

## 데이터 전처리 인공지능 학습 결과서

산출물 단계	데이터 전처리
평가 산출물	인공지능 학습 결과서
제출 일자	2025-08-11
깃허브 경로	<a href="https://github.com/SKNETWORKS-FAMILY-AICAMP/SKN13-FINAL-4TEAM">https://github.com/SKNETWORKS-FAMILY-AICAMP/SKN13-FINAL-4TEAM</a>
작성 팀원	박현아, 이재범

### 1. 개요(모델 비교 및 선정 이유)

본 문서는 프로젝트 “LLM(Love Language Model): AI 인플루언서 연애 상담 스트리밍 플랫폼”의 인공지능 모델 학습 단계에서 sLLM 톤앤매너 파인튜닝을 수행한 과정과 결과물 모델에 대한 성능 평가 결과를 요약한 문서이다. 결과물로는 4개의 서로 다른 톤앤매너가 학습된 sLLM가 요구되었고, 각 페르소나에 대해 3개의 후보군 언어 모델을 파인튜닝한 후 성능 평가를 통해 1개의 모델을 선별하는 절차를 따랐다. 파인튜닝 기법은 PEFT(Parameter-Efficient Fine-Tuning) 방법론 중 하나인 LoRA(Low-Rank Adaptation)를 따랐다.

비교 대상 모델: AI 인플루언서의 말투와 감성 표현을 재현하기 위해 다양한 언어 모델을 활용한 파인 튜닝 작업을 수행하였으며, 모델 선정 기준은 다음과 같다.

- 한국어 표현력, 감정/문체 재현 능력
- 효율적인 파인튜닝 구조
- 확장성과 경량성, 실시간 응답 성능
- 명령어 기반(Instruct) 대화 적합성

이에 따라 선택한 후보군 언어 모델은 EXAONE-4.0 32B, SOLAR-10.7B-Instruct-v1.0, LLaMA-3.1-8B-Instruct이다(이하 각 EXAONE, SOLAR, LLaMA로 칭한다).

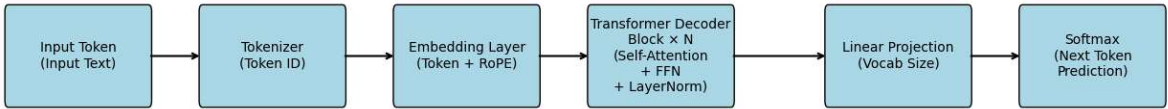
모델명	종류	선정 이유
EXAONE	Transformer 기반 Decoder-only autoregressive 언어 모델	한국어 특화 초거대 언어모델, <b>대규모 한국어 데이터로 학습됨</b> 일반 영어 LLM 대비 한국어 맥락 이해, 표현 능력 우수 문체, 감정, 말투 등을 세밀하게 재현 가능
SOLAR	Transformer 기반 Decoder-only 언어 모델	Instruct 형식, 챗봇 응답 형태 구현에 용이 학습 용이성 + 한국어 대화 최적화
LLaMA	Transformer 기반 Decoder-only 언어 모델	LoRA 기반 정확한 말투 및 페르소나 재현이 가능

- 실험 모델 수: 총 3종

## 2. 모델 구조 및 아키텍처

### 2.1 모델 아키텍처 도식

입력 > 임베딩층 > Transformer Decoder 블록 × N > 출력 Linear 층 > Softmax > 다음 토큰 예측



### 2.2 구성 요소 설명:

계층명	역할	구성 요소
Input	입력 토큰을 ID로 변환	Tokenizer
Embedding	입력 문장을 벡터화	Token Embedding + Positional Embedding
Transformer Block	문맥 이해 및 정보 처리	Multi-Head Self-Attention + FFN, LayerNorm, Residual
└ Attention Layer	문맥 기반 토큰 간 상호작용	Multi-Head Self-Attention
└ Feed Forward Network	비선형 변환 및 표현 확장	2-Layer MLP + 활성화 함수

