로컬 의료 환경을 위한 QLoRA 기반 소형 언어 모델의 미세조정 연구

(A Study on QLoRA-based SLM Fine-tuning for Local Medical Environments)

유영준*, 김웅식**

(Young-Joon Yoo, Woong-Sik Kim)

요 약

본 연구는 의료진의 임상 문서화 부담을 경감하기 위해 의사와 환자 간의 비정형 대화로부터 섹션별 임상 노트를 자동 생성하는 시스템의 성능을 비교 분석하였다. 의료 데이터의 민감성과 데이터 부족 문제를 해결하기 위해 소형 언어 모델(LLaMA 3.2 3B, Qwen3 4B, Gemma3 4B)에 QLoRA 미세조정 기법을 적용하였으며, 이를 거대 모델인 Gemini 2.5 Flash의 In-context Learning 방식과 비교하였다. 실험 결과, 섹션 분류 정확도에서는 미세조정된 Qwen3 4B가 0.735로 가장 우수한 성능을 보였다. 요약 성능의 경우, Gemma3 4B가 Aggregate Score에서 0.613으로 가장 높은 효율성을 입증하였다. 반면 Gemini 2.5 Flash는 ROUGE-2와 BLEURT 지표에서 우위를 보여 문장의 유창성 측면에서 강점을 나타냈다. 본 결과는 보안이 중요한 의료 환경에서 소형 언어 모델의 미세조정이 실용적이고 효과적인 대안이 될 가능성을 확인하였다.

■ 중심어 : 딥러닝 ; 소형 언어 모델 ; 임상 노트 생성

Abstract

This study evaluates the performance of an automated clinical note generation system from unstructured doctor-patient dialogues to alleviate the clinical documentation burden on medical staff. To address the challenges of data sensitivity and scarcity in the medical domain, we applied QLoRA fine-tuning to small language model (SLM) specifically LLaMA 3.2 3B, Qwen3 4B, and Gemma3 4B and compared their performance with the In-context Learning approach of a Large Language Model (LLM), Gemini 2.5 Flash. Experimental results demonstrated that the Fine-tuned Qwen3 4B achieved the highest section classification accuracy of 0.735. In terms of summarization performance, Gemma3 4B proved most efficient with a leading Aggregate Score of 0.613. Meanwhile, Gemini 2.5 Flash showed strengths in linguistic fluency, outperforming others in ROUGE-2 and BLEURT metrics. These findings suggest that Fine-tuning SLM using QLoRA offers a practical and effective alternative for structured clinical documentation in secure medical environments where data privacy is paramount.

■ keywords : Deeplearning ; Small Language Model ; Medical Report Generation

Ⅰ. 서 론

최근 의료 현장에서 임상 문서화에 대한 부담은 의료진의 직무 만족도를 저하시키고 번아웃을 유발하는 결정적인 요인으로 지목되고 있다. 의료진은 환자 진료라는 본연의 업무 외에도 상담 내용을 기록하고 관리하는 행정적 작업에 상당한 시간과 에너지를 할애하고 있으며, 이는 결과적으로 환자와의 소통 질을 저하시키는 원인이 된다[1-3]. 특히 코로나19 팬데믹 이후 텔레케어 서비스가 급격히 확산됨에 따라 비대면 상

* 준회원, 건양대학교 AI소프트웨어융합학과
** 정회원, 건양대학교 인공지능학과
이 논문은 2025년도 건양대학교 학술연구비 지원에 의하여 이루어진 것임
접수일자 : 2025년 12월 30일
수정일자 : 2026년 01월 28일
게재확정일 : 2026년 02월 11일
교신저자 : 김웅식 e-mail : wskim@konyang.ac.kr

담 데이터가 기하급수적으로 증가하였고, 이러한 방대한 대화 기록을 효율적으로 요약하여 임상 노트로 변환해야 할 필요성이 그 어느 때보다 높아졌다[4]. 정확한 문서화는 의료진 간의 원활한 정보 공유와 안전한 환자 진료를 위한 필수 조건이지만, 이를 수동으로 작성하는 방식은 규모의 경제 측면에서 한계가 분명하다.

의료 상담 대화를 자동 요약하는 기술은 일반적인 텍스트 요약과 달리 몇 가지 고유한 난제를 안고 있다. 의사와 환자 간의 대화는 구조화되지 않은 자연어 형태로 진행되며, 한 번의 상담 내에서도 여러 증상과 질환이 복합적으로 논의되는 특성이 있다. 또한, 실제 대화에서는 일상적인 구어체가 사용되지만 최종적인 임상 노트에는 대화에 명시적으로 등장하지 않는 전문 의학 용어를 사용하여 요약해야 하므로 고도의 추상적 요약 능력이 요구된다[5,6]. 무엇보다 의료 데이터의 민감성으로 인한 데이터 부족 문제는 딥러닝 모델의 학습을 어렵게 만드는 주요한 병목 현상으로 작용해 왔다[7,8]. 이러한 한계를 극복하기 위해 자연어 처리 분야에서는 다양한 모델 아키텍처를 활용한 연구가 진행되어 왔다. 초기 연구들은 주로 사전 학습된 트랜스포머 모델을 의료 도메인 데이터에 맞게 미세조정하여 성능을 최적화하는 방식을 취해 왔다[9]. 하지만 최근 LLM의 등장으로 인해 적은 양의 예시만으로도 문맥을 파악하는 In-Context Learning(ICL)이 가능해지면서, 데이터 희소성 문제를 해결할 새로운 대안으로 떠오르고 있다. 특히 GPT-4와 같은 최신 모델은 전통적인 지도 학습 방식의 모델들을 능가하는 생성 능력을 보여주며 임상 노트 자동 생성의 가능성을 증명하고 있다[10,11]. 하지만 LLM을 직접 구축하는 것은 막대한 하드웨어 비용이 발생하며, 외부 API 활용 시에는 민감한 의료 정보 유출 등 보안 취약점이 존재한다. 이러한 한계를 극복하기 위해 최근에는 특정 도메인에 특화된 소형 언어 모델(SLM)을 미세조정하여 폐쇄형 환경에서 효율적으로 활용하려는 연구들이 활발히 진행되고 있다[12-14].

