JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 31 No. 5, pp. 197-205, May 2026
https://doi.org/10.9708/jksci.2026.31.05.197

Fine-tuning and Performance Evaluation of a Korean Essential Medical Domain-Specific Small Language Model Based on QLoRA

JongHwi Song*, Urtnasan Erdenebayar**

*Research Professor, Institute of AI Convergence Science, Yonsei University, Wonju, Korea
**Associate Professor, Division of AI Semiconductor, Yonsei University, Wonju, Korea

[Abstract]

This study proposes a method for fine-tuning a small language model (sLLM) specialized in Korean essential medical knowledge using QLoRA (Quantized Low-Rank Adaptation). We utilized the AI Hub Essential Medical Knowledge QA dataset comprising 17,280 question-answer pairs across four clinical departments—internal medicine, pediatrics, obstetrics and gynecology, and emergency medicine—to fine-tune Qwen2.5-7B-Instruct with 4-bit NF4 quantization and LoRA adapters. The fine-tuned model was trained on a single consumer-grade GPU (RTX 5070 Ti, 16GB VRAM) with only 0.92% of the total parameters being trainable. Experimental results demonstrate that the fine-tuned model achieves significant improvements over the base model: ROUGE-1 increased from 0.131 to 0.291, ROUGE-L from 0.126 to 0.291, and MCQ accuracy on the sampled internal test set (198 out of 1,728) improved from 0.0% to 68.2%, where the base model's 0.0% was primarily attributable to output format mismatch rather than lack of medical knowledge. Furthermore, on the external KorMedMCQA benchmark, the fine-tuned model achieved 58.0% accuracy compared to the base model's 55.0%. These results indicate that parameter-efficient fine-tuning with domain-specific Korean medical QA data can effectively align the output format and language consistency of general-purpose LLMs for Korean medical QA tasks, while the enhancement of medical knowledge itself remains limited and requires further investigation.

▸ **Key words**: QLoRA, Fine-tuning, Korean Medical QA, Small Language Model, Parameter-Efficient Fine-Tuning

• First Author: JongHwi Song, Corresponding Author: Urtnasan Erdenebayar
*JongHwi Song (jh_song@yonsei.ac.kr), Institute of AI Convergence Science, Yonsei University
**Urtnasan Erdenebayar (edenbyra@yonsei.ac.kr), Division of AI Semiconductor, Yonsei University
• Received: 2026. 03. 24, Revised: 2026. 04. 28, Accepted: 2026. 04. 29.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

[요 약]

본 연구는 QLoRA(Quantized Low-Rank Adaptation) 기법을 활용하여 한국형 필수의료 분야에 특화된 소규모 언어모델(sLLM)의 파인튜닝 방법을 제안한다. AI Hub에서 제공하는 필수의료 의학지식 QA 데이터셋 17,280쌍을 활용하여, 내과·소아청소년과·산부인과·응급의학과 4개 진료과에 대한 Qwen2.5-7B-Instruct 모델의 도메인 적응을 수행하였다. 4-bit NF4 양자화와 LoRA 어댑터를 적용하여 전체 파라미터의 0.92%만을 학습하면서도 단일 소비자급 GPU(RTX 5070 Ti, 16GB)에서 효율적으로 파인튜닝이 가능하였다. 실험 결과, 파인튜닝 모델은 베이스 모델 대비 ROUGE-1이 0.131에서 0.291로, ROUGE-L이 0.126에서 0.291로 향상되었으며, 내부 테스트셋에서 균등 샘플링된 198건에 대한 객관식 정답률이 0.0%에서 68.2%로 개선되었다. 베이스 모델의 0.0%는 의학 지식의 부재가 아닌 출력 형식 불일치에 기인한 것으로, 파인튜닝의 효과는 지식 보강과 출력 형식 학습 양 측면에서 나타났다. 외부 벤치마크인 KorMedMCQA에서도 베이스 모델 55.0% 대비 58.0%의 정답률을 달성하였다. 본 연구는 도메인 특화 한국어 의료 QA 데이터를 활용한 파라미터 효율적 파인튜닝이 범용 LLM의 출력 형식 정렬과 한국어 답변 일관성 확보에 효과적임을 보이며, 의학 지식 자체의 향상을 위해서는 추가적인 연구가 필요함을 시사한다.

▸ **주제어**: QLoRA, 파인튜닝, 한국어 의료 QA, 소규모 언어모델, 파라미터 효율적 파인튜닝

I. Introduction

최근 Transformer[1] 아키텍처를 기반으로 한 대규모 언어모델(Large Language Model, LLM)의 발전은 자연어 처리 분야 전반에 걸쳐 혁신적인 성과를 가져왔다[2]. 특히 의료 분야에서는 임상 질의응답, 의학 지식 검색, 진단 보조 등 다양한 응용 가능성이 탐색되고 있으며[3], Instruction Tuning과 같은 기법을 통해 LLM의 지시 수행 능력이 크게 향상되었다[4, 5]. 그러나 범용 LLM은 한국어 의료 지식에 대한 학습이 부족하여, 한국 임상 환경에 특화된 정확한 답변을 생성하는 데 한계가 있다.

한국은 필수의료 분야—내과, 소아청소년과, 산부인과, 응급의학과—의 의료 인력 부족과 접근성 문제가 사회적 과제로 대두되고 있다. 이러한 상황에서 AI 기반 의료 지식 보조 시스템의 개발은 의료 접근성 향상에 기여할 수 있는 잠재력을 갖는다. 그러나 의료 AI의 실용화를 위해서는 한국어 의학 용어와 임상 맥락을 정확히 이해하는 언어모델이 필수적이다. 프롬프트 엔지니어링이나 RAG(Retrieval-Augmented Generation)를 통해 추론 시점에 맥락을 제공하는 접근도 가능하나, 이는 매 질의마다 적절한 근거 문서의 검색과 프롬프트 구성을 필요로 하며, 모델 자체의 한국어 의료 용어 이해와 답변 형식 생성 능력을 근본적으로 개선하지는 못한다. 본 연구는 이러한 추론 시점 기법과 상호 보완적인 접근으로서, 모델 수준의 도메인 적응을 통해 한국 의료 QA에 적합한 기반 역량을 확보하는 것을 목표로 한다.

