

**SK네트웍스 Family AI 과정 10기  
 데이터 전처리 인공지능 학습 결과서**



| **산출물 단계** | 데이터 전처리 |
| --- | --- |
| **평가 산출물** | 인공지능 학습 결과서 |
| **제출 일자** | 2025.08.29 |
| **깃허브 경로** | https://github.com/SKNETWORKS-FAMILY-AICAMP/SKN14-FINAL-6Team |
| **작성 팀원** | 송유나 |

**1. 모델 비교 및 선정 이유**

* 비교 대상 모델:

| **모델명** | **종류** | **선정 이유** |
| --- | --- | --- |
| beomi/Llama-3-Open-Ko-8B | Decoder-only LLM + LoRA 미세조정 | 한국어 지원이 강화된 Llama-3 기반 모델, 사전학습된 코퍼스가 풍부하여 범용적인 한국어 질의응답에 강점 |
| LGAI-EXAONE/EXAONE-3.5-7.8B-Instruct | Decoder-only LLM + LoRA 미세조정 | 중국어·영어 중심이지만 멀티링구얼에 강점, 소규모 환경에서 빠른 학습 가능 |
| Qwen/Qwen2.5-7B-Instruct | Decoder-only LLM + Instruct 튜닝 | 한국어 대화 최적화 모델, 상담 도메인 적합성 높고, HuggingFace trust\_remote\_code 지원으로 커스텀 토크나이저·모델 구조 활용 가능 |

* 실험 모델 수: 총 3종
* 최종 선정 모델: **LGAI-EXAONE/EXAONE-3.5-7.8B-Instruct**

**2. 모델 구조 및 아키텍처**

2.1 모델 아키텍처 도식 (선택사항: 도식 첨부 또는 말로 설명)

* **EXAONE 3.5-7.8B Instruct** 기반, Decoder-only Transformer
* 입력층 → 임베딩층 → 디코더 블록(32개) → 출력층(Logits)
* 챗봇 응답 생성을 위한 **Causal LM 구조**

2.2 구성 요소 설명:

| **계층명** | **역할** | **구성 요소** |
| --- | --- | --- |
| Embedding | 텍스트 입력 벡터화 | Pretrained tokenizer (AutoTokenizer, trust\_remote\_code) |
| Decoder Block | 문맥/의미 표현 학습 | Multi-Head Attention + FFN |
| LoRA Adapter | 효율적 미세조정 | r=8, α=32, dropout=0.1,  대상 모듈: q\_proj, k\_proj, v\_proj, o\_proj |
| Output Layer | 다음 토큰 확률 예측 | LM Head (Softmax) |

**3. 학습 설정 및 하이퍼파라미터**

| **항목** | **값** |
| --- | --- |
| 학습 데이터 수 | 9,842건 (전처리 후) |
| 검증 데이터 수 | 2000건 |
| 에폭(Epoch) 수 | 3 |
| 배치 크기 (Batch Size) | 2 (per device) |
| Gradient Accumulation | 4 step → Effective batch size ≈ 8 |
| 학습률 (Learning Rate) | 2e-5 |
| 옵티마이저 | paged\_adamw\_8bit (AdamW 변형) |
| 손실 함수 | CrossEntropyLoss |
| 정규화/안정화 | Gradient Clipping (0.3),  Warmup Ratio=0.03 |
| 양자화 | 8-bit (BitsAndBytesConfig) |
| 장치 설정 | device\_map="auto", offload\_folder 사용 |

**4. 학습 결과 및 성능 평가**

4.1. 중간 학습 결과 요약

현재 학습 진행시점에서의 로그:

| **지표** | **값** |
| --- | --- |
| Loss | 1.2013 |
| Gradient Norm | 1.1468 |
| Learning Rate | 2e-4 |
| 학습 처리 토큰 수 | 1,692,691 |
| Mean Token Accuracy | 73.6% |

* 초기 학습 단계임에도 불구하고 토큰 단위 정확도가 **73% 이상** 확보됨.
* 손실(Loss)도 1.2 수준으로 빠르게 수렴 중.
* EXAONE 사전학습 모델 위에서 **LoRA 미세조정이 효과적으로 진행**되고 있음을 보여줌.

4.2 해석 및 분석

* **현재까지** 과적합 징후 없음 (Train Loss 안정적으로 하락, Grad Norm 정상)
* **모델 학습 속도**: Token Accuracy가 빠르게 상승 중 → 챗봇 학습에 필요한 맥락 이해가 조기에 정착되는 양상
* **주의사항**: Epoch 2~3 진행 시 Validation Loss plateau 발생 여부 모니터링 필요

**5. 과적합/과소적합 대응**

| **기법** | **설명** | **적용 여부** |
| --- | --- | --- |
| Dropout | 과적합 방지 | O (0.1 사용) |
| 조기 종료 (Early Stopping) | validation loss 모니터링 | 수동 모니터링 |
| 학습률 감소 | Scheduler(plateau) 미적용 → constant lr | X |
| Gradient Checkpointing | 메모리 절약 및 일반화 개선 | O |

