFT Model 테스트

지금까지 PLM에 금융 특화 데이터셋 추가하여 FLM으로 파인튜닝하는 방법과 결과를 검증하는 방법을 알아보았다. 본 구현 사례를 참조한다면 도메인의 분류에 따라, 그리고 업무 크기 정도에 따라 업무에 맞게 관련된 LLM을 다양하게 파인튜닝하여 FLM을 만들어 업무에 사용할 수 있을 것이다.

4.2. 금융 LLM의 활용 분야 및 사례

파인튜닝된 금융 도메인 특화 PLM은 아래와 같이 다양한 금융 분야에서 활용될 수 있으며, 실제 많은 기업에서 도입을 검토하거나 서비스를 하고 있다.

4.2.1. 고객 응대 및 서비스 향상

고객 응대 및 서비스 향상 분야에서는 고객의 문의에 대한 응대, 금융 상품에 대한 설명 등을 통해 고객의 만족도를 향상시킬 수 있다. 자동 응답 시스템은 금융 기관에서는 다양한 고객 문의에 대응해야 한다. LLM을 활용하여 자동 응답 시스템을 구축함으로써, 고객의 질문에 빠르고 정확하게 응답할 수 있다. 이를 통해 고객 서비스의 응답 시간을 단축하고 효율성을 높일 수 있다. 또한 금융 상품 권장 업무에 사용될 수 있다. LLM은 고객의 금융 이력 및 선호도를 학습하여 맞춤형 금융 상품을 추천할 수 있다. 고객에게 개인화된 금융 상품을 제안함으로써 고객 만족도를 높이고 기관의 서비스 품질을 향상시킬 수 있다.

대표적인 활용 사례로는 KB국민카드와 KBpay에서 운영하는 마케팅 이벤트 정보를 일상 대화하듯 쉽고 빠르게 알려주는 ‘이벤트 Q&AI’ 서비스로 RAG 기반 LLM 활용하였고, 인터넷전문은행인 토스도 앱에서 ‘챗GPT에 물어보기’ 기능을 서비스하고 있다(최광민, 2023b). 농협은행의 ‘아르미AI’는 AI 콜봇을 활용해 고객 만족도 조사를 자동화하고, 조사 결과에 대한 통계 추출 및 분석까지 자동으로 진행하며(장갑수, 2023), JP Morgan의 ‘IndexGPT’는 고객 요청에 따라 금융 자산을 분석하고, 투자 의사 결정을 지원하는 AI 투자상담사 서비스를 하고 있다(장갑수, 2023).

4.2.2. 금융 예측 및 트레이딩

금융 예측 및 트레이딩 구축 분야에서는 주가 예측, 금융 뉴스 감성 분석 등을 통해 금융 시장의 동향을 예측하고, 이를 바탕으로 투자 전략을 수립하거나 트레이딩을 수행할 수 있다. 주가 예측은 LLM을 활용하여 주가 예측 모델을 개발할

수 있다. 예를 들어, 모델은 금융 뉴스, 기업 보고서, 시장 동향 등을 종합적으로 분석하여 향후 주가의 추이를 예측한다. 이를 통해 투자자들은 더 나은 의사결정을 할 수 있게 된다. 금융 뉴스 감성 분석은 금융 뉴스의 감성 분석을 통해 시장 참여자들의 감정을 파악할 수 있다. LLM은 금융 뉴스에서 나타나는 감정을 분석하여 투자자들에게 시장 동향에 대한 정보를 제공한다. 긍정적인 뉴스일 경우 주가 상승이 예상되는 등의 정보를 제공하여 트레이딩 전략에 활용할 수 있다.

대표적인 활용 사례로는 미래에셋증권이 챗GPT를 활용해 종목의 시황을 요약하는 서비스를 도입하여 서비스중이다(임지윤, 2023).