본 연구에서는 의료 상담 대화로부터 특정 섹션별 임상 노트를 자동 생성하는 시스템의 성능을 분석하고 비교하고자 한다. 이를 위해 비정형 대화를 적절한 섹션 헤더로 분류하는 동시에, 각 섹션의 특성에 맞는 요약문을 추출하는 프로세스를 구축하였다. 특히 단일 모델의 성능에 의존하기보다 다양한 모델과 비교 분석하였으며, 미세조정한 소형 언어 모델과 프롬프트 기반의 거대 모델 간의 성능 편차를 벤치마킹하였다. 이러한 비교 연구는 개인정보 보호와 실시간 처리가 필수적인 의료 현장의 특수성을 고려하여 제한된 데이터 환경에서 임상 노트 생성의 품질을 극대화할 수 있는 최적의 모델 조합과 전략을 제시함으로써, 의료 현장의 행정 효율화와 진료 품질 향상에 기여할 수 있는 기술적 가능성을 확인하고자 한다.

Ⅱ. 관련 연구

1. 임상 노트 자동 생성 연구 동향

비정형 의료 대화를 구조화된 임상 노트로 변환하기 위한 연구들이 최근 활발히 진행되고 있다. Wang 등[15]은 사전 학습된 언어 모델의 미세조정과 GPT-4의 ICL을 결합한 하이브리드 전략을 통해 요약 성능을 극대화하는 방식을 제안하였으며, Mathur 등[16]은 시맨틱 유사도에 기반하여 최적의 예시를 선별하는 Few-shot 프롬프팅 기법의 효과를 입증하였다. 또한, Tang 등[17]은 BioBART와 같은 의료 특화 모델의 미세조정 결과와 거대 언어 모델의 성능을 다각도로 비교 분석하였고, Sharma 등[18]은 데이터 증강 기법을 활용하여 소규모 의료 데이터셋의 한계를 극복하고자 하였다. Mishra 등[19]은 다양한 트랜스포머 기반 모델들의 앙상블을 통해 생성된 문서의 안정성을 높이는 연구를 수행하였다. 이러한 기존 연구들은 주로LLM의 성능 지표 향상이나 복잡한 모델 앙상블에 초점을 맞추어 왔다. 그러나 이는 높은 연산 자원이 요구되거나 외부 API 활용에 따른 데이터 보안 문제를 수반한다. 본 연구는 이러한 선행 연구들과 달리 보안이 강조

되는 로컬 의료 환경을 상정하여, 저사양 인프라에서도 운용 가능한 SLM 기반의 QLoRA 실무적 적용 가능성을 분석하였다.

2. LLaMA3

Llama 3는 Meta에서 공개한 오픈 소스 파운데이션 모델이다[20]. Dense Transformer 아키텍처를 기반으로 하며, 추론 효율성을 극대화하기 위해 8개의 Key-Value Heads를 사용하는 Grouped Query Attention을 채택하였다[21]. 또한, 더 긴 컨텍스트와 향상된 압축률을 지원하기 위해 128K 토큰 규모의 Vocabulary와 500,000의 ROPE 베이스 주파수를 적용하였다[22]. 본 연구에서 활용한 Llama 3.2 3B 모델은 이러한 Llama 3의 소형 모델이다. 대형 모델로부터 추출된 지식을 사후 훈련 단계에서 활용하여 품질을 개선하였다. 특히 사후 훈련 단계에서는 Supervised Finetuning(SFT)과 Direct Preference Optimization[23]를 결합한 다단계 정렬 기법이 사용되었다. 이를 통해 Llama 3.2 3B와 같은 SLM은 제한된 파라미터 수에도 다양한 태스크에서 높은 성능을 보여준다. 이러한 특성은 온디바이스 환경이나 저전력 서버 환경에서 효율적인 고성능 추론을 가능하게 한다.

3. Qwen3

Alibaba Qwen 팀이 발표한 Qwen3 시리즈는 단일 프레임워크 내에서 Thinking Mode와 Non-thinking Mode를 통합적으로 구현하였다[24]. 본 연구에서 활용한 Qwen3 4B는 엣지 컴퓨팅 및 실시간 응용 분야에 최적화된 모델이다. 해당 모델은 36조 개의 대규모 토큰을 활용한 3단계 사전 훈련 과정을 거쳤다. Qwen3 4B의 성능 극대화를 위해 적용된 Strong-to-Weak Distillation 기법은 대형 모델의 지식을 효율적으로 전이하여, 기존 강화학습 방식 대비 약 10분의 1의 연산 비용만으로도 상위 모델 수준의 추론 정확도를 확보하여 경량 모델이 복잡한 에이전트 및 추론 작업에서 중추적인 역할을 수행할 수 있음을 나타낸다.

