

경제학 분야의 소규모 언어모형 개발과 훈련*

정 재 식** · 박 대 민*** · 박 종 현**** · 홍 소 은*****

논문 초록

본 연구에서는 AI 언어모형의 경제학 분야 적용을 위한 기초 작업 과정과 응용을 실증분석한다. 경제·금융·재정에 특화된 한국어 소형언어모형(Small-scale Large Language Model; sLLM)을 개발하고, 이를 활용한 응용과정을 다룬다. Meta-llama/Llama-3.1-8B-Instruct를 베이스로, 신뢰성 있는 경제·금융 및 재정 데이터와 다양한 훈련방법을 적용한 후 경제분야 데이터 베이스에 활용하여 그 응용성과 범용성을 평가한다. 실증분석 결과, 본 연구의 소형언어모형은 경제학 관련 질문에 대해 그 맥락의 이해도가 향상되었으며, RAG(Retrieval Augmented Generation)에 적용할 경우 개선된 효과를 보여주고 있다. 이는 번역 오류와 문맥 이해 부족 등의 한계를 완화하며 AI Sovereign 문제를 해결할 수 있는 초석이 될 것으로 판단된다.

핵심 주제어: 경제·금융·재정 특화, 소형언어모형 학습, RAG(Retrieval Augmented Generation)

경제학문헌목록 주제분류: C63, C88, O33

투고 일자: 2025. 3. 24. 심사 및 수정 일자: 2025. 5. 26. 게재 확정 일자: 2025. 6. 12.

* 본 논문은 한국경제학회 2025 경제학공동학술대회 'AI(인공지능) 관련 경제환경 변화에 대한 대응' 특별세션에서 발표되었습니다.

** 교신저자, 서강대학교 경제학부 교수, e-mail: cschung@sogang.ac.kr

*** 공동저자, 서강대학교 일반대학원 경제학과 AI 금융 전공 석사, e-mail: qjirrogoa@naver.com

**** 공동저자, 서강대학교 일반대학원 경제학과 AI 금융 전공 석사, e-mail: pjh970731@gmail.com

***** 공동저자, 서강대학교 일반대학원 경제학과 AI 금융 전공 석사, e-mail: wendy3395@naver.com

I. 서 론

언어모형(Language Model)의 발전이 눈부신 속도로 진행되어 거의 모든 산업 분야에서 AI의 잠재성과 활용성이 언급되고 있다. 거대언어모형(Large Language Model; LLM)은 인간의 언어를 이해하고 생성하는 능력을 기계에 부여함으로써 자연어 처리(Natural Language Process)의 한계를 극복하고, 추론(Reasoning)의 가능성을 탐색하고 있다. ChatGPT의 경우, 기본적인 경제학 교과서 질문뿐만 아니라 대학원 고급교재까지 학습되어 있어 경제 및 금융 교육에 긍정적인 역할을 할 것으로 기대된다.

AI 발전에 따른 큰 파도가 금융시장의 변화에 큰 역할을 할 것으로 예상됨에 따라, 경제 및 금융 도메인에 특화된 한국어 언어모형의 개발 및 지속적인 훈련은 AI 시대에 필수불가결한 절차로 예상된다. 이러한 흐름에 발맞춰 최근에는 한국어 언어모형 생태계가 형성되고, 한국어 언어모형의 개발과 연구가 활발히 진행(Openko-llm-leaderboard¹⁾)되고 있다. 그러나 대부분의 언어모형은 영어를 기반으로 훈련되었으며, 한국어 처리는 주로 번역에 의존하는 경우가 많다. 이로 인해 경제 및 금융 관련 질문과 답변에서 번역 오류가 자주 발생하며, 특히 전문 용어 번역에서는 문맥에 맞지 않는 결과가 빈번히 나타난다. 이러한 한계를 극복하고 특정 도메인에 특화된 언어모형의 응용 가능성을 확인하기 위해 본 연구를 진행하였다.

본 연구는 한국어 언어모형 중에서도 경제 및 금융, 특히 재정 분야에 특화된 모형의 개발 과정, 응용 사례, 그리고 결과물의 질적 성과를 다룬다. 도메인에 특화된 언어모형의 성능을 극대화하기 위해서는 훈련 데이터의 범위를 신중히 설정하는 것이 중요하다. LLM을 이용하여 파라미터 수를 100억 개 이상으로 확장하거나 훈련 데이터를 대폭 늘리는 방법도 가능하지만, 대다수의 연구기관과 대학이 이를 위한 컴퓨팅 자원을 확보하는 데는 한계가 있다. 그리고 AI 에이전트의 중요성이 부각되고 있는 상황에서, 상대적으로 소형언어모형(Small-scale Large Language Model; sLLM)을 활용하는 것이 적합하다. 이러한 sLLM으로 특정 작업을 수행할 경우, 적절한 훈련 데이터와 도메인 선택 기준을 마련하는 것이 필요하다. 금융 분야에 특화된 언어모형이 금융 데이터로 충분히 훈련될 수 있는지, 아니면 재정 분

1) <https://huggingface.co/spaces/upstage/open-ko-llm-leaderboard> 참조(2024-12-01).

아까지 포함해야 하는지에 대한 논의가 필요하지만, 이와 관련된 선행 연구는 국내 외적으로 부재한 상황이다.

그리하여 본 연구를 통해 도메인에 특화된 sLLM을 구축할 수 있는 기초 발판을 마련하고자 한다. 활용된 훈련 데이터는 경제 및 금융과 재정 분야이다. 경제 및 금융에 관련한 필수 용어에 대한 학습을 위해 세 가지 카테고리인 한국어에서의 영어로의 번역, 영어에서 한국어로의 번역, 용어에 대한 정의를 학습 데이터로 활용하였다. 더 나아가, 재정 분야에 특화된 데이터를 추가로 학습시킴으로써 동 모형이 도메인에 대한 이해가 확장될 수 있는지 분석하였다. 마지막으로, 활용 가능성을 평가하기 위해 훈련된 언어모형을 경제학 논문에 적용하여 관련한 다양한 질문에 답할 수 있는지 분석하였으며, 그 결과는 긍정적이었다.

본 연구에서는 개방형 sLLM인 Meta-llama/Llama-3.1-8B-Instruct를 활용하였으며, 이는 비용 효율성과 실용성을 고려한 선택이다. 폐쇄형 언어모형 (Openai의 ChatGPT, Claude의 Sonnet 등)은 뛰어난 성능을 제공하지만, 높은 비용을 유발한다. 반면, sLLM은 무료로 사용 가능한 오픈소스 모형이 많아 비용 부담을 효과적으로 줄일 수 있으며, 대규모 데이터를 처리하는 데 필요한 컴퓨팅 자원과 전력 소비를 줄이는 데도 유리하다. 또한, sLLM은 상용화 가능성이 높아 특정 산업에 특화된 모형 설계나 기업 내부 시스템 통합에 적합하며, 멀티모달 데이터 처리와 맞춤형 조정에도 유연하게 대응할 수 있다. 이렇게 경량화된 설계를 바탕으로 빠른 처리 속도와 유연한 배포 옵션을 제공해 사용자 만족도를 높일 수 있다.

본 연구의 활용 가치는 다음과 같이 요약할 수 있다. 첫째, 경제학 관련 기본 질문에 대한 한국어 질문-답변을 통해 AI의 활용을 증진하여 경제학 학습의 효율성을 개선될 것으로 기대된다. 이는 경제와 금융 분야에 익숙하지 않은 사람들도 쉽게 접근하고 경제 이해도를 향상하는 데 활용될 수 있다. 둘째, 언어모형은 단순한 질문과 답변을 넘어 다양한 응용 분야의 기초가 된다. 그러나 AI가 내놓는 여러 답변 등이 정확하지 않은(Hallucination) 문제점을 최소화하여야 한다. 데이터 및 데이터 베이스(임베딩 벡터 DB 포함)가 한국어로 저장된 경우, 질문과 답변의 조율(Alignment)이 필요하다. 세 번째로, 경제와 금융 분야는 상황에 따라 다른 대응을 요구한다. 같은 인플레이션 수준과 실업률이라 할지라도, 정책금리 결정에 있어서는 인상, 인하, 혹은 유지 등 다양한 조치가 취해질 수 있다. 이와 같은 정책적 결정은 해당 국가의 경제 상황과 경제주체의 기대심리 등을 종합적으로 고려하여 내

려져야 한다.

본고는 경제학과 AI의 융합에 필수적인 과정을 기술하고, 그 결과와 한계점을 제시한다. 분석 결과, 경제 및 금융에 특화된 한국어 언어모형의 발전 가능성을 확인할 수 있었다. 이는 한국어로 훈련된 개방형 sLLM을 훈련한 과정과 현장 활용성에 주안점을 두었다. 약 80억 개의 파라미터(Parameter)를 가진 언어모형을 공개적으로 획득할 수 있는 데이터를 이용하여 훈련하였다. 또한, 개발한 모형을 기반으로 RAG에 적용함으로써, 사용자가 원하는 답변을 만들어 내었다는 긍정적인 결과를 얻었다. 이를 기반으로, 컴퓨터공학에서 진행 중인 추론학습에 본 연구의 모형이 활용될 수 있을 것으로 기대된다.

본 연구는 다음과 같이 진행된다. 제Ⅱ에서는 본 연구와 관련한 언어모형의 발전 과정과 도메인 특화 언어모형에 대한 선행연구를 검토한다. 제Ⅲ에서는 모형의 개발 과정에서 사용된 방법론을 각 단계별로 상술한다. 이를 바탕으로, 제Ⅳ에서는 직접 구축한 경제·금융·재정 도메인에 특화된 데이터와 모형의 학습 과정을 설명하고, 답변 성능을 비교하였다. 제Ⅴ에서는 경제·금융·재정 도메인 특화 한국어 sLLM에서 문서검색 능력 및 답변 능력을 실험적으로 확인하여 RAG 방법론의 활용 가능성을 살펴본다.

Ⅱ. 선행 연구

1. 트랜스포머(Transformer) 기반 언어모형

언어모형의 개발은 Vaswani, A. et al. (2017)에서 발표한 트랜스포머의 등장으로 시작되었다는 것이 일반론이다. 트랜스포머는 단어와 문맥의 장기기억(Long Memory) 특징을 자기회귀적 모형으로 분석할 경우 최적화의 문제점이 발생하는 데, 이를 해결하기 위해 고안된 모형이다. 동 모형은 셀프어텐션(Self-Attention) 메커니즘에 의존하여 문장 내의 각 단어 간의 관계를 직접 모형화함으로써 장기기억 의존성을 효과적으로 인식하고 처리할 수 있다. 더욱이, 이러한 어텐션 메커니즘을 다수의 헤드에서 동시에 수행하는 멀티 헤드 어텐션(Multi-Head Attention) 구조를 도입함으로써 다양한 관점에서 단어 간의 관계를 분석할 수 있게 되었다. 이러한 구조로 인해 자연어처리 분야는 크게 발전하기 시작했으며, 대규모 데이터를 효율

적으로 훈련할 수 있는 기반이 되었다.

Devlin, J. et al. (2018)의 BERT가 트랜스포머를 양방향으로 훈련하여 단어의 맥락을 이전과 이후에 나타나는 단어를 통해 이해할 수 있게 되었다. 이것은 이전 모형들이 텍스트를 단방향으로 처리한 것과 차별화되며, 질문 응답, 언어 추론 등의 작업에서 상당한 개선을 이루었다. Brown, T. B. et al. (2020)에 따르면, OpenAI의 GPT-3는 텍스트를 왼쪽에서 오른쪽으로 생성하는 단방향 생성 모형을 기반으로 하며, 모형 파라미터는 1,750억 개에 이른다. 2022년 11월 Chat GPT 출시 이후에는 2024년 1분기에 버전 4.0까지 출시되었으며, 그 이후 9월 12일에 출시된 o1-preview는 고급 추론과 복잡한 문제해결 능력을 갖춘 강력한 모형이다. 이는 특히 인간의 단계적 사고 과정을 모방하여 복잡한 문제를 해결할 수 있는 능력으로 주목받고 있으며, 과학, 수학, 코딩과 같은 고도의 전문 분야에서도 탁월한 성능을 보여준다. 더 나아가, DALL-E(text to image), Code Interpreter, Sora(text to video)라는 응용프로그램이 빠르게 개발되고 있다. 질문에 대한 최적의 답변을 찾기 위한 다양한 프롬프팅(Prompting) 방식도 약 1년의 짧은 기간 동안 개발된 바 있다. 여기에는 질문과 관련된 예시를 단 한 번만 제시하는 원샷 학습(One-Shot Learning), 몇 가지 예시를 제시하는 퓨샷 학습(Few-Shot Learning), 그리고 질문과 관련된 예시를 전혀 제공하지 않는 제로샷 학습(Zero-Shot Learning) 등이 포함되며, 이러한 접근법은 다른 언어모형들에 비해 탁월한 성능을 보여주고 있다. 더 나아가, 인간의 피드백으로부터 얻은 보상을 기반으로 추가적인 훈련을 시키는 RLHF(Reinforcement Learning From Human Feedback) 방식의 개발은 모형이 인간의 의도와 선호를 더욱 잘 파악할 수 있게 하였다.

언어모형의 민주화라는 가치 아래 다수의 오픈소스 모형도 속속 등장하고 있다. Touvron et al. (2023)은 Meta의 LLaMA2(Large Language Model Meta AI)는 파라미터가 공개되어, 사용자들이 다양한 방식으로 모형을 재훈련할 수 있다. LLaMA2는 연구와 상용(商用) 목적을 위해 무료로 제공되며, 70억에서 700억에 이르는 다양한 파라미터 크기로 구성되었다. GPT-3 모형보다 적은 파라미터를 사용했음에도 불구하고 일부 지표에서는 GPT-3를 능가하는 결과를 보였다. 이후에 출시된 Llama 3.1은 언어 이해와 생성 성능이 대폭 개선된 모형으로써 문맥 이해와 복잡한 요청에 대한 답변 생성 능력이 향상되었고, 사용자가 원하는 목적에 맞게 커스터마이징(Customizing)할 수 있는 새로운 프롬프트 포맷을 도입했다. 이렇게 오픈소스

로 활용이 가능한 sLLM은 높은 성능과 효율성을 유지하면서도 접근성과 사용자 맞춤형성이 뛰어나 다양한 분야에서 혁신적인 확장성을 가능하게 한다.