4.2.3. 금융 리서치 및 정보 추출

금융 리서치 및 정보 추출 분야에서는 금융 관련 데이터에서 유익한 정보를 추출하여 금융 리서치에 활용하거나, 고객의 투자 성향에 맞는 금융 상품을 추천할 수 있다. LLM을 활용하여 금융 리서치에 필요한 정보를 효율적으로 추출할 수 있다. 모델은 다양한 소스에서 금융 정보를 수집하고 종합하여 리서치 보고서를 작성한다. 이를 통해 금융 전문가들은 최신 정보에 빠르게 접근하고 더 효과적인 의사결정을 할 수 있다.

대표적인 활용 사례로는 한국투자증권의 AI 기반 리서치 서비스인 ‘AIR 상장지수펀드(ETF·Exchange Traded Fund)’를 서비스(임지윤, 2023)와 DB손해보험의 AI를 활용해 관계 데이터를 학습한 후 협의자와 공모 관계를 파악, 보험사기를 잡아내는 방식을 적용한 서비스이다(김세관, 2022).

4.2.4. 자동화된 금융 문서 처리

자동화된 금융 문서 처리 분야에서는 계약서 분석, 금융 보고서 작성 보조 등을 통해 금융 업무의 효율성을 향상시킬 수 있다. 계약서 분석은 금융 기관에서는 다양한 계약서를 다뤄야 한다. LLM을 활용하여 계약서의 내용을 자동으로 분석하고 필수 정보를 추출할 수 있다. 이는 금융 업무의 효율성을 향상시키고 오류를 줄일 수 있다. 또한 금융 보고서 작성 보조역할을 할 수 있다. 금융 보고서 작성은 전문적인 지식과 시간이 많이 소요되는 작업이다. LLM을 이용하여 금융 보고서 작성을 보조하면, 금융 전문가들은 빠르고 효과적으로 보고서를 작성할 수 있게 된다. 모델은 특정 주제에 대한 정보를 요약하고 필요한 내용을 추출하여 보고서 작성을 지원한다. 대표적인 활용 사례로는 삼성생명의 금융 특화 AI 광학문자인식(OCR) 기반 보험금 지급 업무의 자동화(최광민, 2023a)와 JP Morgan의 ‘COiN’은 AI가 기업 대출 계약서 등을 분석하여 유형을 분류하고 핵심적 문구를 추출함으로써 직원이 서류를 분석하는 시간을 줄이고 정확도를 제고하고 있다(장갑수, 2023).

이와 같이 LLM을 금융 분야에 적용함으로써 다양한 작업들을 자동화하고, 의사결정을 지원하는 효과를 얻을 수 있다는 것을 활용 사례를 통해 알 수 있었다. 이는 금융 기관들에게 경쟁 우위를 제공하고, 빠르게 변화하는 금융 시장에서의 적응력을 향상시킬 수 있을 것이다.

