AI 응답 일관성에 영향을 미치는 프롬프트 요인 분석 보고서

12202039 안정민 12202049 이찬형
12202051 장현우 12224270 전수빈

1. 서론: AI 응답 일관성의 중요성

인공지능(AI) 기술의 발전은 사회 전반에 걸쳐 혁신적인 변화를 가져오고 있으며, 특히 의료, 금융, 교육과 같은 핵심 분야에서 AI는 정보 전달과 의사결정 지원의 중추적인 역할을 수행한다. 이러한 AI의 광범위한 적용은 신뢰성과 효율성에 대한 요구를 증대시키며, '응답의 일관성(consistency)'은 AI 시스템의 핵심 품질 지표로 부상하고 있다.

1.1 AI의 사회적 역할 확대와 일관성의 필수성

AI가 고위험 영역에 깊이 관여함에 따라, 응답의 일관성은 단순한 기술적 우수성을 넘어 사회적 수용과 신뢰 구축의 근본적인 전제 조건이 된다. 일관성 없는 응답은 사용자에게 혼란을 야기하고, 민감한 분야에서는 심각한 의사결정 오류로 이어질 수 있어, AI의 광범위한 채택을 저해하는 요인으로 작용할 수 있다. 따라서 일관성은 AI의 윤리적 배포와 대중적 수용을 위한 필수적인 전제 조건이다.

1.2 일관성 결여로 인한 위험성

AI 응답의 일관성 부족은 사용자 신뢰 형성과 소통 안정성에 결정적인 영향을 미친다. 동일한 질문에 대해 상이한 응답을 제공할 경우, 사용자들은 혼란을 겪고 정보의 신뢰성에 의문을 제기하게 된다. 특히 법률 자문이나 교육 AI와 같이 민감한 분야에서는 일관성 결여가 직접적인 피해나 학습 효과 저해로 이어질 수 있다.

1.3 일관성 향상을 위한 프롬프트 최적화의 필요성

AI가 높은 응답 일관성을 유지하기 위해서는 '질문 구성 방식(prompt design)'에 대한 체계적인 분석이 필수적이다. 프롬프트의 표현 방식에 따라 AI의 응답이 달라질 수 있으므로, 정확하고 신뢰할 수 있는 정보를 제공하기 위해서는 프롬프트를 실험적으로 검증하고 최적화하는 연구가 요구된다. 감정적 프레이밍, 명시성, 문장 톤 등 다양한 요소들이 응답의 방향성과 일관성에 실질적인 영향을 미친다.

1.4 선행 연구: DeepMind의 OPRO와 프롬프트 설계의 패러다임 전환

Google DeepMind의 OPRO 연구는 프롬프트가 LLM 응답 품질과 일관성에 결정적이라는 걸 보여줬다. 이 연구는 저품질 프롬프트가 안 좋은 결과를 내고, 반복적인 최적화가 사람보다 낫다는 것을 증명했다. 즉, 프롬프트 설계가 LLM 성능과 신뢰성을 좌우하는 핵심 요소이며, 프롬프트 엔지니어링을 '예술'에서 '과학'으로 바꾼 계기가 되었다.

1.5 프로젝트 목표

본 프로젝트는 프롬프트 요소가 AI 응답 일관성에 미치는 영향을 분석하고 최적의 프롬프트 설계 방안을 제안하여, AI 시스템의 신뢰성과 예측 가능성 향상, 의사결정 과정의 오류 최소화 및 효율성 극대화, 그리고 프롬프트 엔지니어링의 과학적 기반 강화를 목표로 한다.

---

2. 데이터 생성 및 일관성 측정

본 프로젝트는 AI 응답의 일관성을 정량적으로 평가하기 위해 체계적인 데이터 생성 과정을 수립하고, LLM 출력의 미묘한 의미적 차이를 포착할 수 있는 측정 지표를 도입하였다.

2.1 데이터 생성 과정

본 프로젝트에 사용된 모든 실험 데이터는 직접 생성되었다. 이는 실험 설계의 목적에 맞춰 데이터를 엄격하게 제어하고 맞춤화하기 위함이다. 데이터 생성 과정은 다음과 같은 단계로 진행되었다.

1. 카테고리 정의 및 질문 분할: 실험 설계의 블로킹(Blocking) 요인으로 활용하기 위해 질문 카테고리를 '인성(윤리, 정의 포함)', '창의성', '논리적 추론'의 세 가지로 정의하고 분할하였다.
2. 기본 질문**(Base Question)** 생성: 각 카테고리별로 기본적인 질문들을 개발하였다. 기본 질문의 수는 각 실험에 필요한 데이터 크기에 따라 결정되었다. 예를 들어, 인성 카테고리에서는 "개인의 자유와 사회적 책임 중 무엇이 더 중요한가?", 창의성 카테고리에서는 "좁은 원룸을 넓어 보이게 하는 창의적 인테리어 방법은?", 논리적 추론 카테고리에서는 "아파트 구매 결정을 내릴 때의 논리적 판단 기준은?"과 같은 질문들이 포함되었다.
   기본 질문 데이터 링크: https://github.com/statjhw/DOE/blob/master/rcbd/experiment_design.py
3. **GPT API**를 통한 프롬프트 생성: GPT API를 활용하여 각 실험의 요인 조합에 맞는 프롬프트를 생성하였다. 생성된 프롬프트는 각 요인 조합에 적합한지 정성적인 평가를 거쳤다. <2.2에서 자세한 설명>
4. 응답 생성 및 수집: 생성된 각 프롬프트에 대해 GPT API를 5회 반복 호출하여 여러 응답을 생성하고 수집하였다.
5. 응답 전처리 및 정제: 수집된 응답들은 분석을 위해 필요한 전처리 및 정제 과정을 거쳤다.
6. 의미적 유사도 측정 및 평균 계산: 생성된 하나의 질문(프롬프트)에 대해 5회 반복 생성된 응답들 간의 의미적 유사도를 측정하고, 그 유사도의 평균을 계산하였다. 이 평균 유사도 값은 실험의 주요 반응 변수(bert_multilingual_similarity)로 활용되었다. <2.3에서 자세한 설명>