Jiang, A. Q. et al. (2023)의 Mistral 7B는 공개적으로 사용 가능한 데이터와 기업이 제공한 데이터로만 학습을 진행하여 저작권 문제를 해결했다. LLaMA2 이후 출시된 오픈소스 모형 중에서 가장 높은 성능을 보인 모형으로써 파라미터 개수가 더 많은 LLaMA2 13B를 모든 평가 벤치마크에서 능가했다. Mistral 7B는 더 빠른 추론을 위한 GQA(Grouped Query Attention)와 임의 길이의 시퀀스(Sequence)를 효과적으로 처리할 수 있게 하는 SWA(Sliding Window Attention)를 적용했다. 이러한 아키텍처는 추론과정에서 사용되는 불필요한 작업을 줄이며 복잡한 텍스트 시퀀스를 처리하고 생성하는 모형의 능력을 향상시켰다.

Kim, D. et al. (2023)에서 발표한 Upstage의 SOLAR는 이전까지 사용되던 기존의 복잡한 모형 확장 방법들과 다르게 새로운 기술인 DUS(Depth Up-Scaling)를 사용하여 간단하면서도 효과적인 방법을 사용했다. DUS 방식은 기존 모형의 층을 조정하고 지속적으로 사전 훈련을 적용하여 모형의 성능을 개선하는 깊이 스케일링 방법론이다. 10.7B의 훈련된 파라미터를 통해 복잡한 추론 작업인 자연어처리 작업에서 우수한 성능을 보여준다.

Google Gemini Team(2024)이 개발한 Gemini는 다중 모달 AI 모형으로, 이미지, 오디오, 비디오, 텍스트를 모두 이해할 수 있는 뛰어난 성능을 제공한다. 이는 초거대 규모의 사전학습과 포스트 트레이닝(Post-training)을 통해 언어, 코딩, 멀티모달 추론 등에서 뛰어난 성능을 보이며, 특히 Gemini Ultra는 MMLU(Massive Multitask Language Understanding)와 같은 주요 벤치마크에서 인간 전문가 수준의 성과를 달성했다. 더 나아가, 텍스트와 영상 데이터를 결합하고 그래프 신경망(GNN)을 활용하여 모달 간의 관계를 분석하는 등 복잡한 문제를 해결하는 데 최적화되어 있다. 이후, Google Gemma Team(2024)이 발표한 Gemma는 Gemini를 기반으로 개발한 경량 오픈 모형으로, 20억 및 70억 매개변수 크기로 제공되며 텍스트와 코딩 작업에서 뛰어난 성능을 발휘한다. RoPE 임베딩²⁾, RMSNorm 정규

2) RoPE(Rotary Position Embedding)는 트랜스포머 모형에서 위치 정보를 효율적으로 인코딩하기 위해 설계된 기법으로, 절대 위치 임베딩 대신 회전 변환을 활용하여 상대적 위치 정보를 보존한다. 이 방식은 각 단어 임베딩에 회전 행렬을 적용하여 상대적 위치 관계를 명확히 표현하고, 긴 문맥에서도 효과적인 성능을 발휘한다.

화³⁾, GeGLU 활성화 함수⁴⁾와 같은 최신 기술을 적용하여 효율성과 안전성을 강화했으며, 사전 훈련과 RLHF를 통해 품질과 책임성을 높였다.

2. 도메인 특화 언어모형

최근 활발하게 개발 중인 생성형 도메인 언어모형의 경우에는 조 단위 이상의 단어가 활용되어 훈련되었고, 경제 및 금융 분야에 대해서는 상당량의 문서가 활용된 것으로 추측된다. 세계 경제 흐름에 맞춰 다양한 금융상품이 출현하고, 경제여건을 기술하는 경제용어의 시대적 변천을 고려할 때 경제 및 금융 부문에 특화된 한국어 언어모형의 필요성은 여지가 없다고 판단된다.

2023년 이후 다양한 금융 부문 특화 언어모형이 소개되었다. Wu, S. et al. (2023)의 연구를 통해, 블룸버그는 비공개된 자료인 뉴스, 재무제표, 분석자료 등을 활용한 금융 부문 언어모형인 BloombergGPT를 발표했다. 금융에 훈련되지 않은 여타 모형과의 성능평가를 통해 도메인 특화 언어모형의 중요성을 강조하였다. 비공개된 자료와 270만 달러가 소요된 점을 고려할 때, 해당 모형을 미세 조정하는 것은 불가능하다. 이를 보완한 Yang, H. et al. (2023)의 FinGPT는 저비용으로 학습이 가능하고, 훈련이 용이하도록 다양한 파이프라인을 설계하였다. 기저모형으로 Llama2-7b, falcon-7b 등 여러 오픈소스 모형을 이용했고, LoRA를 이용한 미세 조정을 통해 낮은 사양의 컴퓨터에서도 활용이 가능하도록 했다. 학습한 데이터로는 Yahoo Finance, Facebook, Google Trends 등에서 얻은 금융시장 데이터를 이용하였다. 한국에서도 Donggyu Kim et al. (2022)에서는 금융 특화 언어모형인 KB-BERT를 공개하였다. 위키, 뉴스, 웹 문서, 금융상품 설명서 및 투자 리포트 문서를 사용하였고, 주로 금융 토픽 분류, 금융 특화 감성 분류, 주어진 문서와 사

3) RMSNorm(Root Mean Square Layer Normalization)은 딥러닝 모형에서 정규화를 수행하는 기법으로, 학습 안정성을 높이고 효율성을 개선하기 위해 설계되었다. 기존 LayerNorm과 달리 입력 데이터의 평균을 고려하지 않고, 각 입력의 제곱 평균(Root Mean Square)을 사용하여 정규화를 수행한다. 이는 계산량을 줄이고 특히, 언어모형에서 LayerNorm보다 더 빠르고 효율적으로 작동한다.

4) GeGLU(Gated GELU)는 활성화 함수로, Transformer 모형에서 일반적으로 사용되는 ReLU나 GELU보다 더 효율적이고 정확한 계산을 목표로 설계되었다. 입력 텐서를 두 부분으로 분리한 뒤, 한 부분에 게이팅(Gating)을 적용해 다른 부분과 곱하며, 결과적으로 정보 흐름을 조정한다.

용자 질문을 바탕으로 문서 내 정답을 찾아 제공하는 질의응답에 특화되어 있다.

보다 더 금융 데이터의 복잡성을 효과적으로 처리하기 위해 다중 소스의 정보를 통합하고, 이를 기반으로 정교한 금융 의사결정과 예측을 수행하는 모형들이 발전되었다. Yu, Y. et al. (2024)의 FINCON은 다중 에이전트 시스템을 활용해 계층적 커뮤니케이션 구조와 조건부 가치위험(CVaR) 기반의 리스크 관리 메커니즘으로 투자 결정을 최적화하고, 높은 누적 수익률과 샤프 비율을 기록하며 강력한 금융 의사결정 시스템으로 평가받았다. Ouyang, K. et al. (2024)가 개발한 MANAGER 모형은 멀티모달 데이터를 통합하고 그래프 신경망을 활용하여 비정형 데이터를 효과적으로 분석함으로써, 금융정책 변화와 시장 변동성을 예측하는 데 탁월한 성과를 보여주었다.

경제 및 금융 분야 중에서도 특히, 재정 분야에 특화된 언어모형 연구는 세금 및 금융 규제와 같은 복잡한 재정 정보를 효과적으로 처리하고 응답하도록 설계되었으며, 주로 대량의 특화된 데이터를 수집하고 미세 조정을 통해 발전되었다. Yan Zhong et al. (2024)는 세금 지능형 의사결정 모형개발을 통해 세금 QA를 전문성, 안전성, 만족도로 평가하며, ChatGLM-6B 모형을 미세조정하여 높은 정확도를 보였으나, 데이터 수집 및 검색 시간, 부정확한 정보 생성 등의 한계를 지적했다. Zhiyu Cao et al. (2024)에서는 바젤 III 규정과 같은 금융 규제를 수학적 프레임워크로 변환하고 실행 가능한 코드로 변환하는 작업을 목표로 하며, GPT-4의 우수한 규제 해석 능력을 활용하여 글로벌 금융 리스크 관리 시스템의 효율성을 제고했다. 또한, Kim, Alex G. et al. (2024)은 GPT-4가 재무제표 분석에서 주요 항목을 신속하고 정확하게 요약할 수 있음을 입증하며, 재정 전문가의 분석 수준에 근접한 성능을 보여 재정 특화 언어모형의 잠재력을 강조한 바가 있다.

Ⅲ. 방법론

본절에서는 이렇게 LLM을 경제한 분야의 지식과 말투를 훈련시키기 위한 컴퓨터 분야의 방법론을 기존 연구를 통해 약술한다. 본 연구는 제한된 컴퓨터 자원 환경에서도 효율적인 학습이 가능하도록 PEFT와 LoRA(또는 QLoRA)와 같은 경량화 기법을 활용하였다. 이러한 기법을 바탕으로 언어모형은 SFT(Supervised Fine-Tuning)와 DPO(Direct Preference Optimization) 방식으로 학습되었다. SFT와

DPO는 각각의 목적에 따라 언어모형을 다르게 훈련시키는 방법론이다. 쉽게 설명하자면, SFT는 경제학을 전혀 모르는 고등학생에게 경제학 용어를 차근차근 가르치는 과정이고, DPO는 그 학생에게 경제학 시험을 통해 어떤 답변이 A학점이고 어떤 답변이 F학점인지를 구별할 수 있도록 훈련시키는 과정이다. 더 나아가, 학습된 언어모형이 질문의 문맥을 파악하고, 그와 관련한 외부 문서를 검색해 적절한 답변을 생성할 수 있는지를 평가하기 위해 RAG 기법을 활용하였다.

1. PEFT(Parameter-Efficient Fine-Tuning)

LLM은 7B, 13B, 70B 등 상당히 많은 파라미터를 가지고 있기에 이용목적에 맞게 모형을 학습시키기 위해서는 상당한 양의 컴퓨팅 파워가 요구된다. 매번 전체 모형을 특정 목적에 따라 훈련하는 것은 현실적으로 어렵다. 따라서, LLM을 학습시키기 위해서는 파라미터를 효과적으로 학습시키는 것이 굉장히 중요하며, 그중 한 방법론이 PEFT이다. Liu, H. et al. (2022)가 다룬 PEFT는 사전학습된 모형(Pretrained Model)의 파라미터 중에서 소수만을 효율적으로 업데이트하여 전체적인 계산 부담을 줄인다. 즉, 기존 언어모형에 소수의 레이어를 추가하고, 새로운 레이어만을 훈련하여 새로운 작업(Downstream Job)에 대한 효율적인 미세 조정 방법이다. 이는 모형을 학습하고 저장하는데 필요한 메모리와 저장 공간을 극적으로 줄인다. PEFT는 sLLM의 미세 조정에도 효과적으로 활용되고 있다.

PEFT의 종류로는 Adapter Tuning과 LoRA와 같은 기법이 있다. 이러한 기법들이 훈련된 모형의 성능에 미치는 영향에 대해서는 성능이 유지된다는 주장부터 성능 저하가 발생한다는 주장에 이르기까지 다양한 의견이 존재한다. 특히, 금융과 같은 특정 분야에 이러한 방법론을 적용할 때는 성능 저하의 가능성이 우려되기도 한다. 그러나 몇백억 개의 파라미터를 추정해야 하는 작업에 필요한 장비와 예산의 한계를 고려하면, 이러한 접근법을 선택할 수밖에 없다는 것이 현실이다.

2. LoRA(Low-Rank Adaptation) & QLoRA(Quantized Low-Rank Adaptation)

LoRA는 PEFT의 일종으로, 언어모형을 특정 작업이나 도메인에 효율적으로 적용할 수 있도록 설계된 기법이다. 이 방법은 기존 모형의 모든 파라미터를 재학습

하지 않고, 트랜스포머의 셀프 어텐션 모듈 내 쿼리(Query) 및 값 프로젝트(Value projection) 층에만 소규모의 학습 가능한 구조를 추가함으로써 전체 연산량을 줄이고, 메모리 사용을 크게 줄인다. 즉, 기존의 가중치 행렬은 고정된 채, 업데이트 항(A, B)을 두 개의 저차원 행렬의 곱으로 근사하여 학습하게 된다. 이는 다음과 같은 방식으로 표현된다.

$$W' = W + \Delta W + AB$$

여기서, $W \in \mathbb{R}^{d \times k}$ 는 기존 고정된 가중치 행렬, $A \in \mathbb{R}^{d \times r}$, $B \in \mathbb{R}^{r \times k}$ 는 학습 가능한 low-rank 행렬, $r \ll d, k$ 는 연산량을 줄이기 위한 저차원 조건이다. 즉, LoRA는 기존의 큰 가중치 행렬을 건드리지 않고, $\Delta W = AB$ 형태의 보정 항만 학습함으로써 파라미터 수와 메모리 사용량을 극적으로 줄인다.

실제로 Hu, E. J. et al. (2021)은 LoRA가 RoBERTa, DeBERTa, GPT-2, GPT-3와 같은 다양한 언어모형에서 미세 조정했을 때와 동등하거나 우수한 성능을 보이는 것으로 확인하였다. 또한, 훈련 가능한 파라미터 수가 최대 10,000배 적고, 추가적인 모형의 입출력 시간(Inference Latency)을 줄이는 장점이 있다. 이처럼 LoRA는 새로운 작업에 언어모형을 적응시킬 때, Low-Rank 구조를 활용한 업데이트를 통해 훈련 가능한 파라미터의 수를 줄이면서도 모형 성능을 유지하는 데에 효과적인 방법이다.