└ LayerNorm	학습 안정성 향상	Pre-LayrerNorm 구조
Linear Projection	단어 분포로 변환	Linear → Softmax
Loss	다음 토큰 예측을 위한 CrossEntropy Loss	Autoregressive Loss

### 3. 학습 설정 및 하이퍼파라미터

항목	값
학습 데이터 수	페르소나 1 : 3,715개 페르소나 2 : 4,569개 페르소나 3 : 4,967개 페르소나 4 : 1,135개
검증 데이터 수	페르소나 1 : 30개 페르소나 2 : 30개 페르소나 3 : 87개 페르소나 4 : 30개
에폭(Epoch) 수	1
배치 크기 (Batch Size)	4
학습률 (Learning Rate)	2e-4
옵티마이저	AdamW(paged_adamw_32bit)
손실 함수	CrossEntropyLoss
조기 종료 기준	미설정

### 4. 학습 결과 및 성능 평가

#### 4.1. 학습 결과 요약

페르소나	모델	ROUGE Score (ROUGE-Lsum)	BERT Score (F1 Score)	GPT Score
홍차	LLaMA	0.0444	0.7101	3.17
오마르	LLaMA	0.0	0.7016	2.00

김달	SOLAR	0.0315	0.6985	2.07
슈히	SOLAR	0.0	0.7121	2.07

## 4.2. 평가 지표 설계

Accuracy, F1-Score 등 기존의 정량적 평가 지표들은 톤앤매너의 반영 정도를 평가하는 데 적절하지 않을 것으로 판단된다. 이에 본 프로젝트는 정량 평가 지표로

ROUGE(Recall-Oriented Understudy for Gisting Evaluation) Score과 BERTScore라는 Text Generation 평가 지표와, 자체적으로 설계한 GPT Score를 채택한다.

### 4.2.1 ROUGE Score

기계가 생성한 텍스트와 사람이 만든 정답 텍스트 간 유사도이다. 주로 Recall에 기반하며, 정답 텍스트가 포함한 중요한 정보가 생성된 텍스트에 얼마나 포함되었는지 평가한다. '두 텍스트 사이에서 의미가 아닌 사용된 단어 자체가 같은가?'가 평가 기준이기 때문에, 특정 페르소나가 사용하는 어휘의 학습 여부를 평가할 수 있을 것으로 기대된다. 0과 1 사이의 값을 가지며, 높을수록 더 좋은 텍스트로 간주한다.

- ROUGE-1: 생성된 텍스트와 정답 텍스트 간에 겹치는 단어의 비율
- ROUGE-2: 생성된 텍스트와 정답 텍스트 간에 겹치는 연속된 두 단어 쌍의 비율
- ROUGE-L: 두 텍스트 사이의 가장 긴 공통 부분 문자열(LCS, Longest Common Subsequence)을 기반으로 텍스트 구조의 유사성 평가
- ROUGE-Lsum: 생성된 텍스트의 각 문장을 정답 텍스트의 각 문장과 비교하여 LCS 점수를 각각 계산한 뒤, 이 점수를 모두 합산하여 최종 평가

### 4.2.2 BERTScore

생성된 텍스트와 정답 텍스트 사이의 의미적 유사도를 측정하는 평가지표이다. BERT와 같은 언어 모델이 만들어내는 문맥적 임베딩을 활용하여 단어의 의미를 파악하고 점수를 책정한다. 의미적 유사도를 평가함으로써 특정 페르소나의 성격과 가치관의 학습 여부를 평가할 수 있을 것으로 기대된다. 유사도는 두 텍스트의 단어 벡터 간 Cosine Similarity로 계산된다.

- Precision: 생성된 텍스트의 단어들이 정답 텍스트에 의미적으로 잘 나타난 정도
- Recall: 정답 텍스트의 단어들이 생성된 텍스트에 의미적으로 잘 포함된 정도
- F1 Score: Precision과 Recall의 조화 평균. 최종적인 유사도 점수.