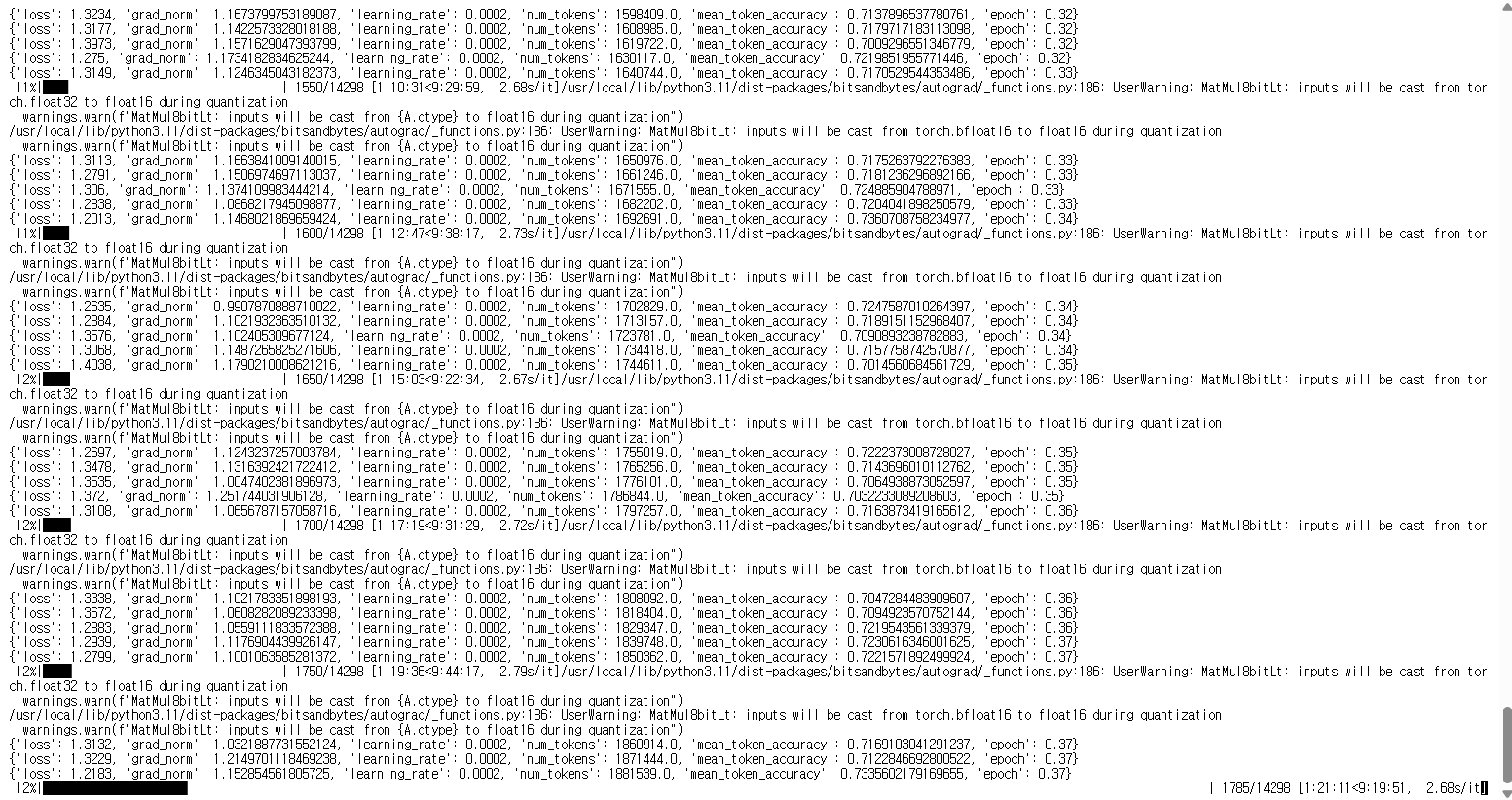
* 결과: 학습/검증 간 loss 차이가 적고 안정적 → 과적합 크게 나타나지 않음

**6. 결론 및 향후 계획**

* 최종 선정 모델: LGAI-EXAONE/EXAONE-3.5-7.8B-Instruct + LoRA 미세조정
* 활용 방안: 고객 응대 챗봇 모델 (정확한 답변 + 상담사 말투 반영)
* 향후 계획:  
  + 데이터 추가 전처리 후 재학습 예정
  + 테스트 데이터로 실사용 환경 평가 예정
  + API 연동, 실사용 환경 테스트 진행

**7. 부록**

* 학습 로그 캡처



* 주요 파라미터 설정:
  + 토크나이저/모델 pre\_trained 파라미터(HuggingFace trust\_remote\_code 지원)

trust\_remote\_code=True

* + 토크나이저 패딩 토큰

tokenizer.pad\_token = tokenizer.eos\_token