4. Gemma3

Google DeepMind에서 개발한 Gemma 3는 차세대 경량 오픈 모델 시리즈이다[25]. 본 연구에 활용한 Gemma 3 4B는 모바일, 노트북 같은 저사양 하드웨어에서 원활한 구동을 목표로 설계되었으며, 이를 위해 혁신적인 아키텍처 개선을 단행했다. 특히 긴 문맥 처리 시 발생하는 KV-캐시 메모리 폭발 문제를 해결하기 위해, Local sliding window self-attention[26]과 Global self-attention[27]을 5:1 비율로 교차 배치하는 Layer Interleaving 구조를 적용하였다. 이를 통해 모델은 성능 저하 없이 최대 128K 토큰의 문맥을 수용하면서도 메모리 사용량을 획기적으로 줄였다. 또한 기존의 Soft-capping 대신 QK-norm을 도입하여 학습 안정성을 높였으며, 지식 증류 기법을 통해 이전 세대보다 훨씬 적은 파라미터로도 대규모 모델와 비슷한 성능을 구현하였다.

5. Gemini 2.5

Google DeepMind에서 개발한 대규모 언어 모델 Gemini 2.5 시리즈는 기존 Gemini 1.5 모델[28]의 아키텍처를 계승 및 발전시켜, Advanced Reasoning, Multimodality, Long Context 처리 능력을 통합적으로 제공한다[29]. 특히 이 모델군은 Sparse Mixture-of-Experts 트랜스포머 아키텍처[30,31]를 기반으로 설계되어, 총 모델 용량과 토큰당 계산 비용을 분리함으로써 높은 성능과 효율성을 동시에 달성하였다.

본 연구에 활용한 Gemini 2.5 Flash는 하이브리드 추론 모델로서, 품질, 비용, 지연 시간 사이의 최적의 균형을 제공하는 것이 특징이다. 이 모델은 사용자가 Thinking Budget을 동적으로 제어할 수 있는 기능을 지원하며, 이를 통해 추론 시간 계산량을 조절하여 복잡한 문제 해결 시 정확도를 획기적으로 높일 수 있다.

6. QLoRA

LLM의 미세 조정은 특정 태스크 최적화에 효과

적이지만, 파라미터 증대에 따른 GPU 메모리 비용이 매우 증가하는 문제가 있다. 특히 본 연구에서 다루는 의료 데이터는 보안상 로컬 환경에서 많이 사용하기 때문에 효율적인 학습 기법인 QLoRA는 필수적이다. QLoRA는 다음과 같은 세 가지 핵심 기술을 통해 모델의 성능을 유지하면서 메모리 사용량을 획기적으로 줄였다[32]. 첫째로 4-bit NormalFloat(NF4) 데이터 타입은 가중치가 정규 분포를 따르는 사전 학습 모델의 통계적 특성을 활용한다. 이는 단순 양자화와 달리 정보 이론적으로 최적화된 비트 할당을 수행하여, 의료 대화의 미세한 문맥적 뉘앙스를 보존하는 데 기여한다. 둘째로 Double Quantization를 통해 양자화 상수를 다시 8-bit로 양자화하여 파라미터당 약 0.37비트의 추가 메모리를 절감한다. 이는 본 연구에서 활용한 SLM이 제한된 VRAM 내에서 더 긴 컨텍스트를 처리할 수 있는 여유 자원을 제공한다. 셋째로 Paged Optimizers를 도입하여 GPU 메모리가 부족할 때 체크포인트를 CPU로 일시 페이징함으로써 메모리 스파이크를 관리한다. 이는 비정형 대화 데이터 학습 중 발생하는 예기치 못한 연산 부하 환경에서도 중단 없는 미세 조정을 가능하게 한다.

Ⅲ. 방법론

본 연구는 의료 데이터의 보안성과 자원 제한적인 로컬 환경에서의 실무적 실효성을 확인하기 위해 연구 설계 단계에서 비교군을 설정하였다. 실제 의료 현장에서 선택 가능한 두 가지 방법인 로컬 모델과 외부 API 활용 사이의 실효성을 대조 분석하는 데 중점을 두었다. 이는 데이터 외부 유출을 차단하는 폐쇄형 인프라 구축과 범용 거대 모델 활용 중 무엇이 임상 문서화 시스템 설계에 더 적합한 대안인지 비교하기 위함이다. SLM 후보군은 구조적 다양성과 추론 효율성을 기준으로 선정하였다. 밀집 트랜스포머 아키텍처의 대표 모델인 LLaMA 3.2 3B, 논리 추론 성능이 검증된 Qwen3 4B, 그리고 최신 멀티모달 아키텍처인 Gemma3 4B를 비교군으로 설정함으로써, 각 모델의 구조적 차이가 특정 도메인 데이터 적응력 및 가중치 수렴에 미치는 영향을 분석하고자 하였다. 파라미터 규모가 매우 큰 LLM인 Gemini 2.5 Flash를 대조군으로 포함시킨 것은 저사양 인프라 환경에서 단순히 모델 크기보다 도메인 특화 데이터에 대한 정밀한 미세조정이 성능 향상에 미치는 영향을 확인하기 위함이다. 이러한 비교는 보안 중심의 로컬 의료 AI 시스템 구축을 위한 기술적 근거가 될 수 있다.

1. SLM 미세조정

첫 번째 파이프라인은 보안과 자원 효율성을 극대화하기 위해서 SLM에 QLoRA 미세조정을 적용한 방식이며, 두 번째 파이프라인은 LLM의 추론 능력을 기준으로 삼기 위해 전략적 프롬프트 엔지니어링을 통한 ICL 방식이다. 두 가지 파이프라인의 비교를 통해 데이터 보안 및 컴퓨팅 자원이 제한된 의료 현장에서 실질적으로 사용 가능한 기술적 프레임워크의 실효성을 고찰한다.