2.2 질문 생성을 위한 프롬프트

2.2.1 RCDB 요인에 맞는 질문 데이터 생성

RCDB 질문 생성기 코드 링크 : https://github.com/statjhw/DOE/blob/master/rcbd/prompt_generator.py

```
[변형 규칙]
1. 질문의 핵심 주제와 의미는 유지해야 함
2. {framing_level} 프레이밍의 특징을 명확히 반영해야 함
3. 한 문장으로 작성 (최대 2문장 허용)
4. 자연스럽고 현실적인 표현 사용
5. 따옴표, 번호, 부가 설명 없이 변형된 질문만 출력

[주의사항]
- 중립적: 감정 없이 객관적으로
- 정서적: 온건한 감정과 개인적 관심 포함
- 자극적: 강한 감정, 비판, 위기감 표현
"""

user_prompt = f"""
다음 질문을 '{framing_level}' 프레이밍 수준에 맞게 변형해주세요.

원본 질문: {base_question}

{framing_level} 프레이밍의 특징을 반영한 변형된 질문을 한 줄로 출력해주세요.
"""
```

<질문 생성 프롬프트>

```
# 프레이밍 정의
FRAMING_DEFINITIONS = {
    "중립적": {
        "description": "감정 유도 없이 사실 전달 중심",
        "characteristics": [
            "객관적이고 중립적인 어조",
            "감정적 단어나 표현 배제",
            "단순하고 직접적인 질문 형태",
            "판단이나 평가 없이 정보 요청"
        ]
    },
    "정서적": {
        "description": "긍정/부정의 온건한 감정 포함",
        "characteristics": [
            "공감이나 이해를 구하는 어조",
            "개인적 관심이나 걱정 표현",
            "온건한 감정적 언어 사용",
            "상황에 대한 개인적 반응 포함"
        ]
    },
    "자극적": {
        "description": "강한 비판, 갈등, 위협 등 포함",
        "characteristics": [
            "강한 감정이나 비판적 어조",
            "극단적이거나 도발적인 표현",
            "갈등이나 대립 상황 암시",
            "긴급성이나 위기감 조성"
        ]
    }
```

<각 감정 프레이밍 별 특징>

<프레이밍 수준 특징 및 생성 질문 예시>

프레이밍 수준	설명	특징	“새로운 인관관계를 형성하는 방법” 기본 질문의 생성 프롬프트
중립적	감정 유도 없이 사실 전달 중심	객관적, 중립적, 단순, 직접적, 정보 요청	새로운 인관관계를 형성하는 방법에는 어떤 것들이 있습니다?
정서적	긍정/부정의 온건한 감정 포함	공감, 이해, 개인적 관심/걱정, 온건한 감정적 언어	새로운 인관관계를 형성하는 혁신적인 방법에 대해 생각해보면, 우리가 서로를 더 잘 이해하고 공감하는 방식에는 무엇이 있을까요?
자극적	강한 비판, 갈등, 위협 등 포함	강한 감정/비판, 극단적/도발적, 갈등/대립, 긴급성/위기감	이제껏 시도하지 않았던 혁신적 방식으로 인간관계를 맺지 않으면, 당신은 고립되고 외면당할지도 모릅니다. 어떻게 해야할까요?

각 감정 프레이밍의 조건들을 명시하고 감정 프레이밍에 맞는 질문을 생성하도록 명시각 감정 프레이밍의 조건을 명시하고, 이에 적합한 질문을 생성하도록 지시하였다.

생성 질문 데이터 : https://github.com/statjhw/DOE/blob/master/rcbd_results/01_prompts_generated.json

2.2.2 2k factorial 요인에 맞는 질문 데이터 생성

2k factorial 질문 생성기 코드 링크: https://github.com/statjhw/DOE/blob/master/2kfactorial/prompt_generator.py

```
        # 역할 부여
        if factor_combination['role_assignment'] == 'with_role':
            role = self.category_roles[category]
            factor_descriptions.append(f"역할: 당신은 {role}입니다")

        # 명시성
        if factor_combination['explicitness'] == 'high':
            factor_descriptions.append("답변 방식: 구체적이고 명확한 예시와 함께 자세히 설명해 주세요")

        # GPT에게 한국어 프롬프트 생성 요청
        system_prompt = """당신은 실험용 프롬프트를 생성하는 전문가입니다.
주어진 조건들을 자연스럽게 통합하여 하나의 완전한 한국어 프롬프트를 만들어주세요.
반드시 한국어로만 작성하고, 마지막에 '한국어로 답변해 주세요.'를 추가해주세요.
프롬프트만 출력하고 다른 설명은 하지 마세요."""
```

<질문 생성 프롬프트 일부(역할, 명시성 요구)>

<각 요인 별 프롬프트 생성 방법>

요인	설명	프롬프트 생성 방법
모델	gpt 3.5-turbo, gpt 4o-mini	프롬프트 상관 x
언어	한국어, 영어	한국어로 생생된 프롬프트를 영어로 번역
역할	역할 부여o, 역할 부여 x	역할 부여o 생성 시 카테고리(인성, 창의성, 논리적 추론)에 대한 전문가 지정 ex) 당신은 “논리적 추론” 전문가 입니다.
명시성 요구	명시성 요구 high, 명시성 요구 low	명시성 요구 high 생성 시 “구체적이고 명확인 예시와 함께 자세히 설명해 주세요"라는 명시적인 답변 형식을 추가

실제 생성 프롬프트 예시(요인 : gpt-3.5-turbo, 한국어, 역할 부여o, 명시성 요구 o) :
“어려움에 처한 친구를 도와야 하는 이유는 무엇인가요**?** 윤리학과 도덕철학을 전공한 전문가로서**,** 구체적이고 명확한 예시를 통해 자세히 설명해 주세요**.** 한국어로 답변해 주세요**.”**

프롬프트에 영향을 미치는 역할과 명시성 요구은 전문가 역할 명시, 명시적인 답변 형식을 통해 질문 프롬프트를 생성하였다.

생성 질문 데이터 : **https://github.com/statjhw/DOE/blob/master/factorial_results/01_full_prompts_generated.json**

2.3 의미 유사도 측정 지표 선택**: BERTScore**

본 프로젝트는 동일한 프롬프트 조건에서 생성된 LLM 응답 간의 의미적 일관성(consistency)을 측정하는 것을 목표로 하였다.이 목표를 달성하기 위해, 기존 방식의 한계를 극복하고 LLM 출력의 특성을 정확히 반영할 수 있는 측정 지표를 신중하게 선택하였다.

2.3.1 기존 방식**(**코사인 유사도**)**의 한계

코사인 유사도는 문장 임베딩 기반으로 응답 유사도를 측정했지만, 어휘 차이나 문장 구조, 문맥 흐름을 제대로 반영하지 못하는 한계가 있습니다. 특히 한국어처럼 어순이 자유로운 언어에서는 미묘한 표현 차이에 민감하게 반응하여 문맥을 고려하지 못해 LLM 출력의 일관성을 평가하기 어려웠습니다.

