인공지능 학습 결과서

2025년 **11**월 **18**일

1. 개요

1.1 배경 및 필요성

- 회의 전문 분석 자동화의 중요성**:** 구두로 진행된 회의에서 핵심 의사결정 사항 및 후속 태스크를 자동 추출하여 생산성을 극대화.
- 기존 범용 **LLM**의 한계**:** 복잡한 회의록 구조, 구어체 비문, 암묵적인 결정 사항 및 태스크 간의 종속성을 추론하는 능력이 부족하여 결과의 신뢰도 저하.

1.2 프로젝트 목표

- 고난도 **Few-Shot** 학습 데이터**(**합성 데이터**)** 생성 및 확보
  - GPT-4 Turbo의 고급 추론 능력을 활용하여 기존에 확보하기 어려웠던 고난도 시나리오 데이터셋(약 2,700건)을 구축.
- **SLLM** 개발 및 목표
  - 경량화 모델을 합성 데이터로 파인튜닝하여, 회의 전문 분석의 정확도 및 일관성을 GPT-4급으로 개선.

2. 데이터 생성 및 전처리

2.1 합성 데이터 생성

- Few-Shot 프롬프트 설계: 고난도 조건(태스크 종속성, 재정의, null 처리 등) 명시.
- 데이터 다양성 확보: DOMAINS / ROLES 확장을 통한 전문 용어 및 상황 다양화.
    - DOMAINS 예시: "IT 개발팀", "제품 기획팀", "마케팅 전략팀", "법무/컴플라이언스팀" 등.
    - ROLES 예시: "PM", "개발자", "디자이너", "QA 엔지니어" 등.
- 사용 모델: GPT-4 Turbo를 활용한 이유 및 비용 효율성 검토.
- 데이터 양: 약 2,700건

2.2 합성 데이터 전처리

① 화자/구어체 정규화 : 실제 STT 전문 적용 시 발생할 수 있는 간투사 제거, 구어체 표준화.

② Chat Completion 형식 : LLM에게 대화 전달하고 JSON 응답을 받기 위한 표준 구조로 변환.

(*[{"role": "user", "content": "..."}]*)

③ 템플릿 JSON 구조 표준화 : 모델이 출력해야 할 회의록 템플릿 구조를 고정된 JSON Schema로 정의해

모델 출력 오류 최소화.

3. 모델 선정 및 구조

3.1 모델 비교 및 초기 선정 배경

본 프로젝트에서는 sLLM을 활용한 회의록 분석 파이프라인 구축을 목표로 하였다. 이에 따라 모델 후보군은 Weights & Biases에서 제공하는 한국어 최신 LLM 모델 리더보드를 기준으로, 15B 미만 모델 중 상위 성능 모델 5종을 1차 후보로 선정하였다.

1차 후보 모델 중 Qwen 2.5-14B-Instruct 모델이 종합 성능 1위를 기록하였으며, 한국어 특화 모델 부문에서는 A.X-4.0-Light 모델이 우수한 성능을 보였다. 그러나 Qwen 2.5-14B-Instruct 모델은 14B 규모로 본 프로젝트의 sLLM 지향 범위를 초과한다고 판단되어, 동일 계열의 경량 모델인 Qwen2.5-1.5B, 3B, 7B 계열로 범위를 재조정하였다.

3.2 경량 모델 성능 검증 및 1차 선정

경량 모델에 대한 성능 검증은 AI-Hub 국회 회의록 데이터셋을 활용하여 요약 성능을 비교 평가하는 방식으로 수행되었으며, 정량적 평가지표로는 BertScore(Precision, Recall, F1) 를 사용하였다.

그 결과, Qwen2.5-1.5B-Instruct 모델이 가장 높은 F1 Score(0.7653)를 기록하여 초기 기준 모델로 선정되었다. 이 단계에서는 모델 경량성 및 추론 효율성을 최우선 고려 요소로 두었다.