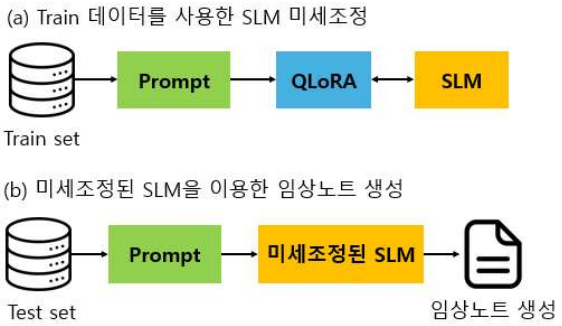

그림 1. SLM 미세조정 파이프라인

그림 1은 SLM의 미세조정 및 임상 노트 생성 시스템의 전체 아키텍처를 나타낸다. 시스템은 크게 학습하는 (a) 과정과 추론하는 (b) 과정으로 설계되었다. (a) 과정은 Train 데이터셋을 활용하여 모델에 도메인 특화 지식을 주입하는 SFT 단계이다. 본 연구에서는 단일 GPU 환경에서의 효율적인 학습을 위해 QLoRA 기법을 사용하였다. 각 SLM을 NF4 타입으로 양자화하여 VRAM 효율성을 극대화하면서 로컬 환경에서의 학습 가능성을 확인하였다. 단일 모델에서 분류와 요약이라는 복합적인 태스크를

수행하기 위해, 어텐션 레이어의 핵심 모듈인 q_proj, k_proj, v_proj, o_proj를 타켓 모듈로 선정하였다. 이를 통해 저사양 로컬 환경에서도 모델이 의료 전문 용어의 분류 체계와 문맥적 뉘앙스를 동시에 학습할 수 있도록 설정하였다. 파라미터 업데이트의 효율성을 위해 Rank(r)는 8, Alpha는 32로 설정하여 모델의 유연성을 확보하였으며, Paged AdamW 8-bit 옵티마이저를 활용하여 학습 중 발생하는 메모리 스파이크를 제어하였다. (b) 과정은 학습이 완료된 모델을 추론하는 단계이다. Test 데이터셋이 입력되면 (a) 과정에서 사용된 것과 동일한 구조의 프롬프트를 모델에 입력한다. 미세조정된 SLM은 학습된 도메인 지식을 바탕으로 비정형 대화 내에서 적절한 섹션 분류를 수행함과 동시에 임상 노트를 생성한다. 기존의 의료 문서화 연구들이 분류 모델과 요약 모델을 따로 실행하여 연산 복잡도를 높였던 것과 달리, 본 시스템은 단일 SLM 내에서 섹션 분류와 요약문 생성이 동시에 생성하여 모델이 대화의 전체 맥락을 유지하면서도 특정 섹션의 핵심 정보를 즉각적으로 구조화하게 함으로써, 추론 속도를 극대화하고 시스템 운영의 복잡성을 감소하도록 설계하였다.

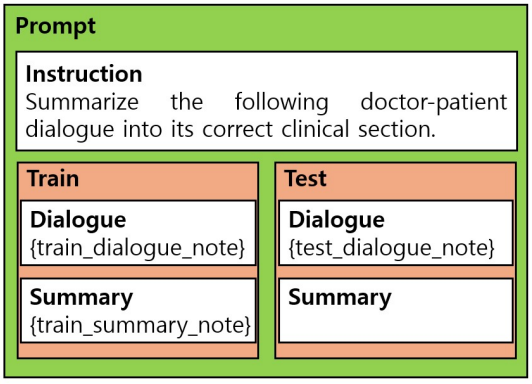

그림 2. SLM 미세조정을 위한 프롬프트 구조

그림 2는 그림 1에서 (a) SFT 단계와 (b) 추론 단계에 적합한 프롬프트 설계를 나타낸다. 프롬프트 상단에 위치한 Instruction은 모델의 역할과 출력 형식을 규정한다. 구체적인 설명을 통해 모델이 단순한 대화 생성이 아닌 의료 요약 및 분류 작업을 하도록 지시한다. 동일한 지시문을 Train, Test 모두 사용하여 추론 단계에서도 학습 시 습득한 지식을 일관되게 유지하도록 설계하였다. SFT는 모델에게 입력값과 정답 쌍을 제공하여 학습시키는 방식이다. Train 블록은 SFT 방식을 고려하여 입력값인 Dialogue와 정답인 Summary를 사용한다. Test 블록은 정답인 Summary를 제외한 Instruction, Dialogue만 모델의 입력으로 사용한다.

2. Gemini를 활용한 Few-shot ICL

SLM의 미세조정 성능을 평가하고 실용성을 비교하기 위해, Gemini 2.5 Flash를 활용한 ICL 파이프라인을 설계하였다. 이는 별도의 가중치 업데이트 없이 프롬프트 내에 포함된 예시와 지시문만으로 모델의 추론 능력을 제어하는 방식이다.

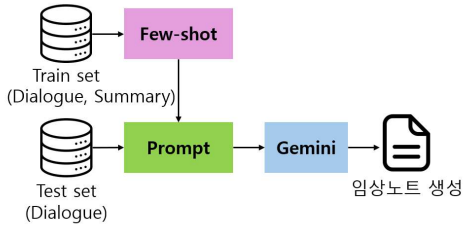

그림 3. Gemini를 활용한 ICL 파이프라인

그림 3은 Gemini 모델의 추론 파이프라인을 나타낸다. SFT 파이프라인과 달리 학습 과정 대신 무작위로 선정된 3개의 Train 데이터셋을 프롬프트에 포함하여, 모델이 실시간으로 의료 대화 요약의 구조와 문맥을 파악하도록 설계하였다. Test 데이터셋이 입력되면, 모델은 제시된 Few-shot 예시의 패턴을 참고하여 임상 노트를 생성한다.

