JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 31 No. 3, pp. 79-87, March 2026
https://doi.org/10.9708/jksci.2026.31.03.079

A Comparative Study on the Explainability and Efficiency of Foundation Models in Prompt-based Sentiment Analysis

Misun Lee*

*Ph.D. Candidate, Department of AI Convergence Engineering, Sejong University, Seoul, Korea

[Abstract]

Recently, large language models (LLMs) have demonstrated strong performance in natural language processing tasks, including sentiment analysis; however, prediction performance and interpretability consistency can vary significantly depending on prompt design. This study conducts a comparative analysis of the sentiment analysis performance and LIME-based explainability of foundation models across different prompt types. Experiments were conducted on Korean app reviews, English IMDB reviews, and the English TweetEval dataset using role-based, context-rich, few-shot, and format-constrained prompts, and the performance of GPT-4o-mini and Gemini 2.5 Flash models was evaluated. In addition, the semantic consistency and explanation stability of prediction results were analyzed both quantitatively and qualitatively using sentence embedding–based cosine similarity and LIME. Experimental results showed that in the binary classification setting (IMDB), performance differences remained within 1.5 percentage points (%p), and differences in explanation consistency were also 0%p, indicating overall limited variation. In contrast, in the three-class settings (APP and Tweet), performance differences of up to approximately 3.0%p were observed. In the tweet domain, differences in explanation consistency were confirmed to be 25.0%p based on LIME agreement and 12.5%p based on overall agreement. This study is meaningful in that it systematically analyzes the impact of prompt design on both performance and explainability in LLM-based sentiment analysis.

▸ **Key words**: Foundation Models, Large Language Models, Sentiment Analysis, Prompt Engineering, Explainable Artificial Intelligence

[요 약]

최근 대규모 언어 모델(Large Language Model, LLM)은 감성 분석을 포함한 자연어 처리 분야에서 우수한 성능을 보이고 있으나, 프롬프트 설계 방식에 따라 예측 성능과 해석 일관성이 크게 달라지는 문제가 존재한다. 본 연구에서는 프롬프트 유형에 따른 파운데이션 모델의 감성 분석 성능과 LIME을 활용한 설명 가능성을 비교 분석하였다. 실험에서는 한국어 앱 리뷰, 영어 IMDB 리뷰, 영어 TweetEval 데이터셋을 대상으로 역할 기반, 문맥 강화, 소수 예제, 형식 제한 프롬프트를 적용하고, GPT-4o-mini와 Gemini 2.5 Flash 모델의 성능을 평가하였다. 또한 문장 임베딩 기반 코사인 유사도와 LIME을 활용하여 예측 결과의 의미적 일관성과 설명 안정성을 정량, 정성적으로 분석하였다. 실험 결과, 이진 분류(IMDB)에서는 성능 차이가 1.5%p 이내였고 설명 일관성 차이도 0%p로 나타나 전반적으로 제한적이었다. 반면 삼중 분류(APP, Tweet)에서는 최대 약 3.0%p의 성능 차이가 나타났으며, 트윗 도메인에서는 설명 일관성 차이가 LIME 기준 25.0%p, 전체 합의율 기준 12.5%p로 확인되었다. 본 연구는 LLM 기반 감성 분석에서 프롬프트 설계가 성능과 해석 가능성에 미치는 영향을 체계적으로 분석하였다는 점에서 의의를 갖는다.

▸ **주제어**: 파운데이션 모델, 대규모 언어모델, 감성 분석, 프롬프트 엔지니어링, 설명가능 인공지능

• First Author: Misun Lee, Corresponding Author: Misun Lee
*Misun Lee (llmss2000@hanmail.net), Department of AI Convergence Engineering, Sejong University
• Received: 2026. 02. 09, Revised: 2026. 02. 23, Accepted: 2026. 03. 03.

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

I. Introduction

최근 자연어 처리 기술은 대규모 언어 모델(Large Language Model, LLM)의 발전에 따라 비약적인 성장을 이루고 있으며, 감성 분석과 같은 텍스트 분류 문제에서도 기존의 규칙 기반 및 통계적 기법을 넘어서는 성능을 보이고 있다[1]. 사전학습 언어 모델과 트랜스포머 기반 구조의 도입은 텍스트 분류 성능 향상에 기여한 것으로 보고되고 있으며[2]. 최근에는 대규모 사전학습 기반으로 다양한 다운스트림 태스크에 활용 가능한 파운데이션 모델이 주목받고 있다[3].

파운데이션 모델은 프롬프트를 통해 태스크가 지정되며, 프롬프트 설계에 따라 모델의 응답과 예측 성능이 달라질 수 있다[4]. 다중 클래스 감성 분석에서는 중립 또는 혼합 감성과 같이 경계가 불명확한 범주로 인해 세밀한 감성 구분이 어려워질 수 있으며, 이에 따라 실험 설정이나 프롬프트 방식에 따라 예측 결과가 달라질 수 있다[5].

최근에는 프롬프트 기반 LLM 감성 분석에서 높은 분류 성능과 함께 예측 근거의 해석 가능성을 확보하려는 연구들이 보고되고 있으며, 프롬프트 설계에 따른 성능 변화와 설명의 일관성을 동시에 분석하려는 시도가 이루어지고 있다[6][7].