LLM의 전체 파라미터를 재학습하는 풀 파인튜닝(Full Fine-tuning)은 수십~수백 GB의 GPU 메모리를 요구하여, 대학 연구실이나 소규모 기관에서는 현실적으로 적용하기 어렵다. 이를 해결하기 위해 Hu 등이 제안한 LoRA(Low-Rank Adaptation)는 사전학습 가중치를 동결한 채 저차원 분해 행렬만을 학습하여 학습 파라미터 수를 대폭 줄이는 방법이다[6]. 나아가 Dettmers 등이 제안한 QLoRA는 4-bit 양자화와 LoRA를 결합하여, 단일 소비자급 GPU에서도 70억 파라미터 이상의 모델을 효과적으로 파인튜닝할 수 있는 기법이다[7].

본 연구에서는 QLoRA를 활용하여 Qwen2.5-7B-Instruct[8] 모델을 AI Hub 필수의료 의학지식 데이터[9]로 파인튜닝하고, 베이스 모델 대비 의료 QA 성능 향상을 정량적으로 평가한다. 또한 외부 벤치마크인 KorMedMCQA를 통해 일반화 성능을 검증한다[10]. 본 연구의 주요 기여는 다음과 같다.

첫째, 한국 필수의료 4개 진료과 데이터를 활용한 Instruction Tuning 기반 의료 특화 sLLM 파인튜닝 파이프라인을 제시하고, 진료과별·질문 유형별 성능 차이를 분석한다. 둘째, QLoRA 기반 파인튜닝을 한국어 의료 도메인에 적용하고, 단일 소비자급 GPU(16GB VRAM) 환경에서의 구체적인 메모리 사용량과 학습 설정을 보고하여

재현 가능한 파이프라인을 제시한다. 셋째, ROUGE 스코어, 객관식 정답률, 외부 벤치마크 등 다각적 평가를 통해 도메인 특화 파인튜닝의 효과를 검증한다.

II. Related Work

1. Parameter-Efficient Fine-Tuning (PEFT)

대규모 언어모델의 도메인 적응을 위한 파라미터 효율적 파인튜닝(Parameter-Efficient Fine-Tuning, PEFT) 기법은 최근 활발히 연구되고 있다. 대표적으로 LoRA는 Transformer의 가중치 행렬 변화가 저차원 구조(low intrinsic rank)를 갖는다는 관찰에 기반하여, 사전학습 가중치를 동결하고 저차원 분해 행렬 쌍(A, B)만을 학습한다[1, 6]. 이를 통해 GPT-3 175B 기준으로 학습 파라미터를 10,000배 줄이면서도 풀 파인튜닝과 동등한 성능을 달성할 수 있다[6].

QLoRA는 LoRA를 한 단계 발전시켜, 4-bit NormalFloat(NF4) 양자화, 이중 양자화(Double Quantization), 페이지드 옵티마이저(Paged Optimizer)를 결합하였다[7]. 이를 통해 65B 파라미터 모델을 단일 48GB GPU에서 파인튜닝할 수 있으며, 16-bit 풀 파인튜닝 대비 성능 손실 없이 메모리 사용량을 크게 절감한다. 양자화 기법으로는 이 외에도 LLM.int8(), GPTQ 등이 있으며, 이들은 추론 효율화에 초점을 맞추는 반면, QLoRA는 학습 과정 전체의 메모리 최적화를 목표로 한다는 점에서 차별화된다[11, 12]. 이 외에도 최근에는 LoRA의 학습 능력을 풀 파인튜닝 수준으로 향상시키기 위해 가중치를 크기와 방향 성분으로 분해하여 학습하는 DoRA[18]나, 그래디언트 자체의 저차원 구조를 활용하여 옵티마이저 상태의 메모리를 절감하는 GaLore[19] 등이 제안되었다. 본 연구에서는 단일 소비자급 GPU(16GB VRAM)에서의 안정적 운용을 우선시하여, 구현 성숙도와 생태계 지원이 가장 높은 QLoRA를 채택하였다.

2. Medical LLMs and Korean Medical Benchmarks

의료 분야 LLM 연구는 주로 영어권에서 MedQA, MedMCQA, PubMedQA 등의 벤치마크를 중심으로 진행되어 왔다[13]. Singhal 등은 Med-PaLM 2를 통해 LLM이 의사 수준의 의료 질의응답 성능에 근접할 수 있음을 보였다[3]. 그러나 Kweon 등은 한국 의료 면허시험 기반의 KorMedMCQA 벤치마크를 공개하며, 한국어 의료 맥락에서의 LLM 평가 필요성을 강조하였다[10]. 그들의 연구에서 영어 의료 데이터로 사전학습한 모델(예: Meditron)이 한국어 의료 벤치마크에서는 유의미한 성능 향상을 보이지 못하는 반면, 한국어 연속 사전학습 모델은 성능이 향상되는 것으로 나타나, 언어 특화 학습의 중요성이 확인되었다.