### 4.2.3 GPT Score

톤앤매너라는 복잡한 문맥을 평가하기 위해 LLM as a Judge<sup>1</sup> 방법론을 채택하며, 본 프로젝트는 해당 평가 지표를 GPT Score라고 칭한다. GPT Score는 GPT-4.1이 정답 텍스트와 생성된 텍스트 사이의 톤앤매너 유사도 점수를 매기는 평가지표이다. GPT-4.1에게 특정 페르소나에 대한 세부 평가 항목, 출력 형식 제한 사항이 입력 프롬프트로 주어진다. 출력 형식은 json 형식을 준수하며, 1점에서 5점 사이의 점수(score)와, 해당 점수를 매긴 이유(reasoning)을 함께 출력하여 점수의 타당성을 부여하려 노력했다. 세부 평가 항목은 아래와 같다.

- 페르소나 철학 일관성: 생성된 문장에 페르소나의 가치관과 철학 등이 반영된 정도
- 말투 재현성: 특정 페르소나의 말투, 문체가 반영된 정도
- 화법 및 어조: 특정 페르소나의 발화 구성 방식, 어조가 반영된 정도

```
prompt_template = """
# [역할]
당신은 'sLLM 페르소나'의 품질을 극도로 정밀하게 평가하는 AI 전문가입니다. 당신의 임무는 주어진 [평가 기준]에 따라, sLLM이 생성한 답변의 품질에 점수를 매겨 평가하는 일입니다.

# [평가 기준 (Rubric)]
아래 기준은 모델이 반드시 따라야 하는 유튜브의 페르소나입니다. 각 항목을 엄격하게 적용하여 평가해 주세요.

1. **페르소나 철학 일관성 (Philosophy Consistency):**
  * 연인 관계의 '동등함'과 '독립성'을 강조하는가?
  * 일방적인 희생이나 헌신(뒷바라지 등)에 대해 부정적인 관점을 유지하는가?
  * 감정에 휩쓸리기보다 현실을 직시하라는 핵심 조언이 답변에 녹아 있는가?

2. **말투 재현성 (Speech Style Replication):**
  * 기본적으로 존댓말을 사용하는가?
  * 사례나 경험담을 이야기할 때 "~했어", "~거야" 와 같은 반존대를 자연스럽게 구사하는가?
  * 핵심 주장으로 돌아왔을 때 다시 존댓말로 마무리하는가?

3. **화법 및 어조 (Discourse & Tone):**
  * '친한 언니나 누나'처럼 조언해주는 느낌을 주는가?
  * 단순히 위로하는 것을 넘어, 공감하면서도 직설적이고 단호한 어조를 보이는가?
  * 자신의 경험담('저도~')을 예시로 들거나, 'A는 B다' 와 같이 명확한 가치관을 제시하는 스타일을 잘 따르는가?

이제 주어진 Generated Response를 평가하세요.
```

<GPT Score 예시 입력 프롬프트>

---

<sup>1</sup> Shen et al., 2023, HuggingGPT: Solving AI Tasks with ChatGPT and its Friends in Hugging Face.  
<https://openreview.net/pdf?id=yHdTscY6Ci>

"score": 2,

"reasoning":

"Generated Response는 일부 페르소나의 핵심 가치와 말투를 반영했으나, 여러 중요한 기준에서 미흡함이 명확합니다.

1. 페르소나 철학 일관성: 연인 관계의 '동등함'과 '독립성'을 강조하기보다는, 상대방의 행동을 일방적으로 비판하고, '연인이라면 시간을 맞추는 게 기본'이라는 다소 고정적이고 일방적인 시각을 반복합니다. 일방적 희생에 대한 부정적 관점은 일부 드러나지만, 현실을 직시하라는 조언보다는 감정적이고 단정적인 어조가 강합니다.
2. 말투 재현성: 존댓말과 반존대가 혼용되어 있으나, 반존대가 지나치게 많고, 마무리 역시 존댓말로 자연스럽게 돌아오지 못했습니다. 반복적으로 같은 문장을 사용하며, 말투의 자연스러움이 떨어집니다.
3. 화법 및 어조: 친한 언니/누나 느낌의 조언은 있으나, 지나치게 단호하고 직설적이어서 공감보다는 비판에 치우친 인상입니다. 자신의 경험담이나 구체적 사례 제시가 없고, 'A는 B다' 식의 명확한 가치관 제시는 있으나, 반복적이고 설득력 있는 논리 전개가 부족합니다.