이에 본 연구에서는 프롬프트 유형에 따른 파운데이션 모델의 감성 분석 성능과 해석 가능성을 종합적으로 비교 분석하였으며, 이를 위해 한국어 앱 리뷰, 영어 IMDB 리뷰, 영어 TweetEval 데이터셋을 대상으로 역할 기반, 문맥 강화, 소수 예제, 형식 제한 프롬프트를 적용하여 LLM 모델(GPT-4o-mini, Gemini 2.5 Flash) 의 Prompt Runner API 기반 응답결과 데이셋을 이용하여 파운데이션모델의 감성 분석 성능을 평가하였다. 또한 문장 임베딩 기반 코사인 유사도와 LIME을 활용하여 예측 결과의 의미적 일관성과 설명 안정성을 정량적, 정성적으로 분석하였다.

본 논문의 구성은 다음과 같다. 2장에서는 LLM 프롬프트 엔지니어링, 감성분석 성능평가 및 XAI 관련 선행연구를 검토하고, 3장에서는 실험 알고리즘과 성능 및 평가방법을 제시한다. 4장에서는 각 모델의 실험결과를 분석하고, 마지막 5장에서는 결론을 제시한다.

II. Related Work

대규모 언어 모델에서 프롬프트 엔지니어링은 별도의 파인튜닝 없이 인컨텍스트 러닝(in-context learning)을 통해 모델의 추론 성능을 조절하는 방법으로 활용되고 있다. Rouzegar and Makrehchi는 역할 기반 프롬프트 설정과 소수 예제(few-shot) 구성의 변화가 텍스트 분류 및 감성 분석 과제에서 모델 성능에 영향을 미칠 수 있음을 실험적으로 분석하였다[8]. 해당 연구에서는 역할 정보의 명시 여부와 예제 제시 방식에 따라 분류 정확도와 예측 안정성이 달라질 수 있음을 보였으며, 이는 프롬프트 설계 방식이 LLM 기반 감성 분석 성능에 중요한 요인임을 시사하고 있다[8]. 또한 Huttula는 문맥 정보와 예제 제시 방식에 따라 LLM의 예측 안정성과 분류 정확도가 달라짐을 실험적으로 분석하였다[9]. 또한 Aqsa and Saeed는 형식 및 제약 기반 프롬프트 설계가 출력 품질과 평가 가능성에 기여함을 보고하고, 다양한 프롬프트 기법을 유형별로 정리하였다[10].

감성분석에서 정확도는 LLM의 응답과 기준 정답 간 의미적 유사도를 산출하는 지표이며, 문장 간 의미적 근접도는 Sentence-BERT(SBERT) 임베딩 벡터의 코사인 유사도 식을 통해 계산된다[11][12].

$$(a,b) = \frac{a \bullet b}{\| a \| \ \| b \|} \qquad (1)$$

코사인 유사도는 평점을 벡터로 생각하고, 2개 벡터 사이의 각도를 계산하고, 그 각도가 적을수록 가까이 있다고 판단하기 때문에 서로 유사하다고 결정하는 방식이다[8]. Lee & Ahn은 이 공식을 협업 필터링 기반 사용자 유사도 계산에 적용하여 평가 정확도를 향상시킬 수 있음을 실험적으로 검증하였다[12].

딥러닝 모델의 해석 가능성(Explainability)은 모델의 신뢰성과 윤리적 투명성을 확보하기 위한 필수 요소로 부각되고 있다[13][14].

대표적인 XAI 기법으로는 Local Interpretable Model-agnostic Explanations (LIME) [14], SHapley Additive exPlanations (SHAP) [15], Integrated Gradients (IG), Layer-wise Relevance Propagation (LRP) 등이 있다 [16]. LIME은 입력 주변의 데이터 공간을 국소적으로 샘플링하여 로컬 대리모델 을 학습하고, 모델의 예측을 근사화한다. 그 목적 함수는 다음과 같이 정의된다 [14].

$$\xi(x) = \arg\min_{g \in G} L(f, g, w_x) + (g) \qquad (2)$$

여기서 L은 예측 오차(loss function), w_x는 데이터 근접도(weighting function), $\Omega(g)$는 모델 복잡도 정규화 항이다.

III. Methodology

본 연구 방법은 Fig. 1과 같이 다음 단계로 진행된다.

3.1 Step 1. Dataset Preparation and Task Definition

본 연구에서는 프롬프트 유형에 따른 LLM 기반 감성 분석 성능과 해석 가능성을 비교하기 위해 한국어 앱 리뷰(3-class), 영어 IMDB 영화 리뷰(2-class), 영어 TweetEval 감성 데이터셋(3-class)을 사용하였다.

Table 1에 제시된 것과 같이, 한국어 앱 리뷰 데이터셋은 웹 크롤링을 통해 수집한 60,000건의 원천 데이터를 대상으로 전처리를 수행하였다. 영어 IMDB 리뷰 데이터셋[17]은 공개된 50,000건의 사전 정제 데이터를 사용하여 최소한의 추가 전처리만 적용하였으며, 영어 TweetEval 데이터셋[18]은 Hugging Face에서 제공되는 59,899건의 원천 데이터를 수집하여 한국어 앱 리뷰 데이터셋과 동일한 전처리 과정을 수행하였다.

수집된 3개 데이터셋은 입력 문장의 의미를 유지한 상태에서 모델 입력 형식에 맞도록 정규화를 수행하였으며, 언어 및 분류 조건에 따른 프롬프트 효과를 일관되게 비교할 수 있도록 구성하였다. 또한 Prompt Runner API 환경에서 안정적인 응답 결과를 도출하기 위해, 3분류 데이터셋(APP 리뷰, TweetEval)은 각 감성 클래스별(긍정, 중립, 부정) 300건씩 총 900건을, 2분류 데이터셋(IMDB)은 긍정 500건과 부정 500건으로 총 1,000건을 추출하였다. 모든 샘플은 클래스 불균형을 배제하기 위해 동일 기준의 균등 샘플링(stratified random sampling) 방식으로 Google Colab 환경에서 구성하였다.