3. Qwen2.5 Model

Qwen2.5는 Alibaba Cloud에서 개발한 대규모 언어모델 시리즈로, 18조 토큰의 사전학습 데이터와 100만 건 이상의 지도 미세조정(SFT) 데이터를 활용하여 학습되었다[8]. 특히 다국어 지원이 우수하여 한국어를 포함한 비영어권 작업에서도 높은 성능을 보인다. Qwen2.5-7B-Instruct는 Open Ko-LLM Leaderboard 등 한국어 벤치마크에서 동일 파라미터 규모의 Llama 3[14]나 Mistral 계열 모델 대비 상위 성능을 기록한 바 있어, 한국어 도메인 적응의 기반 모델로 선택하였다. 7B 파라미터 규모의 Qwen2.5-7B-Instruct는 소비자급 GPU에서 운용 가능한 크기이면서도 강력한 언어 이해 및 생성 능력을 갖추고 있어, 도메인 적응 연구의 기반 모델로 적합하다.

III. Dataset and Preprocessing

1. Dataset Overview

본 연구에서는 AI Hub에서 제공하는 '필수의료 의학지식 데이터'를 활용하였다[9]. 이 데이터셋은 국내 주요 대형병원(서울성모병원, 삼성서울병원, 서울대학교병원, 세브란스병원, 보라매병원)이 참여하여 구축한 의학 지식 질의응답 데이터로, 임상학적 근거가 분명한 의학 지식을 포함한다. 해당 데이터셋은 AI Hub의 데이터 품질 관리 가이드라인에 따라 구축되었으며, 참여 병원 소속 의료 전문가에 의한 검수를 거친 것으로 보고되어 있다[9]. 객관식 문항의 선택지 구성 및 정답 레이블, 서술형 답변의 의학적 정확성은 데이터 구축 단계에서 임상 전문가의 감수를 통해 확보된 것이다. 본 연구에서는 해당 데이터를 별도의 추가 필터링 없이 원본 그대로 활용하였으며, 빈 필드 존재 여부를 확인하여 누락 데이터가 없음을 검증하였다.

총 17,280개의 QA 쌍으로 구성되며, 각 샘플은 질의응답 고유 ID(qa_id), 의료분야(domain), 질문 유형(q_type), 질문(question), 답변(answer)의 5개 필드로 이루어져 있다.

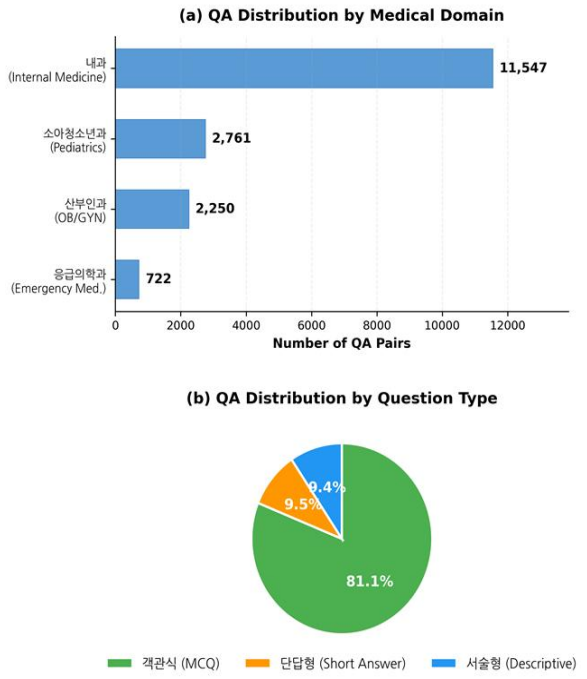

Fig. 1. Distribution of QA pairs: (a) by medical domain, (b) by question type.

Fig. 1의 (a)에서 볼 수 있듯이, 진료과별 QA 분포는 내과가 11,547쌍(66.8%)으로 가장 많고, 소아청소년과 2,761쌍(16.0%), 산부인과 2,250쌍(13.0%), 응급의학과 722쌍(4.2%) 순이다. 내과의 비중이 압도적으로 높은 것은 내과가 포괄하는 질환의 범위가 넓기 때문이다.

Fig. 1의 (b) 파이 차트는 질문 유형별 분포를 나타낸다. 객관식이 81.2%로 대다수를 차지하며, 서술형 9.4%, 단답형 9.5%로 구성되어 있다. 객관식 문항이 지배적인 것은 의학 교육 및 면허시험의 출제 형태를 반영한다.

2. Data Preprocessing

전체 데이터를 학습(Train) 13,824건, 검증(Validation) 1,728건, 테스트(Test) 1,728건으로 8:1:1 비율로 분할하였다. 각 QA 쌍은 Qwen2.5 모델이 사용하는 ChatML 형식의 Instruction 포맷으로 변환하였다[4]. 시스템 프롬프트는 "당신은 한국의 필수의료 분야에 특화된 의학 전문 AI 어시스턴트입니다. 임상 근거에 기반하여 정확하고 신뢰할 수 있는 의학 지식을 제공합니다."로 설정하였다. 변환된 Instruction의 평균 길이는 416 글자였다.

테스트셋의 질문 유형별 분포는 객관식 1,403건, 단답형 164건, 서술형 161건이다.

IV. Proposed Method

1. Base Model and Quantization Setup

베이스 모델로 Qwen2.5-7B-Instruct를 선택하였다[8]. 이 모델은 76억 개의 파라미터를 가지며, 한국어를 포함한 다국어 작업에서 우수한 성능을 보인다.

메모리 효율을 위해 BitsAndBytes 라이브러리를 활용한 4-bit NF4(NormalFloat4) 양자화를 적용하였다[11]. NF4는 정규분포를 따르는 모델 가중치에 대해 정보이론적으로 최적인 데이터 타입이다[7]. 추가로 이중 양자화(Double Quantization)를 적용하여 양자화 상수(quantization constants) 자체도 양자화함으로써 메모리 사용량을 더욱 절감하였다. 연산 정밀도는 bfloat16을 사용하였다.

양자화 적용 후 모델의 GPU 메모리 사용량은 약 7.21 GB로, 16GB VRAM의 소비자급 GPU에서 충분히 운용 가능한 수준이었다.