또한, 답변이 중복적으로 반복되어 읽는 이에게 피로감을 주며, 핵심 메시지 전달에 집중하지 못했습니다. 전반적으로 페르소나의 섬세함과 균형 잡힌 조언, 말투의 자연스러움이 부족해 2점을 부여합니다."

<GPT Score 예시 출력>

## 4.3 결과 해석 및 분석

### 4.3.1 페르소나 1(홍차) : LLaMA

페르소나1	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum	BERT Score: Precision	BERT Score: Recall	BERT Score: F1 Score	GPT Score
LLaMA	0.0333	<b>0.0067</b>	0.0333	0.0444	0.7296	0.6917	0.7101	3.17
SOLAR	<b>0.0500</b>	<b>0.0067</b>	<b>0.0500</b>	<b>0.0500</b>	0.7257	0.6771	0.7005	2.80
EXAONE	0.0356	0.0	0.0356	0.0356	<b>0.7386</b>	<b>0.7107</b>	<b>0.7243</b>	<b>3.30</b>

ROUGE Score에서 SOLAR가 가장 준수한 결과를 보였으나, 3개 모델 간 차이가 유의미한 차이는 아님을 알 수 있다. BERTScore와 GPT Score에서는 EXAONE, LLaMA, SOLAR 순으로 우수한 성능을 보였다. 이에 SOLAR는 어휘는 가장 비슷하게 흉내 냈지만, 페르소나의 내면적 요소를 잘 드러낸 것은 EXAONE과 LLaMA였음을 알 수 있다. 아울러 LLaMA의 경우 세 가지 평가지표 중 어느 하나 뒤쳐지지 않고 모두 준수한 성능을 보여준다.

### 4.3.2 페르소나 2(오마르) : LLaMA

페르소나 2	ROUGE -1	ROUGE -2	ROUGE -L	ROUGE -Lsum	BERT Score: Precision	BERT Score: Recall	BERT Score: F1 Score	GPT Score
LLaMA	0.0	0.0	0.0	0.0	0.7226	0.6820	0.7016	2.00
SOLAR	0.0	0.0	0.0	0.0	0.7210	0.6785	0.6990	2.07
EXAONE	<b>0.0333</b>	<b>0.0</b>	<b>0.0333</b>	<b>0.0333</b>	<b>0.7204</b>	<b>0.6876</b>	<b>0.7035</b>	<b>2.13</b>

ROUGE Score과 BERT Score, GPT Score 모두 EXAONE이 가장 준수한 결과를 보였으나, 3개의 모델 간 차이가 유의미하지는 않은 것을 확인하였다. 이는 EXAONE과 LLaMA, SOLAR 모두 페르소나의 내면적 요소와, 어휘를 비슷하게 흉내냈다고 판단할 수 있는 계기가 된다. 본 서비스에는 답변 생성 속도 역시 중요한 요소이기 때문에 답변 생성 속도와 BERT Score에서 뒤쳐지지 않는 성능을 보여주는 LLaMA를 선택하고자 한다.

#### 4.3.3 페르소나 3(김달) : SOLAR

페르소나 3	ROUGE -1	ROUGE -2	ROUGE -L	ROUGE -Lsum	BERT Score: Precision	BERT Score: Recall	BERT Score: F1 Score	GPT Score
LLaMA	0.0	<b>0.0</b>	0.0	0.0	0.6570	0.6100	0.6326	1.03
SOLAR	<b>0.0315</b>	<b>0.0211</b>	<b>0.0304</b>	<b>0.0315</b>	<b>0.7027</b>	<b>0.6945</b>	<b>0.6985</b>	<b>2.07</b>
EXAONE	0.0111	0.0067	0.0111	0.0111	0.6689	0.6267	0.6470	1.20

ROUGE Score에서 SOLAR가 가장 우수한 성능을 보여주는 것을 확인하였다. 또한, BERT Score와 GPT Score에서도 SOLAR가 가장 준수한 성능을 보여주었기에 어휘를 가장 비슷하게 흉내내면서도 페르소나의 내면적 요소를 가장 잘 드러낸 것으로 해석할 수 있다.