2.3.2 BERT Multilingual 기반 의미 유사도**(BERTScore)**의 강점

코사인 유사도의 한계를 극복하기 위해 BERTScore를 도입했습니다. BERTScore는 한국어를 포함한 다국어 처리가 가능하며, 문맥 정보를 반영하여 의미 중심의 유사도를 계산합니다. 이는 단순한 어휘 일치보다 의미론적 유사성을 기반으로 정밀한 평가를 수행하여 문맥을 더 잘 고려한 유사도 측정이 가능합니다.

2.3.3 예시 비교**:** 단순 코사인 유사도 **vs. BERTScore**

다음 표는 두 문장에 대한 코사인 유사도와 BERTScore의 비교를 보여준다.

- 문장**A**: "타인에게 공감하는 태도가 중요하다고 생각합니다."
- 문장**B**: "저는 다른 사람의 감정을 이해하려고 노력합니다."

평가지표	코사인 유사도	**BERT** 기반모델
유사도 값	0.2233	0.7618

코사인 유사도는 어휘 차이로 인해 의미가 유사한 문장도 낮게 평가할 수 있습니다. 반면, BERTScore는 문맥을 이해하여 의미론적 유사성을 정확하게 측정하므로, LLM 응답의 일관성을 평가하는 데 더 적합합니다.

2.3.4 측정 지표 선택에 대한 결론 및 최종 반응변수 선택
결과적으로 본 프로젝트는 단순 텍스트 유사도 지표로는 감지하기 어려운 표현의 다양성과 의미 차이를 포착하고자 의미 기반 정량 지표인 BERTScore를 사용했습니다. 동일한 프롬프트(질문)에 대해 총 5회 생성된 응답들 간의 의미 유사도(BERTScore)를 측정하고 그 평균값을 계산하여 이를 최종 반응 변수로 설정함으로써 실험의 신뢰도를 높이고자 하였다
최종 데이터 링크 : https://github.com/statjhw/DOE/tree/master

---

3. 실험 1: 감정 프레이밍이 AI 응답 일관성에 미치는 영향 (RCBD 분석)

첫 번째 실험은 감정적 프레이밍이 LLM 응답의 일관성에 미치는 영향을 평가하기 위해 무작위 완전 블록 설계(RCBD)를 활용하였다.

3.1 실험 설계 개요

본 실험의 목적은 처리 요인에 따른 언어 모델의 응답 일관성을 평가하는 것이다.

- 처리 요인**(Treatment):** 감정 프레이밍 수준 **(3**수준**)**:
  - 중립적 표현
  - 정서적 표현
  - 자극적/도발적 표현
- 블록 요인**(Block):** 질문 카테고리 **(3**수준**)**:
  - 인성 (윤리/정의 포함)
  - 창의성
  - 논리적 추론

카테고리 **&** 요인 조합 별 데이터 수

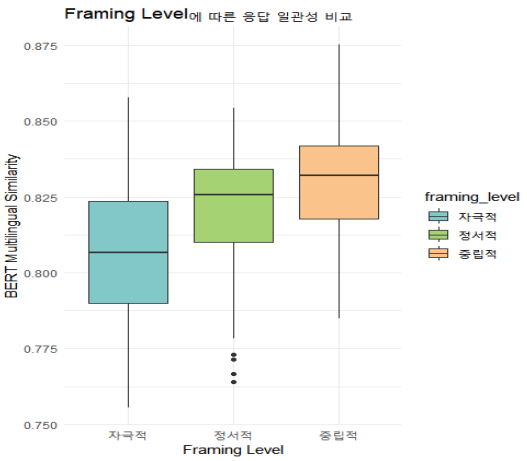
데이터 갯수	중립적	정서적	자극적
인성	42	42	42
창의성	43	43	43
논리적 추론	48	48	48

총 데이터 수는 399개이다.

- 반응 변수**(Response variable)**: bert_multilingual_similarity (응답의 일관성).
- 가설:
    - 귀무가설(H0): 세 가지 감정 프레이밍 수준(중립적/정서적/자극적) 간에 응답 일관성의 평균 차이가 없다 ($\mu_1 = \mu 2 = \mu 3$).
    - 대립가설(H1): 세 가지 감정 프레이밍 수준 간에 응답 일관성의 평균 차이가 존재한다 (not H0).

3.2 프레이밍 수준 효과 시각화 **(Boxplot)**

프레이밍 수준에 따른 응답 일관성 비교를 위한 Boxplot은 다음과 같은 결과를 명확히 보여준다.



- 중립적 프레이밍: 평균적으로 가장 높은 의미 유사도(일관성)를 보였다. 이는 중립적인 프레이밍이 LLM 응답의 일관성을 가장 높게 유지하는 데 효과적임을 시사한다.
- 정서적 프레이밍**:** 중립적 프레이밍보다는 낮지만 자극적 프레이밍보다는 높은 중간 수준의 평균 일관성을 보였습니다. 응답의 분산 역시 자극적 프레이밍보다는 작았으며, 이는 온건한 감정을 포함하는 프레이밍이 AI 응답의 일관성을 다소 저해할 수는 있지만, 자극적인 표현만큼 예측 불가능성을 크게 만들지는 않는다는 것을 시사합니다.
- 자극적 프레이밍: 평균 일관성이 가장 낮았으며, 응답 간의 분산 또한 상대적으로 크게 나타났다. 이는 자극적인 프레이밍이 LLM 응답의 일관성을 저해하고, 예측 불가능한 행동을 유발할 수 있음을 의미한다. 높은 분산은 LLM이 평균적으로 덜 일관적일 뿐만 아니라, 더 불규칙적인 응답을 생성한다는 것을 나타낸다. 이는 중요한 애플리케이션에서 높은 평균 일관성뿐만 아니라 낮은 변동성 또한 신뢰성 확보에 필수적임을 강조한다.

3.3 ANOVA 분석 결과

RCBD에 대한 ANOVA 분석 결과는 다음과 같다.