2. LoRA Adapter Configuration

LoRA 어댑터는 Transformer 블록의 Attention 레이어(q_proj, k_proj, v_proj, o_proj)와 MLP 레이어(gate_proj, up_proj, down_proj) 총 7개 대상 모듈에 적용하였다. 주요 하이퍼파라미터는 Table 1과 같다.

Table 1. LoRA Adapter Configuration

Parameter	Value
LoRA Rank (r)	16
LoRA Alpha (α)	32
LoRA Dropout	0.05
Target Modules	q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj
Bias	none
Task Type	CAUSAL_LM

이 설정에서 학습 가능한 파라미터는 40,370,176개로, 전체 파라미터 4,393,342,464개의 약 0.92%에 해당한다.

3. Training Setup

SFTTrainer를 활용하여 지도 미세조정(Supervised Fine-Tuning)을 수행하였다[5]. 주요 학습 하이퍼파라미터는 Table 2에 정리하였다.

Table 2. Training Hyperparameters

Parameter	Value
Epochs	3
Batch Size (per device)	2
Gradient Accumulation Steps	8
Effective Batch Size	16
Learning Rate	2×10^{-4}
LR Scheduler	Cosine
Warmup Ratio	0.05
Weight Decay	0.01
Max Gradient Norm	1
Max Sequence Length	1,024
Optimizer	Paged AdamW 8-bit
Gradient Checkpointing	Enabled
Precision	bfloat16

Paged AdamW 8-bit 옵티마이저를 사용하여 옵티마이저 상태의 메모리 사용량을 절감하였고, Gradient Checkpointing을 활성화하여 중간 활성화값의 메모리 사용을 최소화하였다.

4. Experimental Environment

본 실험은 Linux(WSL2) 환경에서 수행하였으며, 주요 소프트웨어 및 하드웨어 사양은 Table 3과 같다. GPU는 NVIDIA GeForce RTX 5070 Ti(16GB VRAM)를 사용하였으며, 학습 프레임워크로는 Hugging Face TRL 라이브러리의 SFTTrainer를 활용하였다. 모든 실험의 재현성을 위해 난수 시드(random seed)를 42로 고정하였다.

Table 3. Experimental Environment

Component	Specification
OS	Linux (WSL2)
Python	3.13.5
PyTorch	2.7.1+cu128
CUDA	12.8
GPU	NVIDIA GeForce RTX 5070 Ti
VRAM	16 GB
Base Model	Qwen/Qwen2.5-7B-Instruct
Quantization	4-bit NF4 (QLoRA)
Training Framework	Hugging Face TRL (SFTTrainer)
Random Seed	42

V. Experimental Results

1. Training Loss Curve Analysis

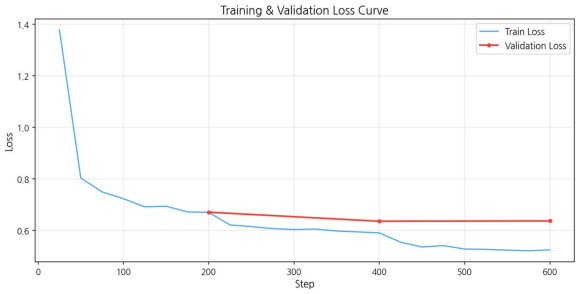

Fig. 2. Training and validation loss curves over training steps.

Fig. 2는 학습 과정에서의 훈련 손실과 검증 손실의 변화를 보여준다. 훈련 손실은 초기 약 1.38에서 시작하여 600스텝 이후 약 0.51까지 지속적으로 감소하였다. 검증 손실은 200스텝 부근에서 약 0.67로 처음 측정된 후, 학습이 진행됨에 따라 완만하게 감소하여 최종적으로 약 0.64에 수렴하였다.

훈련 손실과 검증 손실 사이의 간격은 학습 후반부로 갈수록 다소 벌어지는 경향을 보이나, 검증 손실이 상승 반전하지 않아 심각한 과적합(overfitting)은 발생하지 않은 것으로 판단된다. 이는 LoRA의 저차원 제약과 0.05의 드롭아웃이 정규화 효과를 제공한 것으로 해석할 수 있다.

2. Overall Performance Comparison

베이스 모델과 파인튜닝 모델의 성능을 비교하기 위해, 테스트셋에서 질문 유형별 균등 샘플링을 통해 198개 샘플을 추출하고, 각 모델로 답변을 생성하였다. 전체 테스트셋(1,728건) 대신 샘플링 평가를 수행한 것은, 베이스 모델과 파인튜닝 모델을 동일 GPU에서 순차적으로 추론하는 과정에서의 시간적 제약에 기인한다. 평가 지표로는 텍스트 중첩 기반의 ROUGE 스코어(ROUGE-1, ROUGE-2, ROUGE-L)와 객관식 문항에 대한 정답률(MCQ Accuracy)을 사용하

Table 4. Performance Comparison: Base Model vs. QLoRA Fine-tuned Model

Metric	Base Model	Fine-tuned (QLoRA)	Improvement
ROUGE-1 (Overall)	0.1309	0.2911	+0.1602
ROUGE-2 (Overall)	0.0277	0.0877	+0.0600
ROUGE-L (Overall)	0.1260	0.2911	+0.1651
ROUGE-L (MCQ)	0.2544	0.7146	+0.4602
ROUGE-L (Short Answer)	0.0119	0.0152	+0.0033
ROUGE-L (Descriptive)	0.1118	0.1436	+0.0318
MCQ Accuracy (AIHub Test)	0.00%	68.20%	+68.2%p
MCQ Accuracy (KorMedMCQA)	55.00%	58.00%	+3.0%p