5. 연구결과 및 논의

본 연구에서는 금융 분야에 특화된 LLM의 파인튜닝과 다양한 활용방안에 대하여 제시하였으며

금융 데이터의 특성을 고려한 파인튜닝을 통해 모델의 성능을 향상시키고, 이를 통해 금융 예측, 자동화, 리서치, 고객 응대 등 다양한 작업에 LLM을 적용하는 방안을 모색하였다. 특히 금융 특화 LLM 생성을 위한 데이터 수집 및 전처리 절차로서 금융 특화 데이터셋 선정 및 데이터 전처리 방법과 모델 선정 및 파인튜닝 절차로 사전 훈련된 LLM 모델 선정, 하이퍼파라미터 조정, 파인튜닝 실행 환경 설정, 파인튜닝 성능평가지표와 모델 생성 방법에 대하여 단계별로 상세하게 검증결과와 함께 제시하였다. 또한 금융 분야의 LLM 파인튜닝 방법과 Mistral-7B SLM을 활용한 구현 코드를 제시하여 참조할 수 있도록 하였으며, 특히, 금융분야 적용시 고려사항에 대해서도 제안하고 금융 분야에서 LLM을 활용 분야 및 사례에 대하여 살펴보았다. 예측 정확도 향상을 위해서는 주가 예측 및 금융 뉴스 감성 분석에서 모델의 예측 정확도가 향상되어, 투자자들은 더 정확한 의사결정을 내릴 수 있도록 활용될 수 있으며, 자동화로 인한 업무 효율성 증대측면에서는 은행 업무 자동화에서는 계약서 분석과 금융 보고서 작성이 자동화되어 업무 처리 시간이 단축될 수 있고, 금융 전문가들은 보다 전략적인 업무에 집중할 수 있도록 하는데 적용할 수 있다. 또한 LLM을 활용한 금융 리서치 부분에서는 다양한 정보를 효율적으로 추출하여 리서치 보고서를 작성함으로써 금융 기관의 의사결정과 전략 수립에 도움을 줄 수 있다. 고객 응대 개선측면에서는 자동 응답 시스템과 금융 상품 권장에서는 고객 응대의 품질을 향상시키는 분야에 적용될 수 있음을 제시하였다.

하지만 본 연구에서는 파인튜닝 모델 생성 방법에 대해 주안점을 두고 있기 때문에 몇 가지 한계점이 있다. 데이터의 한정성부분으로 사용된 데이터셋은 생성방법을 가이드하기위한 최소한의 자료 이용으로 인해 모델이 특정 도메인에 완전히 커버할 수 있지 않으므로 더 다양하고 많은 대표적인 데이터셋의 확보가 필요하며, 모델의 일반화 능력이 부족한 경우가 있을 수 있다. 또한 금융 도메인 지식 부재로 모델이 일부 금융 도메인 지식을 학습하지 못한 경우가 있을 수 있기 때문에 도메인 특화된 지식을 더욱 효과적으로 전달하기 위한 방안을 모색해야 한다는 것이다.

향후 연구에서는 더 다양하고 다양한 금융 도메인의 많은 데이터셋을 확보하여 학습하고, 더 다양한 파라미터 튜닝과 심도 있는 모델 구조의 최적화하여 모델의 학습을 강화하는 방향으로 연구를 확장할 필요가 있다. 또한, 금융 도메인 지식을 더욱 강화시키는 방안을 고려하여 모델의 성능을 향상시킬 필요가 있다. 뿐만 아니라, 금융 분야에서의 LLM의 활용을 더 다양한 업무에 확장하는 연구가 필요하다. 예를 들어, 부도 예측, 투자 포트폴리오 최적화, 신용 스코어링 등의 다양한 금융 작업에 대한 탐색을 심화하면서 모델의 실용성을 높일 수 있다. 또한 금융 이외의 제조, 공공 등 다른 산업군으로 확장한 연구도 필요하다.

마지막으로, 연구에서 다루지 못한 윤리적인 측면에 대한 연구도 중요하다. 금융 분야에서의 민감한 정보 다룸에 있어서 개인 정보 보호, 공정한 의사결정 등의 측면을 강화하여 모델의 신뢰성을 높이는 방향으로 연구가 진행되어야 할 것이다.

종합적으로, 본 연구는 LLM을 금융 분야에 적용하는 데 있어서의 가능성과 한계를 탐색하였으며, 향후 특화 도메인 연구에 대한 기초를 제공하는 역할을 수행하고 기업내 금융 서비스에 LLM을 적극적으로 활용할 수 있도록 하는데 의미와 가치가 있다.