모델명 / 평가지표	**Precision**(정밀도)	**Recall**(재현율)	**F1 Score**(조화평균)
Qwen2.5-1.5B-Instruct	0.7560	0.7750	0.7653
Qwen2.5-7B-Instruct-1M	0.7481	0.7740	0.7607
skt/A.X-4.0-Light	0.7385	0.7623	0.7497
Qwen2.5-3B-Instruct	0.7323	0.7459	0.7387

3.3 파이프라인 적용 과정에서의 한계 및 추가 검토 필요성

그러나 실제 회의록 분석 파이프라인에 모델을 적용하는 과정에서, Qwen2.5-1.5B-Instruct 모델은 프롬프트 길이가 증가할수록 출력 안정성이 저하되는 현상이 관찰되었다. 특히 출력 형식 유지 실패, 할루시네이션 발생 빈도 증가, 그리고 instruction following 성능 저하 등의 한계가 확인되었다.

실제 서비스 시나리오에서 요구되는 긴 입력 문맥과 구조화된 출력 요구사항을 안정적으로 충족하기에는 1.5B 규모 모델이 구조적으로 제약이 있음을 의미한다.

3.4 중·대형 경량 모델 확장 실험 및 최종 비교

이에 따라 모델 규모 증가에 따른 성능 변화를 검증하기 위해, 동일 계열의 Qwen2.5-7B-Instruct 모델을 추가로 파인튜닝하여 실험을 확장하였다.

7B 모델은 1.5B 대비 문맥 유지력 및 출력 안정성 측면에서 개선된 결과를 보였으나, 추론 시간 증가의 단점이 확인되었다.

이러한 상황에서, 비교적 최근에 공개된 Qwen3-8B 모델은 최신 아키텍처 기반으로 instruction following 및 장문 문맥 처리 능력 개선이 기대되는 모델로 판단되어 추가 비교 대상 모델로 선정하였다.

3.5 비교 실험 설계 및 평가 기준

본 프로젝트에서는 모델 규모에 따른 성능 차이를 정량적으로 분석하기 위해, Qwen2.5-1.5B, Qwen2.5-7B, Qwen3-8B 모델을 각각 파인튜닝한 후 비교 실험을 수행하였다.

비교 실험은 실제 회의록 분석 파이프라인 적용을 가정하여 설계되었으며, 요약 및 태스크 추출 성능뿐만 아니라 프롬프트 준수도, 출력 구조 안정성, 처리 시간을 함께 고려하였다.

이를 위해 평가 기준은 다음과 같이 구성하였다.

- 정량 평가 : GPTScore 기반 지표
    - Faithfulness
    - Instruction_following
    - Structure_clarity
- 정성 평가 : LLM-as-a-Judge 기반 평가
    - 항목별 1~5점 척도
- 효율성 평가 : 전사본 입력부터 결과 생성까지의 전체 처리 시간

4. 학습 환경

4.1 하드웨어 환경

- 테스트 장비: RunPod NVIDIA A40 기반 GPU 서버
- 프로세서: AMD EPYC 기반 32 vCPU
- 그래픽: NVIDIA A40 48GB
- RAM: 128GB
- 저장 장치: NVMe SSD

4.2 소프트웨어 환경

- 운영 체제: Ubuntu 24.04 LTS
- 개발 언어: Python 3.11
- 딥러닝 프레임워크 : PyTorch, Hugging Face Transformers
- 파인튜닝 프레임워크 : PEFT (LoRA, QLoRA)
- 모델 평가 도구:
    - GPTScore 기반 자동 평가 (Qwen3-14B 평가 모델 사용)
    - LLM-as-a-Judge 기반 정성 평가 (GPT 4o-mini 평가 모델 사용)

4.3 네트워크 환경

- 테스트 서버: 내부 네트워크 환경에서 분리된 전용 네트워크

5. 하이퍼파라미터

본 프로젝트에서는 모델 규모 차이에 따른 성능 비교의 공정성을 확보하기 위해, 파인튜닝 시 공통 하이퍼파라미터를 기본으로 설정하고 모델 규모에 따라 불가피한 항목만 부분 조정하는 방식을 채택하였다.