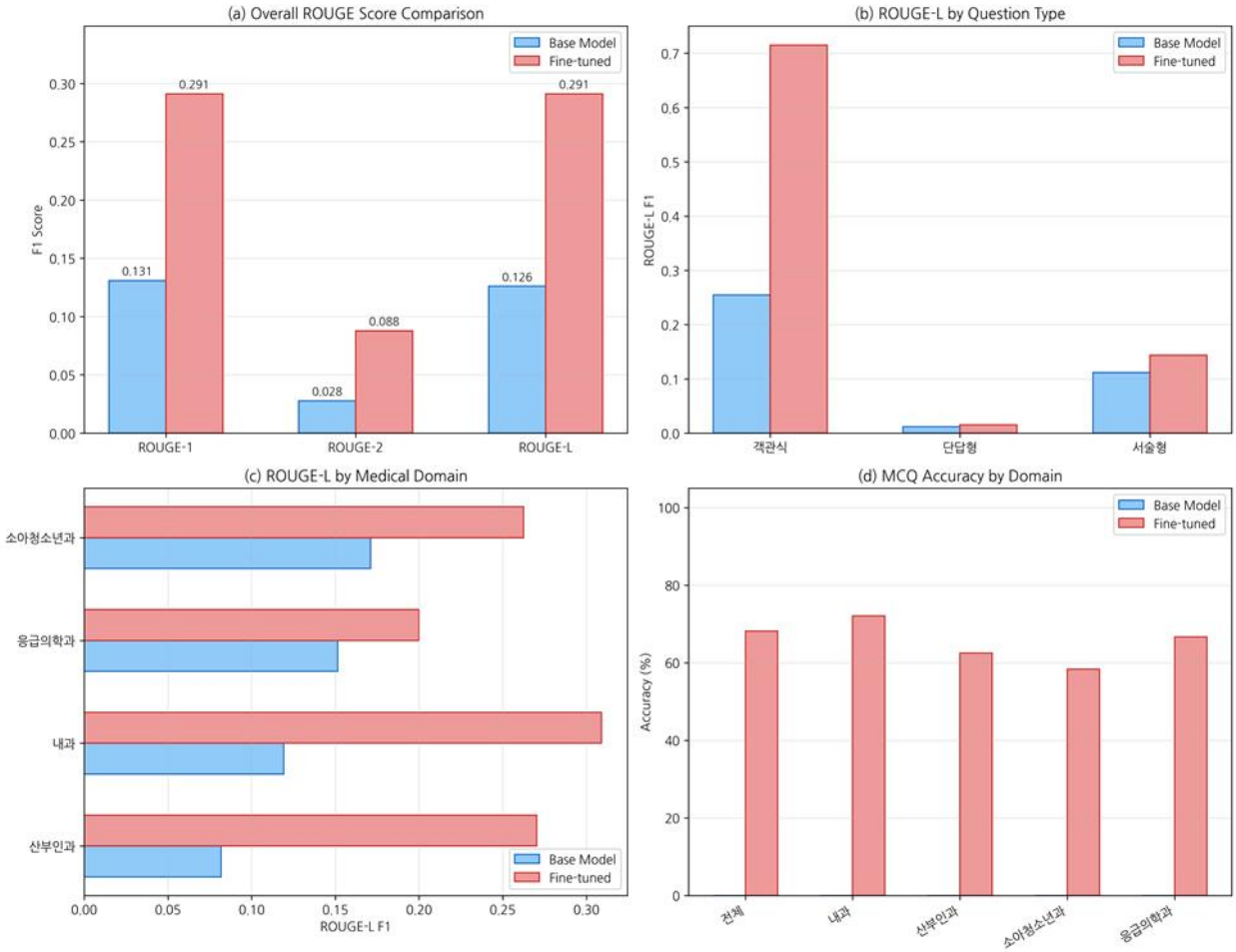

Fig. 3. Base model vs. fine-tuned model performance comparison: (a) overall ROUGE scores, (b) ROUGE-L by question type, (c) ROUGE-L by medical domain, (d) MCQ accuracy by domain.

였다[15]. 전체 결과는 Table 4에 정리하였다. 파인튜닝 모델은 모든 지표에서 베이스 모델을 상회하였다. ROUGE-1은 0.131에서 0.291로 122.3%, ROUGE-L은 0.126에서 0.291로 131.0% 향상되었다. 파인튜닝 모델에서 ROUGE-1과 ROUGE-L이 동일한 값(0.2911)을 보이는 것은, 평가 샘플의 다수를 차지하는 객관식 문항(81.2%)에서 모델이 "1) 아달리무맙"과 같은 짧은 정답 형식의 답변을 생성하기 때문이다. 답변이 소수의 토큰으로 구성될 경우, unigram 일치율(ROUGE-1)과 최장 공통 부분 수열 기반 스코어(ROUGE-L)가 수학적으로 수렴하게 된다. 다만 파인튜닝 후에도 ROUGE 스코어의 절대값이 0.3 미만인 것은, 의료 QA 답변이 동일 정답에 대해서도 다양한 표현이 가능하여 텍스트 중첩 기반 지표의 특성상 낮게 측정되는 경향이 있기 때문이다.

가장 두드러진 개선은 객관식 문항에서 나타났다. 베이스 모델의 AIHub 테스트셋 객관식 정답률은 0.0%로 나타났다. 이는 베이스 모델의 의학 지식이 부재하기 때문이 아니라, 베이스 모델이 "1)", "2)" 같은 선택지 번호 형식 대신 정답 내용을 포함한 장문의 설명을 생성하여 자동 정답 추출이 실패한 것에 기인한다. 따라서 파인튜닝의 효과는 의학 지식의 보강뿐 아니라, 한국 의료 QA에 적합한 출력 형식의 학습에도 기인한 것으로 해석된다. 반면 파인튜닝 모델은 68.2%의 정답률을 달성하였으며, 이는 모델이 한국 의료 면허시험 스타일의 출제 형식을 학습하여 간결한 답변 형식을 생성하게 된 것을 의미한다.

