JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 31 No. 2, pp. 43-50, February 2026
https://doi.org/10.9708/jksci.2026.31.02.043

A Comparative Study on the Performance of GPT-4o-mini, Claude 4 Sonnet, and Gemini 2.5 Flash Models Using the Prompt Runner Framework

Misun Lee*

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

[Abstract]

This study presents a comparative analysis of three large language models (LLMs)—GPT-4o-mini, Claude 4 Sonnet, and Gemini 2.5 Flash—using a novel evaluation framework called Prompt Runner. The framework systematically measures the models' performance across nine linguistic and reasoning prompt types, totaling 90 items. Evaluation criteria include Accuracy, Consistency, Logic, Creativity, and Response Time. Accuracy was computed through Sentence-BERT-based cosine similarity, with Consistency and Logic derived by applying weight factors (0.95, 0.9 respectively). Creativity was assessed based on a weighted sum of Novelty, Diversity, and Fluency (0.5N + 0.3D + 0.2F). The analysis revealed that Claude 4 Sonnet demonstrated superior performance in logical reasoning (0.58) and creativity (0.44), while GPT-4o-mini exhibited faster response times. Gemini 2.5 Flash showed higher performance in accuracy (0.66) and consistency (0.62). Notably, Claude 4 Sonnet achieved the most stable and consistent performance in balancing overall capability and response time, thereby being evaluated as a model that effectively ensures both efficiency and quality. This study systematically identified the characteristics and performance differences of large language models (LLMs) across various prompt types by conducting a comparative analysis of quantitative performance indicators based on each model's API.

▸ **Key words**: Prompt Runner, LLM Evaluation, GPT-4o-mini, Claude 4 Sonnet, Gemini 2.5 Flash

[요 약]

본 연구는 Prompt Runner 프레임워크를 기반으로 GPT-4o-mini, Claude 4 Sonnet, Gemini 2.5 Flash 모델을 대상으로 프롬프트 유형별 성능을 비교하였다. 총 9개 유형, 90문항으로 구성된 프롬프트 데이터셋을 활용하여 정확도(Accuracy), 일관성(Consistency), 논리성(Logic), 창의성(Creativity), 응답 시간(Response Time)을 평가하였다. 정확도는 SBERT 임베딩 기반 코사인 유사도로 산출하였으며, 일관성과 논리성은 각각 정확도에 0.95, 0.9의 가중치를 적용하였다. 창의성은 새로움, 다양성, 유창성의 가중합(0.5N+0.3D+0.2F)으로 산출하였다. 분석 결과 Claude 4 Sonnet은 논리성(0.58)과 창의성(0.44)에서 우수한 성능을 보였으며, GPT-4o-mini는 빠른 응답시간을 나타냈고, Gemini 2.5 Flash는 정확도(0.66)와 일관성(0.62)에서 높은 성능을 보였다. 특히, Claude 4 Sonnet은 전반적인 성능과 응답 시간 간의 균형 측면에서 가장 안정적이고 일관된 성능을 보여, 효율성과 품질을 동시에 확보한 모델로 평가되었다. 본 연구를 통해 LLM 성능평가에서 각 AI 모델의 API 기반 정량적 성능지표를 비교 분석함으로써, 프롬프트 유형별 모델 특성과 성능 차이를 체계적으로 규명하였다.

▸ **주제어**: Prompt Runner, LLM 평가, GPT-4o-mini, Claude 4 Sonnet, Gemini 2.5 Flash

• First Author: Misun Lee, Corresponding Author: Misun Lee
*Misun Lee (llmss2000@hanmail.net), Department of AI Convergence Engineering, Sejong University
• Received: 2026. 01. 07, Revised: 2026. 01. 20, Accepted: 2026. 02. 01.

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

I. Introduction

최근 인공지능(AI)의 급격한 발전으로 대규모 언어모델(Large Language Model, LLM)이 다양한 언어 생성과 추론 과제에서 인간 수준의 성능을 보이고 있다[1][2].

GPT, Claude, Gemini와 같은 대표적인 LLM은 자연어 처리, 질의응답, 요약, 창의적 글쓰기 등 다양한 영역에서 활용되고 있으나, 프롬프트 유형(prompt type)에 따라 결과의 품질과 응답 특성이 상이하다는 점이 주요 한계로 지적된다[3][4].

최근 연구에서는 단순한 정확도 중심 평가에서 벗어나, 정확도(Accuracy), 일관성(Consistency), 논리성(Logicality), 창의성(Creativity) 등의 다차원적 평가 지표를 제안하고 있다[5][6].

특히 의미유사도 기반의 Sentence-BERT 임베딩 코사인 유사도를 활용하여 정확도를 측정하고, 논리성과 일관성을 정확도에 가중치를 부여하여 산출하는 방식을 제시하였다[7][8].

창의성 점수 C는 새로움(Novelty, N), 다양성(Diversity, D), 유창성(Fluency, F)을 통합한 가중합 지표로[9][10], 본 연구에서는 다음과 같은 가중치값을 정의하였다. C = 0.5N +0.3D + 0.2F 이때 N, D, F의 개념화와 다차원적 창의성 평가 정의는 유창성, 다양성(유연성),새로움(독창성)을 포함하는 다차원적 창의성 구성요인과 이를 통합 지수로 평가하는 접근을 보여주는 연구들[9][10]에 근거하였다.