3.2 Step 2. Prompt Design and Construction

프롬프트 설계 단계에서는 감성 분류 태스크에 대한 지시 방식의 영향을 비교하기 위해 역할 기반(role-based), 문맥 강화(context-rich), 소수 예제(few-shot), 형식 제한(format-constrained)의 네 가지 프롬프트 유형을 정의하였다.

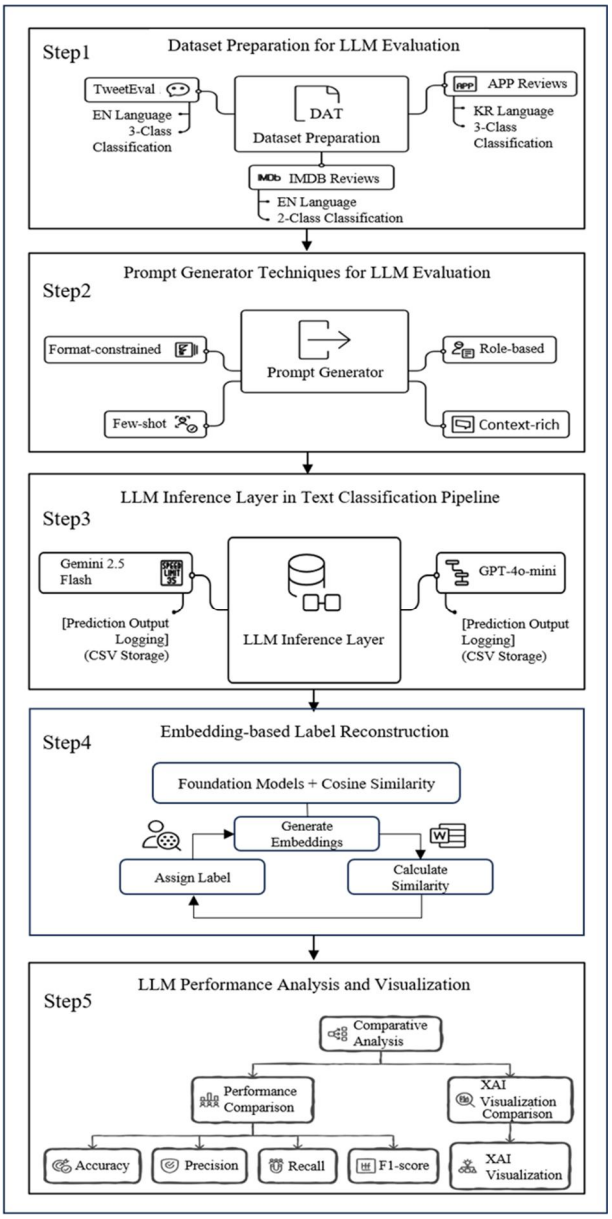


Fig. 1. Implementation algorithm for prompt-based LLM sentiment analysis

모든 프롬프트는 감성 분류 결과와 함께 입력 리뷰에서 직접 인용한 근거 구문을 줄바꿈 없는 단일 라인 형식으로 출력하도록 설계하였으며, 고정된 출력 포맷을 통해 결과를 CSV파일 필드에 그대로 저장할 수 있도록 구성하였다. 또한 해석 가능성 분석을 위해 근거 구문은 반드시 입력 문장에 존재하는 연속된 텍스트에서 발췌되도록 제한하고, 플레이스홀더 출력은 허용하지 않았다.

동일한 입력 문장과 모델 환경에서 프롬프트 유형만을 변경하여 실험을 수행함으로써 프롬프트 설계 방식에 따른 응답 특성과 성능 차이를 분석하기 위함이다.

Table 1. Average Response Time Results by Model

Dataset	Original Data Size	Sampling Criterion	Final Sample Size	Class Distribution
Korean App Review Dataset (Web-crawled and preprocessed)	Total 60,000 samples (Neutral : 20,000 Positive : 20,000 Negative : 20,000)	Uniform sampling of 300 samples per sentiment class	900	Neutral 300 Positive 300 Negative 300
English IMDB Movie Review Dataset (Publicly available, preprocessed)	Total 50,000 samples (Positive : 25,000 Negative : 25,000)	Uniform sampling of 500 samples per sentiment class	1,000	Positive 500 Negative 500
English TweetEval Sentiment Dataset (Hugging Face, preprocessed)	Total 59,899 samples (Negative : 11,377 Neutral : 27,479 Positive : 21,043)	Uniform sampling of 300 samples per sentiment class	900	Neutral 300 Positive 300 Negative 300

3.3 Step 3. LLM Inference and Response Collection

설계된 프롬프트를 기반으로 GPT-4o-mini와 Gemini 2.5 Flash 모델을 활용하여 감성 분석을 수행하였다. 각 모델은 Google Colab CPU 환경의 Prompt Runner API를 통해 동일한 실험 조건에서 실행되었으며, 모델의 출력 결과를 구조화된 형태로 수집하였다. 이 과정에서 모델별, 프롬프트 유형별, 언어 도메인별 응답 결과를 데이터셋 단위의 CSV 파일로 저장하여 이후 성능 평가 및 해석 분석에 활용하였다.

