컴퓨터교육학회 논문지 2025년 제28권 제12호
https://doi.org/10.32431/kace.2025.28.12.001

프롬프트 형식이 LLM의 강건성에 미치는 영향 분석*

The Impact of Prompt Formats on the Robustness of LLMs

이승현† · 이영호††
Seunghyun Lee† · Youngho Lee††

요약

본 연구는 프롬프트 형식(Markdown, JSON, YAML, XML)이 국어 및 영어 환경에서 LLM의 강건성에 미치는 영향을 분석하였다. 이를 위해 해결 정확성(SA), 응답 일관성(RC), 성능 안정성(PSF)을 측정하고 이를 종합한 실용 강건성 지수(PRI)를 통해 모델의 다차원적인 성능을 평가했다. 분석 결과 첫째, 고성능 플래그십 모델들은 프롬프트 형식 변화에 민감하지 않았지만 경량화 모델과 일부 모델은 형식에 크게 의존하는 특성을 보였다. 둘째, 모든 환경에서 절대적으로 우수한 프롬프트 형식은 존재하지 않으며 최적의 형식 또한 국어와 영어 과업 환경에서 다르게 나타났다. 셋째, PRI는 모델의 성능을 다각도로 평가하며 여러 차원에서 균형 잡힌 모델을 효과적으로 식별하는 도구임을 입증했다. 본 연구는 LLM의 강건성이 모델 성능, 과업의 언어, 프롬프트 형식의 복합적인 상호작용에 의해 결정된다는 점을 정량적으로 입증했으며 교육적 요구와 다양한 학습 환경 및 활용 목적을 고려한 최적의 모델과 프롬프트 형식을 선택하는 데 필요한 실증적인 판단 기준과 방향성을 제시했다는 점에서 의의가 있다.

주제어 대규모 언어 모델, 프롬프트 엔지니어링, AI 강건성, 프롬프트 형식, 모델 평가, 소버린 AI

ABSTRACT

This study investigates the impact of prompt format(Markdown, JSON, YAML, XML) on the robustness of Large Language Models(LLMs) in both Korean and English contexts. We evaluated the multidimensional performance of various models by measuring three key metrics: Solution Accuracy(SA), Response Consistency(RC), and Performance Stability(PSF). These metrics were then synthesized into a comprehensive Practical Robustness Index(PRI). The analysis yielded the following key findings. First, high-performance flagship models were largely unaffected by changes in prompt format, whereas lightweight and other models exhibited a strong dependency on the format. Second, no single prompt format proved universally superior, and the optimal format differed between the Korean and English contexts. Third, the PRI was validated as an effective tool for identifying models that maintain a strong balance across the different performance dimensions. This study quantitatively demonstrates that the robustness of an LLM is determined by the complex interaction among model performance, task language, and prompt format, and it is significant in providing empirical criteria and guidance for selecting the optimal model and prompt format that take into account diverse educational demands, various learning environments, and application purposes.

Keywords Large Language Models(LLM), Prompt Engineering, AI Robustness, Prompt Format, Model Evaluation, Sovereign AI

†정회원 대구교육대학교 교육대학원 AI교육전공 박사과정
††정회원 대구교육대학교 컴퓨터교육과 조교수 (교신저자)

논문투고 2025년 07월 07일
심사완료 2025년 09월 30일
게재확정 2025년 10월 22일
발행일자 2025년 12월 31일

* 본 논문은 2025년 한국연구재단의 지원을 받아 수행된 연구임(NRF-2025S1A5A8006876)

1. 서론

대규모 언어 모델(LLM)은 단순한 정보 검색을 넘어 정교한 추론과 분석을 요구하는 지식 집약적 과업(Knowledge-Intensive Task)을 수행하는 단계로 진화하고 있다. 교육에서도 초기에는 단순히 LLM을 정보 검색이나 수업 자료 생성에 활용하는 수준에 머물렀지만 최근에는 교육 플랫폼 설계, 평가 도구 개발, 개별화 학습 지원 등 더욱 복잡하고 고도화된 과업으로 그 활용 범위가 확대되고 있다. 이에 따라 LLM 활용의 교육적 효과를 극대화하고 LLM을 활용하여 개발한 도구의 성능과 신뢰성을 확보하기 위한 노력도 함께 중요해지고 있으며 이러한 맥락에서 프롬프트 엔지니어링에 대한 연구가 주목을 받고 있다. 프롬프트 엔지니어링은 LLM의 극대화하기 위해 특정 작업에 최적화된 지시문을 전략적으로 설계하여 출력 결과를 안내하는 방법을 의미한다[1]. 초기의 프롬프트 엔지니어링에 대한 연구가 단순히 질문의 내용이나 구조의 최적화에 집중했다면 최근에는 프롬프트 형식이 LLM의 성능에 미치는 영향을 분석하는 방향으로 그 범위가 확장되고 있으며 프롬프트의 구조적 형식 변화가 모델의 추론 성능을 결정하는 변수임을 밝혀낸 He et al.(2024)의 연구를 통해 이러한 경향을 확인할 수 있다[2].