5.1 공통 하이퍼파라미터 설정

세 모델(Qwen2.5-1.5B, Qwen2.5-7B, Qwen3-8B)에 공통적으로 적용한 주요 설정은 다음과 같다.

항목	값
학습 방식	PEFT 기반 LoRA (QLoRA)
per_device_train_barch_size	4
gradient_accumulation_steps	4
optimizer	AdamW
lr_scheduler_type	cosine
warmup_ratio	0.05
num_train_epochs	4
bf16	True
fp16	False
gradient checkpointing	적용
평가 기준	validation loss

5.2 모델별 조정 하이퍼파라미터

모델 규모 증가에 따른 메모리 사용량 및 문맥 처리 특성을 고려하여, 일부 하이퍼파라미터는 모델별로 다음과 같이 조정하였다.

항목	**1.5B / 7B**	**8B**
learning_rate	1e-4	5e-5
max sequence length	512	1024
answer max length	256	384

특히 Qwen3-8B 모델의 경우, 회의 전문 및 지시 프롬프트를 보다 충분히 반영하기 위해 입력 토큰 길이를 확장하였다.

6. 학습 과정 및 성능 평가

6.1 학습 과정 (PEFT-LoRA 기반 파인튜닝)

본 프로젝트의 모든 모델 학습은 PEFT-LoRA 방식을 기반으로 수행되었다. Base 모델의 파라미터는 동결하고, LoRA 어댑터 모듈만을 학습하는 방식으로 진행하여 효율성을 극대화 하였다.

6.1.1 학습 데이터 및 입력 구성

학습 데이터는 회의 전문 시나리오를 반영한 chat-completion 형식의 구조로 구성되었다.

- system/user 메시지를 프롬프트로 사용
- assistant 메시지를 정답으로 설정
- 출력은 안건 및 태스크를 JSON 구조로 생성하도록 유도

Qwen3-8B 모델의 경우, 회의 전체 맥락을 보다 충분히 반영하기 위해 최대 입력 길이를 1024 토큰으로 확장하였다.

6.1.2 학습 안정성 및 수렴 특성

- 초기 단계에서 loss가 빠르게 감소
- 이후 epoch 진행에 따라 완만하게 수렴
- 전형적인 LoRA 파인튜닝 수렴 패턴 확인

또한 gradient checkpointing을 적용함으로써, 모든 모델에서 OOM 없이 전체 학습 epoch를 안정적으로 완료하였다.

6.2 평가 지표 정의 및 평가 방법

6.2.1 평가 지표 정의

회의록 분석 결과의 품질을 다각도로 평가하기 위해 다음 세 가지 지표를 공통 평가 기준으로 사용하였다.

Faithfulness	- 생성된 결과가 입력 전사본의 내용에 근거하여 작성되었는지 - 입력에 존재하지 않는 정보(할루시네이션)의 여부 중점 측정
Instruction Following	- 모델이 주어진 지시사항(요약, 주요 안건, 태스크 추출 등)을 이해하고 요구된 형식과 내용을 따랐는지
Structure Clarity	- 출력 결과가 JSON 포맷을 유지하며, 항목 간 구분이 명확하고 일관성 있게 구성되었는지

해당 지표들은 단순한 문장 유사도 기반 평가로는 측정하기 어려운 실제 서비스 관점의 출력 품질을 반영하기 위해 선정되었다

6.2.2 GPTScore 기반 정량 평가

정량 평가는 GPTScore를 활용하여 수행하였다. GPTScore는 입력 회의 전사문과 생성 결과를 함께 고려하여, 각 평가 지표별로 출력 결과의 상대적 적합도를 로그 확률 기반 점수로 산출한다.

본 프로젝트에서는 Qwen3-14B 모델을 GPTScore 평가 모델로 사용하였다. 선정 이유는 다음과 같다.