그러나 대부분의 기존 연구는 단일 모델 중심의 정적 평가에 머물러 있으며, 여러 LLM을 동일 프롬프트로 비교하는 체계적인 분석은 부족하다.

이에 본 연구는 GPT-4o-mini, Claude 4 Sonnet, Gemini 2.5 Flash를 대상으로 Prompt Runner 프레임워크를 적용하여 첫 번째 프롬프트 유형별 응답 정확도, 일관성, 논리성, 창의성을 정량평가하고 두 번째 응답시간을 포함한 모델별 종합 성능을 비교 분석하였다.

본 논문의 구성은 다음과 같다. 2장에서는 LLM 성능평가와 프롬프트 엔지니어링 관련 선행연구를 검토하고, 3장에서는 Prompt Runner 기반 실험 알고리즘과 평가방법을 제시한다. 4장에서는 각 모델의 실험결과를 분석하고, 마지막 5장에서는 연구의 시사점과 향후 연구 방향을 제시한다.

II. Related Work

2.1 LLM Evaluation Approaches

대규모 언어모델(LLM)의 평가 연구는 초기의 정확도(Accuracy) 중심 접근에서 벗어나, 일관성(Consistency), 논리성(Logicality), 창의성(Creativity) 등을 포함한 다차원적 프레임워크로 발전하고 있다.

Chiang et al.(2024)은 LLM 평가에서 단일 점수 대신 다차원적, 다중지표 평가 프레임워크와 기존 지표들을 종합적으로 정리하였으며[5], Joshi(2025)는 Evaluation of Large Language Models: Review of Metrics, Applications, and Methodologies에서 정확도·일관성·논리성의 가중치 기반 통합 평가 메트릭(Weighted Metric Aggregation) 구조를 제시하였다[6].

Joshi(2025)는 여러 평가 지표를 가중합하는 통합 평가 프레임워크를 제안하였으며, 본 연구에서는 이를 참고하여 정확도, 일관성, 논리성 지표에 각각 1.0, 0.95, 0.9의 가중치를 부여하고, 다음과 같은 종합 평가식을 정의한다.

$$Score_{LLM} = 1.0A + 0.95C + 0.9L \qquad (1)$$

이 중 정확도 A는 LLM의 응답과 기준 정답 간 의미적 유사도를 산출하는 지표이며, 문장 간 의미적 근접도는 Sentence-BERT(SBERT) 임베딩 벡터의 코사인 유사도 식을 통해 계산된다[7][8].

$$(a,b) = \frac{a \bullet b}{\| a \| \ \| b \|} \qquad (2)$$

코사인 유사도는 평점을 벡터로 생각하고, 2개 벡터 사이의 각도를 계산하고, 그 각도가 적을수록 가까이 있다고 판단하기 때문에 서로 유사하다고 결정하는 방식이다. 코사인 유사도는 두 문장 임베딩 벡터의 각도 기반 의미 유사도를 정량화하는 방식으로, 값이 1에 가까울수록 문맥적 일치도가 높음을 의미한다.

Lee & Ahn(2023)은 이 공식을 협업 필터링 기반 사용자 유사도 계산에 적용하여 평가 정확도를 향상시킬 수 있음을 실험적으로 검증하였다[8].

Suprun et al.(2025)은 GPT-4o, Claude 4 Opus, Gemini 2.5 Pro 모델을 비교하여 각 모델의 논리적 일관성과 문체적 창의성의 상호 보완 관계를 실험적으로 검증하였다[11]. Claude 모델은 논리성과 창의성에서 우수하였으며, Gemini 모델은 정확도와 일관성에서 강점을 보였

다. 이 결과는 LLM의 성능을 단일 지표가 아닌 다차원적 평가 프레임워크로 통합해야 함을 시사한다.

본 연구에서는 창의성(C)을 정량적으로 평가하기 위해 새로움(Novelty), 다양성(Diversity), 유창성(Fluency)의 세 가지 구성요소[9][10]로 세분화하고, 각 요소의 상대적 중요도를 반영한 새로운 가중평균식을 정의하였다. 이를 통해 창의성 평가의 객관성과 분석의 정밀도를 높였다.

$$C = 0.5N + 0.3D + 0.2F \quad (3)$$

이 식은 창의성의 하위 구성요소를 정량적으로 조합하여 평가하는 방식으로, Joshi(2025)가 제시한 다중지표 평가 구조와 논리적으로 일치한다. 따라서 최근 LLM 평가 연구는 정확도, 일관성, 논리성, 창의성 등 다차원 평가체계(Multi-dimensional Evaluation Framework)로 발전하고 있다[6].

2.2 Prompt-based Assessment Frameworks

프롬프트 엔지니어링은 LLM의 평가 품질에 결정적인 영향을 미치는 요소로, 입력 구조의 설계 방식에 따라 응답의 정확도와 논리성이 달라진다. OpenAI(2023)의 GPT-4 Technical Report는 다양한 벤치마크와 사용자 프롬프트 코퍼스에서 few-shot, chain-of-thought 등 여러 프롬프트 설계에 따른 성능 차이를 분석하고, 지시 준수성 향상을 정량적으로 보고하였다[2].