Model <- aov(bert_multilingual_similarity ~ framing_level + category)

	Df(자유도)	**Sum Sq**	**Mean Sq**	**F value**	**P value**
framing Level	2	0.03082	0.015408	41.79	<2e-16
Category	2	0.02098	0.010492	28.46	2.86e-12
Residuals	394	0.14526	0.000369		

분석 결과, framing_level의 P-value는 <2e-16으로, 유의수준 α=0.01보다 훨씬 작게 나타났다. 이는 귀무가설(세 가지 프레이밍 수준 간 응답 일관성의 평균 차이가 없다)을 기각하며, 감정 프레이밍 방식에 따라 LLM의 응답 일관성이 통계적으로 매우 유의미하게 달라진다는 것을 의미한다. 또한, 블록 요인인

category 역시 P-value가 2.86e-12로 유의미한 영향을 미치는 것으로 나타났는데, 이는 RCBD 설계가 질문 카테고리 간의 변동을 성공적으로 분리했음을 보여준다.

3.4 모델 적합성 검정

ANOVA 분석 결과의 유효성을 확인하기 위해 모델 적합성 검정을 수행하였다.

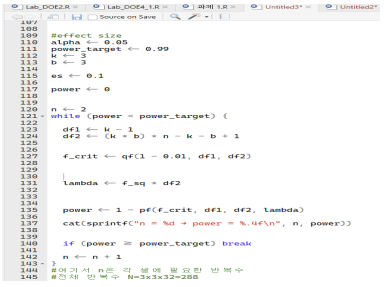
- 정규성 검정: 잔차의 Q-Q 플롯은 직선 형태에 가깝게 분포하였고, 잔차의 히스토그램은 대칭적인종 모양에 가까워 정규성 가정을 만족하는 것으로 확인되었다.
- 선형성 및 등분산성 검정: 'Fitted Values vs Residuals' 플롯에서 점들 사이에 특정한 곡선 패턴이 나타나지 않아 선형성 가정을 만족하였다. 또한, 잔차가 0을 중심으로 특정 구간에 몰려있지 않고 전체 구간에 걸쳐 고르게 분포하여 등분산성 가정을 만족하는 것으로 확인되었다.
- 처리 요인 및 블록에 따른 잔차 분포: 각 프레이밍 수준별 잔차 분포와 각 카테고리 블록별 잔차 분포를 확인한 결과, 눈에 띄는 이상치가 관측되지 않았으며, 모든 수준과 블록에서 잔차가 유사한 범위 내에서 분포하여 등분산성 가정이 전반적으로 만족됨을 재확인하였다.

이러한 모델 적합성 검증은 ANOVA 분석 결과의 신뢰성과 일반화 가능성을 높인다. 통계적 가정을 충족함으로써, 프롬프트 프레이밍과 AI 일관성 간의 인과 관계에 대한 추론이 통계적으로 타당하고 신뢰할 수 있음을 입증한다.

3.5 효과 크기 및 표본 크기 적절성

실험의 통계적 검정력을 확인하고 결과의 안정성을 확보하기 위해 적절한 표본 크기를 계산하였다.

- 검정력 계산 알고리즘: 목표 검정력 0.99 달성을 위해 F-분포의 비중심성 모수를 활용하여 표본 크기(n)를 점진적으로 증가시키며 측정합니다.



```
#effect size
alpha <- 0.05
power_target <- 0.99
k <- 3
b <- 3

es <- 0.1

power <- 0

n <- 2
while (power < power_target) {

  df1 <- k - 1
  df2 <- (k * b) * n - k - b + 1

  f_crit <- qf(1 - 0.01, df1, df2)

  lambda <- f_sq * df2


  power <- 1 - pf(f_crit, df1, df2, lambda)
  cat(sprintf("n = %d → power = %.4f\n", n, power))

  if (power >= power_target) break
  n <- n + 1
}
#여기서 n은 각 셀에 필요한 반복수
#전체 반복수 N=3x3x32=288
```

```
n = 2 → power = 0.0376
n = 3 → power = 0.0742
n = 4 → power = 0.1208
n = 5 → power = 0.1753
n = 6 → power = 0.2354
n = 7 → power = 0.2988
n = 8 → power = 0.3636
n = 9 → power = 0.4280
n = 10 → power = 0.4905
n = 11 → power = 0.5500
n = 12 → power = 0.6057
n = 13 → power = 0.6571
n = 14 → power = 0.7039
n = 15 → power = 0.7460
n = 16 → power = 0.7834
n = 17 → power = 0.8165
n = 18 → power = 0.8453
n = 19 → power = 0.8703
n = 20 → power = 0.8918
n = 21 → power = 0.9102
n = 22 → power = 0.9257
n = 23 → power = 0.9389
n = 24 → power = 0.9499
n = 25 → power = 0.9591
n = 26 → power = 0.9667
n = 27 → power = 0.9730
n = 28 → power = 0.9782
n = 29 → power = 0.9825
n = 30 → power = 0.9859
n = 31 → power = 0.9887
n = 32 → power = 0.9910
```

현재 399개의 샘플 데이터는 목표 검정력 0.99를 달성하는 데 필요한 288개를 초과하므로, 이 실험은 통계적으로 충분히 안정적이고 신뢰할 수 있습니다. 이는 제2종 오류의 위험을 최소화하고, 실제 효과 부재의 강력한 증거를 제공합니다.

3.6 사후 분석 (Tukey's HSD Test)

ANOVA 분석 결과 귀무가설이 기각되었으므로, 각 프레이밍 수준 간의 구체적인 평균 차이를 파악하기 위해 Tukey's HSD 사후 분석을 수행하였다.

	차이	하한값	상한값	p 조정값(보정p_값)
정서적-자극적	0.013733685	0.008194195	0.01927318	0.0000000
중립적-자극적	0.02122691	0.015683201	0.02676218	0.0000000
중립적-정서적	0.007489006	0.001949516	0.01302850	0.0045127

Tukey's HSD 테스트 결과, '정서적-자극적', '중립적-자극적', '중립적-정서적' 모든 쌍에서 p-조정값(p-adjusted value)이 유의수준 α=0.01보다 현저히 작게 나타났다. 이는 세 가지 프레이밍 수준 간의 평균 일관성 차이가 무작위로 발생한 것이 아니라 통계적으로 매우 유의미한 차이임을 의미한다. 반응 변수의 범위가 0.75에서 0.88 사이의 좁은 범위에서 움직이는 수치임을 고려할 때, 이러한 평균 차이는 통계적으로 의미 있는 것으로 판단된다.

3.7 RCBD 결론 요약

ANOVA 분석 결과, 세 가지 감정 프레이밍 수준 간에 응답 일관성의 평균 차이가 통계적으로 유의미하게 존재하며, 귀무가설은 기각되었다. 사후 분석을 통해 각 수준 간에도 유의미한 차이가 확인되었다.