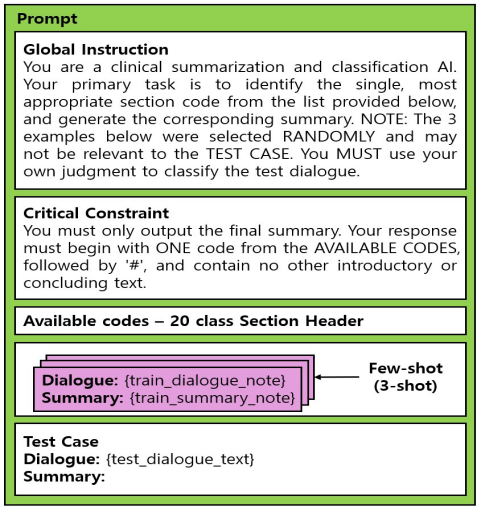

그림 4. Gemini를 활용한 ICL 프롬프트 구조

그림 4는 모델의 추론을 정밀하게 지시하기 위해 4개의 블록으로 프롬프트를 설계하였다. Global Instruction은 모델의 목적을 명시한다. 임상 요약 및 분류 역할을 지시하여 의료 전문 지식을 바탕으로 추론하도록 유도한다. 또한, 제공된 리스트에서 가장 적절한 섹션 코드를 식별하고 요약문을 생성하라는 구체적인 과업을 정의한다. Critical Constraint은 출력 구조를 일관성있게 하기 위해, 구조적 제약 조건을 명시한다. 섹션 분류와 요약문 사이에 구분자 '#'을 포함하고 부가적인 텍스트를 제외한 결과만 출력하도록 지시한다. Available codes에서는 분류 해야할 섹션 헤더의 종류를 명시한다. ICL 프롬프트는 Few-shot을 사용하기 때문에 Train 데이터셋에서 무작위로 선정된 3개를 예시로 제시한다. 마지막으로 Test Case는 임상 노트 생성을 위해 Test 데이터셋의 Dialogue를 제시한다.

Ⅳ. 실 험

1. Datasets

MTS-Dialog는 1700쌍의 의사-환자 대화 데이터셋이다[33].

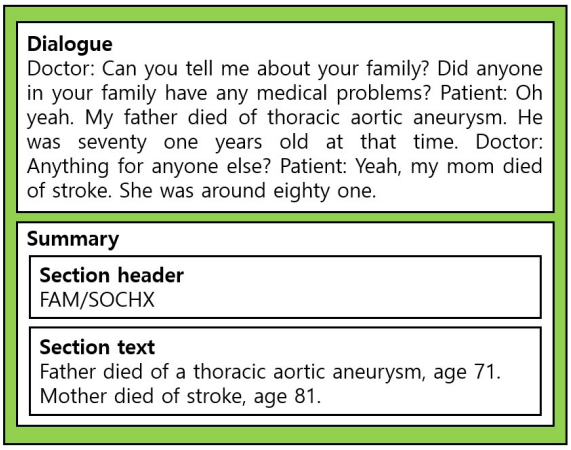

그림 5. MTS-Dialog 데이터 구조

그림 5는 본 데이터셋의 구조를 나타낸다. Dialogue는 환자의 가족력, 증상, 과거 병력 등에 대한 의사와 환자 간의 대화로 구성된다. 섹션 헤더는 대화의 주제를 사전에 정의된 임상 섹션 코드로 분류된다. 표 1은 총 20개의 섹션 헤더 종류를 나타낸다. 예시의 대화는 가족력 및 사회생활력에 해당하므로 FAM/SOCHX 섹션으로 분류되어 있다. 섹션 텍스트는 대화문 내용을 의료진이 판독하기 쉽게 요약된 텍스트이다.

표 1. Section Header 종류

Section Header	
Family History/Social History (fam/sochx)	History of Present Illness (genhx)
Past Medical History (pastmedicalhx)	Chief Complaint (cc)
Past Surgical History (pastsurgical)	allergy
Review of Systems (ros)	medications
assessment	exam
diagnosis	disposition
plan	Emergency Department Course (edcourse)
immunizations	imaging
Gynecologic History (gynhx)	procedures
other_history	labs

학습에 사용한 데이터 개수는 Train, Validation, Test 각각 1201, 100, 200개를 사용하였다.

2. Models

본 연구에서는 제안된 태스크를 수행하기 위해 오픈 소스 SLM과 폐쇄형 LLM을 모두 활용하였다. SLM은 LLaMA3.2 3B, Qwen3 4B, 그리고 Gemma3 4B를 선정하였으며, 효율적인 파라미터 미세조정을 위해 QLoRA 기법을 적용하여 학습을 진행하였다. 폐쇄형 LLM인 Gemini 2.5 Flash는 ICL 방식으로 출력을 생성하고 평가를 수행하였다. 표 2는 SLM 학습에 사용한 하이퍼파라미터를 나타낸다.