본 연구는 실제 API 기반의 반복 실험 환경을 고려하여 접근성이 안정적이고 저지연 응답 특성을 갖는 경량 모델인 GPT-4o-mini와 Gemini 2.5 Flash를 비교 대상으로 선정하였다. 두 모델은 동일한 프롬프트 세트에 대해 반복 호출이 가능하며, 실험 시간과 비용 측면에서 안정적인 비교 환경을 제공한다. 또한 원본 데이터셋은 대규모로 구성되어 있으나, API 기반 추론 환경의 제약(응답 지연, 호출 제한, 실험 재현성)을 고려하여 클래스 균형 기반의 샘플링 평가셋을 구성하였으며, 이를 통해 모든 모델에 동일한 프롬프트와 평가 조건을 적용하여 비교의 공정성과 실험 재현성을 확보하였다.

3.4 Step 4. Performance Evaluation Using Semantic Similarity

성능 평가 및 XAI 시각화 실험은 Google Colab Pro 환경의 NVIDIA A100 GPU(40GB VRAM)에서 수행하였으며, 한글 도메인은 KLUE-BERT, 영어 도메인은 DistilBERT 인코더 기반 파운데이션 모델을 사용하였다. 해당 인코더 모델은 각 도메인의 원본 대규모 데이터셋(한글 APP 약 6만 건, 영어 IMDB 약 5만 건)에 대해 사전 성능 및 자원 효율성 비교 실험을 수행 후 가장 안정적인 결과를 보인 모델 기준으로 선정하여, 언어 도메인별 성능평가의 일관성과 계산 효율을 동시에 확보하였다. 감성 분석 성능은 식(1)을 통해 예측 결과와 기준 정답 간의 의미적 유사도를 기반으로 평가하였으며, 이를 위해 문장 임베딩 기반 코사인 유사도를 적용하였다. 이를 통해 단순 정답 일치 여부를 넘어 의미적 일관성을 정량적으로 평가하고, 프롬프트 유형 및 분류 환경에 따른 모델 성능 차이를 비교 분석하였다.

3.5 Step 5. Explainability and Stability Analysis

마지막 단계에서는 인간 중심의 외부적 시각적 해석에 효과적인 로컬 기반 XAI 기법인 식(2) LIME을 적용하여, 모델 예측 결과와 프롬프트 유형에 따른 설명의 안정성과 일관성을 분석하였다. 동일한 입력에 대해 생성된 설명 결과를 비교함으로써, 프롬프트 설계가 모델의 해석 가능성에 미치는 영향을 정성적, 정량적으로 평가하였으며, 이를 통해 성능뿐만 아니라 설명 신뢰성 측면에서도 프롬프트 효과를 종합적으로 분석하였다.

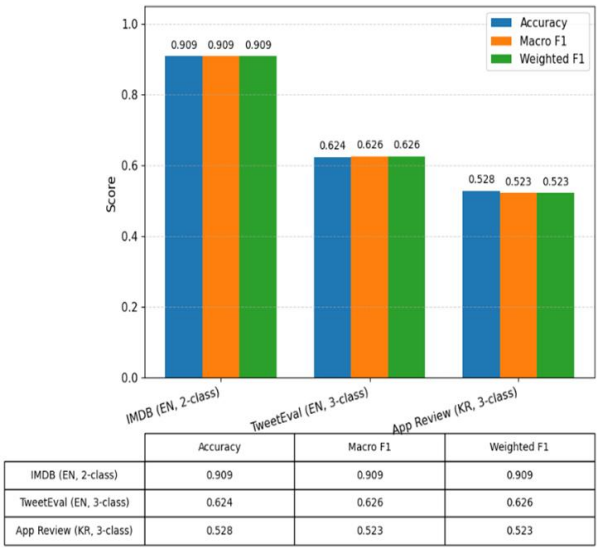
IV. Results and analysis

4.1 Sentiment analysis performance evaluation

본 연구에서는 감성 분석 성능 평가와 XAI 시각화 결과를 구분하여 비교 분석하였다. Table 2는 GPT-4o-mini 모델에 대해 프롬프트 유형별 및 도메인별 문장 임베딩 코사인 유사도를 기반으로 산출한 감성 분석 성능 지표를 나타낸다. 실험 결과, 2분류 환경인 IMDB 영화 리뷰 데이터셋이 3분류 환경인 한글 APP 리뷰와 TweetEval 데이터셋 대비 전반적으로 높은 성능을 보였다. 이는 TweetEval의 경우 중립 클래스의 의미적 경계 모호성과 코사인 유사도 기반 분리 한계에 기인하며, 한글 APP 리뷰에서는 형태소 기반 감정 표현의 불명확성과 중립 클래스 혼선이 주요 원인으로 분석된다. 또한 한글 도메인과 영어 도메인에 대한

3분류 성능을 비교한 결과, 전반적으로 영어 도메인에서의 성능이 미세하게 우수한 경향을 보였다. 한편 프롬프트 유형별 성능 차이는 크지 않게 나타나 전반적으로 균일한 결과가 도출되었으며, 이는 안정적인 프롬프트 설계와 일관된 제약조건 적용을 통해 모델 간 성능 편차가 제한적으로 나타났음을 시사한다.