이러한 발견을 바탕으로, 프롬프트 작성 시 중립적인 어조를 사용하는 것이 일관되고 유용한 답변을 얻는 데 효과적임을 제안한다. 자극적인 표현이나 감정적인 어투는 일관적이지 않은 답변을 초래할 수 있으므로 피하는 것이 바람직하다.

---

4. 실험 2: 복합 요인 및 상호작용이 AI 응답 일관성에 미치는 영향 (2^4 완전 요인 설계 with blocking)

두 번째 실험은 여러 프롬프트 관련 요인들의 주효과 및 교호작용이 AI 응답 일관성에 미치는 영향을 탐색하기 위해 2^4 완전 요인 설계를 활용하였다.

4.1 실험 설계 개요

본 실험은 2^4 완전 요인 설계와 무작위 완전 블록 설계(RCBD)를 결합한 구조로 설계되었다.

- 요인(Factors) - 4개 요인, 각 2수준:
  - model: GPT-3.5-turbo vs. GPT-4o-mini
  - prompt_language: 한국어(korean) vs. 영어(english)
  - explicitness (명시성 요구): 높음(high) vs. 낮음(low)
  - role_assignment (역할 부여): 있음(with_role) vs. 없음(no_role) 총 요인 조합 수: 2^4 = 16가지 고유 조합.
- 블록(Block) 요인: 질문 카테고리(category)를 블록으로 설정하였으며, 이는 인성, 창의성, 논리적 추론의 세 가지 수준으로 구성된다.
- 반응 변수: bert_multilingual_similarity (응답 간 유사도).
- 반복(Replicates): 각 블록 내 각 요인 조합에 대해 총 7회의 반복 실험을 수행하였다.
- 모델 코딩: 분석을 위해 각 요인의 수준은 -1(낮은 수준)과 1(높은 수준)로 코딩되었다.
- 가설:
  - 주효과 가설 (각 요인별):

- H0: 모델(model)은 응답 일관성에 유의미한 영향을 미치지 않는다. (H1: 유의미한 영향을 미친다.)
- H0: 프롬프트 언어(prompt_language)는 응답 일관성에 유의미한 영향을 미치지 않는다. (H1: 유의미한 영향을 미친다.)
- H0: 역할 부여(role_assignment)는 응답 일관성에 유의미한 영향을 미치지 않는다. (H1: 유의미한 영향을 미친다.)
- H0: 명시성 요구(explicitness)은 응답 일관성에 유의미한 영향을 미치지 않는다. (H1: 유의미한 영향을 미친다.)

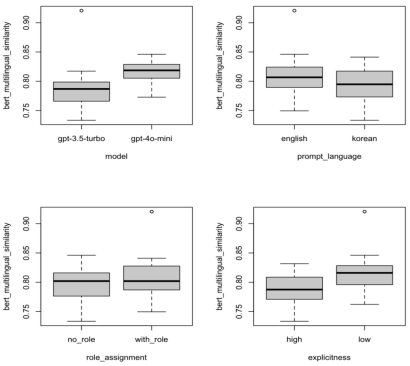
○ 교호작용 가설 **(**주요 교호작용**)**:

- H0: 모델과 역할 부여(model:role_assignment) 간에는 응답 일관성에 대한 유의미한 교호작용이 없다. (H1: 유의미한 교호작용이 있다.)
- H0: 모델과 명시성 요구(model:explicitness) 간에는 응답 일관성에 대한 유의미한 교호작용이 없다. (H1: 유의미한 교호작용이 있다.)
- H0: 프롬프트 언어, 역할 부여, 명시성 요구(prompt_language:role_assignment:explicitness) 간에는 응답 일관성에 대한 유의미한 3차 교호작용이 없다. (H1: 유의미한 3차 교호작용이 있다.)

○ 그 외 모든 2차, 3차, 4차 교호작용에 대해서도 유사한 귀무가설을 설정한다.

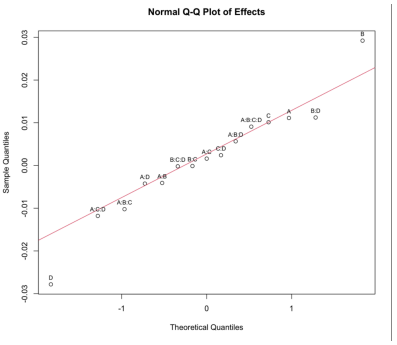
4.2 사전 탐색**:** 주효과 및 교호작용 효과

각 요인의 개별 및 상호작용 효과를 초기 확인하고자, 2^4 완전 요인 설계를 3수준의 블록 조건 하에 2회 반복 수행하여 사전 분석을 진행한다 (데이터 수 = 16 * 3 * 2 = 96)



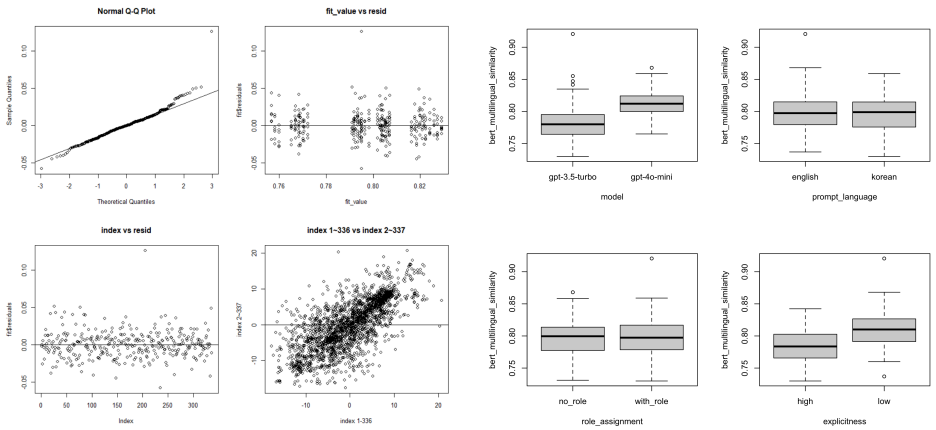
초기 Boxplot을 통해 각 요인(role_assignment, model, explicitness, prompt_language)이 bert_multilingual_similarity에 미치는 개별적인 영향을 시각적으로 확인하였다.

```
> eff_vec
          A          B          C          D
[1,] 0.0111276 0.02922907 0.01012344 -0.0278197
         A:B         A:C          B:C
[1,] -0.004096525 0.001599855 -0.0001352027
         A:D        B:D         C:D       A:B:C
[1,] -0.004248122 0.01123775 0.002395484 -0.01023164
        A:B:D       A:C:D         B:C:D     A:B:C:D
[1,] 0.00565028 -0.01180327 -0.0001981029 0.009078069
```



사전 분석은 **정규 확률 플롯(Normal Probability Plot of Effects)**을 통해 진행되었습니다. 분석 결과, 요인 B(모델)와 D(명시성 요구)의 주효과가 이론적 정규 분포에서 현저히 벗어나, 이들이 AI 응답 일관성에 가장 큰 영향을 미치는 유의미한 요인임을 강력하게 시사합니다. 또한, B(모델)와 D(명시성 요구) 간의 교호작용을 포함하여 잠재적으로 유의미한 여러 요인 조합의 상호작용 효과가 발견되었습니다. 이 사전 분석은 향후 전체 ANOVA 결과 해석의 방향성을 제시하고, 주요 요인과 상호작용을 효율적으로 식별하는 데 기여했습니다.