Joshi(2025)는 엔트로피 기반 안정성 지표와 도메인별 메트릭 상관관계를 활용한 Adaptive Weight Assignment 알고리즘을 통해 복합 평가 파이프라인을 제안하였다[6]. 또한 프롬프트의 논리적 순서(logical prompt sequence)가 모델의 사실성(factuality) 평가에 영향을 미친다는 점을 실험적으로 입증하였다[12].

Prompt Runner Framework는 이러한 기존 연구의 평가 흐름을 통합하여 동일한 프롬프트 세트를 여러 모델에 병렬 적용하고, 각 모델의 정확도, 일관성, 논리성, 창의성 및 응답시간(Response Time)을 정량적으로 비교할 수 있도록 평가 알고리즘을 설계하였다.

III. Methodology

본 연구는 다양한 LLM의 프롬프트 유형별 성능을 체계적으로 비교하기 위해 Prompt Runner 프레임워크 기반의 다중 LLM 평가 알고리즘을 제안한다.

제안된 프레임워크는 프롬프트 데이터셋과 모델 응답을 입력으로 받아 정확도, 일관성, 논리성, 창의성, 응답시간을 평가하는 구조로 구성되며, 전체 알고리즘의 절차는 Fig. 1과 같다.

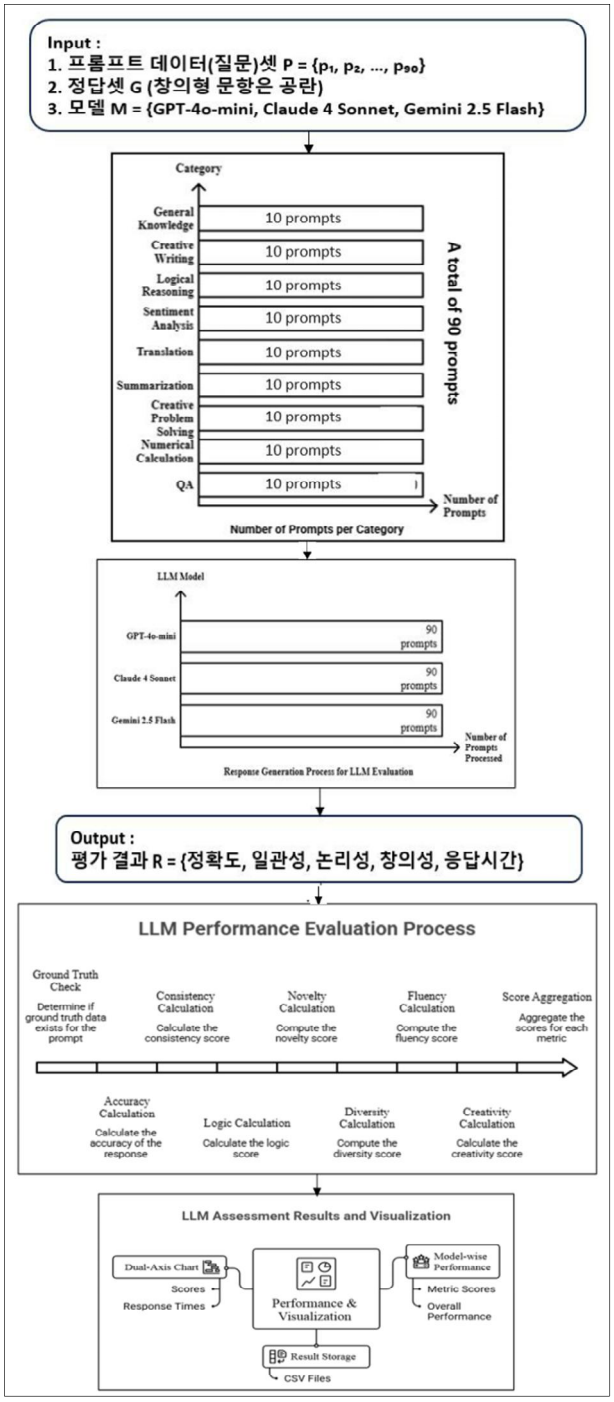


Fig. 1. Prompt runner-based multi-LLM evaluation algorithm

3.1 Data collection

본 연구에서 구축한 질문 프롬프트 데이터셋은 질문 복잡도와 인지 부하 수준을 반영하여 3단계 난이도 구조로

설계하였다. 1단계는 단일 사실 회상 기반의 일반 상식형 질문으로 구성되었으며, 2단계는 두 개 이상의 정보 결합 또는 단계적 논리 처리가 요구되는 다단계 추론형 질문으로 구성하였고, 3단계는 특정 조건, 제약 사항 또는 상황 맥락을 포함하여 응답 생성 시 추가적인 제약 처리가 필요한 조건 기반 질문 유형으로 구성하였다. 각 카테고리별 질문은 이러한 난이도 특성을 반영하여 다양성을 확보하도록 설계되었다.