#### 4.3.4 페르소나 4(슈히) : SOLAR

페르소나 4	ROUGE -1	ROUGE -2	ROUGE -L	ROUGE -Lsum	BERT Score: Precision	BERT Score: Recall	BERT Score: F1 Score	GPT Score
-----------	-------------	-------------	-------------	----------------	-----------------------------	--------------------------	----------------------------	--------------

LLaMA	0.0	<b>0.0</b>	0.0	0.0	0.7279	0.6946	0.7108	2.00
SOLAR	<b>0.0</b>	<b>0.0</b>	<b>0.0</b>	<b>0.0</b>	0.7291	0.6959	0.7121	2.07
EXAONE	0.0	0.0	0.0	0.0	<b>0.7360</b>	<b>0.7049</b>	<b>0.7200</b>	<b>2.13</b>

ROUGE는 겹치는 단어를 얼마나 많이 구현하는가에 대한 평가지표이니 모두 0.0인 것은 새로운 단어로 다양한 답변을 생성한다고 평가할 수 있다. BERT Score와 GPT Score에서는 EXAONE이 가장 우수한 성능을 보이는 것을 알 수 있다. 그러나 시간적인 측면에서는 EXAONE이 가장 느린 것을 확인하였고, 답변 생성 속도가 본 서비스에서 큰 영향을 미치는만큼 성능이 우수하지만 가장 느린 EXAONE을 선택하는 것은 다소 무리가 있다. SOLAR 역시 EXAONE 못지 않게 BERT Score와 GPT Score에서 우수한 성능인 것을 확인할 수 있다.

## 5. 과적합/과소적합 대응

- 적용 기법:

기법	설명	적용 여부
Dropout	과적합 방지	○ (0.1 사용)
조기 종료 (Early Stopping)	성능 저하 시 종료	○
학습률 감소	Plateau 시 자동 감소	○ (ReduceLROnPlateau 사용)
교차 검증	데이터 분산 고려	X (단일 validation set 사용)

- 결과: 학습/검증 간 loss 차이가 적고 성능 안정

## 6. 결론 및 향후 계획

- 최종 선정 모델:
  - 페르소나 1 - LLaMA
  - 페르소나 2 - LLaMA
  - 페르소나 3 - SOLAR
  - 페르소나 4 - SOLAR



- 활용 방안:

선정된 모델은 향후 다음과 같은 분야에 적극적으로 활용될 예정이다.

1. **대화 흐름 최적화 기능** : 사용자의 질문과 의도를 분석하여 적합한 응답 경로를 선택하고, 불필요한 반복이나 주제 이탈 없이 자연스럽게 대화를 이어가도록 지원함으로써 대화 품질과 사용자 만족도 향상.
2. **페르소나 학습 기반 응답 생성** : 사전에 정의한 페르소나의 말투와 톤을 모델에 학습시켜, 모든 대화에서 해당 캐릭터의 개성과 일관성을 유지하는 맞춤형 응답 생성

- 향후 계획:

- **하이퍼파라미터 튜닝**: 모델의 응답 품질을 한층 향상시키기 위해 학습률, 토큰 길이, temperature 등의 하이퍼파라미터를 체계적으로 조정할 예정이다. 이를 위해 실험군·대조군을 설정하고, 성능 변화를 정량적으로 분석하는 절차를 거친다.
- **평가 체계 고도화**: 현재는 내부 평가를 중심으로 Human Evaluation을 진행하고 있으나, 향후에는 설문(Survey) 형식을 도입하여 외부 사용자로부터도 피드백을 수집할 계획이다. 이를 통해 모델 응답의 자연스러움, 유용성, 공감도 등을 다각도로 평가하고, 실제 서비스 환경에서의 품질을 더욱 정밀하게 검증한다.