Fig. 2는 GPT-4o-mini 모델의 도메인별 평균 성능 지표를 시각화하여 정량적 비교 결과를 직관적으로 제시하였다. 또한 Table 3과 Fig. 3은 Gemini 2.5 Flash 모델의 성능 평가 결과를 나타내며, GPT-4o-mini와 동일하게 IMDB 영화 리뷰 도메인에서 가장 우수한 성능을 보였다. 다만 Gemini 2.5 Flash 모델은 전반적으로 GPT-4o-mini 대비 소폭 높은 성능 우위를 나타내고 있음을 확인할 수 있다.



	Accuracy	Macro F1	Weighted F1
IMDB (EN, 2-class)	0.909	0.909	0.909
TweetEval (EN, 3-class)	0.624	0.626	0.626
App Review (KR, 3-class)	0.528	0.523	0.523

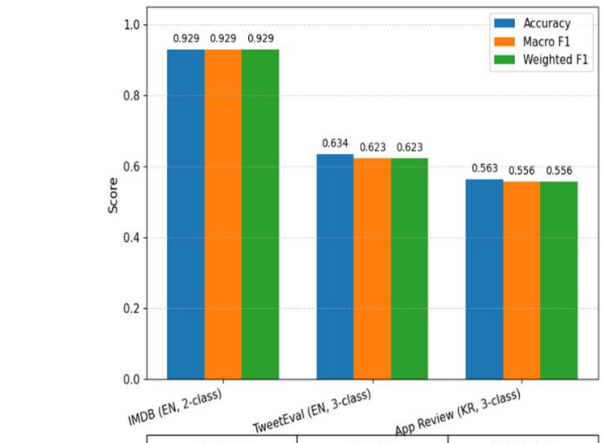
Fig. 2. Mean performance of GPT-4o-mini sentiment models

4.2 LIME-based explainability analysis

본 절에서는 프롬프트 유형에 따른 감성 분석 결과를 설명 가능성(XAI) 관점에서 보완적으로 분석하기 위해 LIME을 활용하였다. 설명 가능성 평가는 성능 지표와 달리, 모델 예측에 대한 국소적 설명의 신뢰도와 인간 중심 해석 가능성에 초점을 두었다.

Table 2. Experimental Results of Cosine Similarity-Based Sentiment Analysis Performance of the GPT-4o-Mini Model According to Prompt Types and Domain

APP Review Domain			
Prompt Type	Accuracy	Macro-F1	Weighted-F1
role_based	0.530	0.528	0.528
context_rich	0.532	0.528	0.528
few_shot	0.514	0.506	0.506
format_constrained	0.534	0.531	0.531
Tweet Sentiment Domain			
Prompt Type	Accuracy	Macro-F1	Weighted-F1
role_based	0.633	0.635	0.635
context_rich	0.622	0.626	0.626
few_shot	0.606	0.605	0.605
format_constrained	0.636	0.639	0.639
IMDB Movie Review Domain			
Prompt Type	Accuracy	Macro-F1	Weighted-F1
role_based	0.908	0.908	0.908
context_rich	0.912	0.912	0.912
few_shot	0.909	0.909	0.909
format_constrained	0.906	0.906	0.906



	Accuracy	Macro F1	Weighted F1
IMDB (EN, 2-class)	0.929	0.929	0.929
TweetEval (EN, 3-class)	0.634	0.623	0.623
App Review (KR, 3-class)	0.563	0.556	0.556

Fig. 3. Mean performance of Gemini 2.5 Flash sentiment models

Table 3. Experimental Results of Cosine Similarity-Based Sentiment Analysis Performance of the Gemini 2.5 Flash Model According to Prompt Types and Domain

APP Review Domain			
Prompt Type	Accuracy	Macro-F1	Weighted-F1
role_based	0.567	0.562	0.562
context_rich	0.567	0.562	0.562
few_shot	0.560	0.547	0.547
format_constrained	0.557	0.551	0.551
Tweet Sentiment Domain			
Prompt Type	Accuracy	Macro-F1	Weighted-F1
role_based	0.638	0.629	0.629
context_rich	0.644	0.635	0.635
few_shot	0.629	0.615	0.615
format_constrained	0.624	0.612	0.612
IMDB Movie Review Domain			
Prompt Type	Accuracy	Macro-F1	Weighted-F1
role_based	0.925	0.925	0.925
context_rich	0.930	0.930	0.930
few_shot	0.938	0.938	0.938
format_constrained	0.923	0.923	0.923

먼저, Fig. 4-6의 LIME 시각화 결과는 각 도메인에서 확신도(confidence)가 가장 높은 상위 1개 샘플을 대표 사례로 선정하여 제시하였다. 이는 모든 프롬프트 유형의 시각화를 나열하는 대신, LIME 설명이 실제로 직관적인 해석 근거를 제공하는지를 정성적으로 확인하기 위함이다. 시각화 결과는 감성 판단에 기여한 주요 단어의 상대적 중요도를 강조함으로써, 모델 예측 과정의 외부적 해석 가능성을 보여준다.

Fig. 4는 APP 도메인에서 GPT-4o-mini 모델에 context-rich 프롬프트를 적용한 대표 사례에 대한 LIME 기반 감성 분석 시각화를 나타낸다. 시각화 결과, 모델은 “불편할”, “때가”와 같은 중립적 상황 표현에 상대적으로 높은 가중치를 부여하고 있으며, 해당 단어들은 긍,부정 감성을 직접적으로 드러내지 않으면서 문맥 전반의 중립성을 형성하는 요소로 작용하며, 이로 인해 모델의 중립 감성 예측이 인간 관점에서도 합리적으로 해석 가능함을 보여주고 있다.