정답 데이터셋은 질문 데이터셋의 객관적 평가 정보 검색 기반 지식 또는 일반 상식 추론을 통해 단일 정답 도출이 가능한 항목에 대해서만 Ground Truth 값을 정의하였고, 창의적 글쓰기(카테고리 2)와 창의 문제 해결(카테고리 7)의 경우 응답의 주관성 및 다양성으로 인해 단일 정답 정의가 어려운 특성이 있어, 해당 항목의 정답 필드는 공백 처리하여 창의성 지표를 구현하였다. 9개 카테고리는 1. 일반 상식, 2. 창의적 글쓰기, 3. 논리적 추론, 4. 감성 분석, 5. 번역 요청, 6. 요약 요청, 7. 창의 문제 해결, 8. 수치/계산형, 9. 질의 응답(지식 검색형)로 구성된 질문 프롬프트 10문항씩 총 90문항의 프롬프트 데이터셋을 구축한 이후, 질문에 대한 정답 데이터셋을 9개의 카테고리 중 2. 창의적 글쓰기와 7. 창의 문제 해결는 정답은 공백으로 처리하였다. 구축 프롬프트 각 카테고리별 질문을 바탕으로 GPT-4o-mini, Claude 4 Sonnet, Gemini 2.5 Flash 각 AI 모델별로 API를 통해 응답결과 데이터를 수집하였다.

	A	B	C
1	No	Category	Kor
2	1	일반 상식	대한민국의 수도는 어디인가요? 단답형으로 대답 및 부연설명 금지
3	2	일반 상식	태양은 어떤 종류의 별인가요? 단답형으로 대답 및 부연설명 금지
4	3	일반 상식	인간의 정상 체온은 몇 도인가요? 단답형으로 대답 및 부연설명 금지
5	4	일반 상식	지구의 위성은 무엇인가요? 단답형으로 대답 및 부연설명 금지
6	5	일반 상식	물은 몇 도에서 끓나요? 단답형으로 대답 및 부연설명 금지
7	6	일반 상식	대한민국 국화는 무엇인가요? 단답형으로 대답 및 부연설명 금지
8	7	일반 상식	인구가 가장 많은 대륙은 어디인가요? 단답형으로 대답 및 부연설명 금지
9	8	일반 상식	세계에서 가장 긴 강은 어디인가요? 단답형으로 대답 및 부연설명 금지
10	9	일반 상식	최초의 인공위성 이름은 무엇인가요? 단답형으로 대답 및 부연설명 금지
11	10	일반 상식	한글을 창제한 왕은 누구인가요? 단답형으로 대답 및 부연설명 금지
12	11	창의적 글쓰기	"가을의 향기"를 주제로 짧은 시를 써보세요. 500자 이내로 내용이 자연스럽고 완결되게 작성
13	12	창의적 글쓰기	인간과 인공지능이 대화하는 짧은 소설을 작성하세요. 500자 이내로 내용이 자연스럽고 완결
14	13	창의적 글쓰기	미래의 학교를 상상해 묘사해 보세요. 500자 이내로 내용이 자연스럽고 완결되게 작성
15	14	창의적 글쓰기	'시간 여행'을 소재로 한 짧은 스토리를 만드세요. 500자 이내로 내용이 자연스럽고 완결되게
16	15	창의적 글쓰기	꿈속의 세계를 묘사하는 문단을 써보세요. 500자 이내로 내용이 자연스럽고 완결되게 작성

Fig. 2. Prompt runner question dataset

	A	B	C	D	E
1	No	Category	Kor	GroundTruth	EvalType
2	1	일반 상식	대한민국의 수도는 어디인가요? 단답형으로 대답 및 부연설명 금지	서울	factual
3	2	일반 상식	태양은 어떤 종류의 별인가요? 단답형으로 대답 및 부연설명 금지	G형 주계열성	factual
4	3	일반 상식	인간의 정상 체온은 몇 도인가요? 단답형으로 대답 및 부연설명 금지	36.0℃ ~ 37.7℃	factual
5	4	일반 상식	지구의 위성은 무엇인가요? 단답형으로 대답 및 부연설명 금지	달	factual
6	5	일반 상식	물은 몇 도에서 끓나요? 단답형으로 대답 및 부연설명 금지	100℃	factual
7	6	일반 상식	대한민국 국화는 무엇인가요? 단답형으로 대답 및 부연설명 금지	무궁화	factual
8	7	일반 상식	인구가 가장 많은 대륙은 어디인가요? 단답형으로 대답 및 부연설명 금지	아시아	factual
9	8	일반 상식	세계에서 가장 긴 강은 어디인가요? 단답형으로 대답 및 부연설명 금지	아마존강	factual
10	9	일반 상식	최초의 인공위성 이름은 무엇인가요? 단답형으로 대답 및 부연설명 금지	스푸트니크 1호(Sputnik 1)	factual
11	10	일반 상식	한글을 창제한 왕은 누구인가요? 단답형으로 대답 및 부연설명 금지	세종대왕	factual
12	11	창의적 글쓰기	"가을의 향기"를 주제로 짧은 시를 써보세요. 500자 이내로 내용이 자연스럽고 완결되게 작성		creative
13	12	창의적 글쓰기	인간과 인공지능이 대화하는 짧은 소설을 작성하세요. 500자 이내로 내용이 자연스럽고 완결되게 작성		creative

Fig. 3. Prompt runner answer dataset

위 Fig. 2는 Prompt runner 사용자 질문 데이터셋이고, Fig. 3은 Prompt runner 사용자 정답 데이터셋으로 총 9개의 카테고리 중 창의성에 관한 질문 카테고리 '창의적 글쓰기', '창의 문제 해결'에 대한 정답은 공백으로 처리 하였다. 그리고 Fig. 4는 각 모델별 API를 통한 Prompt runner 질문에 대한 응답결과 중 대표로 GPT-4o-mini에 대한 Prompt runner 결과를 보여주고 있다. 위 결과에는 각 AI 모델별 각 카테고리에 대한 질문 프롬프트 데이터셋에 대한 응답결과와 응답시간과 수행 상태를 도출할 수 있도록 각 AI 모델별로 Prompt runner 구축하였다.