프롬프트 엔지니어링의 기술적 토대는 Vaswani et al.(2017)이 제안한 어텐션 메커니즘에서 마련되었다[3]. 이후 Brown et al.(2020)의 연구에서 제안한 제로샷(Zero-shot), 원샷(One-shot), 퓨샷(Few-shot)으로 대표되는 인컨텍스트 학습((In-context Learning, ICL)을 통해 프롬프트 엔지니어링에 대한 연구가 본격화되기 시작했다[4]. Wei et al.(2022)는 모델의 사고 과정을 단계별로 유도하는 사고의 연쇄(Chain-of-Thought, CoT) 기법을 소개하며 프롬프트의 구조적 설계가 모델의 잠재적 추론 능력을 극대화할 수 있음을 밝혔다[5]. Sclar et al.(2023)은 단일 형식의 프롬프트로 모델의 성능을 평가하던 기존 방법론의 한계를 지적하며 의미적으로 동일한(semantically equivalent) 다양한 프롬프트 형식에 따른 모델의 강건성(Robustness)을 측정하는 새로운 접근법인 FORMATSPREAD 방식을 제안하였다[6]. 하지만 Wei et al.(2022)의 연구는 특정한 프롬프팅 기법의 효과성에 대한 논의에 집중했으며 Sclar et al.(2023)의 연구는 표면적 · 구문론적 변화에만 초점을 맞춘 연구라는 한계가 있다. 또한, 프롬프트 구조 자체에 초점을 맞춘 He et al.(2024)의 연구 역시 영어로 된 과업만을 중심으로 진행되어 한국어와 같은 교착어(agglutinative language)의 구조적 특성이 다양한 프롬프트 형식 자체의 변화와 어떤 상호작용을 일으키는지에 대한 체계적 연구는 부족한 실정이다.

이에 본 연구는 선행 연구의 한계를 극복하고자 한국의 대학수학능력시험 국어와 영어 문항을 데이터셋으로 설정하여 프롬프트의 구조적 차이가 LLM의 강건성에 미치는 영향을 분석하고자 한다. 구체적으로는 다음과 같은 방식으로 연구를 진행하고자 한다. 먼저 네 가지 주요 프롬프트 형식(Markdown, JSON, YAML, XML)이 모델의 해결 정확성에 미치는 영향을 분석하고자 한다. 그리고 동일 프롬프트 반복에 대한 모델의 응답 일관성과 프롬프트 형식 변화에 따른 성능 안정성을 분석하고자 한다. 마지막으로 세 가지 분석 결과를 종합한 실용 강건성 지수(Practical Robustness Index, PRI)를 산출하여 프롬프트 형식에 대한 LLM의 강건성을 다각도로 평가하고 종합적인 통찰을 얻고자 한다. 본 연구는 교육용 AI 도구 개발 및 LLM의 교육 현장 적용 시 고려해야 할 강건성의 요소를 종합적으로 분석하고 이를 바탕으로 다양한 교육적 요구와 학습 환경에 부합하는 전략적 LLM 도입의 방향성을 제시하고자 한다.

2. 선행연구

2.1. AI 강건성

Chander et al.(2025)는 AI 강건성(Robustness)의 의미를 예측하지 못한 입력이나 변화하는 조건 속에서도 AI 시스템이 안정적이고 일관된 성능을 발휘할 수 있는 능력이라고 정의하였다[7]. AI 강건성은 신뢰할 수 있는 AI 시스템을 구축하는 핵심 요소로 작용한다. 또한 안전성, 공정성, 신뢰성을 보장하여 다양한 분야에서 AI 적용을 확대하고 사용자와 사회 전체의 신뢰를 얻는 데도 필수적인 요소라고 할 수 있다[8].