4.3 데이터 탐색 및 적합성 검정



<적합성 검정> **<boxplot>**

최종 분석에 앞서, 최종적으로 수집된 데이터를 바탕으로 복잡한 2^4 완전 요인 모델의 적합성을 검증하기 위해 잔차 진단 플롯을 분석했습니다. (잔차의 정규 Q-Q 플롯, Index vs. Residuals, Fitted Values vs. Residuals). 분석 결과, 잔차의 정규성**,** 등분산성**,** 선형성 등 모델의 기본 가정이 전반적으로 충족됨을 확인하여 모델이 데이터에 잘 적합되었음을 검증했습니다

또한, 박스플롯을 통해 각 요인 수준별 응답 일관성(bert_multilingual_similarity)의 분포와 중심 경향을 시각적으로 탐색했습니다.

4.4 ANOVA 분석 결과

실험 2에 대한 전체 ANOVA 분석 결과는 다음과 같다.

각 블록 내 각 요인 조합에 대해 총 7회의 반복 실험을 수행하였다. (총 데이터 수 : 16 * 3 * 7=336)

model <- aov(bert_multilingual_similarity~ category + prompt_language * model * role_assignment * explicitness)

```
Response: bert_multilingual_similarity
                                                Df   Sum Sq  Mean Sq  F value    Pr(>F)
model                                            1 0.077455 0.077455 223.0422 < 2.2e-16 ***
prompt_language                                  1 0.000527 0.000527   1.5169  0.219052
role_assignment                                  1 0.000347 0.000347   0.9987  0.318376
explicitness                                     1 0.055459 0.055459 159.7014 < 2.2e-16 ***
category                                         2 0.000602 0.000301   0.8672  0.421103
model:prompt_language                            1 0.000087 0.000087   0.2509  0.616762
model:role_assignment                            1 0.001550 0.001550   4.4628  0.035420 *
prompt_language:role_assignment                  1 0.000012 0.000012   0.0350  0.851693
model:explicitness                               1 0.002366 0.002366   6.8139  0.009473 **
prompt_language:explicitness                     1 0.001176 0.001176   3.3875  0.066625 .
role_assignment:explicitness                     1 0.000703 0.000703   2.0242  0.155789
model:prompt_language:role_assignment            1 0.000002 0.000002   0.0067  0.935017
model:prompt_language:explicitness               1 0.000270 0.000270   0.7762  0.378985
model:role_assignment:explicitness               1 0.000608 0.000608   1.7515  0.186637
prompt_language:role_assignment:explicitness     1 0.001763 0.001763   5.0755  0.024947 *
model:prompt_language:role_assignment:explicitness 1 0.000026 0.000026   0.0757  0.783377
Residuals                                      318 0.110431 0.000347
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

유의미한 주효과 **(P-value < 0.01)**:

- model (모델): F-value 223.0422, P-value < 2.2e-16 (***)로, LLM 모델 선택이 응답 일관성에 매우 유의미한 영향을 미친다. 이는 '모델은 응답 일관성에 유의미한 영향을 미치지 않는다'는 귀무가설을 기각한다.
- explicitness (명시성 요구): F-value 159.7014, P-value < 2.2e-16 (***)로, 프롬프트의 명시성 요구 또한 응답 일관성에 매우 유의미한 영향을 미친다. 이는 '명시성 요구은 응답 일관성에 유의미한 영향을 미치지 않는다'는 귀무가설을 기각한다.

유의미하지 않은 주효과 **(P-value > 0.05)**:

- prompt_language (프롬프트 언어): P-value 0.219052. 이는 '프롬프트 언어는 응답 일관성에 유의미한 영향을 미치지 않는다'는 귀무가설을 기각하지 못한다.