3. Detailed Analysis by Question Type and Medical Domain

Fig. 3(a)는 전체 ROUGE 스코어의 비교를 시각적으로 보여준다. ROUGE-1, ROUGE-2, ROUGE-L 모든 지표에서 파인튜닝 모델이 베이스 모델을 크게 상회한다.

Fig. 3(b)의 질문 유형별 ROUGE-L 분석에서, 객관식 유형은 베이스 모델 0.254에서 파인튜닝 후 0.715로 가장 큰 향상을 보였다. 이는 객관식 문항의 답변이 정형화된 짧은 형식이어서, 파인튜닝을 통한 출력 형식 학습 효과가 극대화된 결과로 해석된다. 반면 단답형은 0.012에서

0.015로, 서술형은 0.112에서 0.144로 상대적으로 적은 향상을 보였다. 단답형의 낮은 ROUGE-L은 정답이 짧은 단어나 구로 이루어져 있어, 모델이 정답과 동일한 표현을 생성하지 못하면 스코어가 급격히 떨어지는 ROUGE 메트릭의 특성에 기인한 것으로 판단된다.

Fig. 3(c)의 진료과별 ROUGE-L 분석에서는 4개 진료과 모두에서 파인튜닝 모델이 베이스 모델을 상회하였다. 특히 내과와 소아청소년과에서 높은 향상을 보였는데, 이는 이 두 진료과의 학습 데이터 비중이 각각 66.8%와 16.0%로 높기 때문으로 분석된다.

Fig. 3(d)의 진료과별 MCQ 정답률에서, 전체 평균 68.2%를 달성하였으며, 내과 약 72%, 산부인과 약 63%, 소아청소년과 약 58%, 응급의학과 약 66%의 정답률을 보였다. 내과의 정답률이 가장 높은 것은 학습 데이터에서 내과의 비중이 가장 높은 것과 일치하는 결과이다. 데이터 불균형이 모델 성능에 미치는 영향을 구체적으로 살펴보면, 학습 데이터 비중이 가장 높은 내과(66.8%)에서 MCQ 정답률 약 72%로 가장 높은 성능을 보인 반면, 비중이 가장 낮은 응급의학과(4.2%)에서는 약 66%에 그쳤다. 이는 학습 데이터 양과 모델 성능 사이에 양의 상관관계가 존재함을 시사하며, 특히 데이터가 부족한 진료과의 답변 신뢰성이 상대적으로 낮을 수 있음을 의미한다. 본 연구에서 오버샘플링이나 클래스별 손실 가중치 조절과 같은 불균형 해소 기법을 적용하지 않은 것은, 원본 데이터의 자연 분포를 그대로 반영한 상태에서의 파인튜닝 효과를 측정하는 것이 본 연구의 일차적 목적이었기 때문이다. 불균형 해소 기법 적용에 따른 성능 변화는 향후 연구에서 체계적으로 비교할 필요가 있다.

4. External Benchmark Evaluation (KorMedMCQA)

일반화 성능을 검증하기 위해 한국 의사 면허시험 기반 외부 벤치마크인 KorMedMCQA의 의사 시험(doctor) 분할에서 200문항을 무작위 샘플링하여 평가하였다[10]. 베이스 모델은 55.0%, 파인튜닝 모델은 58.0%의 정답률을 보여, 3.0%p의 소폭 향상을 확인하였다. 다만, 200문항 샘플 기준에서의 3.0%p 향상은 통계적으로 유의미한 차이로 단정하기 어려우며, 이는 파인튜닝 데이터(AI Hub)와 평가 데이터(KorMedMCQA)의 도메인 차이와 문항 난이도 분포의 상이함에 기인하는 것으로 판단된다. 따라서 외부 벤치마크에서의 성능 향상은 제한적이며, 이를 강건한 일반화 성능의 근거로 해석하기보다는 탐색적 결과로 이해하는 것이 적절하다.

내부 테스트셋(68.2%)에 비해 외부 벤치마크에서의 향상 폭이 작은 것은, KorMedMCQA가 실제 의사 면허시험 문항으로 구성되어 학습 데이터의 도메인과 난이도 분포가 상이하기 때문으로 분석된다. 또한 Kweon 등이 지적한 바와 같이, 한국어 의료 맥락은 미국 USMLE 기반 벤치마크와 상당한 차이를 보이므로, 한국 의료에 특화된 추가 학습 데이터의 확보가 성능 향상의 관건이 될 것으로 판단된다[10].

5. Qualitative Case Analysis

파인튜닝의 효과를 질적으로 확인하기 위해, ROUGE-L 향상이 가장 큰 사례를 분석하였다. 내과 영역의 "크론병의 관해 유도를 위해 면역 반응 조절에 주로 사용되는 약물"에 대한 5지선다형 문항에서, 베이스 모델은 정답인 "아달리무맙"을 포함하는 장문의 설명을 생성하였으나 정답 번호를 명시하지 않았다(ROUGE-L = 0.0). 반면 파인튜닝 모델은 "1) 아달리무맙"이라는 간결하고 정확한 답변을 생성하여 ROUGE-L = 1.0을 달성하였다. 이는 파인튜닝을 통해 모델이 의료 QA의 답변 형식 관습을 학습한 것을 보여주는 대표적 사례이다.

파인튜닝 효과의 의미론적 측면을 검증하기 위해, 서술형 및 단답형 답변 30건을 대상으로 Claude Sonnet 4를 평가자(judge)로 활용한 LLM 기반 의미론적 평가를 수행하였다. 의학적 정확성(medical accuracy), 완전성(completeness), 관련성(relevance)을 5점 척도로 평가한 결과를 Table 5에 정리하였다.