Li et al.(2023)의 연구에 따르면 AI 강건성은 데이터 분포 변화, 알고리즘에 대한 적대적인 공격, 시스템에 허용되지 않은 입력 등 다양한 문제의 영향을 받는다[9]. 즉 모델 학습을 위한 데이터 준비 단계에서부터 알고리즘 설계, 테스트와 검증, 시스템 운영에 이르는 전 단계에 걸쳐 강건성 저하에 대한 대비가 이루어져야 신뢰성 있는 AI 시스템을 구축할 수 있다. AI 시스템의 강건성을 평가하는 대표적인 방법 중 적대적 공격(Adversarial Attacks)은 입력 데이터에 손실 함수를 추가하여 모델의 취약점을 평가하는 방법이다. 이 방법은 모델의 취약점을 의도적으로 공격하여 강건성이 저하되는 측면을 파악하고 보완하기 위해 사용된다. 입력 변형(Input Perturbations)은 변형된 입력에 대한 모델 민감도를 평가하는 방법이다. 실제 사용자가 경험할 수 있는 자연스러운 변화에 대한 모델의 안정성을 측정한다는 점에서 실용적 의미가 큰 방법이다. 모델 수준 테스트(Model-Level Tests)는 하이퍼파라미터를 수정하거나 데이터에 분포 변화를 일으켜 이것이 모델의 성능에 어떤 영향을 미치는지 분석하는 방법이다[10]. 이 밖에도 이상 탐지(Anomaly Detection), 적대적 규제(Adversarial Regularization), 테스트 케이스 생성(Test Case Generation)을 통한 평가 방법 등이 사용된다. 최근에는 사람의 개입 없이도 AI가 자동으로 여러 가지 공격을 조합하여 평가하는 AutoAE(Automatically Constructed Attack Ensemble) 방식도 활발하게 사용되고 있다[11]. AI

시스템의 강건성에 대한 평가는 이러한 평가 기법들을 통합적이고 체계적으로 적용하는 과정을 통해 효과적으로 이루어질 수 있다.

2.2. 프롬프트 형식

프롬프트 형식은 사용자의 의도와 제공되는 정보 및 조건들을 LLM이 명확히 구분하여 처리하도록 돕는 요소이다. 본 연구에서는 네 가지 프롬프트 형식(Markdown, JSON, YAML, XML)을 선정하고 이들 형식의 구조적 차이가 LLM의 강건성에 미치는 영향을 분석하였다. 평가에 활용한 네 가지 프롬프트 형식의 구조는 Fig. 1 과 같다.

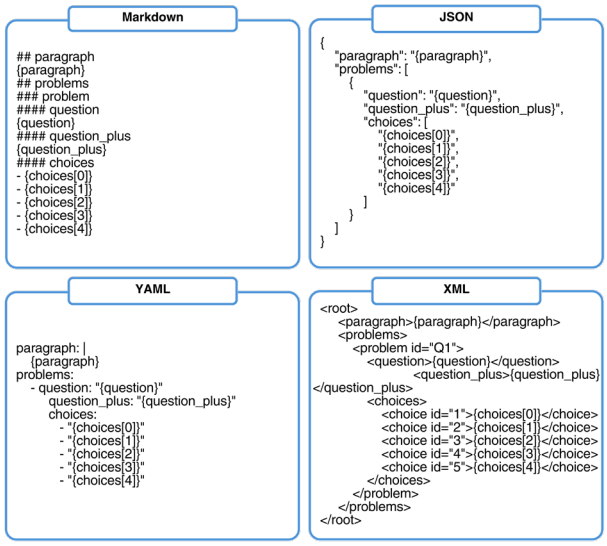

Figure 1. Prompt Template Used in the Study

Markdown은 헤더(#), 리스트, 인용구, 코드 블록을 이용한 구조를 가진 경량 마크업 언어이다. 자연어(Natural Language)와 유사한 서식으로 간결하고 직관성이 높으며 가독성이 뛰어나다. 또한 정보의 논리적 흐름을 표현하는 데도 효과적이다. 하지만 엄격한 키-값 구조가 없어 기계적 파싱에 한계가 있으며 복잡한 계층 구조를 표현하는 데 어려움이 있다[12]. JSON(JavaScript Object Notation)은 중괄호와 대괄호를 사용한 키-값 구조로 이루어진 텍스트 기반의 데이터 교환 포맷이다[13]. 명확한 구조 덕분에 기계적 파싱이 쉬우며 대부분의 프로그래밍 언어에서 지원되어 범용성이 높다. 하지만 구문 오류에 매우 민감하여 사소한 실수에도 전체 구조가 손상될 수 있다는 한계가 존재한다. YAML(YAML Ain't Markup Language)은 사람이 읽고 편집하기 쉽게 설계된 데이터 직렬화 포맷이다. 들여쓰기를 통해 프롬프트 구조를 표현하는 것이 특징이며 간결한 키-값 구조와 리스트 표기로 가독성이 높다[14]. 단, 들여쓰기나 공백에 민감하여 기계적 파싱 실패로 이어질 위험이 있다. XML(eXtensible Markup Language)은 데이터를 저장, 전송, 재구성하기 위한 마크업 언어의 일종이다. 문서의 구조와 내용을 태그로 표현하며 명시적인 태그 계층과 속성을 통해 스키마 검증이 가능해 입력 형식 일관성을 철저히 보장한다는 장점이 있다. 하지만 태그가 많아 코드가 장황해질 수 있으며 이에 따라 기계적 파싱 속도가 저하될 수 있다는 단점이 있다[15].