- role_assignment (역할 부여): P-value 0.318376. 이는 '역할 부여는 응답 일관성에 유의미한 영향을 미치지 않는다'는 귀무가설을 기각하지 못한다.
- category (블록 요인): P-value 0.421103. 블록 요인인 카테고리는 통계적으로 유의미한 영향을 미치지 않는 것으로 나타났는데, 이는 일반적인 블록 요인의 역할과는 다소 차이가 있으나, 처리 효과의 유효성에는 영향을 미치지 않는다.

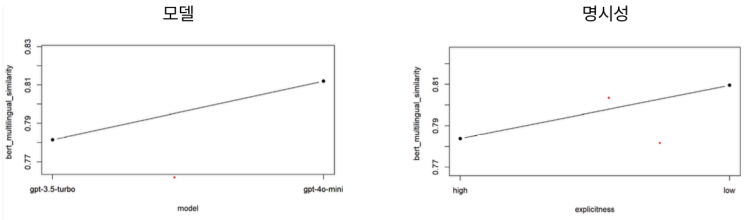
유의미한 교호작용 효과 **(P-value < 0.05)**:

- model:role_assignment (모델:역할 부여): P-value 0.035420 (*)로, 역할 부여의 효과가 사용된 LLM 모델에 따라 달라짐을 나타낸다. 이는 '모델과 역할 부여 간에는 유의미한 교호작용이 없다'는 귀무가설을 기각한다.
- model:explicitness (모델:명시성 요구): P-value 0.009473 (**)로, 명시성 요구의 효과가 사용된 LLM 모델에 따라 달라짐을 나타낸다. 이는 '모델과 명시성 요구 간에는 유의미한 교호작용이 없다'는 귀무가설을 기각한다.
- prompt_language:role_assignment:explicitness (프롬프트 언어:역할 부여:명시성 요구): P-value 0.024947 (*)로, 3차 교호작용이 유의미하게 나타나, 프롬프트 언어의 효과가 역할 부여와 명시성 요구 수준에 따라 복합적으로 변화함을 시사한다. 이는 해당 3차 교호작용에 대한 귀무가설을 기각한다.

유의미하지 않은 교호작용 효과: 그 외의 2차, 3차, 4차 교호작용은 통계적으로 유의미하지 않았다.

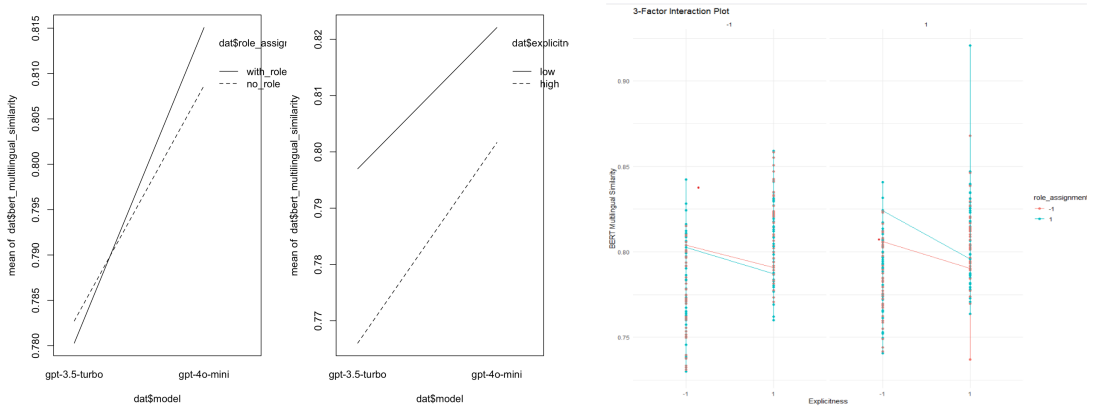
4.5 주효과 분석 **(**시각화**)**

주효과 플롯은 model과 explicitness 요인의 영향을 시각적으로 보여준다.



- 모델: GPT-40-mini는 GPT-3.5-turbo에 비해 전반적으로 더 높은 응답 일관성을 보였다.
- 명시성 요구: 낮은 명시성 요구은 높은 명시성 요구보다 전반적으로 더 높은 응답 일관성을 이끌어냈다.

4.6 교호작용 분석 **(**시각화**)**



교호작용 플롯은 model:role_assignment, model:explicitness, 그리고 3차 교호작용 prompt_language:role_assignment:explicitness에 대한 시각적 정보를 제공한다. 이 플롯들은 한 요인의 효과가 다른 요인의 수준에 따라 어떻게 변화하는지를 보여주며, 통계적으로 유의미한 교호작용의 존재를 뒷받침한다

이러한 교호작용의 존재는 프롬프트 최적화가 단순한 가산적 과정이 아님을 시사한다. 예를 들어, 역할 부여가 응답 일관성에 미치는 영향은 특정 LLM 모델에 따라 다르게 나타날 수 있다. 즉, GPT-4o-mini에서는 역할 부여가 매우 효과적일 수 있지만, GPT-3.5-turbo에서는 영향이 미미할 수 있다는 것이다. 이러한 상호작용은 LLM 내부에서 프롬프트 구성 요소들이 일관성에 미치는 영향을 어떻게 조절하는지에 대한 복잡한 인과 관계를 시사한다.

4.7 요인 설계 결론 요약

본 실험은 복합적인 요인들이 AI 응답 일관성에 미치는 영향을 확인하는 데 목적이 있었다. 주요 발견은 다음과 같다.

- 강력한 주효과: 명시성 요구(explicitness)과 모델(model) 요인이 AI 응답의 일관성에 강력한 주효과를 미치는 것으로 확인되었다. 이는 해당 요인에 대한 귀무가설을 기각한다.
- 미미한 효과: 역할 부여(role_assignment)와 프롬프트 언어(prompt_language)는 응답 일관성에 미미한 주효과를 미치므로, 특정 맥락에서는 프롬프트 설계의 주요 고려 사항에서 제외될 가능성이 있다. 이는 해당 요인에 대한 귀무가설을 기각하지 못한다.
- 교호작용 발견: 모델과 역할 부여, 모델과 명시성 요구 간의 유의미한 교호작용이 발견되었다. 또한, 프롬프트 언어, 역할 부여, 명시성 요구 간의 3차 교호작용도 유의미하게 나타나, 프롬프트 요소들의 복잡한 상호작용이 일관성에 영향을 미침을 보여준다. 이는 해당 교호작용에 대한 귀무가설을 기각한다.

---

5. 종합 논의 및 주요 발견

본 프로젝트는 AI 응답 일관성이라는 중요한 주제에 대해 두 가지 체계적인 실험을 통해 심층적인 분석을 수행하였다. 실험 1(RCBD)은 감정 프레이밍의 영향을, 실험 2(2^4 완전 요인 설계)는 모델, 명시성 요구, 언어, 역할 부여와 같은 복합 요인 및 그 상호작용의 영향을 탐색하였다.

5.1 실험 결과 종합

실험 1의 결과는 프롬프트의 감정적 프레이밍이 AI 응답 일관성에 지대한 영향을 미침을 명확히 보여주었다. 특히, 중립적인 프레이밍이 가장 높고 안정적인 일관성을 유도하는 반면, 감정적이거나 자극적인 언어는 LLM 출력의 예측 불가능성을 증가시키는 것으로 나타났다. 이는 언어적 톤이 LLM의 응답 안정성에 미치는 심오한 영향을 강조한다. 실험 1에서 설정한 귀무가설("세 가지 감정 프레이밍 수준 간에 응답 일관성의 평균 차이가 없다")은 기각되었으며, 이는 감정 프레이밍이 응답 일관성에 유의미한 영향을 미친다는 것을 입증한다.

실험 2에서는 LLM '모델' 선택과 프롬프트 '명시성 요구'이 일관성에 가장 강력한 주효과를 미치는 요인임을 확인하였다. GPT-40-mini가 GPT-3.5-turbo보다 전반적으로 높은 일관성을 보였으며, 낮은 명시성 요구은 일관성 향상에 기여했다. 이는 '모델'과 '명시성 요구'에 대한 귀무가설을 기각한다.