참고문헌(References)

[국내 문헌]

김세관. (2022, 10월 18일). 보험 AI 인공지능, 적용 기술과 활용 사례. 머니투데이. https://www.thedatahunt.com/trend-insight/ai-in-insurance

인공지능신문. (2023, 12월 16일). '7B의 전쟁'... LLM보다 치열해진 SLM 시장. 인공지능신문. https://www.aitimes.com/news/articleView.html?idxno=155898

임지윤. (2023, 5월 2일). 챗GPT發 증권가 'AI 열풍'… "리서치·투자도 척척". 한국금융신문. https://www.fntimes.com/html/view.php?ud=2023042903280383 0dd55077bc2_18

장갑수. (2023, 10월 31일). ChatGPT의 탄생과 진화, 생성형 AI, 은행 비즈니스 생산성 제고에 기여. 파이낸셜포커스. http://ffnews.co.kr/detail.php?number=4768&thread=28r38

정천수, 정지환. (2020). 포스트 코로나19 언택트 시대 대응을 위한 AI 챗봇 구축방법에 관한 연구, *한국IT서비스학회지*, 19(4), 31-47. https://doi.org/10.9716/KITS.2020.19.4.031

정천수. (2023a). 하이브리드 AI 챗봇 구현을 위한 RPA연계 방안 연구, *정보처리학회논문지/소프트웨어 및 데이터 공학*, 12(1), 41-50. https://doi.org/10.3745/KTSDE.2023.12.1.41

정천수. (2023b). E2E 비즈니스 프로세스 자동화를 위한 하이퍼오토메이션 플랫폼 적용방안 및 사례연구, *경영정보학연구*, 25(2), 31-56. https://doi.org/10.14329/isr.2023.25.2.031

정천수. (2023c). 전통적인 챗봇과 ChatGPT 연계 서비스 방안 연구, *한국정보기술응용학회지*, 3(4), 11-28. https://doi.org/10.21219/jitam.2023.30.4.001

정천수. (2023d). LLM 애플리케이션 아키텍처를 활용한 생성형 AI 서비스 구현: RAG모델과 LangChain 프레임워크 기반, *지능정보연구*, 29(4), 129-164. https://dx.doi.org/10.13088/jiis.2023.29.4.129

최광민. (2023, 10월 25일a). 삼성생명, "보험업계 문서 자동화 혁신!"...사람 손 필요 없는 업스테이지 AI OCR 솔루션 'DocAI'로 문서 자동화 달성. 인공지능신문. https://www.aitimes.kr/news/articleView.html?idxno=29203

최광민. (2023, 10월 25일b). KB국민카드, LLM 기반 고객 경험 혁신 위한 '이벤트 Q&AI' 베타서비스 출시. 인공지능신문. https://www.aitimes.kr/news/articleView.html?idxno=29202

[국외 문헌]

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258.*

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., ... & Fiedel, N. (2023). Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240), 1-113.

Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2024). Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems, 36.*

Dilmegani, C. (2023, June 21). Large Language Models: Complete Guide in 2023. AIMultiple. Retrieved January 12, 2024, from https://research.aimultiple.com/large-language-models/

Google DeepMind. (2024, February 21). Gemma: Open Models Based on Gemini Research and Technology. Google DeepMind. Retrieved February 21, 2024, from https://storage.googleapis.com/deepmind-media/gemma/gemma-report.pdf

Greyling, C. (2023, October 20). Large Language Model (LLM) Disruption of Chatbots. Cobusgreyling.com. Retrieved January 16, 2024, from https://cobusgreyling.medium.com/large-language-model-llm-disruption-of-chatbots-8115fffadc22

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

Jeong, C. S. (2023). A Study on the Implementation of Generative AI Services Using an Enterprise Data-Based LLM Application Architecture. *Advances in Artificial Intelligence and Machine Learning,* 3(4). 1588-1618. https://dx.doi.org/10.54364/AAIML.2023.1191

Jeong, C. S. (2024). Fine-tuning and Utilization Methods of Domain-specific LLMs. *arXiv preprint arXiv: 2401.02981.*