표 2. 모델 학습 하이퍼파라미터

Hyperparameter	Value
Max sequence Length	512
Batch Size	4
Learning Rate	5e-5
Optimizer	Paged AdamW(32-bit)
Epochs	10
Weight Decay	0.001
BF16	True
LoRA Rank	8
LoRA Alpha	32
Target Modules	q, k, v, o_proj
LoRA Dropout	0.05
Quantization Type	NF4
Double Quantization	Enabled
Gradient Accumulation Steps	4
Warmup Ratio	0.03
LR Scheduler	Cosine

실험에 사용한 하드웨어는 NVIDIA RTX 3060 12GB이다. 입력 데이터의 Max sequence Length는 512로 설정하였다. Batch Size는 4로 지정하여 하드웨어 부하를 최소화하였다. Learning Rate는 5e-5로 설정하여 안정적인 학습을 유도하였으며, Optimizer로는 메모리 효율이 높은 Paged AdamW(32-bit)를 사용하였다. 전체 Epochs는 10으로 설정하였다. 과적합 방지를 위한 Weight Decay 0.001과 연산 정밀도를 위한 BF16 True 설정을 적용하였다. QLoRA의 핵심 파라미터인 LoRA Rank는 8, LoRA Alpha는 32로 지정하여 학습 효율을 높였으며, Target Modules로 q_proj, k_proj, v_proj, o_proj를 선정하여 어텐션의 핵심 레이어에 학습을 집중시켰다. 또한 LoRA Dropout 0.05를 통해 일반화 성능을 강화하였고, Quantization Type으로 NF4를 선택하고 Double Quantization을 사용하여 저사양 환경에서도 대규모 모델의 미세 조정이 가능하도록 최적화하였다. 마지막으로 Gradient Accumulation Steps를 4로 설정하여 실질적으로 16의 배치 사이즈 효과를 내도록 하였으며, 0.03의 Warmup Ratio와 Cosine 방식의 LR Scheduler를 결합하여 안정적인 수렴을 달성하고 모델의 일반화 성능을 확보하였다.

3. Results

표 3. 모델 평가 결과

	미세조정 LLaMA 3.2 3B	미세조정 Qwen3 4B	미세조정 Gemma3 4B	Gemini 2.5 Flash
Classification				
Accuracy	0.725	**0.735**	0.680	0.640
Summarization				
ROUGE-1	0.396	0.401	**0.407**	0.381
ROUGE-2	0.170	0.172	0.159	**0.185**
ROUGE-L	0.340	**0.341**	0.339	0.317
BERT Score	**0.897**	0.896	**0.897**	0.892
BLEURT	0.537	0.536	0.534	**0.550**
Aggregate Score				
	0.610	0.611	**0.613**	0.608

표 3은 평가 지표를 통한 결과를 나타낸다. ROUGE-1과 ROUGE-2는 각각 단일 단어와 연속된 두 단어의 중복을 측정하는 n-gram 기반 지표이다. ROUGE-1은 내용의 포괄성을, ROUGE-2는 요약의 유창성을 평가하며 특히 단일 문서 요약에서 인간의 판단과 높은 상관관계를 나타낸다. ROUGE-L은 최장 공통 부분 수열을 활용하여 단어의 불연속적 일치를 허용하되 문장 내 순서를 반영하여 별도의 n-gram 길이 지정 없이도 문장 구조를 효과적으로 포착하고 짧은 요약문 평가에서 우수한 성능을 보인다[34]. BERTScore는 사전 훈련된 BERT의 Contextual Embeddings을 기반으로 후보 문장과 참조 문장 간의 토큰별 코사인 유사도를 계산하고, 이를 Greedy Matching을 통해 결합하여 의

미론적 등가성을 측정하는 자동 평가 지표이다[35]. BLEURT는 BERT 기반의 학습형 텍스트 생성 평가 지표로, 합성 데이터 사전 학습을 통해 낮은 상관관계와 데이터 편향 문제를 해결하였다[36]. 평가 지표의 종합적인 계산을 위해 ROUGE-1 F1, BERTScore F1, BLEURT-20의 평균을 통해 Aggregate Score를 계산한다[37].

실험 결과, 섹션 분류 정확도는 미세조정된 Qwen3 4B가 0.735로 가장 높은 정확도를 보였다. Gemini 2.5 Flash는 0.640으로 상대적으로 낮은 분류 성능을 보였는데, 이는 특정 도메인에 특화된 섹션 체계에 적응하기 위해서는 미세조정이 큰 영향을 주는 것을 나타낸다. ROUGE-1에서는 Gemma 3 4B가 0.407로 가장 우수하였으나, ROUGE-L에서는 Qwen3 4B가 0.341로 근소한 우위를 점하였다. ROUGE-2에서는 0.185, BLEURT에서는 0.550로 Gemini 2.5 Flash가 유의미한 차이로 높은 점수를 달성했다. BERTScore는 0.897로 LLaMA 3.2 3B와 Gemma3 4B가, Aggregate Score는 0.613로 Gemma3 4B가 근소한 차이로 우위를 점하였다. 실험 결과, SLM이 파라미터가 훨씬 큰 Gemini보다 높은 성능을 달성한 이유는 특정 도메인 데이터로의 미세조정이 ICL 방식보다 정교한 도메인 적응을 수행함을 시사한다. 특히, 의료 상담의 비정형적 특성과 복잡한 섹션 체계를 학습하는 데 있어, 방대한 파라미터를 가진 거대 모델의 추론 능력보다 도메인 특화 데이터셋에 대한 정밀한 학습이 성능 향상에 유의미한 영향을 준다는 것을 확인하였다.