QLoRA는 사전 학습된 모형을 양자화(Quantization)하여 메모리 사용량을 줄이고 추가적인 효율성을 얻는 방식이다. 양자화는 언어모형을 더 작은 비트(Bit)로 압축해 제한된 메모리 자원을 가진 하드웨어에서도 모형의 활용을 가능하게 한다. 이를 LoRA 파라미터의 효율적 미세 조정과 결합해 성능을 극대화한다. Dettmers et al. (2023)은 QLoRA로 미세 조정된 sLLM이 기존 SoTA(State Of The Art) 모형의 성능을 능가할 수 있음을 입증했다.

3. SFT

SFT는 지도학습 데이터를 이용하여 사전 학습된 언어모형을 더욱 발전시키는 방법론이다. 이 방식에서 사용되는 데이터는 명확한 지시(Instruction), 구체적인 입

력(Input), 그리고 그 입력에 대응하는 레이블(Output)을 포함한다. 이를 통해, 모형은 복잡한 패턴과 상황을 효과적으로 학습하고, 특정 작업에 적합하도록 파라미터가 조정됨으로써 세밀한 이해와 반응 능력이 향상된다. 결과적으로, 모형은 다양한 지시사항에 대해 정확한 이해와 적합한 반응을 생성함으로써 특정 작업의 요구사항을 만족시키고 최적화된 성능을 제공한다. 특히 본 연구에서는, 이러한 SFT 과정에서 기존 언어모형의 전체 파라미터를 조정하지 않고, LoRA 기법을 활용하여 트랜스포머의 셀프 어텐션 모듈 내 쿼리 및 값 프로젝션 층에 삽입된 저차원 행렬만을 학습 대상으로 삼았다. 이는 대규모 언어모형의 안정성과 성능을 유지하면서도, 인간 선호에 기반한 세밀한 조정을 가능하게 한다.

4. DPO

Rafailov, R. (2023)가 발표한 DPO는 언어모형을 인간의 피드백을 바탕으로 미세 조정하는 과정을 간소화한 기법이다. 기존 강화 학습 방식이 복잡한 학습 목표 설정과 보상 모형 구축에 많은 인적 자원을 필요로 한 것과 달리, 이런 복잡성을 대폭 줄인다. 즉, 보상 모형이나 별도의 강화 학습 과정 없이도 이항교차 엔트로피(Binary Cross Entropy) 목적 함수⁵⁾를 사용하여 언어모형을 직접 최적화할 수 있도록 한다. 이는 인간의 선호도 데이터를 기반으로 모형을 학습하기 위해, 참조 정책(Reference Policy)과 최적 정책(Target Policy) 간의 확률 비율을 활용하여 보상 함수를 단순화한 것이다. 이 확률 비율은 로그 값을 사용해 계산되며, 모형이 선호되는 답변을 생성하도록 학습 방향을 조정한다.

요컨대, DPO 학습의 과정을 정리하면 다음과 같다. 우선, 참조 정책(Π_{ref})을 사용하여 각 프롬프트(Prompt)에 대해 완성된 문장, 선호하는 답변(y_{Chosen}), 거절하는 답변($y_{Rejected}$)를 샘플링하고, 인간의 선호도로 레이블링하여 선호도 데이터셋 D 를 구성한다. 그리고 주어진 참조 정책 Π_{ref} 과 선호도 데이터셋 D 그리고 원하는 스케일링 하이퍼파라미터(Scaling Hyperparameter) β 로 DPO 선호강도(Temperature scaling)를 설정하고, 이에 따라 언어모형 Π_{θ} 를 최적화하여 L_{DPO}

5) 이항교차 엔트로피(Binary Cross Entropy) 목적 함수는 예측값과 실제값 간의 차이를 측정해 모형의 성능을 평가하는 데 사용되는 함수다. 이 함수는 예측 확률이 실제 레이블(0 또는 1)과 얼마나 가까운지를 계산하며, 값이 작을수록 예측이 정확하다는 것을 의미한다.

$(\Pi_{\theta}; \Pi_{ref})$ 를 최소화하는 방식이다:

$$L_{DPO}(\pi_{\theta}; \pi_{ref}) = E_{(x, y_{chosen}, y_{reject}) \in \mathcal{D}} [-\log \sigma(\beta(\log \pi_{\theta}(y_{chosen}|x) - \log \pi_{\theta}(y_{reject}|x) - \log \pi_{ref}(y_{chosen}|x) + \log \pi_{ref}(y_{reject}|x)))]$$

이를 통해, 선호되는 답변은 강화하고 거절하는 답변은 억제하며, 효율적이고 실용적인 미세 조정 방법을 제공한다. DPO 학습 역시 SFT와 마찬가지로 LoRA 기법을 활용하여 학습 가능한 파라미터만 학습했다.

5. RAG(Retrieval-Augmented Generation)

RAG는 언어모델의 한계를 보완하기 위한 접근법으로, 모델이 학습한 정보 외의 외부 지식(External Knowledge), 최신 정보나 특정 도메인에 특화된 데이터를 활용하여 답변의 정확성을 높이는 방법이다. 특히 Lewis et al. (2020)은 사전학습된 LLM과 외부 지식 소스를 결합하여, 모델이 학습하지 않은 정보에 기반한 보다 정확한 응답 생성을 가능하게 하는 RAG 구조를 제안하였다. 이는 LLM의 거짓정보 생성을 줄일 수 있는 방법으로도 주목받았으며, 이후 보안이 요구되는 문서와 언어 모델을 연결하는 중간 매개 기술로 확장되고 있다.

RAG는 일반적으로 문서 분할, 임베딩, 검색, 생성의 네 단계로 구성된다. 질문(Input)을 임베딩 벡터로 변환한 뒤, 외부 지식으로 구축된 벡터 데이터베이스에서 관련 문서를 검색하고, 이를 조건(Condition)으로 활용하여 언어모델이 최종 답변을 생성한다. 이는 SFT, DPO와 같이 모델 내부의 특정 층을 재학습하거나 수정하는 것이 아니라, Retriever와 Generator를 결합한 아키텍처 수준의 연동으로 외부 지식을 실시간으로 활용하는 구조이다. 이를 통해, 학습 시점 이후의 정보를 반영하지 못한다는 언어모델의 정적인 한계를 극복할 수 있도록 하는 보완책이다.

(1) Contextual RAG

Contextual RAG는 기존 RAG의 단점을 극복하기 위해 Anthropic社에 의해 설계된 기술로, 문서의 텍스트 조각(Chunk)에 문서 전체의 맥락(Context)을 추가하여

검색과 응답의 정확성을 높인다. 기존 RAG는 텍스트를 작은 조각으로 나누어 효율적으로 검색할 수 있도록 설계되었지만, 개별 조각이 충분한 맥락이 포함되지 않는 경우, 적절한 정보를 검색하지 못하거나 의미를 잘못 전달할 위험이 있다. Contextual RAG는 이러한 문제를 해결하기 위해, 각 조각에 해당 조각이 문서에서 어떤 의미인지 설명하는 맥락을 추가한다. 예컨대, “회사의 매출은 전 분기 대비 3% 증가했다.”라는 텍스트에 “이 조각은 2023년 2분기 ACME Corp의 SEC 보고서에서 발췌한 내용으로, 전 분기 매출은 3억 1,400만 달러였다”라는 문맥을 추가한다. 이는 Claude와 같은 LLM을 사용해 자동으로 수행되며, 이를 통해 조각별 의미와 문맥의 손실을 줄일 수 있다.

또한, Contextual BM25 기술을 활용해 TF-IDF (Term Frequency-Inverse Document Frequency)⁶⁾ 기반의 정확한 텍스트 매칭을 강화함으로써, 검색 실패율을 대폭 감소시킬 수 있다. 여기서 말하는 BM25는 정보 검색(IR, Information Retrieval) 시스템에서 텍스트 검색을 위한 랭킹 함수로, 문서와 쿼리 간의 관련성을 측정하는 데 사용하는 기술이다. 결과적으로, Contextual RAG는 단순한 임베딩 검색과 비교하여 검색 정확도를 크게 향상시킨다.

(2) Rerank

Rerank는 Contextual RAG의 검색 결과를 정제하는 기술로, 초기 검색 단계에서 반환된 조각들 중 가장 관련성이 높은 내용을 재평가하여 선택한다. 이 과정에서 사용자의 질문과 각 조각의 관련성을 다시 점수화하며, 상위 K개의 조각만을 선별한다. Cohere와 같은 Rerank 모형을 활용해 각 조각에 점수를 부여하고, 점수가 높은 조각만 모형에 전달함으로써 검색 결과의 품질을 크게 향상시킨다. 이를 통해 처리해야 할 조각 수를 줄이고, 응답 생성 과정의 비용과 대기 시간을 줄일 수 있다. 특히, 문맥이 포함된 조각(Contextualized Chunks)을 사용하는 경우 성능이 더욱 향상되며, 검색 실패율을 최대 67%까지 감소시키는 것으로 나타났다.

6) TF-IDF는 특정 문서에서 단어의 중요도를 평가하는 자연어처리 기법으로, 단어 빈도(TF: 특정 문서에서 단어가 나타나는 횟수)와 역문서 빈도(IDF: 전체 문서 수를 해당 단어가 포함된 문서 수로 나눈 값에 로그를 취한 값)의 곱으로 계산된다. 특정 문서에서 자주 등장하지만, 전체 문서에서는 드물게 나타나는 단어에 높은 가중치를 부여하여 해당 단어의 상대적인 중요성을 측정한다.

IV. 경제 · 금융 · 재정 특화 한국어 언어모형 개발

경제 및 금융 특화 한국어 언어모형의 개발 과정은 <Figure 1>에서와 같이 두 가지 단계로 설명된다. 1단계에서는 한국어로 사전학습된 모형인 Meta-llama/Llama-3.1-8B-Instruct에 SFT 학습을 진행한다. 이는 경제 및 금융에 특화된 SFT 데이터를 이용하였다. 2단계에서는 SFT 모형에 경제 및 금융 분야에 특화된 데이터를 기반으로 DPO 학습을 진행하여 EconLlama3.1-8B를 생성한다. 3단계에서는 재정 분야에 특화된 데이터를 활용하여 SFT 학습을 진행하고, 마지막 단계에서는 재정 분야에 특화된 DPO 학습을 진행하여 최종 모형인 EconFisLlama3.1-8B를 구축한다.

본 연구의 파인튜닝 실험은 Ubuntu 22.04 환경에서 수행되었으며, 하드웨어 사양은 Intel Xeon Gold 6226R 프로세서(32코어, 64스레드), 시스템 메모리 256GB, NVIDIA RTX A6000 GPU 3대(각 48GB VRAM)로 구성되었다. 모형 학습은 Python 3.12 및 PyTorch 프레임워크를 기반으로 구현되었으며, CUDA 11.7 환경에서 실행되었다. 학습은 총 4단계로 구성되었으며, 각 단계별 소요 시간은 1단계 1시간, 2단계 30분, 3단계 1시간, 4단계 30분 소요되었다.

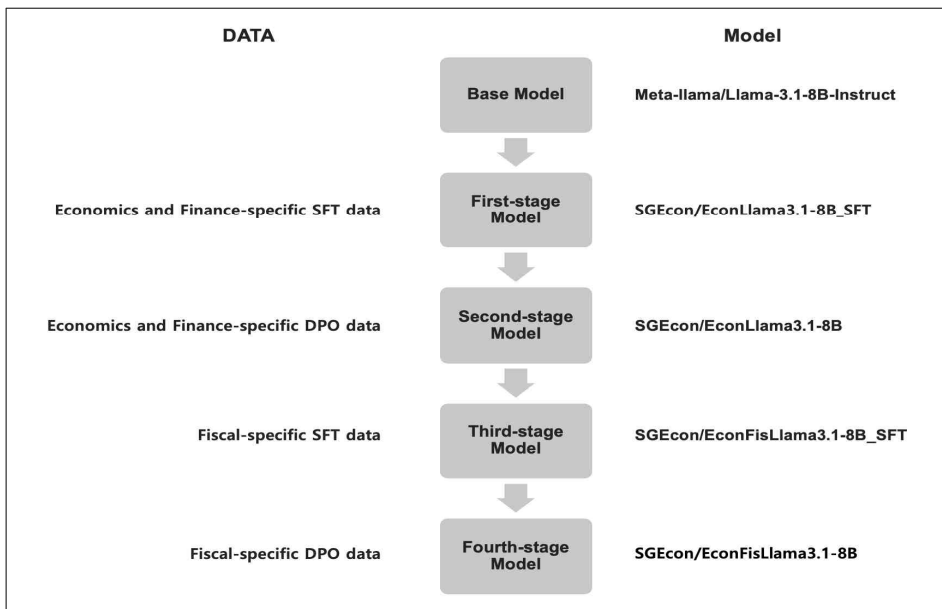
경제 및 재정 분야에 특화된 단계를 별도로 설정한 이유는 두 가지 주요 고려사항을 반영한 결과이다. 첫째, 도메인 또는 태스크(Task)에 맞춰 점진적으로 파인튜닝하는 전략이 학습 성능 향상에 기여할 수 있기 때문이다. Haoran Xu et al. (2021)은 NLP 모형을 대상으로 모형 구조나 학습 목표를 변경하지 않고도, 단계적으로 학습 데이터를 조정하는 점진적 파인튜닝(Gradual Fine-tuning) 방식이 단일 단계 학습보다 우수한 도메인 적응 성능을 달성할 수 있음을 실증적으로 제시하였다. 비록 해당 연구는 LLM이 아닌 소형 NLP 모형을 기반으로 하지만, 본 연구는 이러한 전략을 착안하여 LLM 환경에서 학습 효율성과 안정성을 높이기 위한 실험적 구조를 설계하였다.