흥미롭게도, '프롬프트 언어'와 '역할 부여'는 개별적으로는 유의미한 주효과를 보이지 않아 해당 귀무가설을 기각하지 못했지만, 다른 요인들과의 유의미한 교호작용에 관여하여 그 영향이 맥락에 따라 달라짐을 시사하였다. 이는 '모델과 역할 부여', '모델과 명시성 요구', 그리고 '프롬프트 언어, 역할 부여, 명시성 요구' 간의 교호작용에 대한 귀무가설을 기각한다.

5.2 프롬프트 설계 및 **AI** 일관성에 대한 다층적 이해

본 프로젝트의 통합된 발견들은 프롬프트 설계가 AI 응답 일관성에 미치는 영향에 대한 포괄적인 이해를 제공한다.

- 언어적 톤과 명확성의 중요성: 실험 결과는 질문의 표현 방식(감정적 톤)과 지시의 명확성(명시성 요구)가 AI 응답 일관성의 근본적인 결정 요인임을 일관되게 보여준다. 중립적인 톤을 사용하고 구체적인 예시를 요구하는 등 과도한 지시를 피하는(낮은 명시성 요구) 프롬프트는 LLM의 예측 가능한 동작을 유도하는 데 효과적이다.
- 모델별 일관성 특성: 다른 LLM 모델들은 그 자체로 일관성 수준에서 내재적인 차이를 보인다. GPT-40-mini가 GPT-3.5-turbo보다 전반적으로 높은 일관성을 나타낸다는 점은, 모델 선택이 일관성 최적화를 위한 주요한 선택임을 시사한다.
- 교호작용 효과의 미묘함: model:role_assignment, model:explicitness, prompt_language:role_assignment:explicitness와 같은 유의미한 교호작용의 발견은 프롬프트 최적화가 단순한 개별 요소의 합이 아님을 강조한다. 특정 프롬프트 요소(예: 역할 부여 또는 언어)의 효과는 보편적이지 않고, 다른 요인들, 특히 사용되는 LLM 모델과 명시성 요구수준에 따라 달라졌다.

최적의 프롬프트 설계는 중립적 언어, 적절한 명시성, 모델 적응, 맥락적 상호작용 이해를 포괄하는 다차원적 접근을 통해 AI 시스템 신뢰성을 높이는 증거 기반 프롬프트 엔지니어링을 발전시킨다.

---

6. 한계점 및 향후 연구 방향

본 프로젝트는 AI 응답 일관성에 영향을 미치는 핵심 요인들을 실증적으로 탐구하였으나, 모든 연구와 마찬가지로 내재적인 한계점을 지니며, 이는 향후 연구의 중요한 방향성을 제시한다.

6.1 현재 프로젝트의 한계점

- 제한적인 프롬프트 조건: 본 실험에서 다룬 프레이밍, 명시성 요구 등의 조건들이 실제 사용 환경에서 나타나는 다양한 변수들을 모두 포괄하지 못했다는 점이 한계로 지적된다. 이는 실제 AI 응답 일관성에 영향을 미칠 수 있는 더 많은 요인들이 존재할 수 있음을 의미한다.
- 특정 **LLM** 모델에 대한 집중: 본 프로젝트는 GPT-3.5-turbo와 GPT-40-mini 모델에 초점을 맞추어 진행되었다. 따라서 본 프로젝트의 발견이 다른 LLM 아키텍처나 미래 버전의 모델에 직접적으로 일반화될 수 있을지는 추가적인 검증이 필요하다.
- 특정 작업 유형에 대한 제한: 정의된 질문 카테고리(인성, 창의성, 논리적 추론)는 잠재적인 AI 작업의 일부만을 대표한다. 프롬프트 요인의 영향은 다른 전문 분야나 복잡한 대화에서는 다르게 나타날 수 있다.

6.2 향후 연구 방향

LLM 연구의 역동적이고 빠르게 진화하는 환경을 고려할 때, 본 프로젝트는 추가적인 탐구를 위한 아이디어를 제공한다.

6.2.1 평가 지표의 다변화

응답 일관성 외에도 정확성, 논리적 타당성, 사용자 선호도 등 다양한 평가 지표를 추가적으로 도입하여 LLM 응답의 품질을 다각적으로 평가할 수 있다. 이는 LLM 출력의 품질에 대한 보다 포괄적인 이해를 제공할 것이다.

6.2.2 다양한 **LLM** 및 다국어 실험으로 확장

향후 연구에서는 GPT-4(또는 GPT-40-mini) 외의 다른 LLM 모델들을 테스트하고, 언어별 특성 분석을 포함한 다국어 실험을 수행해야 한다. 이는 연구 결과의 보편성과 적용 범위를 확장하여, 다양한 모델과 언어적 맥락에서 프롬프트 엔지니어링 원칙에 대한 보다 일반화된 이해를 제공할 것이다.

이는 본 프로젝트가 최종 결론이 아닌, 지속적인 연구와 프롬프트 엔지니어링 전략을 위한 시작 단계임을 시사한다.

7. 결론

본 보고서는 신뢰할 수 있는 AI 시스템 구축을 위해 AI 응답 일관성에 영향을 미치는 핵심 요인들을 실증적으로 탐구하였다. 면밀하게 설계된 일련의 실험을 통해, 감정 프레이밍, 명시성 요구, LLM 모델 선택을 포함한 프롬프트 특성들이 AI 생성 응답의 일관성에 유의미한 영향을 미친다는 것을 입증하였다.

구체적으로, 실험 1은 중립적 프레이밍이 높은 일관성을 달성하는 데 가장 중요하며, 감정적이거나 도발적인 언어는 예측 불가능하고 덜 일관적인 출력을 초래한다는 것을 보여주었다. 이는 "세 가지 감정 프레이밍 수준 간에 응답 일관성의 평균 차이가 없다"는 귀무가설을 기각하는 결과이다. 실험 2는 모델 선택과 프롬프트 명시성 요구의 강력한 주효과를 추가로 강조하며, 일관성을 결정하는 데 있어 이들의 근본적인 역할을 나타냈다. 이는 '모델'과 '명시성 요구'이 응답 일관성에 유의미한 영향을 미치지 않는다는 귀무가설을 기각하는 결과이다. 결정적으로, 유의미한 교호작용의 발견은 최적의 프롬프트 설계가 미묘하고 복잡한 문제이며, 특정 요인의 효과가 다른 요인에 따라 달라진다는 점을 강조한다. 이는 해당 교호작용에 대한 귀무가설을 기각하는 결과이다.

본 프로젝트는 프롬프트 엔지니어링에 대한 이해를 제공한다. 이러한 발견에서 도출된 권고안은 실제 사용자들이 AI 시스템의 신뢰성과 예측 가능성을 향상시키기 위한 실질적인 전략을 제공할 수 있길 기대한다.