V. 결 론

본 연구는 의료 데이터의 민감성과 자원 제한성을 동시에 해결하기 위해서 의료 대화 구조를 반영한 QLoRA 기반의 SLM 미세조정 방법론을 제안하고 LLM의 ICL 방식과 비교 분석하여 보안 환경에서의 실무적 활용 가능성을 확인하였다. 또한 파이프라인에 맞는 프롬프트 구조를 설계하였다. 실험 결과, 특정 도메인에 특화된 섹션 분류에서는 미세조정된 Qwen3 4B 모델이 0.735의 정확도를 기록하며 가장 우수한 성능을 보였으며, 이는 범용 모델인 Gemini 2.5 Flash보다 높은 수치로 나타났다. 요약 성능 측면에서도 Gemma3 4B가 Aggregate Score 0.613으로 가장 높은 종합 점수를 달성하며 SLM의 효율성을 입증하였다. Gemini 2.5 Flash는 ROUGE-2와 BLEURT 지표에서 상대적으로 높은 점수를 기록하여, 문장의 유창성과 논리적 추론 능력에서는 거대 모델 특유의 강점을 보여주었다. 결론적으로, 의료 상담 데이터와 같이 보안성이 최우선시되고 자원 접근성이 제한된 환경에서는 QLoRA를 활용한 SLM의 미세조정이 섹션 분류와 구조화된 요약에서 실용적인 대안이 될 수 있음을 확인하였다. 본 연구의 결과는 보안 환경에 특화된 저비용, 고효율 임상 문서화 시스템 구축을 위해 활용될 수 있을 것이다. 향후 연구에서는 실제 의료 현장의 전문의 피드백을 수용하여 요약문의 임상적 타당성을 확인하고, 미세조정 알고리즘을 더욱 고도화하여 재현 가능성과 신뢰성을 확보할 예정이다.

REFERENCES

[1] M. W. Friedberg, P. G. Chen, K. R. Van Busum, F. Aunon, C. Pham, J. Caloyeras, and M. Tutty, “Factors affecting physician professional satisfaction and their implications for patient care, health systems, and health policy,” *Rand health quarterly*, vol. 3, no. 4, pp. 1, 2014.

[2] S. Babbott, L. B. Manwell, R. Brown, E. Montague, E. Williams, M. Schwartz, and M. Linzer, “Electronic medical records and physician stress in primary care: results from the MEMO Study,” *Journal of the American Medical Informatics Association*, vol. 21, no. e1, pp. e100-e106, 2014.

[3] B. G. Arndt, J. W. Beasley, M. D. Watkinson, J. L. Temte, W. J. Tuan, C. A. Sinsky, and V. J. Gilchrist, “Tethered to the EHR: primary care physician workload assessment using EHR event log data and time-motion observations,” *The Annals of Family Medicine*, vol. 15, no. 5, pp. 419-426, 2017.

[4] S. Garfan, A. H. Alamoodi, B. B. Zaidan, M. Al-Zobbi, R. A. Hamid, J. K. Alwan, and F. Momani, “Telehealth utilization during the Covid-19 pandemic: a systematic review,” *Computers in biology and medicine*, vol. 138,

pp. 104878, 2021.

[5] S. Corby, J. A. Gold, V. Mohan, N. Solberg, J. Becton, R. Bergstrom, and J. S. Ash, “A sociotechnical multiple perspectives approach to the use of medical scribes: a deeper dive into the scribe-provider interaction,” *AMIA Annual Symposium Proceedings*, pp. 333, Washington D.C., USA, Mar. 2020.

[6] N. Naik, B. M. Hameed, D. K. Shetty, D. Swain, M. Shah, R. Paul, and B. K. Somani, “Legal and ethical consideration in artificial intelligence in healthcare: who takes responsibility?,” *Frontiers in surgery*, vol. 9, pp. 862322, 2022.

[7] Y. Dai, H. Li, C. Tang, Y. Li, J. Sun, and X. Zhu, “Learning low-resource end-to-end goal-oriented dialog for fast and reliable system deployment,” *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 609-618, Online (Seattle, Washington), USA, Jul. 2020.

[8] A. Goldberger, L. Amaral, L. Glass, J. Hausdorff, P. C. Ivanov, R. Mark, and H. E. Stanley, “PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals,” *Circulation*, vol. 101, no. 23, pp. e215-e220, 2000.

[9] G. Michalopoulos, K. Williams, G. Singh, and T. Lin, “MedicalSum: A guided clinical abstractive summarization model for generating medical reports from patient-doctor conversations,” Findings of the Association for Computational Linguistics: *EMNLP 2022*, pp. 4741-4749, Abu Dhabi, United Arab Emirates, Dec. 2022.

[10] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, and D. Amodei, “Language models are few-shot learners,” *Advances in Neural Information Processing Systems*, pp. 1877-1901, Online (Vancouver, Canada), Dec. 2020.

[11] E. Perez, D. Kiela, and K. Cho, “True few-shot learning with language models,” *Advances in Neural Information Processing Systems*, pp. 11054-11070, Online, Dec. 2021.

[12] 김가현, 김도국, “소형 언어 모델의 특정 도메인에서의 파인튜닝과 RAG 의 성능 비교-질의응답과 감정분석을 중심으로,” *스마트미디어저널*, 제14권, 제6호, 50-59쪽, 2025년 6월

[13] 김승주, 정세훈, 심춘보, “한국어 문서 요약 효율화를 위한 Gemma2 모델 미세조정 및 프롬프트 튜닝,” *스마트미디어저널*, 제14권, 제8호, 81-90쪽, 2025년 8월

[14] 정도윤, 김남호, “sLLM 을 이용한 맞춤형 다층 설문지 형식의 진로적성검사 시스템 개발 연구,” *스마트미디어저널*, 제14권, 제8호, 42-49쪽, 2025년 8월

[15] J. Giorgi, A. Toma, R. Xie, S. Chen, K. An, G. Zheng, and B. Wang, “WangLab at MEDIQA-Chat 2023: Clinical Note Generation from Doctor-Patient Conversations using Large Language Models,” *Proc. of the 5th Clinical Natural Language Processing Workshop*, pp. 323-334, Toronto, Canada, Jul. 2023.