둘째, 경제 및 금융과 재정 도메인 간에는 문장 구조, 용어 체계, 서술 스타일 등에서 뚜렷한 차이가 존재한다는 점에 주목하였다. 예컨대, 경제 및 금융 데이터는 뉴스 기사나 저널 등에서 추출되어 간결하고 용어 중심적인 문장이 주를 이루는 반면, 재정 데이터는 국가기관(한국재정정보원)에서 발행한 보고서 성격의 자료로, 정책적 문맥과 수치 정보가 혼재된 복잡한 구조를 가진다. 이러한 차이는 모형이

각 도메인별 문장 특성을 구분하여 학습할 필요성을 제기한다. Zhang et al. (2019)는 이러한 문제의식과 맥락을 공유하며 NMT (Neural Machine Training) 모형에서 도메인 간 유사도에 기반한 커리큘럼 학습 (Curriculum Learning) 전략을 통해 모형 품질을 향상시킬 수 있음을 보였다. 특히, 분포가 상이한 데이터를 무작위로 섞는 대신 유사한 도메인부터 점진적으로 확장하는 방식이 더 효과적이며, 다양한 문체와 구조를 가진 텍스트를 처리할 때 순차적 학습이 더 안정적인 성능을 제공함을 실험적으로 입증하였다. 이를 통해, 본 연구는 이와 같은 선행 연구들을 기반으로, LLM 환경에서도 순차적 도메인 학습 전략이 유효할 수 있음을 검증하고자 하였다.

더 나아가, 경제 및 금융 분야 중에서도 특히 재정에 특화된 데이터를 활용하여 언어모형을 구축하려는 이유는, 코로나 이후 경제와 재정 간의 밀접한 상호작용이 더욱 심화되었기 때문이다. 코로나에 대응하기 위해 주요 국가들이 시행한 재정 정책은 장기적인 재정 건전성에 영향을 미쳤다. 특히, 정부의 적극적인 재정 지출은 국가채무비율의 증가를 초래하였으며, 이는 재정 건전성을 약화시켜 재정 운용에 부담으로 작용하고 있다. 이러한 배경에서 재정 분야의 정확한 분석과 예측은 점점 더 중요해지고 있다.

〈Figure 1〉 Construction Process of EconFisLlama3.1-8B



1. 데이터

(1) 1단계: SFT 학습 시 사용한 경제 및 금융 특화 한국어 데이터

언어모형의 성능은 사용되는 데이터의 질에 크게 의존한다. 그리하여 이러한 과정을 거쳐 최종적으로 학습 데이터는 47,403개, 테스트 데이터는 22,717개로 구성되었다. 구축한 데이터의 설명과 출처는 아래의 <Table 1>에 기술되어 있다.

언어모형 훈련을 위해 다음과 같은 3가지 형태의 데이터를 활용하였다. 경제 및 금융 용어가 포함된 한국어 문장의 영어로의 번역, 경제 및 금융 용어가 포함된 영어 문장의 한국어로의 번역, 경제 및 금융 관련 용어 정의 설명이 해당한다. 한글-영어 번역은 총 세 단계의 과정을 거쳐 번역을 수행하였다. 먼저, 해당 용어가 실제 한국어 사용자 환경에서 어떻게 사용되는지를 네이버 지식백과의 번역을 활용하였다. 두 번째로, 해당 백과사전에서 적절한 번역어를 찾을 수 없는 경우에는 관련 용어가 사용된 특정 분야의 논문 제목이나 초록의 영어 번역을 참조하였다. 그럼에도 적절한 번역어를 찾기 어려운 예외적인 경우에 한해서는 GPT-3.5 Turbo 및 GPT-4를 보조적으로 활용하였다. 물론, GPT 계열 언어모형의 한국어-영어 번역 성능은 복잡한 문맥이나 구조적 연결이 요구되는 고난도 문장에서는 여전히 한계를 보이지만, 단순 문장이나 전문 용어 수준에서는 일정 수준 이상의 정확성과 일관성을 확보할 수 있다는 점에서 제한적으로 활용⁷⁾하였다. 따라서, 본 연구는 한-영 번역 데이터 품질의 향상을 위해 체계적인 검증 절차를 바탕으로 sLLM의 번역 성능을 제고하고자 하였다. 그리고 경제 및 금융 용어 정의에 대해서는 원본 데이터 중 1~3문장만 추출해 한국어 문장을 생성하고, 영어로 번역한 결과를 영어 번역 문장으로 구성하였다. <Table 2>에서 학습한 데이터의 예시를 확인할 수 있다. ‘지시(Instruction)’에는 사용자가 요청한 작업에 대한 설명을 제공하고 있다. 예를 들어, 번역이나 경제학 용어 설명에 해당하는 작업인지를 명확하게 밝힌 후, 질문하였다. ‘출력(Output)’은 사용자가 요청한 작업에 대한 답이다. ‘입력(Input)’은 모형의 작업

7) Lee, M. (2025)의 연구에 따르면, GPT 계열 중 GPT-4o와 같은 GPT 계열의 LLM이 한국어-영어 번역에서 기존 상용 NMT(Google Translate)보다 우수한 성능을 보인다는 실증적 결과를 제시하였다. 특히, 문맥적 이해가 중요한 한-영 번역에서는 단계별 프롬프트(CAP)를 사용할 경우 LLM의 성능이 더욱 향상되었음을 입증하였다.

수행에 있어서 주어진 지시를 어떻게 처리해야 하는지에 대한 추가적인 지침을 제공하는 부가적인 명령어이다. ‘출처(Source)’는 참조한 자료의 출처를 기재하였다.

〈Table 1〉 Train Data Descriptions

Data	Explanation	Source
700 Key Economic Terms by the Bank of Korea	This publication was issued by the Bank of Korea, the nation's central bank, for the purpose of financial and economic education. It contains key terminology essential for understanding the Bank's main functions—such as monetary policy, the real economy, financial stability, and payment and settlement systems—as well as current economic and financial terms that help readers grasp broader economic and financial trends.	https://www.bok.or.kr/portal/bbs/B0000249/view.do?nttId=235017&menuNo=200765
Glossary of Financial Terms	This is a financial glossary listed on FINE, the financial consumer information portal operated by the Financial Supervisory Service (FSS) of Korea.	https://fine.fss.or.kr/fine/fncTip/fncDicary/list.do?menuNo=900021
KDI Glossary of Economic Terms	This is a reference resource provided by the Economic Information Center under the Korea Development Institute (KDI), offering explanations of terms across various fields such as economics, finance, and business management.	https://eiec.kdi.re.kr/material/wordDic.do
Hankyung Dictionary of Economic Terms	This is an economics dictionary compiled by The Korea Economic Daily, the parent company of the Korea Economic Media Group.	https://terms.naver.com/list.naver?cid=42107&categoryId=42107
Glossary of Terms by Ministry of SMEs and Startups	This is a collection of specialized terms related to SMEs, published by the Ministry of SMEs and Startups, a government agency of South Korea.	https://terms.naver.com/list.naver?cid=42103&categoryId=42103
Glossary of Accounting and Tax Terms	This glossary, written by Professor Ko Sung-sam, is a specialized dictionary of terms in the fields of accounting and taxation.	https://terms.naver.com/list.naver?cid=51737&categoryId=51737
Word Index for Mankiw's Principles of Economics (8th Edition)	Authored by N. Gregory Mankiw and translated by Kyunghwan Kim and Jongseok Kim, this book contains key terminology essential for understanding economics.	Mankiw's Principles of Economics (8th Edition)

〈Table 2〉 Economic & Finance Specialized SFT Data Example

Input	Output	Instruction	Source	Task
The nominal exchange rate refers to the rate at which one country's currency can be exchanged for another country's currency.	The nominal exchange rate is the exchange rate between the currencies of two countries.	You are an AI assistant that helps people find information.	Mankiw's Principles of Economics	ko-en sentence translation
Please translate the following English sentence into Korean. The nominal exchange rate is the exchange rate between the currencies of two countries.	The nominal exchange rate refers to the rate at which one country's currency can be exchanged for another country's currency.	You are an AI assistant that helps people find information.	Mankiw's Principles of Economics	en-ko sentence translation
Please explain the nominal exchange rate.	The nominal exchange rate refers to the rate at which one country's currency can be exchanged for another country's currency.	You are an AI assistant that follows instructions very well. You will be given a task, and you must generate a detailed and lengthy response.	Mankiw's Principles of Economics	Explanation

(2) 2단계: DPO 학습 시 사용한 경제 및 금융 특화 한국어 데이터

본 연구의 DPO 학습에는 응답의 참조 여부와 정답 여부를 기준으로 구분된 응답 쌍이 활용되었다. ‘선호되는 응답(Chosen)’은 사전 정의된 정답 데이터 또는 외부 문서를 명시적으로 참조한 응답으로 구성되었으며, 이는 모형이 보다 정확하고 신뢰할 수 있는 응답을 선호하도록 유도하는 DPO의 목적에 부합한다. 반면, ‘선호되지 않는 응답(Rejected)’은 동일한 질문에 대해 GPT를 제외한 LLM이 자체적으로 생성한 응답으로 구성되었다. 이는 외부 문서나 데이터를 참고하지 않고 답변을 생성한 것이기에 일반적으로 정답 여부나 개념적 일관성이 떨어질 가능성이 높다. 이러한 응답 쌍은 단순한 성능 비교를 위한 것이 아니라, 다양한 품질 수준의 응답을

〈Table 3〉 Economic & Finance Specialized DPO Data Example

Question	Chosen	Rejected	System
Please translate the following English sentence into Korean. The principle of tax law, known as the Tax Law Principle, refers to the principle that all laws imposing and collecting taxes must be enacted by the National Assembly.	The Tax Law Principle refers to the principle that all laws concerning the imposition and collection of taxes must be enacted by the National Assembly.	The principle of tax law requires that all legislation related to the imposition and collection of taxes must be approved by the National Assembly.	You are an AI assistant that helps people find information.
Please explain what issuance at appraised value means in the context of bonds.	Issuance at appraised value in the context of bonds refers to issuing bonds at their face value (par value) when the bond's coupon rate is equal to the market (effective) interest rate at the time of issuance. In such cases, the bond is neither issued at a discount nor at a premium, because the present value of future cash flows matches the bond's face value. This type of issuance is also commonly referred to as issuance at par.	"Issuance at appraised value is a method by which the government issues bonds to raise funds while simultaneously reducing the fiscal deficit. When the government receives favorable evaluations from credit rating agencies, it can offer bonds at lower interest rates, allowing it to borrow money at a reduced cost."	You are an AI assistant that follows instructions very well. You will be given a task, and you must generate a detailed and lengthy response.
Please translate the following Korean sentence into English. An interim audit refers to an audit conducted during the fiscal year.	An interim audit is an audit that takes place during the interim period.	"Interim Audit"	You are an AI assistant that helps people find information.

포함시킴으로써 모형이 특정 유형의 응답 패턴에 과도하게 과적합되는 것을 방지한다. 또한, 다양한 LLM이 생성한 응답은 문법적으로 자연스럽고 유창하지만, 문서 기반 정답성과 개념 정합성이 부족할 수 있기 때문에 사용자 입장에서 선호되지 않

는 응답의 현실적인 사례로 작용할 수 있다. 요컨대, DPO 학습에 있어 선호되지 않는 응답이 문법적으로는 자연스러우나 내용적으로 부적절한 응답일수록, 선호되는 응답과의 대비가 더욱 분명해져 DPO 학습의 효과를 극대화하는 데 기여할 수 있다. 즉, 이는 보다 일반화된 선호 판단 능력을 학습할 수 있도록 설계된 전략적 구성이다.

2단계 DPO 학습에서는 1단계 SFT에서 사용된 데이터를 제외한 새로운 데이터셋을 활용하여 학습을 진행하였다. 본 단계에는 총 1,182개의 학습 쌍과 297개의 테스트 쌍이 포함되었다. 이는 각 학습 단계의 목적이 상이하기 때문에, SFT와는 구분된 데이터로 DPO 학습을 수행함으로써 모형의 성능을 보다 효과적으로 최적화하기 위함이다.

DPO 학습에 사용된 ‘선호되는 응답’은 1단계 SFT에서와 마찬가지로 〈Table 1〉에 제시된 공식 문서 및 정답 데이터를 기반으로 생성되었으며, ‘선호되지 않는 응답’은 Google의 Gemini-1.5-8b-pro, Ollama의 Mistral-nemo 12b, 그리고 Gemma-7b 모형을 활용하여 외부 문서 참조 없이 생성한 응답을 사용하였다. 이와 같은 응답 쌍 구성은 모형이 실제 사용자 선호를 보다 정밀하게 반영할 수 있도록 돕는 핵심 요소로 작용한다.

(3) 3단계: SFT 학습 시 사용한 재정 분야에 특화된 데이터

3단계 학습에서는 앞선 단계에서 경제 및 금융 분야에 특화되도록 학습한 모형을 기반으로, 재정 분야와 관련된 444개의 SFT 데이터로 모형을 재정 분야에 특화되도록 하였다. 재정정보원에서 발간한 보고서를 기반으로 〈Table 4〉와 같은 데이터를 구축하였다. 세금, 예산, 재정 규제 등과 같은 재정 분야는 국가 정책의 기본틀을 형성하고, 금융시장 이해관계자들의 재정적 의사결정에 깊은 영향을 미친다. 따라서, 모형이 재정 분야의 질문에 대해 신뢰성 있고 정확한 답변을 제공할 수 있도록 미세 조정하는 것은 경제 및 금융 도메인에서의 활용 가능성을 확대하는 데 핵심적이다.