Table 5. LLM-based Semantic Evaluation (5-point scale, n=30)

Metric	Base Model	Fine-tuned	Δ
Medical Accuracy	3.30	2.50	-0.80
Completeness	2.93	2.43	-0.50
Relevance	3.83	3.30	-0.53

베이스 모델이 서술형·단답형 문항에서 파인튜닝 모델보다 높은 의학적 정확성 점수를 기록한 것은 주목할 만하다. 이는 베이스 모델이 장문의 상세한 설명을 생성하여 정답 내용을 더 풍부하게 포함하는 반면, 파인튜닝 모델은 학습 데이터의 간결한 답변 형식에 맞춰 핵심만 짧게 출력하면서 일부 세부 사항을 누락하거나 간략화한 것에 기인한다. 한편, 베이스 모델은 다수의 샘플에서 한국어에서 중국어로 언어가 전환(code-switching)되는 현상이 관찰된 반면, 파인튜닝 모델은 전체 샘플에서 한국어 답변을 일관되게 유지하였다. 이 결과는 본 연구의 파인튜닝이 의학 지식의 새로운 습득보다는 한국어 의료 QA에 적합한 출력 형식의 정렬과 언어 일관성의 확보에 주된 효과가 있었음을 시사한다.

VI. Conclusion

본 연구에서는 QLoRA를 활용하여 한국형 필수의료 분야에 특화된 sLLM 파인튜닝을 수행하고, 그 효과를 정량적·정성적으로 검증하였다. 주요 연구 성과를 정리하면 다음과 같다.

첫째, AI Hub 필수의료 의학지식 데이터 17,280쌍을 활용한 Instruction Tuning을 통해, Qwen2.5-7B-Instruct 모델의 한국 의료 QA 성능을 유의미하게 향상시켰다. ROUGE-L은 131.0%(0.126→0.291), 내부 객관식 정답률은 68.2%p(0.0%→68.2%) 개선되었다.

둘째, 전체 파라미터의 0.92%만을 학습하는 QLoRA 기법을 한국어 의료 도메인에 적용하여, 단일 소비자급 GPU(RTX 5070 Ti, 16GB VRAM)에서 약 7GB의 메모리로 파인튜닝이 가능함을 확인하였다. 이는 Dettmers 등[7]이 제시한 QLoRA의 메모리 효율성을 한국어 의료 도메인에서 재확인한 것으로, 구체적인 학습 설정과 메모리 사용량을 보고함으로써 후속 연구의 재현성을 확보하였다.

다만, 본 연구에서 제시한 모델은 연구 목적의 프로토타입으로서, 실제 임상 환경에 직접 적용하기에는 한계가 있다. 모델의 MCQ 정답률이 68.2%로, 약 31.8%의 오답이 발생하며, 이는 의료 현장에서 부정확한 정보 제공으로 이어질 수 있다. 따라서 본 모델을 포함한 의료 AI 시스템은 반드시 의료 전문가의 감독(human oversight) 하에 보조적 수단으로만 활용되어야 하며, 독립적 진단이나 처방 도구로 사용되어서는 안 된다.

셋째, 외부 벤치마크(KorMedMCQA)에서 3.0%p의 향상을 확인하여, 학습 도메인 외 의료 지식에 대한 일정 수준의 일반화 가능성을 보였다.

그러나 본 연구에는 몇 가지 한계가 존재한다. 단답형 및 서술형 문항에서의 ROUGE-L 향상이 상대적으로 작았으며, 이는 ROUGE 메트릭 자체의 한계와 서술형 답변의 다양성에 기인한다. 또한 진료과 간 데이터 불균형(내과 66.8%)이 존재하여, 데이터가 적은 응급의학과(4.2%)에서는 상대적으로 낮은 성능을 보였다. 아울러, 본 연구의 평가는 테스트셋 1,728건 중 유형별 균등 샘플링된 198건에 대해 수행되었으며, 이는 전체 테스트셋의 약 11.5%에 해당하여 결과의 대표성에 한계가 있다. 전체 테스트셋에 대한 평가는 향후 보완이 필요하다. KorMedMCQA 평가에서의 3.0%p 향상 역시 200문항 샘플 기준으로, 통계적 유의성 검증이 추가로 요구된다. 또한, 다른 베이스 모델과의 비교 실험이나 LoRA 하이퍼파라미터에 대한 ablation study를 수행하지 못한 점도 본 연구의 한계이다.

향후 연구에서는 다음과 같은 방향을 모색할 수 있다. 첫째, BERTScore, G-Eval 등 의미론적 평가 지표를 추가로 도입하여 서술형 답변의 품질을 보다 정확하게 측정할 필요가 있다. 둘째, DPO(Direct Preference Optimization)나 RLHF(Reinforcement Learning from Human Feedback) 기반의 정렬(Alignment) 학습을 통해 답변의 안전성과 신뢰성을 강화할 수 있다[5, 16]. 셋째, RAG(Retrieval-Augmented Generation) 기법과의 결합을 통해, 최신 의학 가이드라인을 실시간으로 반영하는 시스템 구축을 탐색할 수 있다[17]. 넷째, 진료과 간 데이터 불균형을 해소하기 위한 데이터 증강(data augmentation) 기법의 적용이 필요하다.

아울러, 의료 분야에서 LLM을 활용할 경우 환각(hallucination)으로 인한 부정확한 정보 생성이 환자 안전에 직접적인 위험을 초래할 수 있다. 이를 완화하기 위해 RAG 기반의 근거 문서 연동을 통해 모델 출력의 사실 근거를 확보하고, 생성 답변에 대한 신뢰도 점수 산출 및 불확실성이 높은 답변에 대한 자동 경고 메커니즘의 도입이 필요하다. 궁극적으로 의료 AI 시스템은 의료진의 감독(human oversight) 하에 보조 도구로 활용되어야 하며, 독립적 의사결정 수단으로 사용되어서는 안 된다.