- 회의 전사문과 파이프라인 결과를 동시에 입력 받아 평가하는 구조상, 충분한 문맥 처리 용량이 필요
- 평가 대상보다 상위 용량 모델을 사용하여, 작은 모델의 한계가 평가 자체에 반영되는 위험을 완화
- 다수 샘플에 대해 반복 실행하므로, 로컬/서버 환경에서 실행 가능한 모델이 필요

GPTScore는 모델 간 상대 비교 지표로 활용되었으며, 점수의 절대값보다는 모델 간 점수 차이 및 경향성 분석에 초점을 두었다.

6.2.3 LLM-as-a-Judge 기반 정성 평가

정성 평가는 LLM-as-a-Judge 방식을 적용하여 수행하였다. LLM-as-a-Judge는 평가 모델이 사람 평가자의 역할을 수행하여, 각 출력 결과를 기준 지표에 따라 1~5점 척도로 판단하는 방식이다.

본 프로젝트에서는 GPT-4o-mini 모델을 평가자로 사용하였다. 선정 이유는 다음과 같다.

- 반복 평가 수행 시 응답 시간 및 비용 측면에서 효율적
- 지시문 및 평가 기준 해석 능력 : faithfulness / instruction_following / structure_clarity와 같은 기준을 텍스트로 설명하고 점수화해야 함

LLM-as-a-Judge는 GPTScore로는 포착하기 어려운 출력의 자연스러움, 지시사항 충실도, 구조적 안정성을 보완적으로 평가하는 역할을 수행하였다.

6.2.4 평가 방식 및 모델 구성 요약

본 프로젝트에서는 두 평가 방식을 병행함으로써 다음을 달성하고자 하였다.

- GPTScore를 통한 정량적·모델 독립적 비교
- LLM-as-a-Judge를 통한 사람 평가에 근접한 정성 판단
- 단일 지표 편향을 방지하고, 실제 파이프라인 적용 적합성을 종합적으로 분석

또한 평가 대상 모델보다 상위 규모의 모델을 평가 모델로 사용함으로써, 자기 평가에 따른 편향을 최소화하고자 하였다.

- 생성 모델 : 2.5-1.5B / 2.5-7B / 3-8B
- 정량 평가 모델(GPTScore) : 3-14B
- 정성 평가 모델(LLM-as-a-Judge) : 상용 대형 모델 (GPT-4o-mini)

이를 통해 평가의 객관성 및 신뢰성을 확보하고자 하였다.

6.3 모델별 성능 비교 결과

6.3.1 평가 데이터 구성

모델의 회의 전문 분석 성능을 보다 현실적으로 평가하기 위해, 성격이 상이한 두 유형의 전사본 데이터를 평가 데이터로 구성하였다.

- 합성 회의 전사본
    - 자체 합성한 시나리오를 팀원들이 직접 녹음하여 생성한 전사문
    - 비교적 구조화되고 정제된 발화 내용
    - 모델이 의도된 사용 시나리오 하에서 안정적으로 동작하는지 확인 위함
- 스크럼 전사본
    - 외부 팀에서 진행한 정기 스크럼 회의 녹음 전사본
    - 발화 중 중복, 비문 등 비교적 비정형적이고 자연 발화에 가까운 데이터
    - 실제 현업 환경에서 발생할 수 있는 입력 조건을 가정하여 평가하기 위함

이와 같이 두 가지 성격의 전사본을 함께 사용함으로써, 본 연구는 모델이 정제된 입력과 실제 환경 입력 모두에서 어느 정도의 성능을 유지하는지를 비교 분석하고자 하였다.

6.3.2 정량 평가 결과 **(GPTScore)**

GPTScore는 faithfulness, instruction_following, structure_clarity의 3개 항목에 대해 산출되었으며, 본 평가에서는 값이 0에 가까울수록(덜 음수일수록) 상대적으로 양호한 결과로 해석하였다.