[16] Y. Mathur, S. Rangreji, R. Kapoor, M. Palavalli, A. Bertsch, and M.R. Gormley, “SummQA at MEDIQA-Chat 2023: In-Context Learning with GPT-4 for Medical Summarization,” *arXiv preprint arXiv:2306.17384*, pp. 1-13, Jun. 2023.

[17] X. Tang, A. Tran, J. Tan, and M. Gerstein, “GersteinLab at MEDIQA-Chat 2023: Clinical Note Summarization from Doctor-Patient Conversations through Fine-tuning and In-context Learning,” *arXiv preprint arXiv:2305.05001*, pp. 1-11, May 2023.

[18] A. Sharma, D. Feldman, and A. Jain, “Team Cadence at MEDIQA-Chat 2023: Generating, Augmenting and Summarizing Clinical Dialogue with Large Language Models,” *Proc. of the 5th Clinical Natural Language Processing Workshop*, pp. 228-235, Toronto, Canada, Jul. 2023.

[19] P. Mishra and R.T. Desetty, “NewAgeHealthWarriors at MEDIQA-Chat 2023 Task A: Summarizing Short Medical Conversation with Transformers,” *Proc. of the 5th Clinical Natural Language Processing Workshop*, pp. 414-421, Toronto, Canada, Jul. 2023.

[20] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, and R. Ganapathy, “The llama 3 herd of models,” *arXiv e-prints*, vol. abs/2407.21783, pp. 1-92, Jul. 2024.

[21] J. Ainslie, J. Lee-Thorp, M. De Jong, Y. Zemlyanskiy, F. Lebrón, and S. Sanghai, “Gqa: Training generalized multi-query transformer models from multi-head checkpoints,” a*rXiv preprint*, vol. abs/2305.13245, pp. 1-6, May 2023.

[22] W. Xiong, J. Liu, I. Molybog, H. Zhang, P. Bhargava, R. Hou, and H. Ma, “Effective long-context scaling of foundation models,” *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4643-4663, Mexico City, Mexico, Jun. 2024.

[23] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn, “Direct

preference optimization: Your language model is secretly a reward model," *Advances in Neural Information Processing Systems*, pp. 53728-53741, New Orleans, USA, Dec. 2023.
[24] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, and Z. Qiu, "Qwen3 technical report," *arXiv preprint*, vol. abs/2505.09388, pp. 1-85, May 2025.
[25] G. Team, A. Kamath, J. Ferret, S. Pathak, N. Vieillard, R. Merhej, and S. Iqbal, "Gemma 3 technical report," *arXiv preprint*, vol. abs/2503.19786, pp. 1-52, Mar. 2025.
[26] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," *arXiv preprint*, vol. abs/2004.05150, pp. 1-16, Apr. 2020.
[27] M. T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," *arXiv preprint*, vol. abs/1508.04025, pp. 1-11, Aug. 2015.
[28] G. Team, P. Georgiev, V. I. Lei, R. Burnell, L. Bai, A. Gulati, and B. O. Batsaikhan, "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context," *arXiv preprint*, vol. abs/2403.05530, pp. 1-104, Mar. 2024.
[29] G. Comanici, E. Bieber, M. Schaekermann, I. Pasupat, N. Sachdeva, I. Dhillon, and M. Velic, "Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities," *arXiv preprint*, vol. abs/2507.06261, pp. 1-118, Jul. 2025.
[30] N. Du, Y. Huang, A. M. Dai, S. Tong, D. Lepikhin, Y. Xu, and C. Cui, "Glam: Efficient scaling of language models with mixture-of-experts," *International Conference on Machine Learning (ICML)*, pp. 5547-5569, Baltimore, Maryland, USA, Jun. 2022.
[31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, pp. 5998-6008, Long Beach, California, USA, Dec. 2017.
[32] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "Qlora: Efficient finetuning of quantized llms," *Advances in Neural Information Processing Systems*, pp. 10088-10115, New Orleans, USA, Dec. 2023.
[33] A. B. Abacha, W. W. Yim, Y. Fan, and T. Lin, "An empirical study of clinical note generation from doctor-patient encounters," *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 2291-2302, Dubrovnik, Croatia, May 2023.
[34] C. Y. Lin, "Rouge: A package for automatic evaluation of summaries," *Text Summarization Branches Out*, pp. 74-81, Barcelona, Spain, Jul. 2004.
[35] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," *arXiv preprint*, vol. abs/1904.09675, pp. 1-20, Apr. 2019.
[36] T. Sellam, D. Das, and A. Parikh, "BLEURT: Learning robust metrics for text generation," *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7881-7892, Online, Jul. 2020.
[37] A. Pu, H. W. Chung, A. Parikh, S. Gehrmann, and T. Sellam, "Learning compact metrics for MT," *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 751-762, Online and Punta Cana, Dominican Republic, Nov. 2021.

저 자 소 개

유영준(준회원)
2025년 건양대학교 인공지능학과 졸업
2025년~현재 : 건양대학교 AI소프트웨어융합학과 석사과정

<주관심분야 : 딥러닝, LLM, 생성형AI>

김웅식(정회원)
1989년 2월 : 인하대학교 정보공학과 (공학석사)
2007년 2월 : 인하대학교 컴퓨터공학과 (공학박사)
2006년 3월~현재 : 건양대학교 인공지능학과 교수

<주관심분야 : 딥러닝, 머신러닝, 의료공학>