〈Table 4〉 Fiscal Specialized SFT Data Example

Input	Output	Instruction	Source
What categories are included under non-tax revenue?	Non-tax revenue refers to income in the general and special accounts that is not derived from taxes. It includes charges for administrative services, fees for the use of government property, and revenues such as fines and penalties.	You are a public finance expert in the Republic of Korea.	Volume 1 of the 2024 Key Fiscal Statistics
What are social security contributions?	Social security contributions are revenue collected in the form of insurance premiums by the government to fund social security programs that support the stable livelihood of its citizens.	You are a public finance expert in the Republic of Korea.	Overview of Korea's National Budget 2024
In fiscal policy, how are public guaranteed debts distinguished from other one-time guarantees?	Public guaranteed debt refers to a liability in which the guarantor ensures the repayment of the principal and interest on an existing obligation. In contrast, other one-time guarantees are distinguished by allowing future access to funds through loan limits and borrowing agreements.	You are a public finance expert in the Republic of Korea.	Guide to Fiscal Statistics
What is the purpose of the Youth Job Creation Support program?	Social security contributions are revenue collected in the form of insurance premiums by the government to fund social security programs that support the stable livelihood of its citizens.	You are a public finance expert in the Republic of Korea.	FIS Issue & Focus No. 23-1 "Linked Management of Tax Expenditures"
What is the difference between the government budget and a fund?	The budget is a unified financial plan in which all revenues are pooled into a common general fund and all expenditures are made in accordance with the principle of budget unity. In contrast, a fund is financed through specific sources such as contributions or levies and is used for particular policy purposes, with its revenues and expenditures directly linked to each other.	You are a public finance expert in the Republic of Korea.	Monthly Public Finance Report
What is a supplementary budget, and when is it formulated?	A supplementary budget is a revised budget established after the original budget has been approved, in response to unavoidable circumstances that arise afterward. It is formulated during the fiscal year when additional funding is needed due to such unforeseen reasons.	You are a public finance expert in the Republic of Korea.	Ministry of SMEs and Startups - Innovative Startup Commercialization Fund (Loan) : Program Overview
How are tax revenues classified according to their purpose in Korea?	In Korea, tax revenues are classified into ordinary taxes, which are imposed to cover the general expenditures of the central or local governments, and earmarked taxes, which are levied for specific expenditure purposes.	You are a public finance expert in the Republic of Korea.	2024 Comprehensive Performance Plan

〈Table 5〉 Fiscal Specialized DPO Data Example

Question	Chosen	Rejected	System
What factors should be considered when setting strategic objectives?	When setting strategic objectives, it is essential to consider national policy goals, the mission and vision of the institution, and their logical alignment. Objectives should be formulated with concrete and specific content.	In addition to macroeconomic goals—such as economic growth, price stability, job creation, balance of payments, and improvement of social welfare—factors like fiscal soundness, social equity, and sustainability must also be taken into account. Moreover, external environmental factors including international developments, technological changes, and shifts in societal demands should be reflected. Expectations and forecasts of market participants, as well as the feasibility of policy implementation, must also be carefully considered.	As an expert in macroeconomics, finance, and fiscal policy, I emphasize that setting strategic goals requires a comprehensive and balanced perspective that integrates internal priorities with external realities.
What is the purpose of general account expenditures?	General account expenditures cover the routine spending of government ministries and agencies. They are also used to supplement shortfalls in special accounts or funds and to provide financial support to local governments and public institutions in the form of grants, subsidies, contributions, or investments.	In addition, general account expenditures are used to carry out the central government's core functions and to implement various programs aimed at promoting public welfare and improving the quality of life for citizens.	As an expert in macroeconomics, finance, and fiscal policy, I emphasize that setting strategic goals requires a comprehensive and balanced perspective that integrates internal priorities with external realities.
Why does the inclusion of non-fiscal activities such as financial operations lead to an overestimation of fiscal size in consolidated finance?	In consolidated finance, only the operating surplus of enterprise special accounts is included as revenue, while loan repayments (loan income) are excluded. As a result, non-fiscal activities such as financial operations are mistakenly included in the fiscal scope, leading to an overestimation of the government's fiscal size.	This issue arises primarily due to the lack of a clear definition of fiscal boundaries, which leads to financial market transactions being misclassified as fiscal activities, or the flows of assets and liabilities in financial institutions being incorrectly treated as fiscal revenues and expenditures. Furthermore, discrepancies in the timing and methodology of consolidated fiscal calculations can result in improper double-counting, exacerbating the distortion in the measurement of actual fiscal activity.	As an expert in macroeconomics, finance, and fiscal policy, I emphasize that setting strategic goals requires a comprehensive and balanced perspective that integrates internal priorities with external realities.

(4) 4단계: DPO 학습 시 사용한 재정 특화 한국어 데이터

4단계에서는 재정 분야 관련 질문에 대한 언어모형의 응답 품질을 높이기 위해 267개의 학습 데이터와 69개의 테스트 데이터를 활용하여 DPO 기법을 적용한 학습을 수행하였다. 선호되는 응답 데이터는 재정정보원 보고서를 기반으로 생성되었고, 국가재정 정책, 세금, 예산 분배 등 재정 분야의 다양한 질문에 적합하도록 설계되었다. 선호되지 않은 응답 데이터는 2단계에서 사용된 모형인 Google의 gemini-1.5-8b-pro, Ollama의 Mistral-nemo 12b, gemma-7b에서 생성된 응답이다.

2. 학습 결과

위와 같은 단계로 개발된 모형의 성능을 비교하기 위해 총 3가지 모형을 활용하였다. 본 연구에서 사용한 베이스 모형(Llama-3.1-8b-instruct), 경제·금융 특화 모형(EconLlama3.1-8B), 그리고 경제·금융·재정 특화 모형(EconFisLlama3.1-8B)이다. 모형들의 성능 비교를 위해 학습된 경제·금융·재정 관련 용어를 바탕으로 대학생 수준의 경제·금융·재정 지식을 평가할 수 있는 50개 문항을 동일하게 제시하였다. 해당 문항들은 사전에 다수의 후보 문항을 구성한 뒤, 경제·금융·재정 분야와 관련하여 경제학적으로 명확한 함의를 지닌 문항들을 중심으로 최종 선별했다. 선정 과정에서는 경제학 및 관련 분야 전문가의 검토와 의견 수렴을 거쳐 평가의 객관성과 학문적 타당성을 반영하였다. 즉, 경제학적으로 유의미한 문항들로 구성하여 평가 문항의 내용 타당성과 신뢰도를 확보하고, 세 모형 간의 성능을 공정하게 비교할 수 있는 기준으로 활용하고자 하였다. 본고에서 선정한 단계별 학습이 모형의 전문성을 얼마나 향상시켰는지 파악하고, 도메인 특화 학습의 유효성을 검증할 수 있다. 이후, GPT-4o를 활용한 동일한 평가 기준과 후처리 평가(프롬프트 기반 채점)를 통해 각 모형의 성능을 비교 분석하였다.

최근 자연어 생성(Natural Language Generation, NLG) 분야에서는 인간 평가자의 대리(Surrogator)로 LLM을 활용하는 자동 평가 방식이 주목받고 있다. Liu et al. (2023)의 G-Eval 연구는 GPT 기반 언어모형이 생성 응답의 품질을 정량적으로 평가할 수 있는 프레임워크를 제시하고, 기존 자동 평가 지표 대비 인간 평가와의 상관도가 더 높다는 점에서 그 타당성을 입증하였다. 또한, Wang et al. (2025)의 연구

에서는 코드 생성, 번역, 요약 업무에서 LLM이 평가한 결과(LLM-as-a-judge)가 인간 평가와 최대 81% 이상 정합되는 수준의 피어슨 상관계수를 기록하였고, 일부 작업에서는 기존 최고 성능의 정량 지표인 ChrF++⁸⁾보다도 우수한 평가 일관성을 보였음이 입증되었다. 이처럼 최근에는 전통적인 학습 평가 지표들로는 LLM 답변의 성능을 다면적으로 평가할 수 없기에 연구 시점에서 가장 성능이 좋은 LLM이 다른 언어모형들을 평가하고 피드백하는 방식을 활용한다. 이를 바탕으로, 본 연구는 Liu et al. (2023) 연구에서 공개한 코드를 참고하여 GPT-4o-mini⁹⁾를 평가 도구로 활용하였다. 아래 <Table 6>은 답변을 평가하기 위한 검증 데이터 질문 예시이다.

<Table 6> Example Questions for Evaluating Economic and Financial Reasoning

Question	Llama-3.1-8b-instruct	EconLlama3.1-8B	EconFisLlama3.1-8B
What impact does the changing proportion of deposit money have on the financial system?	Changes in the share of deposit money are an important factor in the financial system. When the proportion of deposit money in the economy decreases, it may reflect an improvement in financial maturity and an increase in lending, which can positively contribute to the productivity of goods and services and support economic growth. However, if the share of deposit money increases, it may lead to an excessive accumulation of bank deposits, potentially causing negative outcomes such as financial imbalances or even financial crises.	Changes in the share of deposit money can have the following impacts on the financial system: 1. Interest Rate Determination: As the share of deposit money increases, interest rates may decline. This is because banks can more easily replace deposit money with liquidity, making it easier to raise financial capital within the financial system. 2. Financial Capital Markets: Fluctuations in the share of deposit money can influence investment activity in financial capital markets. A higher proportion of deposit money may lead to more active investment in capital markets, while a lower share may result in reduced investment activity.	Changes in the share of deposit money can have various impacts on the financial system. When the proportion of deposit money increases, the liquidity of financial institutions improves, and public confidence in banks tends to strengthen. This, in turn, reduces the risk of bank insolvency and enhances the overall stability of the financial system. As a result, financial institutions can utilize a greater portion of deposits for funding purposes, thereby supporting economic activity. Improved deposit liquidity also enables banks to secure more capital, facilitating increased financial intermediation. Additionally, changes in the share of deposit money can influence government policy. When deposit money rises, the government may respond with policy adjustments aimed at managing deposit levels, thereby reinforcing the effectiveness of its policy tools.

8) 기계번역 품질 평가를 위해 고안된 지표로, 문자(character) 단위 n-gram F1 점수를 기반으로 하며, 단어(word) 단위 n-gram까지 포함하여 평가 정확도를 향상시킨 확장형 지표이다.

9) 수행 시점 기준 가장 우수한 공개 모형이다.

What are the benefits of fund-of-funds for diversifying an investor's portfolio?	Fund-of-funds (FoFs) offer significant benefits for diversifying an investor's portfolio, as they prevent concentrated exposure to specific industries or companies. 1. Risk Diversification: FoFs provide investors with access to a wide range of industries and asset classes, helping to reduce total portfolio risk and enhance long-term sustainability. 2. Reduced Concentration: By limiting dependence on a particular sector or company, FoFs help decrease portfolio concentration and improve overall balance. 3. Improved Profitability: Through diversified investments across various asset types, FoFs support consistent returns and contribute to the profitability of the portfolio.	The benefits of fund-of-funds (FoFs) in diversifying an investor's portfolio are as follows: 1. Reduced Investment Risk: FoFs help reduce risk by including a wide range of assets within a single portfolio. Instead of investing in each asset individually, investors can gain diversified exposure through a single FoF. 2. Increased Investor Returns: By allowing investment across various asset classes, FoFs can enhance potential returns. For example, investors can simultaneously gain exposure to both stocks and real estate, which may improve overall profitability. 3. Greater Stability of Returns: FoFs contribute to the stability of investment returns by diversifying across multiple assets, reducing reliance on the performance of any single asset class.	The benefits of fund-of-funds (FoFs) in diversifying an investor's portfolio are as follows: 1. Ease of Diversification: FoFs simplify portfolio diversification by reducing the complexity of asset allocation. Investors can gain exposure to multiple asset classes without having to purchase each asset directly. 2. Maximization of Profitability: Through professional investment strategies and risk management, FoFs enable investors to maximize their potential returns. 3. Enhanced Risk Management: FoFs provide expert-level risk management, helping to reduce portfolio volatility and enhance overall stability. 4. Improved Asset Management Capability: By leveraging the investment and risk management expertise of FoFs, investors can improve their own ability to manage assets effectively and make more informed financial decisions.
--	--	--	--

본 연구에서 GPT-4o를 활용한 프롬프트 평가 기준은 다음과 같다. 첫째, 응답이 경제 개념과 이론을 명확하게 이해하고 있는지, 둘째, 주어진 질문이나 상황에 경제 원리를 정확하게 적용했는지, 셋째, 답변이 명확하고 간결한지, 넷째, 경제 개념의 오류가 포함되지 않았는지 이렇게 네 가지 기준으로 평가를 하였으며, 특히 개념적 오류가 있는 경우 더 큰 감점을 부여했다. 이와 같은 기준으로 모형의 답변에 대해 10점 만점으로 소수점 2자리까지 점수를 부여했으며, 간단한 평가 의견도 함께 제시하도록 하였다. 아래 <Table 7>은 이에 따른 평가 점수이며, 평가 의견은 생략하였다.

(Table 7) Comparison of Model Performance (Average Scores) Evaluated by GPT-4o

Llama3.1-8b	EconLlama3.1	EconFisLlama3.1
6.95	7.15	7.23

우선, 본 모형이 도메인과 관련한 용어에 대한 정의의 정확도에 있어서 뚜렷한 개선을 보였다. 예를 들어, “예금통화(Deposit Money)의 비중 변화와 금융 시스템 안정성의 관계”에 있어 베이스 모형은 “예금통화 감소가 대출 증가와 경제성장에 긍정적 영향을 미친다.”라고 답하였는데, 이는 다소 애매한 답변임을 확인할 수 있다. 반면, 본 모형은 “예금통화 비중이 증가하면 금융기관의 유동성이 개선되어 신뢰성이 향상되며, 부실 위험 감소를 통해 금융 시스템 안정성에 기여한다.”라는 명확하고 일관된 정의를 제시함으로써, 개념적 타당성과 신뢰성을 확보하였다. 또한 예금통화 비중과 정부 정책의 연관성을 설명함으로써, 금융기관 차원의 논의를 넘어 분석을 확장하고 있다. 이는 다른 모형들이 유동성이나 금리 등 개별적인 요소에 집중하는 것과 달리, 보다 폭넓고 구조적인 시각에서 금융시스템을 해석하고 있음을 보여준다.