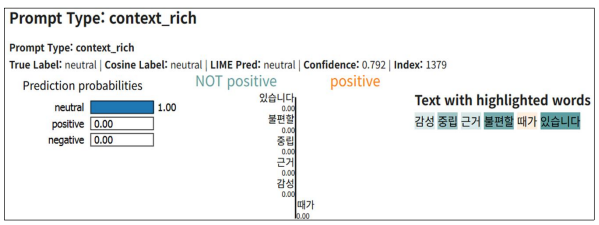


Fig. 4. LIME-based visualization of sentiment analysis using the context-rich prompt with the GPT-4o-mini model on the app domain

실험결과 Table 4-6은 각 도메인별로 프롬프트 유형별 상위 확신도 1개 샘플을 기준으로 산출한 XAI 신뢰도 지표를 평균화(Average, Overall)하여, GPT-4o-mini와 Gemini 2.5 Flash 모델을 비교한 분석표이다.

Table 4는 분석 결과, 두 모델 모두 LIME 기반 설명과 실제 라벨 간 일치율은 8.3%로 동일하게 나타났으며, 코사인 유사도 기반 일치율은 GPT-4o-mini가 66.7%, Gemini 2.5 Flash가 58.3%로 GPT-4o-mini가 상대적으로 높은 설명 일관성을 보였다. 반면, LIME 개선도 측면에서는 Gemini 2.5 Flash가 소폭 우위를 보여, APP 도메인에서는 모델 간 설명 신뢰도의 특성이 상이함을 확인할 수 있다.

Table 4. Quantitative XAI Comparison Analysis of GPT-4o-mini and Gemini 2.5 Flash in the APP Domain

Metric	GPT-4o-mini (APP)	Gemini 2.5 Flash (APP)	Difference (Gemini − GPT)
Match (LIME = True)	8.3%	8.3%	0% (Identical)
Match (Cosine = True)	66.7%	58.3%	−8.4 p ↓ (Decreased)
LIME Improvement (Δ%)	−58.4%	−50.0%	+8.4 p↑ (Improved)
Overall Agreement (%)	66.7%	58.3%	−8.4 p ↓ (Decreased)

Fig. 5는 IMDB 도메인에서 Gemini 2.5 Flash 모델에 few-shot 프롬프트를 적용한 대표 사례에 대한 LIME 기반 감성 분석 시각화를 나타낸다. 시각화 결과, “good”, “good time”과 같은 명확한 긍정 감성 단어가 높은 가중치로 강조되며, 이는 이진 분류 환경에서 모델이 감성 결정에 직접적으로 기여하는 핵심 단어를 중심으로 일관된 설명을 제공함을 보여준다.

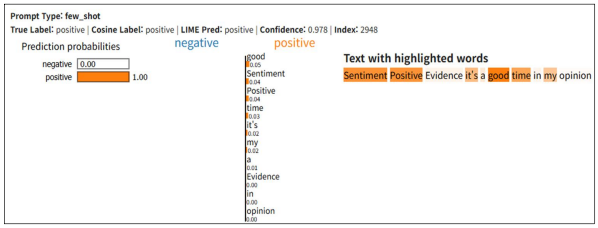


Fig. 5. LIME-based visualization of sentiment analysis for the few-shot prompt using the Gemini 2.5 Flash model on the imdb domain

Table 5는 분석 결과, 두 모델 모두 LIME 기반 설명과 실제 라벨 간 일치율과 코사인 유사도 기반 일치율이 각각 100%로 동일하게 나타났으며, LIME 개선도 및 전체 합의율에서도 차이가 관찰되지 않았다. 이는 이진 분류 환경인 IMDB 도메인에서 프롬프트 유형이나 모델에 관계없이 설명 가능성이 매우 안정적으로 유지됨을 의미한다.

Table 5. Quantitative XAI Comparison Analysis of GPT-4o-mini and Gemini 2.5 Flash in the IMDB Domain

Metric	GPT-4o-mini (IMDB)	Gemini 2.5 Flash (IMDB)	Difference (Gemini − GPT)
Match (LIME = True)	100.0%	100.0%	0% (Identical)
Match (Cosine = True)	100.0%	100.0%	0% (Identical)
LIME Improvement (Δ%)	0.0%	0.0%	0 p (No Change)
Overall Agreement (%)	100.0%	100.0%	0 p (No Change)

Fig. 6은 Tweet 도메인에서 GPT-4o-mini 모델에 format-constrained 프롬프트를 적용한 대표 사례에 대한 LIME 기반 감성 분석 시각화를 나타낸다. 시각화 결과, “listening”, “been”, “think”와 같은 상황 서술 중심의 표현이 중립 감성 판단에 주요 근거로 강조되며, 색상 그라데이션을 통해 특정 감성 방향으로의 강한 편향 없이 중립 판단이 형성됨을 확인할 수 있다. 이는 짧고 비정형적인 트윗 문장에서 모델이 명시적 감성어보다는 문맥적 단서를 기반으로 감성을 판단함을 보여준다.