3. 연구방법

3.1 데이터셋

본 연구에서는 2025학년도 대학수학능력시험(이하 수능)의 국어와 영어 문항을 기반으로 데이터셋을 구축하였다. 모델의 강건성을 신뢰도 있게 평가하기 위해서는 심층적 독해 능력과 논리적 추론 능력을 동시에 측정할 수 있는 고도로 표준화된 데이터셋이 필요하다. 또한 평가 결과가 특정 언어에 국한된 결과가 아닌 한국어와 영어로 된 과업 모두에서 보편성 갖는 결과인지 검증하기 위해서는 두 가지 언어로 구성된 데이터셋이 필요하다. 특히, 수능 영어는 그 자체로 복합적인 언어 능력을 요구하는 특징을 가지고 있다. 일부 지시문과 보기, 선택지 모두가 영어로만 이루어진 문항을 제외하고 대부분의 문항은 지시문과 선택지는 한국어, 보기는 영어로 구성되어 있다. 이러한 구성은 LLM이 단순히 영어 문항을 이해하고 해석하는 수준을 넘어 두 언어의 맥락을 동시에 이해하며 문제를 해결하는 고차원적인 추론 능력을 평가할 수 있도록 돕는다.

최종적으로 본 연구는 국어(화법과 작문 포함) 45문항과 영어(듣기 평가 제외) 28문항을 네 가지 프롬프트 형식에 맞게 변환하여 평가에 활용하였다. 문항은 '지문, 지시문, (필요시) 보기, 선택지'의 구조로 이루어져 있으며 보기에는 그림이나 그래프와 같은 시각 자료가 포함되기도 한다. 이 경우 해당 자료의 내용을 텍스트로 상세히 기술하여 LLM이 모든 정보를 참조하고 문제를 해결할 수 있도록 하였다.

3.2 평가 모델

평가 모델은 연구 시점에서 API를 통해 접근 가능하며 학계와 산업계에서 널리 사용되고 있는 모델을 우선적으로 고려하였다. 이에 따라 OpenAI의 GPT 계열 모델 7개, Google의 Gemini 계열 모델 7개, Anthropic의 Claude 계열 모델 4개, 총 18개의 해외 주요 모델을 선정하였다. 이와 더불어 소버린 AI(Sovereign AI)의 중요성을 인식하여 한국어 및 한국 문화에 특화된 성능을 확보하였으며 영어와 코드 관련 데이터도 균형 있게 학습한 Naver의 HyperCLOVAX(HCX) 계열 모델 2개를 추가로 선정하였다[16]. 그 결과 최종적으로 20개의 모델을 선정하였으며 계열별 모델 리스트는 Table 1 과 같다.

Table 1. List of LLM Model Used in the Study

OpenAI (GPT Family)	Google (Gemini Family)
GPT-4.1 GPT-4.1-Mini GPT-4.1-Nano GPT-4o GPT-4o-Mini GPT-4-Turbo GPT-3.5-Turbo	Gemini-2.5-Pro Gemini-2.5-Flash Gemini-2.0-Flash Gemini-2.0-Flash-Lite Gemini-1.5-Pro Gemini-1.5-Flash Gemini-1.5-Flash-8b
Anthropic (Claude Family)	**Naver (HyperCLOVAX (HCX) Family)**
Claude-3.7-Sonnet Claude-3.5-Sonnet Claude-3-Opus Claude-3-Haiku	HCX-003 HCX-DASH-001

모델 선정 시 각 계열 내에서 플래그십(flagship) 모델, 경량화(lightweight) 모델, 파생(derivation) 모델 등 다양한 하위 모델들을 폭넓게 포함하였다. 이러한 구성은 단순히 계열 간의 성능을 비교하는 것을 넘어 동일 계열 내에서도 모델의 크기나 최적화 방향과 같은 세부적인 특성이 강건성에 미치는 영향을 세밀하게 관찰할 수 있다는 장점이 있다. 또한 LLM 전반에 걸쳐 나타나는 성능 변화 양상을 포착하고 그 원인을 다각도로 추론하는 기반을 마련할 수 있다는 장점이 있다.