- Qwen3-8B : 두 전사본(합성 전문 / 실제 스크럼) 모두에서 가장 높은 (0에 가장 가까운) 점수대를 기록
- Qwen2.5-7B : 합성 전문에서 상대적으로 양호하나, 실제 스크럼 전사본에서는 점수가 더 하락
- Qwen2.5-1.5B : 두 전사본 모두에서 점수 분포가 유사하며, 전반적으로 7B/8B 대비 낮은 수준

모델	데이터 유형	**Faithfulness**	**Instruction Following**	**Structure Clarity**
Qwen2.5-1.5B	합성 회의 전사본	-0.69	-0.71	-0.72
Qwen2.5-1.5B	스크럼 전사본	-0.70	-0.71	-0.74

Qwen2.5-7B	합성 회의 전사본	-0.48	-0.49	-0.52
Qwen2.5-7B	스크럼 전사본	-0.62	-0.64	-0.68
Qwen3-8B	합성 회의 전사본	-0.41	-0.42	-0.44
Qwen3-8B	스크럼 전사본	-0.43	-0.44	-0.45

즉, GPTScore 기준으로는 Qwen3-8B > Qwen2.5-7B > Qwen2.5-1.5B 순으로 평가 기준에 더 부합하는 출력을 생성하는 것을 확인하였다.

6.3.3 정성 평가 결과 **(LLM-as-a-Judge)**

LLM-as-a-Judge 평가는 GPTScore와 동일한 세 지표를 기준으로 1~5점 척도로 수행하였다. 본 평가에서는 점수뿐 아니라, 평가 모델이 제시한 근거를 함께 검토하여 실제 오류 양상(할루시네이션, 지시 미준수, 구조 문제 등)을 확인하였다.

- Qwen3-8B : 두 전사본 모두에서 높은 점수를 기록했으며, 특히 할루시네이션과 구조 유지 측면에서 일관성이 높게 평가됨
- Qwen2.5-7B : 합성 전문에서는 안정적이나, 실제 스크럼에서는 일부 항목에서 기한 생성(due/due_date) 관련 할루시네이션이 지적됨
- Qwen2.5-1.5B : structure_clarity는 높게 평가되었으나, 평가 근거에서 반복적으로 기한 생성에 대한 할루시네이션이 지적되며 faithfulness 점수가 낮게 평가됨

모델	데이터 유형	**Faithfulness**	**Instruction Following**	**Structure Clarity**
Qwen2.5-1.5B	합성 회의 전사본	1/5	3/5	5/5
Qwen2.5-1.5B	스크럼 전사본	1/5	1/5	5/5
Qwen2.5-7B	합성 회의 전사본	5/5	4/5	5/5
Qwen2.5-7B	스크럼 전사본	3/5	4/5	5/5
Qwen3-8B	합성 회의 전사본	5/5	5/5	5/5
Qwen3-8B	스크럼 전사본	5/5	4/5	5/5

여러 결과에서 공통적으로, structure_clarity가 높게 나온 것에 비해 faithfulness가 낮게 나오는 사례가 존재했다. 따라서 본 평가에서는 세 지표를 단일 값으로 합산하지 않고, 지표별로 구분하여 해석하였다.

6.4 처리 시간 및 효율성 분석

각 모델의 효율성을 비교하기 위해 전사본 입력부터 최종 결과 생성까지의 전체 처리 시간을 측정하였다.

- Qwen2.5-1.5B : 평균 처리 시간이 가장 짧으나 출력 품질 및 안정성 측면에서 한계 존재
- Qwen2.5-7B : 처리 시간이 크게 증가. 전사본 길이에 따라 최대 지연 시간 발생
- Qwen3-8B : 7B 대비 처리 시간 증가 폭이 제한적이며 출력 품질 대비 처리 시간 효율이 상대적으로 우수

모델	데이터 유형	입력 길이**(char)**	총 추론 시간 **(sec)**

Qwen2.5-1.5B	합성 회의 전사본	36.831	1947
Qwen2.5-1.5B	스크럼 전사본	36.069	1847
Qwen2.5-7B	합성 회의 전사본	105.559	1947
Qwen2.5-7B	스크럼 전사본	125.545	1847
Qwen3-8B	합성 회의 전사본	79.248	1947
Qwen3-8B	스크럼 전사본	166.943	1847

이를 통해 Qwen3-8B 모델이 성능 대비 효율성 측면에서 가장 균형 잡힌 특성을 보임을 확인하였다.