결과 미리보기 (상위 20건)

No	Category	Model	Prompt	Response	Time	Status
1	일반 상식	GPT-4o-mini	대한민국의 수도는 어디인가요? 단답형으로 대답 및 부연설명 금지	서울입니다.	2.165	Success
2	일반 상식	GPT-4o-mini	태양은 어떤 종류의 별인가요? 단답형으로 대답 및 부연설명 금지	G형 주계열성 (G-type main-sequence star)입니다.	2.160	Success
3	일반 상식	GPT-4o-mini	인간의 정상 체온은 몇 도인가요? 단답형으로 대답 및 부연설명 금지	36.5도에서 37.5도 사이입니다.	0.791	Success
4	일반 상식	GPT-4o-mini	지구의 위성은 무엇인가요? 단답형으로 대답 및 부연설명 금지	달입니다.	1.323	Success
5	일반 상식	GPT-4o-mini	물은 몇 도에서 끓나요? 단답형으로 대답 및 부연설명 금지	100도 섭씨입니다.	1.275	Success
6	일반 상식	GPT-4o-mini	대한민국 국화는 무엇인가요? 단답형으로 대답 및 부연설명 금지	무궁화입니다.	0.564	Success

Fig. 4. GPT-4o-mini response results from the prompt runner API

3.2 Prompt runner evaluation framework

Table 1은 본 연구에서 제안한 Prompt Runner 기반 다중 LLM 평가 프레임워크의 단계별 절차를 요약한 것으로 실험은 동일한 프롬프트 데이터셋(총 90문항, 9개 범주)을 기반으로 수행되었으며, 각 프롬프트는 세 가지 모델(GPT-4o-mini, Claude 4 Sonnet, Gemini 2.5 Flash)에 동일하게 적용하였다. 모든 API 응답결과는 CSV 파일로 저장되었으며, 평균 응답시간도 함께 도출하였다.

Table 1. Step-by-Step Process of the Prompt Runner Evaluation Framework

Step	Implementation method
Step 1	Prompt dataset preparation
Step 2	API response generation
Step 3	Response analysis and scoring (A Accuracy, C Consistency, L Logicality)
Step 4	Creativity metric computation (N Novelty, D Diversity, F Fluency)
Step 5	Performance aggregation and visualization

Prompt Runner API 저장된 응답 결과와 사용자 구축한 정답셋을 비교하여 정확도(Accuracy), 일관성(Consistency), 논리성(Logic) 을 평가하고, 식 (1)의 가중합식을 적용하여 모델별 종합 성능 점수를 산출하였다.

또한, 창의성(Creativity)은 창의적 글쓰기 및 창의 문제해결 두 개의 카테고리에서 새로움(Novelty), 다양성(Diversity), 유창성(Fluency)을 기준으로 정량성능평가를 위해 식 (3)에 따라 가중평균으로 계산하였다.

창의성 지표의 가중치는 새로움(N)에 0.5, 다양성(D)에 0.3, 유창성(F)에 0.2를 부여하여 창의성의 핵심적인 분별력을 높여 모델 간의 성능을 극대화 시키고, 마지막으로 각 모델의 종합 평가 결과를 정량지표와 시각화로 모델별 성능 차이를 명확하게 구현하였다.

IV. Results and analysis

4.1 Experimental design

본 연구에서는 Prompt Runner 프레임워크를 기반으로 하여, 세 가지 모델(GPT-4o-mini, Claude 4 Sonnet, Gemini 2.5 Flash)의 성능을 정량적으로 비교분석하였다. 본 연구 수행 시점에는 GPT-5.1, Gemini 3.0 등 최신 모델의 API 접근이 제한적이거나 비공개 상태였기 때문에 접근성이 확보되고 실제 서비스 환경에서 활용도가 높고 안정적인 API 접근이 가능한 경량 및 중간급 모델 계열을 대상으로 실험을 설계하였다. GPT-4o-mini와 Gemini 2.5 Flash는 저지연 응답을 중심으로 설계된 경량 모델이며, Claude 4 Sonnet은 상대적으로 향상된 추론 성능을 제공하는 중간급 모델로서 모두 실시간 대화형 응용 환경에서 활용이 가능하다는 공통점을 같는다. 이러한 모델선정은 동일한 실험 환경에서 반복 측정이 가능하도록 하여 실험 조건의 일관성을 유지하고, 연구 결과의 공정성, 재현성 및 실험 안정성을 확보하기 위한 목적에 기반하였다. 또한 모든 실험은 Google Colab CPU 환경에서 수행되었었으며, 프롬프트 데이터셋 구축, API 응답 수집, 정답셋 비교분석, 창의성 평가, 종합 성능 시각화의 다섯 단계로 수행되었다.