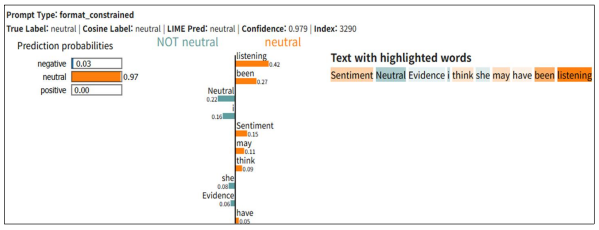


Fig. 6. LIME-based visualization of sentiment analysis for the format-constrained prompt using the GPT-4o-mini model on the tweet domain

Table 6은 분석 결과, Gemini 2.5 Flash 모델은 LIME 기반 설명 일치율과 코사인 유사도 기반 일치율에서 각각 GPT-4o-mini 대비 향상된 성능을 보였으며, 전체 합의율 역시 70.8%로 상대적으로 높게 나타났다. 이는 비정형, 단문 구조의 Tweet 도메인에서 Gemini 2.5 Flash가 감성 판단에 기여하는 핵심 단서를 보다 일관되게 설명함을 의미한다.

Table 6. Quantitative XAI Comparison Analysis of GPT-4o-mini and Gemini 2.5 Flash in the Tweet Domain

Metric	GPT-4o-mini (IMDB)	Gemini 2.5 Flash (IMDB)	Difference (Gemini − GPT)
Match (LIME = True)	33.3%	58.3%	+25.0% ↑ (Improved)
Match (Cosine = True)	58.3%	75.0%	+16.7% ↑ (Improved)
LIME Improvement (Δ%)	−25.0%	−16.7%	+8.3 p ↑ (Improved)
Overall Agreement (%)	58.3%	70.8%	+12.5 p ↑ (Improved)

종합적으로 이진 분류(IMDB) 환경에서는 프롬프트 유형이나 모델에 관계없이 설명 신뢰도가 안정적으로 유지된 반면, 삼중 분류(APP, Tweet) 환경에서는 프롬프트 유형과 모델에 따라 설명 일관성의 차이가 뚜렷하게 나타났다.

V. Conclusion

본 연구에서는 프롬프트 유형에 따른 파운데이션 모델의 감성 분석 성능과 LIME 기반 설명 가능성을 비교 분석하였다. 한국어 APP 리뷰, IMDB 영화 리뷰, TweetEval 데이터셋을 대상으로 역할 기반, 문맥 강화, 소수 예제, 형식 제한 프롬프트를 적용하고, GPT-4o-mini와 Gemini 2.5 Flash 모델을 API 기반 동일 환경에서 평가하였다.

실험 결과, 이진 분류(IMDB)에서는 프롬프트 유형이나 모델 차이에 따른 성능 차이가 1.5%p 이내였고 설명 일관성 차이도 0%p로 나타나 전반적으로 안정적인 예측과 해석이 가능함을 확인하였다. 반면, 삼중 분류 환경(APP 리뷰, TweetEval)에서는 최대 약 3.0%p의 성능 차이가 나타났으며, 특히 트윗 도메인에서는 설명 일관성 차이가 LIME 기준 25.0%p, 전체 합의율 기준 12.5%p로 확인되었다. 또한 중립 클래스가 포함된 경우 의미적 경계의 모호성이 성능 저하와 설명 불안정성의 주요 요인으로 작용하였다. 이러한 성능 및 설명 일관성 차이는 단순 모델 또는 도메인 차이보다는 언어 구조, 형태소 복잡성, 중립 감성 표현 방식 등 언어적 특성의 영향이 함께 작용한 결과로 해석된다. 한편, LIME 기반 분석을 통해 감성 분석 성

능이 높더라도 설명의 신뢰도가 항상 보장되지는 않으며, 다중 클래스 환경에서는 프롬프트 설계와 모델 선택이 설명 가능성에 중요한 영향을 미침을 확인하였다.

본 연구는 실제 API 기반 실행 환경에서 프롬프트 설계가 LLM 감성 분석 성능과 해석 가능성에 미치는 영향을 통합적으로 분석하였다는 점에서 의의를 가지며, 모델 선택과 프롬프트 전략 수립을 위한 실증적 근거를 제시하였다. 또한 각 원본 데이터셋이 대규모 데이터임에도 불구하고, API 응답 지연 및 실험 재현성을 고려하여 클래스 균형 기반 샘플링 평가셋을 구성함으로써 모델 간 비교의 공정성과 실험 안정성을 확보하였다. 이러한 샘플 기반 평가에서도 모델 및 프롬프트 유형에 따른 성능 및 설명 일관성의 구조적 차이가 명확히 확인되었다는 점에서 본 연구 결과는 유의미한 비교 근거를 제공한다.

다만 본 연구는 제한된 모델과 샘플 기반 평가셋을 대상으로 수행되었다는 한계를 가진다. 향후 연구에서는 병렬 API 호출이 가능한 확장 실험 환경을 활용하여 대규모 데이터셋에 대한 검증을 수행하고, 인간 평가 기반의 설명 신뢰도 분석을 병행함으로써 LLM 기반 감성 분석의 설명 가능성을 보다 정교하게 평가할 필요가 있다. 특히 한국어 데이터에서 상대적으로 정확도가 낮게 나타난 샘플에 대한 심층 오류 분석을 통해, 언어적 특성과 프롬프트 설계가 성능에 미치는 영향을 구체적으로 규명하고 개선 방향을 도출할 필요가 있다. 또한 최신 파운데이션 모델 및 다양한 언어 도메인을 포함한 후속 비교 연구를 통해, 프롬프트 설계와 성능효율성 및 설명 가능성 간의 상호 관계를 체계적으로 규명할 수 있을 것으로 기대된다.