다만, 복잡한 경제 논리 전개나 고차원적 추론 능력을 확보하기는 어려운 것으로 판단된다. 예컨대, 복잡한 경제 모형 간 상호작용, 정책 변화의 비선형적 영향, 거시경제 지표 간 상관관계 해석 등처럼 고차원적 문제에서는 베이스 모형 대비 유의미한 개선을 관찰하기 힘들었다. 실제로, “재간접펀드가 투자자 포트폴리오 다각화에 주는 이점”에 대해 베이스 모형과 도메인 특화 모형 모두 ‘포트폴리오 수익성 확보’라는 피상적이고 직접적 인과관계가 불명확한 답변에 머물렀다. 즉, 용어 정의 위주의 학습으로는 재간접펀드의 구조적 특징, 유동성 관리, 다양한 자산 편입을 통한 위험 분산 효과 등과 같은 다면적 메커니즘을 충분히 이해하고 설명하기 어렵다는 것이다.

이를 극복하기 위해서 다음과 같은 개선 방향을 모색할 수 있다. 우선, 단순한 용어 정의를 넘어 사례 연구, 실제 경제 정책 시행 사례, 각종 경제 상황별 시나리오 등을 포함한 다양한 학습 데이터를 확보해야 한다. 또한, 경제 분야의 심층적인 이론, 학술 논문, 규제 보고서, 전문가 논평 등 전문 문헌을 적극적으로 활용함으로써 복합적 문제해결 능력을 제고해야 한다. 이를 통해, 궁극적으로 도메인 특화 모형을 단순한 용어 정의 수준에 머무르지 않고, 경제적인 추론 능력과 정책 분석

능력을 갖춘 전문화된 도메인 특화 모형으로 발전시키는 데 기여할 것이다.

또한, 본 연구에서 제안한 모형은 GPT-4o-mini와 비교하여 응답 품질에서는 정량적으로 일정 수준의 격차가 존재하지만, 일부 질문에서는 상용 모형에 준하거나 더 나은 성능을 보이기도 하였다. 가령, “손해율이 보험 상품의 가격 책정과 보험사의 수익성에 미치는 영향은 무엇인가?”와 같은 질문에서는 GPT-4o-mini보다 더 심층적이고 경제적 함의를 잘 반영한 응답을 생성한 바 있다. GPT-4o-mini는 손해율과 보험료, 수익성 간의 관계를 중심으로 “손해율이 높을 경우 보험료 인상이 필요하며, 이는 장기적으로 수익성에 악영향을 미친다”고 간결하게 서술하였다. 반면, 본 연구에서 학습된 모형은 “높은 손해율은 보험료 상승으로 이어지며 수요 감소 및 자본 감소를 초래하여 신용성과 지속 가능성에 부정적 영향을 미친다”고 응답하였는데, 이는 거시경제학적 함의를 포괄하는 분석을 제시하였다. 이는 제한된 데이터로 학습된 경량화 모형(sLLM)임에도 불구하고, 정책적인 파급 효과와 시장 구조 전반의 흐름을 포착하는 거시적 통찰력을 일부 학습할 수 있음을 시사하고 있다.

또한, 응답 속도와 비용 측면에서도 본 연구에서 개발한 소형언어모형이 실용적 이점을 보인다. 로컬 환경(A100 GPU)에서 평균 1~2초 이내의 응답을 생성하였으며, 클라우드 기반 API를 사용하는 상용 모형과 달리 별도의 호출 비용이 발생하지 않기 때문에 비용 효율성 측면에서도 강점을 가진다. 이러한 실증적 결과는 상용 LLM 대비 성능 격차가 존재함에도 불구하고, 비용과 제한적인 리소스의 한계를 극복하면서 일반인도 특정 분야에 맞춤형 도형을 사용할 수 있다는 선택적 대안으로 활용될 수 있음을 시사한다. 향후 보다 다양한 도메인과 문항에 대한 성능 평가 및 고도화를 통해 실질적인 활용 가능성을 더욱 확대할 수 있을 것으로 기대된다.

V. 경제·금융·재정 특화 한국어 언어모형에서의 RAG 활용

언어모형의 유용성과 필요성에 관련하여 많은 사람들이 동의하고 있지만, 의문점도 역시 존재한다. 대표적인 의문점은 역시 할루시네이션일 것이다. 언어모형은 주어진 문제에 대해 반드시 답변하도록 훈련되어 있고, 훈련된 데이터는 특정 시기를 기점으로 한 과거형 데이터이다. 따라서 학습되지 않은 데이터에 대한 답변을 요구하거나, 업데이트가 필요한 데이터에 대해서는 잘못된 답변을 제시한다. 이를 극복하기 위한 방안으로 RAG의 필요성이 대두되고 있다. 본 연구에서 다루는 경제, 금

용, 재정 분야는 정확한 답변에 대한 신뢰성이 매우 중요한데, 이를 위해서 RAG를 접목하고자 한다. 실제로 RAG 사용할 때, 관련 질문에 대해 저장된 문서를 잘 찾아오는지(Retrieve) 문서검색 능력을 RAG와 Contextual RAG를 대상으로 비교하였다. 추가적으로 본절에서는 개발한 모형의 검색된 문서를 활용한 답변 생성 능력도 확인하였다.

〈Table 8〉 RAG/Contextual RAG Source Data

Title	Author	Publish_year
What Explains Households' Default Probability?: Empirical Evidence from Administrative Data in Korea	Young-Il Kim, Kyung-Hee Jin	2018
Causes of the Increase in the Predictability of the Nominal Term Spread for Future Economic Activity	Ki-Beom Kim, Bon-Il Ku	2019
Analysis on Changes in Determinants of Financial Stress in Emerging Economies before and after Global Financial Crisis: Focusing on Foreign Capital Flows	Young Sik Jeong, Deokki Ko	2020
Estimation of Financial Stability Index and Evaluation of Forecasting Macroeconomic Downside Risk	Dong Jin Lee, Tae-Kyoon Um, Junghwan Mok	2020
A Study on Determinants of Household Debt Repayment Using Hazard Model: Focus on Middle and Elderly Household over 45 Years Old	Ming-Chang Li, Hong-Joo Jung, Moon-Yun Hwang	2020
A Study on Determinants of Corporate Rehabilitation Proceedings by Using Binary Logistic Regression	Hyun-Tak Lee	2021
The Impact of Central Bank Communication on Market Interest Rates	Won Kee Lee, Sammo Kang	2021
The Federal Reserve's Action to primary Dealers as a Lender of Last Resort during the Financial Crisis of 2007~2009: Implication for Bank of Korea	Seungyeon Won	2021
Fiscal Sustainability of National Pension and Generational Equity	Young Jun Chun	2022
A Theoretical Analysis about the Impact of BigTechs' Entrance on the Competition in the Financial Industry: Focusing on the Competition and Information Production in Credit Markets	Kyeong-Hoon Kang	2023
Assessment of Korea's Fiscal Ability to a Financial Crisis: Focusing on Fiscal Space	Jinwook Hur	2023
State-Dependent Effects of Monetary Policy: The Role of Occasionally Binding Constraints	Hye Mi You	2024

1. 벡터 데이터베이스 구축

RAG를 하기 위해서는 벡터 데이터베이스에 관련 문서를 저장하는 것이 필요하다. 본절에서는 한국금융학회(KMFA)의 2019~2024년까지 거시·통화·국제금융 부문, 금융정책 부문 금융연구 우수논문상 12편을 사용하여 참조할 문서 데이터를 확보하였다. 자세한 논문 정보는 위의 <Table 8>과 같다.

RAG와 Contextual RAG를 위한 벡터 데이터베이스를 구축하기 위해서 Anthropic 공식 github에 있는 코드¹⁰⁾를 활용하였다. 임베딩 모형은 프랑스어, 독일어, 일본어, 스페인어, 한국어 등 다양한 언어에서 OpenAI v3 large, Cohere multilingual v3와 같은 모형보다 더 우수한 성능을 보인다는 VoyageAI의 ‘voyage-multilingual-2’¹¹⁾를 사용했고, ReRank 모형은 Anthropic의 Contextual RAG 실험

<Table 9> RAG VectorDB Format Example

```
{
  {
    "doc_id": "The Federal Reserve's Action to primary Dealers as a Lender of Last Resort during the Financial Crisis of 2007~2009: Implication for Bank of Korea,"
    "author": "Seungyeon Won",
    "publish_year": "2021",
    "original_uuid": "00969dc3a1f84cebb0749d97b440fab7",
    "content": "Full text of The Federal Reserve's Action to primary Dealers as a Lender of Last Resort during the Financial Crisis of 2007~2009: Implication for Bank of Korea,"
    "chunks": {
      {
        "chunk_id": "The Federal Reserve's Action to primary Dealers as a Lender of Last Resort during the Financial Crisis of 2007~2009: Implication for Bank of Korea_chunk_0",
        "original_index": 0,
        "author": "Seungyeon Won",
        "publish_year": "2021",
        "content": "The Federal Reserve's Action to primary Dealers as a Lender of Last Resort during the Financial Crisis of 2007~2009: Implication for Bank of Korea_chunk_0_document",
      },
      ....
    }
  }
}
```

10) <https://github.com/anthropics/anthropic-cookbook/blob/main/skills/contextual-embeddings/guide.ipynb>

11) <https://blog.voyageai.com/2024/06/10/voyage-multilingual-2-multilingual-embedding-model/>

〈Table 10〉 Contextual RAG VectorDB Format Example

```
{
  {
    "doc_id": "The Federal Reserve's Action to primary Dealers as a Lender of Last Resort during the Financial Crisis of 2007~2009: Implication for Bank of Korea,",
    "author": "Seungyeon Won",
    "publish_year": "2021",
    "original_uuid": "00969dc3a1f84cebb0749d97b440fab7",
    "content": "Full text of The Federal Reserve's Action to primary Dealers as a Lender of Last Resort during the Financial Crisis of 2007~2009: Implication for Bank of Korea,",
    "chunks": {
      {
        "chunk_id": "The Federal Reserve's Action to primary Dealers as a Lender of Last Resort during the Financial Crisis of 2007~2009: Implication for Bank of Korea_chunk_0",
        "original_index": 0,
        "author": "Seungyeon Won",
        "publish_year": "2021",
        "content": "The Federal Reserve's Action to primary Dealers as a Lender of Last Resort during the Financial Crisis of 2007~2009: Implication for Bank of Korea_chunk_0_document",
        "context": "The Federal Reserve's Action to primary Dealers as a Lender of Last Resort during the Financial Crisis of 2007~2009: Implication for Bank of Korea_chunk_0_context",
      },
      ....
    }
  }
}
```

을 참고하여 다양한 형태의 데이터 처리에 유용하고, 긴 문서 처리에 강점이 있는 Cohere의 ‘rerank-multilingual-v3.0¹²⁾’ 모형을 사용했다. RAG와 Contextual RAG의 벡터 데이터베이스를 구축하기 위한 데이터 형태는 다음과 같다.

〈Table 9〉와 〈Table 10〉에서 볼 수 있듯이, RAG와 Contextual RAG의 주요 차이점은 문맥(Context) 처리 방식에 있다. Contextual RAG는 검색 성능을 향상시키기 위해 문서 조각의 내용뿐만 아니라, 언어모형을 활용해 해당 문서 조각의 문맥을 전체 문서와 비교 분석한 뒤 이를 벡터 데이터베이스의 메타데이터(Metadata)에 추가한다. 이를 바탕으로, 이러한 방법이 실제로 검색 능력을 개선하는지 실험 데이터를 통해 검증하고자 했다.

12) <https://docs.cohere.com/v2/docs/rerank-2>

2. 문서 검색 능력 검증

RAG와 Contextual RAG의 검색 능력을 실험적으로 검증하기 위해서 GPT-4o를 활용하여 각 문서 조각마다 질문-답변을 만들었고, 이후 질문-답변을 직접 확인하며 유의미한 질문들을 다시 선별하는 작업을 거쳤다. 질문의 유형은 하나의 문서 조각에서만 질문을 추출하는 단순 질문, 그리고 두 개의 문서 조각을 합쳐서 좀 더 복잡한 질문을 추출하는 복합 질문으로 구성하였다. 그 결과, 단순 질문 480개, 복합 질문 40개를 선정하였다. 단순 질문과 복합 질문의 예시는 <Table 11>과 같다.

<Table 11> Question Type Example

Type	Question	Answer	Source
Simple Question	What is the difference between South Korea's national debt (D1) and general government debt (D2), and how is D3 defined?	Korea's national debt (D1) refers to the officially announced figure by the government and is distinct from general government debt (D2). D3 is defined as the sum of general government debt (D2) and the debt of non-financial public enterprises, excluding the debt of financial public enterprises.	Assessment of Korea's Fiscal Ability to a Financial Crisis: Focusing on Fiscal Space_chunk_4
Complex Question	What is the difference between South Korea's national debt (D1) and general government debt (D2), and how is D3 defined? Also, how does South Korea's national debt ratio in 2020 compare to that of major countries, and what is particularly noteworthy when comparing it to OECD countries?	Korea's national debt (D1) refers to the figure officially announced by the government and is distinct from general government debt (D2). D3 is defined as the sum of general government debt (D2) and the debt of non-financial public enterprises, excluding the debt of financial public enterprises. As of 2020, Korea's national debt-to-GDP ratio remained lower than that of most major countries, with the exception of a few such as Estonia, Luxembourg, and New Zealand. However, it is noteworthy that the pace of increase in Korea's national debt ratio during the 2010s exceeded the average of OECD countries, highlighting a significant upward trend despite the relatively low level of national debt.	Assessment of Korea's Fiscal Ability to a Financial Crisis: Focusing on Fiscal Space_chunk_4, Assessment of Korea's Fiscal Ability to a Financial Crisis: Focusing on Fiscal Space_chunk_5

질문-답변을 기반으로 검색 능력을 검증하기 위해 질문에 따라 벡터 데이터베이스에서 문서를 가져오는 과정을 평가하였다. 단순 질문의 경우, 가장 높은 점수를 기록한 문서 조각이 질문의 출처와 일치하는지 여부를 기준으로 점수를 부여하였다. 복합 질문의 경우에는 상위 2개의 문서 조각을 대상으로 완전 일치 여부와 부분 일치 여부를 기준으로 점수를 나누어 평가하였다.