ACKNOWLEDGEMENT

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the National Program in Medical AI Semiconductor(2024-0-0097) supervised by the IITP(Institute of Information & Communications Technology Planning & Evaluation) in 2026

REFERENCES

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, et al., "Attention Is All You Need," in Advances in Neural Information Processing Systems 30 (NeurIPS), pp. 5998-6008, 2017. DOI: 10.48550/arXiv.1706.03762

[2] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, et al., "Language Models are Few-Shot Learners," in Advances in Neural Information Processing Systems 33 (NeurIPS), pp. 1877-1901, 2020. DOI: 10.48550/arXiv.2005.14165

[3] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung,

et al., "Large Language Models Encode Clinical Knowledge," Nature, Vol. 620, pp. 172-180, Jul. 2023. DOI: 10.1038/s41586-023-06291-2

[4] J. Wei, M. Bosma, V. Zhao, K. Guu, A. W. Yu, B. Lester, et al., "Finetuned Language Models Are Zero-Shot Learners," in Proc. of the International Conference on Learning Representations (ICLR), 2022. DOI: 10.48550/arXiv.2109.01652

[5] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, et al., "Training Language Models to Follow Instructions with Human Feedback," in Advances in Neural Information Processing Systems 35 (NeurIPS), pp. 27730-27744, 2022. DOI: 10.48550/arXiv.2203.02155

[6] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-Rank Adaptation of Large Language Models," in Proc. of the International Conference on Learning Representations (ICLR), 2022. DOI: 10.48550/arXiv.2106.09685

[7] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLoRA: Efficient Finetuning of Quantized LLMs," in Advances in Neural Information Processing Systems 36 (NeurIPS), 2023. DOI: 10.48550/arXiv.2305.14314

[8] A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, et al., "Qwen2.5 Technical Report," arXiv preprint arXiv:2412.15115, 2024. DOI: 10.48550/arXiv.2412.15115

[9] AI Hub, "Essential Medical Knowledge Data," [Online]. Available: https://www.aihub.or.kr/aihubdata/data/view.do?aihubDataSe=data&dataSetSn=71875

[10] S. Kweon, B. Choi, M. Kim, R. W. Park, and E. Choi, "KorMedMCQA: Multi-Choice Question Answering Benchmark for Korean Healthcare Professional Licensing Examinations," arXiv preprint arXiv:2403.01469, 2024. DOI: 10.48550/arXiv.2403.01469

[11] T. Dettmers, M. Lewis, Y. Belkada, and L. Zettlemoyer, "LLM.int8(): 8-bit Matrix Multiplication for Transformers at Scale," in Advances in Neural Information Processing Systems 35 (NeurIPS), 2022. DOI: 10.48550/arXiv.2208.07339

[12] E. Frantar, S. Ashkboos, T. Hoefler, and D. Alistarh, "GPTQ: Accurate Post-Training Quantization for Generative Pre-Trained Transformers," in Proc. of the International Conference on Learning Representations (ICLR), 2023. DOI: 10.48550/arXiv.2210.17323

[13] D. Jin, E. Pan, N. Oufattole, W.-H. Weng, H. Fang, and P. Szolovits, "What Disease does this Patient Have? A Large-scale Open Domain Question Answering Dataset from Medical Exams," Applied Sciences, Vol. 11, No. 14, p. 6421, 2021. DOI: 10.3390/app11146421

[14] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, et al., "The Llama 3 Herd of Models," arXiv preprint arXiv:2407.21783, 2024. DOI: 10.48550/arXiv.2407.21783

[15] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," in Text Summarization Branches Out: Proc. of the ACL-04 Workshop, pp. 74-81. Barcelona, Spain, Jul. 2004.

[16] R. Rafailov, A. Sharma, E. Mitchell, S. Ermon, C. D. Manning, and C. Finn, "Direct Preference Optimization: Your Language Model is Secretly a Reward Model," in Advances in Neural Information Processing Systems 36 (NeurIPS), 2023. DOI: 10.48550/arXiv.2305.18290

[17] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in Advances in Neural Information Processing Systems 33 (NeurIPS), pp. 9459-9474, 2020. DOI: 10.48550/arXiv.2005.11401

[18] S.-Y. Liu, C.-Y. Wang, H. Yin, P. Molchanov, Y.-C. F. Wang, K.-T. Cheng, and M.-H. Chen, "DoRA: Weight-Decomposed Low-Rank Adaptation," in Proc. of the International Conference on Machine Learning (ICML), pp. 32100-32121, 2024. DOI: 10.48550/arXiv.2402.09353

[19] J. Zhao, Z. Zhang, B. Chen, Z. Wang, A. Anandkumar, and Y. Tian, "GaLore: Memory-Efficient LLM Training by Gradient Low-Rank Projection," in Proc. of the International Conference on Machine Learning (ICML), pp. 61121-61143, 2024. DOI: 10.48550/arXiv.2403.03507

Authors

JongHwi Song received the B.S., M.S. and Ph.D. degrees in Computer Science and Engineering from Yonsei University, Korea, in 2012, 2015 and 2023, respectively. Dr. Song joined the Institute of AI Convergence Science at Yonsei University, Wonju, Korea, in 2025. He is currently a Research Professor at the Institute of AI Convergence Science, Yonsei University. He is interested in text mining, large language models (LLMs), and on-device AI.

Urtnasan Erdenebayar received his B.S. degree in Computer Science from Huree University, Ulaanbaatar, Mongolia, in 2007, and his M.S. degree in Electronic Engineering from Inha University, Incheon, Korea, in 2010.

He earned his Ph.D. in Biomedical Engineering from Yonsei University, Seoul, Korea, in 2018. From 2018 to 2024, he served as a Postdoctoral Researcher and Research Professor at Yonsei University and Wonju Severance Christian Hospital. He is currently a faculty member in the Department of AI Semiconductor at Yonsei University. His research interests include AI systems, on-device AI, edge AI, and medical AI.