Step 1에서는 총 90문항으로 구성된 프롬프트 데이터셋(9개 범주)을 CSV 형식으로 구축하였다. 각 문항은 '일반 상식', '논리적 추론', '창의적 글쓰기', '창의 문제 해결' 등의 9개의 카테고리로 분류되었으며, Prompt Runner를 통해 각 모델의 입력 데이터로 사용되었다.

Step 2에서는 각 프롬프트를 세 모델의 공식 API(OpenAI, Anthropic, Google)를 통해 호출하여 모델별 응답(Response)과 응답시간(Response Time)을 수집하였다. 수집된 응답은 CSV 파일로 저장되었으며, 평균 응답시간과 성공률을 계산하였다. 그 결과는 Table 2에 요약되어 있다.

Step 3에서는 수집된 응답을 사전에 작성한 사용자 정답셋(Ground Truth)과 비교하여 정확도(Accuracy), 일관성(Consistency), 논리성(Logic)을 산출하였다. Sentence-BERT(all-MiniLM-L6-v2) 임베딩 모델을 이용하여 응답과 정답 간의 코사인 유사도(Cosine Similarity)를 계산하였으며, 일관성과 논리성은 정확도의 비율에서 식(1)을 적용하여 산출하였다.

Step 4에서는 창의성(Creativity) 평가를 별도로 수행하였으며, 창의적 카테고리(창의적 글쓰기, 창의 문제 해결)를 대상으로 답의 새로움(Novelty), 다양성(Diversity), 유창성(Fluency)을 계산하였으며, 창의성 점수는 식 (3)의 가중평균식을 적용하여 산출하였다. 창의성 세부 분석 및 모델별 성능평가 결과는 Table 3에 제시하였다.

Step 5에서는 모든 세부 지표(정확도, 일관성, 논리성, 창의성, 응답시간)를 통합하여 모델별로 각각 종합 성능을 비교분석하였다. 각 모델별 비교분석 결과는 Table 4와 Fig. 5에 제시하였으며, 또한 응답 시간은 모델별 질문에 대한 응답결과 산출한 평균 ± 표준편차로 제시하였다. 이를 통해 각 모델의 응답 특성과 성능 차이를 시각적으로 비교하였다.

4.2 Experiment result

본 연구의 실험 결과는 Table 2-4에 제시되어 있다. 세 모델은 동일한 90개 프롬프트를 기반으로 평가되었으며, 응답시간(Response Time), 정확도(Accuracy), 일관성(Consistency), 논리성(Logic), 창의성(Creativity)을 핵심 지표로 분석하였다.

Table 2. Average Response Time Results by Model

API Model	GPT-4o-mini	Claude 4 Sonnet	Gemini 2.5 Flash
Total Questions	90	90	90
Successful Responses	90	90	90
Failed Responses	0	0	0
Success Rate (%)	100.0	100.0	100.0
Average Response Time (s)	1.975	4.623	6.224

Table 2는 각 모델의 평균 응답시간과 성공률을 보여준다. 모든 모델이 100%의 성공률을 기록하여 모든 프롬프트에 정상적으로 응답하였으며, GPT-4o-mini가 평균 1.98초로 가장 빠른 응답속도를 보였다. Claude 4 Sonnet은 4.62초, Gemini 2.5 Flash는 6.22초로 응답시간이 순차적으로 증가하였다. 이 결과는 모델별 연산 특성에 따른 응답 효율성의 차이를 나타내고 있다. 또한 응답 결과에 대한 질문 프롬프트의 일반상식 카테고리에서 "세계에서 가장 긴 강은 어디인가요? 단답형으로 대답 및 부연설명 금지" 질문 프롬프트에서 Gemini 2.5 Flash와 Claude 4 Sonnet는 "나일강"이라고 응답했고 GPT-4o-mini만 "아마존강"이라고 응답하였고, 또한 논리적 추론 카테고리 "3보다 큰 수는 4다. 4보다 큰 수는 5다. 현재 예시에서 3보다 큰 수는? 50자 이내로 작성"에서 Claude 4 Sonnet에서는 "주어진 예시에서 3보다 큰 수는 4와 5입니다.", Gemini 2.5 Flash는 "4", GPT-4o-mini는 "3보다 큰 수는 4, 5, 6, 7, ... 등 무한히 많습니다." 이라고 응답했다. 위 예시와 같이 같은 질문에서 모델마다 다르게 응답하는 결과 확인으로 모델마다 각기 다른 성능의 차이가 있음을 확인 할 수 있다.

Table 3은 정확도, 일관성, 논리성, 창의성의 네 가지 주요 성능지표를 비교한 결과이다. Gemini 2.5 Flash는 정확도(0.66)와 일관성(0.62)에서 가장 높은 값을 기록하였고, Claude 4 Sonnet은 논리성(0.58)과 창의성(0.44)에서 우수한 성능을 보였다. GPT-4o-mini는 전반적으로 낮은 점수를 보였으나, 응답시간 측면에서 가장 효율적인 모델로 평가되었다.