〈Table 12〉 Retrieve Test Results

	RAG	Contextual RAG
Simple Question	87.92%	82.08%
Complex Question (Exact)	25.00%	37.50%
Complex Question (Partial)	57.50%	61.25%

〈Table 12〉에서 알 수 있듯이, 단순 질문에는 RAG가 더 나은 점수를 기록한 반면, 복합 질문에서는 Contextual RAG가 더 우수한 성능을 보였다. 단순 질문의 경우에 Contextual RAG가 검색 과정에서 문맥 정보를 제공하는 것이 오히려 검색 결과에 혼란을 주었기 때문으로 해석할 수 있다. 반면, 복합 질문에서는 문맥이 검색 과정에 있어서 정확도를 높인 것을 확인할 수 있었다. 추가적으로, NBER 문서 3편을 대상으로 단순 질문 성능을 평가한 결과에서도, RAG가 90.95%, Contextual RAG가 87.24%의 성능을 보여 이는 비단 언어의 차이로 인한 문제만은 아니라는 것 역시도 확인할 수 있었다.

본 연구를 통해, 경제, 금융, 재정처럼 특정 도메인에 RAG를 적용하기 위해서는 Contextual RAG가 적합하다는 잠정적인 결과를 도출해 낼 수 있다. 가령 금리 인상이라는 단일 행동을 결정하기 위해 인플레이션, 실업률, 가계 부채 등 다양한 경제 지표들을 고려해야 한다. 이러한 도메인의 복잡성을 고려할 때, 문맥 정보를 활용하여 복잡한 질문에 대해서 보다 정확한 검색 결과를 제공할 수 있는 Contextual RAG가 더 유용할 것으로 판단된다.

다만, Contextual RAG가 절대적인 점수에서 부분 일치 61.25%, 완벽 일치 37.5%라는 결과로 여전히 한계점이 존재한다는 것을 알 수 있다. 따라서, 검색 성능을 더 향상시키기 위한 한국어 경제·금융·재정 도메인에 적용할 수 있는 도메인 특화 임베딩 모형개발이 필요하다. 실제로 VoyageAI는 금융 뉴스, 공시, 재무 보고서 등 다양한 금융 데이터를 효과적으로 처리하도록 금융 특화 임베딩 모형인

‘voyage-finance-2¹³⁾’를 개발하였다. 동 모형은 11개의 금융 질문-답변 데이터셋에 대해 타 모형에 비해 우수한 성과를 보였으며, 실제로 OpenAI v3 large보다 평균 7%, Cohere English v3보다 평균 12% 더 높은 성능을 보였다. 이는 RAG를 구현할 때 도메인 특화 임베딩이 RAG의 성능을 높일 수 있다는 것을 방증한다.

3. RAG 답변 성능 확인

질문 유형에 따른 RAG의 성능 비교에 이어, 검색 결과를 기반으로 언어모형이 얼마나 정확하게 답변을 생성할 수 있는지 평가하였다. 이를 위해 RAG와 Contextual RAG 모두 동일한 문서 조각을 완벽히 일치하도록 검색한 상태에서 질문에 대한 답변을 생성하도록 설정했다. 이는 검색 단계의 성능이 아닌, 검색된 문서를 바탕으로 질문에 적합한 답변을 생성하는 능력에 초점을 맞추기 위함이다.

본절에서는 복합 질문 7개를 중심으로 진행했고, 질문에 대한 정답은 GPT-4o-mini를 통해 생성했다. 비교 대상 모형은 본 연구의 베이스 모형(Llama-3.1-8b-instruct), 경제·금융에 특화된 모형인 EconLlama3.1-8b, 경제·금융·재정 특화 학습 모형인 EconFisLlama3.1-8b을 활용했다. RAG와 Contextual RAG 모두 동일한 문서 조각을 검색했고, 복합 질문을 예시로 사용하고 있기에 Contextual RAG를 중점적으로 사용하여 답변 생성 능력을 평가하였다.

〈Table 13〉에서 보면 알 수 있듯이, sLLM 기반의 답변은 주어진 문서를 활용하긴 했으나, 불필요한 내용도 같이 답변하는 모습을 보였다. GPT-4o-mini는 질문에 대해 정확하고 간결하게 필요한 답변을 생성하는 반면에, 나머지 sLLM 기반 답변은 반복적인 구문 및 불필요한 내용이 많이 추가된 것을 관찰할 수 있었다. 다만, sLLM이더라도 학습할수록 질문이 요구하는 내용을 답변에 더 잘 반영하였다.

그러나 본 연구에서 구축한 데이터로 학습한 모형인 EconLlama와 EconFinLlama의 답변은 질문과는 상관없는 월세 거주 가구에 대한 설명이 명시되기도 하였다. 이는 질문과 그에 해당하는 맥락이 함께 제시되면서 나타나는 결과로, 주어진 질문에 대해 보다 연관성 있는 답변을 정확하게 도출하도록 학습하는 것이 추가적인 연구 과제로 판단된다.

13) <https://blog.voyageai.com/2024/06/03/domain-specific-embeddings-finance-edition-voyage-finance-2/>

〈Table 13〉 RAG Answer Example generated by LLM/sLLM

Question	Answer (GPT-4o-mini)	Llama-3.1-8b	EconLlama3.1	EconFisLlama3.1
What trends were observed in the differences in delinquency risk between wage-earning and non-wage-earning households? What is the relationship between the number of owned homes and delinquency risk, and how is this connected to the capacity to absorb economic shocks?	There is a general trend in which higher income levels are associated with lower delinquency risk. Among higher-income households (3rd to 4th income quintiles), wage-earning households showed lower delinquency risk compared to non-wage-earning households. However, in lower-income households (1st to 2nd income quintiles), wage-earning households exhibited higher delinquency risk than non-wage-earning households. Households with no homes or fewer owned homes tended to have higher delinquency risk. This is likely because owning more	Differences in delinquency risk were observed across income quintiles. The gap in delinquency risk between wage-earning and non-wage-earning households varied by income level. In households within the 3rd to 4th income quintiles, wage-earning households had lower delinquency risk compared to non-wage-earning households. In contrast, among households in the 1st to 2nd income quintiles, wage-earning households exhibited higher delinquency risk than their non-wage-earning counterparts. The more homes a household owned, the lower the delinquency risk tended to be. Households living in apartments had lower delinquency risk than those living in detached houses. Households with limited capacity to absorb financial shocks were more likely to experience	It can be observed that higher income levels are associated with lower delinquency risk. The effect of employment type on delinquency risk varies across income quintiles: in higher-income households, wage earners tend to have lower delinquency risk, whereas in lower-income households, wage earners show higher delinquency risk. Households that own more homes, live in apartments, or reside in owner-occupied or long-term lease (jeonse) housing tend to have lower delinquency risk. In contrast, those living in detached houses or similar types of housing have higher delinquency risk compared to apartment dwellers. Households residing in monthly rental housing are more	Among lower-income households (1st to 2nd income quintiles), wage-earning households showed higher delinquency risk compared to non-wage-earning households. In contrast, among higher-income households (3rd to 4th income quintiles), wage-earning households had lower delinquency risk than non-wage-earning households. Additionally, delinquency risk was lower among households that owned more properties, lived in apartments, or resided in owner-occupied or long-term lease (jeonse) housing. This indicates that the relationship between the number of owned homes and delinquency risk is closely tied to a household's capacity to absorb economic shocks. However, households living in monthly rental housing showed higher delinquency

	properties implies greater asset availability, which enhances a household's capacity to absorb shocks from changes in economic conditions.	delinquency. Additionally, households with insufficient labor supply potential, limited ability to adjust consumption, low collateral or borrowing capacity, and inadequate insurance coverage showed higher levels of delinquency risk.	likely to experience delinquency than those in owner-occupied or jeonse arrangements.	risk compared to those in owner-occupied or jeonse housing. This is likely because monthly rent payers face a continuous financial burden, making it more difficult to absorb economic shocks such as income loss.
--	--	--	---	--

물론, 위 결과는 제한된 데이터를 기반으로 한 실험이므로 일반화에는 한계가 있지만, sLLM이 적절하고 양질의 데이터를 학습한 경우에는 학습되지 않은 모형보다 검색된 문서를 질문에 맞게 효과적으로 활용하는 모습을 확인할 수 있었다. 따라서 도메인 특화 학습이 이루어진 sLLM은 RAG 환경에서 유용하게 활용될 가능성이 높다.

일부에서는 LLM이 sLLM에 비해 우월하다고 주장할 수 있다. 실제로 최근의 LLM은 MoE(Mixture of Experts)¹⁴⁾ 구조 등을 통해 방대한 도메인에 걸쳐 높은 수준의 성능을 보이고 있으며, 본 논문의 실험에서도 GPT-4o-mini가 sLLM 기반 RAG 시스템보다 더 정확하고 간결한 답변을 생성한 바 있다. 그러나 Gururangan et al. (2020)에 따르면 문제 구조가 복잡하고, 세부 도메인 간의 개념적 미묘함과 문맥 해석의 정밀함이 중요한 분야에서는 단일 LLM의 범용성만으로 이를 충분히 포괄하기 어려운 한계가 있다. 경제학 역시 이러한 특징을 가지고 있다. 이러한 문제를 효과적으로 해결하기 위해서는 각 도메인에 특화된 다수의 에이전트(Agent)가 상호작용하는 멀티에이전트 구조가 요구된다. 이 경우, LLM은 막대한 연산 자원과 비용 부담으로 인해 실용성이 떨어질 수 있으며, 반면 도메인 특화형 sLLM은 경량 구조로써 학습 및 운영 비용이 낮고 오프라인 환경에서도 유연하게 활용될 수 있어, 보다 현실적이고 효과적인 대안이 될 수 있다. 따라서 경제학과 같은 복잡한 문제를 해결하기 위해서는 sLLM을 훈련하는 것이 더 유리할 수 있다. 이러한 점을

14) 입력에 따라 서로 다른 전문가 모형(서브 네트워크) 중 일부만 선택적으로 활성화하여 계산 효율성과 모형 표현력을 동시에 높이는 신경망 구조이다.

종합할 때, 도메인 특화 언어모형은 RAG 환경에서 범용 LLM에 비해 비용·자원 측면에서 현실적인 대안이 될 수 있으며, 본 연구의 실험 결과에서도 LLM을 제외하면 도메인 학습된 sLLM이 가장 안정적인 응답 성능을 보였다는 점에서 그 가능성이 간접적으로 확인되었다.

VI. 결 론

본 연구는 금융·재정 분야에 특화된 한국어 sLLM (EconFisLlama3.1-8B)을 개발하고, 도메인 특화 학습의 가능성과 한계를 검토하였다. 도메인 특화 학습 기법을 응용하여 해당 모형을 훈련하였으며, 그 결과 기존 베이스 모형에 비해 용어 정의 및 기본 개념 이해 능력에서 개선된 성과를 확인하였다. 특히, 경제·금융 분야의 전문적 응답을 제공할 수 있는 잠재력을 입증하며, 정책적 맥락이나 특정 금융 용어에 대한 보다 일관적이고 신뢰성 있는 답변을 생성할 수 있음을 보여주었다.

한국 경제의 지속가능한 성장을 위해서는 금융 부문의 발전이 필수적이며, 특히 언어모형 개발이 그 핵심 원동력으로 주목받고 있다. 영어권에서 개발된 언어모형을 번역하여 활용하는 것도 가능하지만, 대부분의 고성능 모형은 데이터의 비공개성, 높은 컴퓨팅 자원 및 인력 투입 비용, 그리고 일부 경우에는 사용량 제한 등의 문제를 안고 있다. 이에 따라, 공개된 언어모형을 여러 금융 분야에 적용하는 것이 현재의 추세이다. 영어 중심의 웹 환경에서 비영어권 언어모형의 성능 격차가 확대될 가능성이 높아지는 가운데, 본 연구는 경제 및 금융 분야에 최적화된 한국어 sLLM 개발의 필요성에 주목하였다. 특히, 코로나 이후 재정 정책이 국가 경제에서 차지하는 비중이 확대됨에 따라 재정 분야의 학습을 통해 보다 전문적이고 도메인 특화된 언어모형을 구축할 필요성을 확인하였다. 이러한 배경을 바탕으로, 경제·금융·재정 분야에 특화된 한국어 언어모형을 개발하였다.

본 연구는 한국은행 경제용어, 다양한 경제·금융 용어사전, 맨큐의 경제학, 재정정보원 보고서 등 공개적으로 접근 가능한 자료를 활용하였다. 용어 번역에는 GPT-3.5 Turbo, GPT-4 등을 사용했으며, 학습 데이터는 한국어와 영어 번역 문장, 경제·금융 용어 설명, 그리고 질문-답변 형식의 데이터를 포함한다. 경제·금융과 재정 분야에서의 SFT와 DPO 학습은 총 네 단계에 걸쳐 각각 진행되었다. 이러한 단계별 학습은 모형의 성능을 확인하는 동시에 전문 도메인 지식을 효과적으

로 학습하는 데 기여할 것으로 판단된다.