Appendix A. Representative prompt structure (tweet domain)

본 연구에서 사용한 프롬프트는 재현 가능성을 위해 대표 도메인(Tweet 3-class)의 구조만 요약하여 제시한다. 실제 API 실행 코드와 자동화 스크립트는 연구 환경에 종속적인 구현 요소이므로 본 논문에는 포함하지 않았다.

A-1 Role-based

Role: English Twitter sentiment classifier
Labels: Positive / Negative / Neutral
Mixed emotion rule: dominant tone selection
Severe issue priority → Negative
Output format:
Sentiment=<label>;Evidence=<quoted phrase>

A-2 Context-rich

Context-sensitive interpretation
Neutral only if no emotional tone
Mixed rule & severe issue rule applied
Same constrained output format

A-3 Few-shot

3-5 labeled tweet examples
Dominant sentiment rule
Severe issue override rule
Same structured output format

A-4 Format-constrained

Explicit sentiment selection rules
Strict one-line output
Evidence must be quoted from input
No placeholders allowed

REFERENCES

[1] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chen, and J. Gao, “Deep learning-based text classification: A comprehensive review,” ACM Computing Surveys, vol. 54, no. 3, Article 62, pp. 1-40, 2021. DOI: 10.1145/3439726

[2] C. Sun, X. Qiu, Y. Xu, and X. Huang, “How to fine-tune BERT for text classification?,” in Proc. 18th China National Conf. Chinese Computational Linguistics (CCL 2019), pp. 194-206, Kunming, China, Oct. 2019. DOI: 10.1007/978-3-030-32381-3_16

[3] R. Bommasani et al., “On the opportunities and risks of foundation models,” arXiv preprint arXiv:2108.07258, 2022. DOI: 10.48550/arXiv.2108.07258

[4] T. B. Brown et al., “Language models are few-shot learners,” arXiv preprint arXiv:2005.14165, 2020. DOI: 10.48550/arXiv.2005.14165

[5] L. Zhao, J. Zhou, J. Cao, and W. Zhu, “EMSA: Explainable multilingual sentiment analysis models providing sentiment analysis across multiple languages,” PLOS ONE, vol. 20, no. 11, e0333508, Feb. 2025. DOI: 10.1371/journal.pone.0333508

[6] S. Stilwell and D. Inkpen, “Explainable prompt-based approaches for sentiment analysis of movie reviews,” in Proc. 37th Canadian Conf. Artificial Intelligence (Canadian AI 2024), Guelph, Ontario, Canada, May 2024.

[7] T. Thogesan, A. Nugaliyadde, and K. W. Wong, “Integration of explainable AI techniques with large language models for enhanced interpretability for sentiment analysis,” arXiv preprint arXiv:2503.11948, 2025. DOI: 10.48550/arXiv.2503.11948

[8] H. Rouzegar and M. Makrehchi, “The impact of role design in in-context learning for large language models,” arXiv preprint arXiv:2509.23501, 2025. DOI: 10.48550/arXiv.2509.23501

[9] A. Huttula, Advanced prompt engineering: Systematic approaches

to enhance LLM performance, Master's thesis, Jamk University of Applied Sciences, Jyväskylä, Finland, June 2025. Available: https://www.theseus.fi/handle/10024/896207

[10] A. Aqsa, A. Aslam, and M. Saeed, "Efficient prompt engineering: Techniques and trends for maximizing LLM output," in Proc. Int. Conf. Computer Science and Intelligent Systems (ICCSIS 2025), Malaysia, Apr. 2025. DOI: 10.5281/zenodo.15186123

[11] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in Proc. 2019 Conf. Empirical Methods in Natural Language Processing (EMNLP 2019), pp. 3982–3992, Hong Kong, Nov. 2019. DOI: 10.18653/v1/D19-1410

[12] M. Lee and H. Ahn, "Improvement of recommendation system using attribute-based opinion mining of online customer reviews," J. Korea Soc. Computer and Information, vol. 28, no. 12, pp. 259–266, Dec. 2023. DOI: 10.9708/jksci.2023.28.12.259

[13] A. Mathew, P. Amudha, and S. Sivakumari, "Deep learning techniques: An overview," in Proc. Adv. Mach. Learn. Technol. Appl., pp. 599–608, May 2020.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS 2017), vol. 30, pp. 5998–6008, 2017.

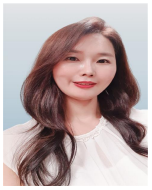
[15] A. Mohammed and R. Kora, "A comprehensive review on ensemble deep learning: Opportunities and challenges," J. King Saud Univ.-Comput. Inf. Sci., vol. 35, no. 2, pp. 757–774, Feb. 2023.

[16] 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," arXiv preprint arXiv:2106.09685, June 2021. DOI: 10.48550/arXiv.2106.09685

[17] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," Proc. 49th Annual Meeting of the Association for Computational Linguistics, pp. 142-150, 2011.

[18] F. Barbieri, J. Camacho-Collados, L. Espinosa-Anke, and L. Neves, "TweetEval: Unified benchmark and comparative evaluation for tweet classification," Proc. Findings of EMNLP 2020, pp. 1644-1650, 2020.

Authors



Misun Lee received a master's degree in engineering from the Graduate School of Business IT at Kookmin University in 2020 and is pursuing a doctoral degree in computer engineering at the Department of AI Convergence Engineering at Sejong University. She has taught at Kunsan National University and Dankook University and currently teaches at Hansung University and Soonchunhyang University. Her research interests include AGI, XAI, Agentic AI, quantum computing, recommendation systems, and natural language processing.