Table 3. Model Performance Evaluation Results

Model	Time	Accuracy	Consistency	Logicality	Average
GPT-4o-mini	1.975	0.609	0.579	0.548	0.578
Claude 4 Sonnet	4.623	0.647	0.614	0.582	0.615
Gemini 2.5 Flash	6.224	0.655	0.622	0.590	0.622
Model	Novelty	Diversity	Fluency	Creativity Score	
GPT-4o-mini	0.070	0.868	0.466	0.388	
Claude 4 Sonnet	0.101	0.933	0.523	0.435	
Gemini 2.5 Flash	0.075	0.932	0.488	0.414	

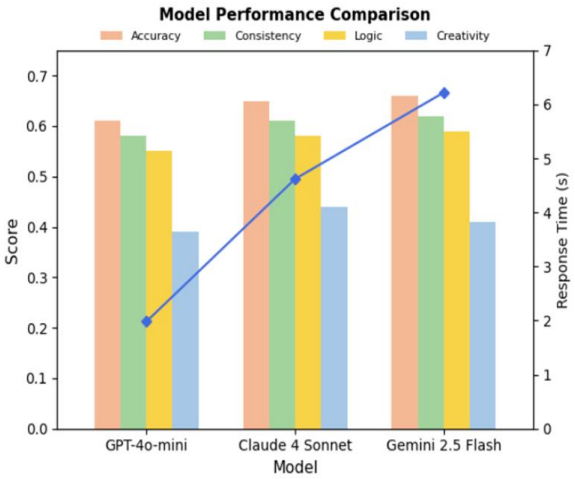


Fig. 5. Comparative analysis of model performance and response time

Fig. 5은 위 지표들을 종합하여 시각화한 결과로, 막대그래프는 네 가지 성능지표(Accuracy, Consistency, Logic, Creativity)이고, 꺾은선그래프는 응답시간을 나타낸다. 이를 통해 각 모델의 성능과 응답속도의 상대적 차이를 직관적으로 확인할 수 있다. GPT-4o-mini는 가장 빠른 응답속도를 나타냈고, Claude 4 Sonnet은 성능과 응답시간의 균형 면에서 가장 안정적인 결과를 보였으며, Gemini 2.5 Flash는 정확도와 일관성에서 강점을 보였다.

Table 4. Experimental Results of Model Performance

Model	Response Time (Mean ± SD, s)	Accuracy	Consistency	Logic	Creativity Score
GPT-4o-mini	1.98 ± 2.45	0.61	0.58	0.55	0.39
Claude 4 Sonnet	4.62 ± 3.93	0.65	0.61	0.58	0.44
Gemini 2.5 Flash	6.22 ± 8.47	0.66	0.62	0.59	0.41

Table 4의 종합 비교 결과, Claude 4 Sonnet이 논리성과 창의성 측면에서 두드러진 성능을 보이며 전반적인 평가에서 가장 우수한 모델로 분석되었다. 이는 Prompt Runner 프레임워크가 LLM의 다양한 성능 특성을 정량적으로 비교 평가하는 데 효과적인 접근임을 입증한다.

V. Conclusion

1. Academic, practical implications

본 연구의 학술적 시사점은 다음과 같다.

첫째, 본 연구는 기존의 단일 정확도 중심 평가에서 벗어나, 정확도(Accuracy), 일관성(Consistency), 논리성(Logic), 창의성(Creativity), 응답시간(Response Time)을 포함한 다차원적 LLM 평가 프레임워크를 제안하였다. 이를 통해 기존 연구에서 간과된 정량적 성능을 통합적으로 비교분석할 수 있는 체계적 방법론을 제시하였다.

둘째, 제안된 Prompt Runner 프레임워크는 실제 API 기반 환경에서 자동화된 성능 측정이 가능하다는 점에서 재현성과 실험 안정성을 확보하였다. 특히 동일한 프롬프트 세트를 기반으로 다양한 LLM의 응답 특성을 정량적으로 비교함으로써 LLM 간 상대적 성능 차이를 명확히 규명하였다.

셋째, Claude 4 Sonnet, GPT-4o-mini, Gemini 2.5 Flash의 비교를 통해 모델별 특성과 강점을 구체적으로 제시하였다. Claude 4 Sonnet은 논리성과 창의성에서, Gemini 2.5 Flash는 정확도와 일관성에서, GPT-4o-mini는 응답속도에서 우수한 성능을 보였다. 이러한 결과는 LLM 선택 시 활용 목적에 따른 최적 모델 선정의 근거를 제공한다.

본 연구의 실무적 시사점은 다음과 같다.

첫째, 제안된 평가 체계는 기업이나 연구기관에서 다수의 LLM을 비교검증 시 활용 가능한 표준화된 평가 기준으로 적용될 수 있다.

둘째, 창의성 지표를 정량적으로 정의하여 모델의 언어 생성 품질을 객관적으로 평가함으로써, 교육, 콘텐츠 제작, 마케팅 등 창의적 응용 분야에서 LLM의 효율적 활용 방향을 제시하였다.

셋째, 본 연구의 평가 프로세스는 다양한 API 기반 모델에 확장 적용이 가능하여, AI 모델의 품질관리(QA)와 서비스 수준 검증(SLA) 지표로 활용될 수 있다.