네 단계로 구축된 최종 모형을 중심으로 베이스 모형, 경제·금융·재정 특화 모형과의 답변 비교를 진행하였다. 그 결과, 베이스 모형에 비해 도메인 특화 지식을 학습한 최종 모형이 경제학 용어 및 개념 정의에 있어서 정확도가 개선되었다. 이는 특정 도메인에 대한 추가 학습이 모형 성능 향상에 실질적으로 기여한다는 사실을 방증한다. 다만, 고차원적 추론 능력을 요구하는 질문에서는 여전히 한계가 존재하였다. 단순 정의를 넘어 다양한 유형의 데이터와 복합적 학습 방법을 통해 추론 능력을 제고하는 것이 AI 시대에서의 전략임을 확인할 수 있었다.

더 나아가, 본고에서 개발한 최종 모형의 할루시네이션 문제를 보완하여 성능을 제고하고자 RAG를 시도하였다. 한국금융학회(KMFA) 논문 12편을 기반으로 벡터 데이터베이스를 구축하여 복합 질문을 생성하고, Contextual RAG를 활용하여 GPT 4o-mini, 베이스 모형, 경제·금융 특화 모형, 최종 모형 간의 답변 성능 비교를 진행하였다. 그 결과, 최종 모형이 가장 적절하게 문서를 참고하여 질문에 답변하였다. 이는 특히 경제·금융과 같은 전문성과 정확성이 요구되는 분야에 있어서 RAG의 활용 가치를 확인할 수 있었다.

특히, 본 연구에서는 sLLM을 사용하였는데, sLLM은 비용 효율적이고, 특정 도메인에 최적화된 학습과 조정을 통해 적은 자원으로도 유의미한 성능을 발휘할 수 있다. LLM 대비 낮은 하드웨어 요구 사항과 비용으로 전문적인 답변 생성이 가능하다. 그리고 경제·금융·재정과 같은 특정 도메인에 특화된 데이터를 추가 학습 시킴으로써 높은 도메인 적합성을 확보할 수 있는 것처럼, 특정 목적이나 필요에 맞게 맞춤형 서비스를 제공할 수 있다.

또한 sLLM은 다양한 환경에서 응용 가능성을 확대할 수 있으며, 특히 로컬 환경에 저장 및 운영이 가능하다는 점에서 특정 조건에서의 데이터 보안이나 독립적인 운용에 유리한 특성을 지닌다. RAG를 로컬에 구축하여 활용하는 경우에 외부 서버와의 통신 없이 내부에서 정보를 검색하고 생성할 수 있어, 온디바이스(On-Device) 환경에서 보안 요구가 높은 상황에 적용 가능성이 있다는 점에서 잠재적 이점을 가진다. 따라서, sLLM은 단순히 성능만을 고려하는 것이 아니라, 비용, 개인화, 확장성을 동시에 충족할 수 있는 현실적인 대안이 될 수 있다.

특히, 본 연구에서 활용한 sLLM은 Meta에서 제공하는 LLaMA로, 공개된 LLaMA 3.1의 Community License에 따르면 Meta가 부여한 권리에 따라 LLaMA

를 자유롭게 사용, 복제, 배포, 복사가 가능하다. 이는 LLaMA를 로컬 환경에서 자유롭게 사용할 수 있는 근거가 된다. 이를 바탕으로 LLaMA는 외부 API와의 통신 없이 독립적으로 운용할 수 있으며, 사용자 데이터가 외부 서버에 저장되지 않도록 구성할 수 있어, 보안과 프라이버시 측면에서 유리한 조건을 제공한다. 이러한 점은 sLLM이 다양한 보안 요구 환경에서 응용될 수 있는 기술적 근거로 작용한다. 이러한 개방성과 자율성 덕분에 개발자 간의 협업과 공유 문화를 촉진시키는 기폭제 역할을 하며, Open-llm-leaderboard와 같이 커뮤니티 중심의 모형 성능 평가 및 공개 경쟁 플랫폼이 활발하게 운영되는 기반이 되고 있다. 더 나아가, sLLM을 이용한 본 연구의 접근 방식은 특정 상용 LLM에 대한 종속성을 완화함과 동시에, 보다 더 유연하고 분산적인 언어모형 개발 생태계를 조성하는 데 기여한다.

본고는 경제·금융·재정 분야에 익숙하지 않은 일반인이나 대학생들이 활용할 수 있는 경제학 분야 sLLM의 개발 과정, 데이터 학습, 그리고 응용 결과를 분석하는 데 목적을 두었다. 특히, 경제학 분야에서 한국어 언어모형 개발의 기초를 마련했다는 점에서 의의가 있으며, 이를 기반으로 도메인 특화 언어모형이 교육, 정책 분석, 금융 의사결정 지원 등 다양한 분야에서 실질적이고 폭넓게 활용될 것으로 기대된다. 이러한 발전은 많은 사람이 금융 지식을 활용할 수 있는 포용적인 금융 생태계 구축에 도움을 주며, AI와 금융의 교차점을 점차적으로 확장해 나가는 데 기여할 것으로 판단된다.

■ 참 고 문 헌

1. Alex G. Kim, Maximilian Muhn, and Valeri V. Nikolaev, "Financial Statement Analysis with Large Language Models," Chicago Booth Research Paper, Fama-Miller Working Paper, 2024.
2. Azhar, Faisal, Chris Bamford, Florian Bressand, Diego de las Casas, Devendra Singh Chaplot, Naman Goyal, Edouard Grave, Eric Hambro, Gautier Izacard, Albert Q. Jiang, Armand Joulin, Marie-Anne Lachaux, Thibaut Lavril, Gianna Lengyel, Guillaume Lample, L  lio Renard Lavaud, Arthur Mensch, Aurelien Rodriguez, Alexandre

- Sablayrolles, H. Touvron, T. Lavril, G. Izacard, X. Martinet, M. -A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, LLaMA: Open and Efficient Foundation Language Models, arXiv preprint arXiv:2302.13971, 2023.
3. Brown, T. B., B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language Models are Few-Shot Learners," arXiv preprint arXiv:2005.14165, 2020.
 4. Cao, Z., and Z. Feinstein, "Large Language Model in Financial Regulatory Interpretation," arXiv preprint arXiv:2405.06808, 2024.
 5. Dettmers, T., A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLoRA: Efficient Finetuning of Quantized LLMs," arXiv preprint arXiv:2305.14314, 2023.
 6. Devlin, J., M. -W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv preprint arXiv:1810.04805, 2018.
 7. Gemini Team, Google, "Gemini: A Family of Highly Capable Multimodal Models," 2024.
 8. Gemma Team, Google DeepMind, "Gemma: Open Models Based on Gemini Research and Technology," 2024.
 9. Gururangan, S., A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith, Don't Stop Pretraining: Adapt Language Models to Domains and Tasks, arXiv preprint arXiv:2004.10964, 2020.
 10. Hu, E. J., Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-Rank Adaptation of Large Language Models," arXiv preprint arXiv:2106.09685, 2021.
 11. Jiang, A. Q., A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M. -A. Lachaux, P. Stock, T. Le Scao, T. Lavril, T. Wang, T. Lacroix, and W. El Sayed, "Mistral 7B," arXiv preprint arXiv:2310.06825, 2023.
 12. Kim, Donggyu, Dongwook Lee, Jangwon Park, Sungwoo Oh, Sungjun Kwon, and Inyong Lee, "KB-BERT: Training and Application of Korean Pre-trained Language Model in Financial Domain," Journal of Korea Intelligent Information Systems Society, 2022.
 13. Kim, D., C. Park, S. Kim, W. Lee, W. Song, Y. Kim, H. Kim, Y. Kim, H. Lee, J. Kim, C. Ahn, S. Yang, S. Lee, H. Park, G. Gim, M. Cha, H. Lee, and S. Kim, "SOLAR 10.7B: Scaling Large Language Models with Simple yet Effective Depth Up-Scaling," arXiv preprint arXiv:2312.15166, 2023.
 14. Liu, H., D. Tam, M. Muqeeth, J. Mohta, T. Huang, M. Bansal, and C. Raffel, "Few-Shot Parameter-Efficient Fine-Tuning is Better and Cheaper than In-Context Learning," arXiv preprint arXiv:2205.05638, 2022.
 15. Liu, Y., D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu, "G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment," arXiv preprint arXiv:2302.13922, 2023.
 16. Lee, M., Y. Noh, and S. J. Lee, "A Testset for Context-Aware LLM Translation in

- Korean-to-English Discourse Level Translation,” Proceedings of the 31st International Conference on Computational Linguistics (COLING), pp.1632–1646, 2025.
17. Lewis, P., E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, Retrieval-augmented Generation for Knowledge-intensive NLP Tasks, arXiv preprint arXiv:2005.11401, 2020.
18. Ouyang, K., Y. Liu, S. Li, R. Bao, K. Harimoto, and X. Sun, “Modal-adaptive Knowledge-enhanced Graph-based Financial Prediction from Monetary Policy Conference Calls with LLM,” arXiv preprint arXiv:2403.16055, 2024.
19. Rafailov, R., A. Sharma, E. Mitchell, S. Ermon, C. D. Manning, and C. Finn, “Direct Preference Optimization: Your Language Model is Secretly a Reward Model,” arXiv preprint arXiv:2305.18290, 2023.
20. Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L., Kaiser and I. Polosukhin, “Attention Is All You Need,” arXiv preprint arXiv:1706.03762, 2017.
21. Wang, R., J. Guo, C. Gao, G. Fan, C. Y. Chong, and X. Xia, “Can LLMs Replace Human Evaluators? An Empirical Study of LLM-as-a-Judge in Software Engineering,” arXiv preprint arXiv:2502.06193, 2025.
22. Wu, S., O. Irsoy, S. Lu, V. Dabrovolski, M. Dredze, S. Gehrmann, P. Kambadur, D. Rosenberg, and G. Mann, “BloombergGPT: A Large Language Model for Finance,” arXiv preprint arXiv:2303.17564, 2023.
23. Xu, H., S. Ebner, M. Yarmohammadi, A. S. White, B. Van Durme, and K. Murray, Gradual Fine-tuning for Low-resource Domain Adaptation, arXiv preprint arXiv:2103.02205, 2021.
24. Yang, H., X.-Y. Liu, and C. D. Wang, “FinGPT: Open-Source Financial Large Language Models,” arXiv preprint arXiv:2306.06031, 2023.
25. Yu, Y., Z. Yao, H. Li, Z. Deng, Y. Cao, Z. Chen, J. W. Suchow, R. Liu, Z. Cui, Z. Xu, D. Zhang, K. Subbalakshmi, G. Xiong, Y. He, J. Huang, D. Li, and Q. Xie, “FinCon: A Synthesized LLM Multi-agent System with Conceptual Verbal Reinforcement for Enhanced Financial Decision Making,” arXiv preprint arXiv:2407.06567, 2024.
26. Zhang, X., P. Shapiro, G. Kumar, P. McNamee, M. Carpuat, and K. Duh, Curriculum Learning for Domain Adaptation in Neural Machine Translation, In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp.1903-1915), Association for Computational Linguistics, 2019.
27. Zhong, Yan, Dennis Wong, and Kun Lan, “Tax intelligent Decision-making Language Model,” IEEE Access, PP.1-1, 2024.
28. https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard
29. <https://blog.voyageai.com/2024/06/03/domain-specific-embeddings-finance-edition-voyage-finance-2/>
30. <https://blog.voyageai.com/2024/06/10/voyage-multilingual-2-multilingual-embedding-model/>
31. <https://docs.cohere.com/v2/docs/rerank-2>
32. <https://github.com/anthropics/anthropic-cookbook/blob/main/skills/contextual-embeddings/guide.ipynb>

Development and Training of a Small-scale Large Language Model (sLLM) for Economics*

Chae-Shick Chung** · Daemin Park*** ·
Jonghyun Park**** · Soeun Hong*****

Abstract

This study investigates the foundational processes and applications of AI language models in the field of economics. It focuses on developing a Korean Small-scale Large Language Model (sLLM) specialized in economics, finance, and fiscal policy, while exploring its practical applications. Using Meta-Llama/Llama-3.1-8B-Instruct as the base model, the sLLM is developed by applying reliable datasets in economics, finance, and fiscal policy, along with various training methods such as SFT and DPO. Its applicability and versatility are evaluated through its implementation on economic databases. Empirical analysis indicates that the proposed sLLM exhibits enhanced contextual understanding in response to economic queries, with additional improvements observed when integrated with RAG (Retrieval-Augmented Generation). These findings suggest that the model successfully addresses challenges such as translation errors and limited contextual comprehension, laying a foundation for tackling AI sovereignty issues.

Key Words: economics, finance, fiscal policy specialization, Small-scale Large Language Model (sLLM) Training, RAG (Retrieval Augmented Generation)

JEL Classification: C63, C88, O33

Received: March 24, 2025. Revised: May 26, 2025. Accepted: June 12, 2025.

* This paper was presented at the special session on “Response to changes in the economic environment related to AI (Artificial Intelligence)” at 2025 Korea’s Allied Economic Associations Annual Meeting.

** Corresponding Author, Professor, Department of Economics, Sogang University, 35 Baekbeom-ro, Mapo-gu, Seoul 04107, Korea, Phone: +82-2-705-8704, e-mail: cschung@sogang.ac.kr

*** Co-Author, Master’s Degree in AI Finance, Department of Economics, The Graduate School, Sogang University, 35 Baekbeom-ro, Mapo-gu, Seoul 04107, Korea, e-mail: qjrogoa@naver.com

**** Co-Author, Master’s Degree in AI Finance, Department of Economics, The Graduate School, Sogang University, 35 Baekbeom-ro, Mapo-gu, Seoul 04107, Korea, e-mail: pjh970731@gmail.com

***** Co-Author, Master’s Degree in AI Finance, Department of Economics, The Graduate School, Sogang University, 35 Baekbeom-ro, Mapo-gu, Seoul 04107, Korea, e-mail: wendy3395@naver.com