7. 결론

7.1 결과 요약

본 프로젝트는 회의록 분석 파이프라인에 적용 가능한 sLLM을 구축하기 위해 모델 규모가 서로 다른 세 가지 LLM(Qwen2.5-1.5B, Qwen2.5-7B, Qwen3-8B)을 대상으로 파인튜닝 및 성능 비교를 수행하였다. 모든 모델은 동일한 파이프라인 구조와 평가 기준을 적용하였으며, 정제된 시나리오 기반 전사본과 실제 회의 전사본을 함께 사용하여 실제 사용 환경을 고려한 성능 평가를 진행하였다. 정량 평가(GPTScore)와 정성 평가(LLM-as-a-Judge)를 종합한 결과, 모델 규모가 증가할수록 출력의 사실성, 지시문 준수도, 구조 안정성이 전반적으로 개선되는 경향이 확인되었다.

또한 베이스 모델과 파인튜닝 모델 간의 직접적인 정량 비교보다는, 파이프라인 적용 가능 여부와 출력 안정성에 초점을 두고 평가를 진행하였다. 베이스 모델의 경우, 실제 실행 과정에서 JSON 파싱 실패, 안건/태스크 미생성, 발화 내용의 단순 나열 등으로 인해 평가 지표 산출 자체가 어려운 사례가 다수 발생하였다. 반면, 파인튜닝 모델은 사전 정의된 출력 구조를 안정적으로 유지하며 회의록 요약, 안건 정리, 후속 태스크 생성이라는 목적 기능을 일관되게 수행하였다.

이를 통해 본 평가에서는 파인튜닝이 단순한 성능 향상 뿐 아니라, 실제 파이프라인 적용 가능성을 확보하는 데 주요한 역할을 했음을 확인하였다.

7.2 최종 모델 선정 및 근거

Qwen3-8B 파인튜닝 모델을 본 프로젝트의 최종 적용 모델로 선정하였다.

- 정량 평가(GPTScore)에서 세 지표 모두 두 유형의 전사본에서 가장 안정적인 점수 분포를 기록함
- 정성 평가(LLM-as-a-Judge)에서도 할루시네이션 발생이 상대적으로 적고, JSON 구조 유지 및 지시문 준수 측면에서 일관된 결과를 보임
- 실제 스크럼과 같이 비정형적이고 노이즈가 많은 입력에서도 출력 품질 저하 폭이 상대적으로 작아 우수한 성능을 확인함
- 7B 모델 대비 처리 시간 증가 폭이 제한적이어서, 성능 대비 효율성 측면에서 균형 잡힌 선택지로 판단됨

7.3 한계점

1. 실제 기업 내부 회의 내용을 수집하지 못하여 평가 데이터의 규모가 제한적인 한계가 존재한다. 따라서, 실제 기업 회의 내용에 대해 테스트 진행이 필요하다.
2. LLM-as-a-Judge 평가의 경우, 평가 모델의 판단 특성이 결과에 영향을 미칠 수 있다는 한계가 존재한다. 따라서, Qwen2.5-7B, Qwen3-8B, Qwen3-14B 등 여러 모델에 대해 테스트를 진행하였고, 유사한 결과가 나오는 것을 확인하였다. 평가 모델의 판단 특성에 영향을 받는 한계점을 극복하지 못했지만, 다양한 평가 모델에 대해 결과를 확인함으로써 대상 모델의 성능을 다방면으로 확인하여 LLM-as-a-Judge 평가의 유효성을 검증하였다.