2. Limitations and future research

본 연구의 한계점은 다음과 같다.

첫째, 본 연구는 연구 수행 시점에서 공개된 모델(GPT-4o-mini, Claude 4 Sonnet, Gemini 2.5 Flash)만을 대상으로 비교하였다. 향후 연구에서는 GPT-5.1, Gemini 3.0 등 최신 모델이 공개될 경우, 동일한 프레임워크를 적용하여 성능을 재평가할 필요가 있다.

둘째, 본 연구는 API 응답 기반 정량 분석에 초점을 맞추었기 때문에, 모델의 언어적 다양성이나 맥락 이해력과 같은 정성적 평가 요소는 일부 반영되지 않았다. 향후 연구에서는 인간 평가자(Human Evaluator)의 주관적 평가와 자동화된 지표를 병행하여, 보다 정교한 하이브리드 평가체계를 구축할 필요가 있다.

셋째, 본 연구에서 사용한 프롬프트 데이터셋은 연구자가 직접 구성한 90문항으로 제한되어 있어, 특정 유형의 질문이나 표현 방식에 대한 편향이 포함되었을 가능성을 배제할 수 없다. 또한 문항 수가 상대적으로 제한적이므로, 본 연구 결과를 다양한 실제 활용 환경으로 일반화하는 데에는 한계가 존재한다. 향후 연구에서는 대규모 공개 벤치마크 데이터셋 및 다국어 프롬프트를 포함한 확장된 평가를 통해 결과의 일반화 가능성을 추가적으로 검증할 필요가 있다.

넷째, 본 연구에서 사용한 정확도, 일관성, 논리성 및 창의성 지표의 가중치는 선행연구를 참고하여 설정하였으나, 해당 값들은 하나의 실험적 가정에 기반한 것으로 최적 가중치에 대한 체계적인 검증은 수행되지 않았다. 또한 평균 성능 지표를 기반으로 모델 간 상대적 특성을 비교하였으나, 가중치 변화에 따른 결과 민감도 분석(sensitivity analysis) 및 통계적 유의성 검증이 포함되지 않아 절대적 성능 우열을 확정적으로 판단하는 데에는 한계가 있다. 향후 연구에서는 가중치 설정에 대한 명확한 추가 실험을 통해 평가 프레임워크의 안정성과 재현성을 보다 체계적으로 검증할 필요가 있다.

REFERENCES

[1] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, “LLaMA: Open and Efficient Foundation Language Models,” arXiv preprint arXiv:2302.13971, Cornell University, February 2023. DOI: 10.48550/arXiv.2302.13971

[2] OpenAI, J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Amanatides, et al., “GPT-4 Technical Report,” arXiv preprint arXiv:2303.08774, OpenAI, March 2023. DOI: 10.48550/arXiv.2303.08774

[3] Anthropic, “Introducing the Next Generation of Claude,” Anthropic Official Announcement, March 4, 2024. Available at: https://www.anthropic.com/news/claude-3-family

[4] K. Kavukcuoglu, “Gemini 2.5: Our Most Intelligent AI Model,” Google DeepMind Official Blog, March 25, 2025. Available at: https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/#gemini-2-5-thinking

[5] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, and Y. Wang, “A Survey on Evaluation of Large Language Models,” ACM Transactions on Intelligent Systems and Technology, Vol. 15, No. 3, Article No. 39, pp. 1–45, March 2024. DOI: 10.1145/3641289

[6] S. Joshi, “Evaluation of Large Language Models: Review of Metrics, Applications, and Methodologies,” Preprints.org, April 2025. DOI: 10.20944/preprints202504.0369.v1

[7] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks,” Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 3982-3992, Hong Kong, Nov. 2019. DOI: 10.18653/v1/D19-1410

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

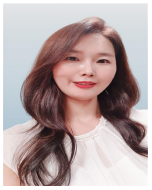
[9] Shehla Naz, Iftikhar Ahmed Baig, "Measurement of Creativity Index at Tween Level," Pakistan Social Sciences Review, Vol. 6, No. 2, pp. 392–400, Apr.–Jun. 2022. DOI: 10.35484/pssr.2022(6-II)34

[10] Joy Desdevies, "The paradox of creativity in generative AI: high performance, human-like bias, and limited differential evaluation," Frontiers in Psychology, Vol. 16, Article 1628486, Aug. 2025. DOI: 10.3389/fpsyg.2025.1628486

[11] A. Suprun, I. Tvoroshenko, V. Gorokhovatskyi, and O. Yakovleva, “Development and Research of a Method for the Combined Use of Large Language Models for Text Generation,” International Journal of Academic and Applied Research (IJAAR), Vol. 9, Issue 10, pp. 249–263, October 2025. ISSN: 2643-9603

[12] Z. Xie, "Order Matters in Hallucination: Reasoning Order as Benchmark and Reflexive Prompting for Large-Language-Models," arXiv preprint arXiv:2408.05093, August 2024. DOI: 10.48550/arXiv.2408.05093

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 is currently serving as a Part-Time Lecturer with Dankook University, Hansung University, and Soonchunhyang University. Her research interests include AGI, XAI, Agentic AI, quantum computing, recommendation systems, and natural language processing.