Jeong, J. H. and Jeong, C. S. (2022). Ethical Issues with Artificial Intelligence (A Case Study on AI Chatbot & Self-Driving Car), *International Journal of Scientific & Engineering Research*, 13(1). 468-471.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. D. L., ... & Sayed, W. E. (2023). Mistral 7B. *arXiv preprint arXiv:2310.06825.*

Mayank, S. (2023, June 30). Generative AI: Empowering Innovation with its Astonishing Capabilities. Retrieved January 15, 2024, Shuru Technologies. from https://shurutech.com/innovating-with-generative-ai/

Nijkamp, E., Pang, B., Hayashi, H., Tu, L., Wang, H., Zhou, Y., ... & Xiong, C. (2022). Codegen: An open large language model for code with multi-turn program synthesis. *arXiv preprint arXiv:2203.13474.*

OpenAI. (2024, February 16). Creating video from text. OpenAI. Retrieved February 16, 2024, from https://openai.com/sora

Raschka, S. (2023, May 20). Fine-tuning LLMs Efficiently with Adapters. Ahead of AI. Retrieved January 15, 2024, from https://magazine.sebastianraschka.com/p/finetuning-llms-with-adapters

Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., ... & Mann, G. (2023). Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564.*

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., ... & Wen, J. R. (2023). A survey of large language models. *arXiv preprint arXiv:2303.18223.*

Abstract

Domain-specialized LLM: Financial fine-tuning and utilization method using Mistral 7B

Cheonsu Jeong*

The recent release of pre-trained general-purpose LLMs (Large Language Models) has been active, but research and methods for generating domain-specific fine-tuned LLMs are lacking. This study explores approaches to the fine-tuning and utilization of domain-specific LLMs and presents the latest trends in LLMs, foundation models, and pre-training of LLMs, as well as methods for domain-specific LLM fine-tuning. In particular, because the use of language models in the financial sector is important, we specifically present the selection and pre-processing methods of financial-specific datasets, model selection and fine-tuning procedures, and considerations for financial-specific LLM fine-tuning. We discuss the construction of domain-specific vocabularies considering the characteristics of financial data and considerations for security and compliance. In the study of the application and utilization of LLM fine-tuning, we present the procedure for generating a real insurance finance domain LLM using the SLM (Small Language Model) Mistral 7B and the implementation procedure, and present cases for various financial fields. Through this, this study explores the possibility of applying LLMs to the financial domain field and proposes limitations and improvement directions, thereby presenting future research directions. Therefore, this study contributes to the application and development of natural language processing technology in the business domain field, and at the same time presents the direction of LLM utilization in various industrial fields, thereby having the meaning and value of enabling the active use of LLMs in financial services and various industries within companies.

Key Words : Financial Domain LLM, SLM(Small Language Model), PLM(Pre-trained Language Model), FLM(Fine-tuning Language Model), PEFT

Received : January 12, 2024 Revised : February 27, 2024 Accepted : February 29, 2024
Corresponding Author : Cheonsu Jeong

* Corresponding Author: Cheonsu Jeong
SAMSUNG SDS AI Automation Team
Olympic-ro 125, Songpa-gu, Seoul 05510, Korea
Tel: +82-2-6155-3114, E-mail: csu.jeong@samsung.com

저 자 소 개

정천수
현재 SAMSUNG SDS AI Automation Team에서 부장으로 재직 중이며 고려대학교에서 컴퓨터공학 석사학위와 국민대학교에서 경영정보시스템 박사학위를 취득하였다. 다수의 AI 프로젝트 구축 PM을 하였으며 컴퓨터정보학회논문지, 인터넷정보학회논문지, 정보시스템연구, 지식경영연구, 한국IT서비스학회지, 정보처리학회논문지, 정보기술응용연구, 지능정보연구, Information Systems Review, AAIML 등을 포함한 다수의 국내.외 저널에 논문을 게재한 바 있으며 주요 관심분야는 Generative AI, LLM, Hyperautomation, Digital Transformation, Conversational AI, Machine Learning, Big Data 등이다.