3.3 하이퍼파라미터 설계

본 연구에서는 응답의 무작위성과 다양성을 제어하고 일관성을 확보하기 위해 Table 2 와 같이 하이퍼파라미터(Hyperparameter)를 설정하였다.

Table 2. Hyperparameter Setting Value

Hyperparameter	Set Value
temperature	0 (0.01 for HCX models)
top_p	0.1
top_k	API Default

temperature 값은 확률 분포의 창의성과 무작위성을 조절하여 다음 토큰을 선택할 때 영향을 주는 값이다[17]. temperature 값이 0에 가까울수록 동일한 응답을 출력할 가능성이 높아진다[18]. 본 연구에서는 응답의 일관성을 극대화하기 위해 temperature 값을 0으로 설정하였다. 단, HCX 모델의 경우 API가 허용하는 최솟값인 0.01로 설정하였다. top_p 값은 누적 확률이 top_p 값 이상이 되는 가장 가능성 높은 토큰들의 집합에서 다음 토큰을 샘플링하는 데 사용된다[19]. 본 연구에서는 응답의 집중도를 높이고 예측 가능성을 최대화하기 위해 top_p 값을 0.1로 설정하였다. 이는 가장 유력한 토큰 후보만을 고려하도록 하여 temperature를 낮게 설정한 효과와 함께 응답의 일관성을 강화하는데 기여한다. top_k 값은 다음 토큰을 선택할 때 고려하는 가장 가능성이 높은 토큰의 수를 k개로 제한한다[19]. 이 값은 일부 모델(Gemini 계열)에서만 명시적 설정이 가능하므로 실험의 공정성을 확보하고 모델 본연의 성능을 평가하기 위해 특정 값으로 통일하는 대신 모델별 기본값으로 설정하였다.

3.4. 평가 지표

본 연구에서 LLM의 성능을 다각도로 평가하기 위해 기존 강건성의 개념을 확장하여 설계한 평가 프레임워크는 Table 3 과 같다. 강건성은 일반적으로 외부 요인 변화에 대한 모델의 일관성(Consistency)과 안정성(Stability)을 중심으로 논의된다. 그러나 이러한 기술적 관점은 모델이 안정적으로 틀린 답변을 내놓을 때도 강건하다고 평가하는 근본적인 한계를 가진다. 이는 실제 사용 환경에서 모델의 유용성을 결정하는 요소인 정확성(Accuracy)을 간과하기 때문이다. 따라서 본 연구에서는 일관성과 안정성에 더불어 정확성을 강건성의 개념에 포함하고 이를 실용적 강건성으로 명명하였다. PRI는 이 세 가지 차원을 종합적으로 평가함으로써 모델의 실질적인 신뢰도와 유용성을 보다 균형 있게 측정할 수 있다. 각 지표는 특정 모델(m), 특정 문항(i)에 대해 계산되며 4개의 프롬프트 형식(f)과 3회의 반복 실험(k)을 통해 얻은 데이터를 기반으로 한다.

Table 3. LLM Evaluation Metrics Framework

Dimension	Metric	Description
Accuracy	Solution Accuracy (SA)	Accuracy for each individual prompt format
Consistency	Response Consistency (RC)	Reliability of identical answers under repeats
Stability	Performance Stability across Formats (PSF)	Uniformity of accuracy across prompt formats
Robustness	Practical Robustness Index (PRI)	Composite robustness score over SA_{avg}, RC_{avg}, and PSF

해결 정확성(Solution Accuracy, SA)은 특정 프롬프트 형식에 대한 모델의 정답률을 나타내는 지표로서 (1) 과 같은 방법으로 산출된다.

$$SA_{m,f} = \frac{1}{3N}\sum_{k=1}^{3}\sum_{i=1}^{N} S_{m,i,f,k} \quad (1)$$

$$S_{(m,i,f,k)} = \begin{cases} 1 \text{ if } correct \\ 0 \text{ if } incorrect \end{cases}$$

산출된 정답률을 이용하여 네 가지 프롬프트 형식 간 해결 정확성의 전반적인 차이는 크루스칼-월리스 검정(Kruskal-Wallis Test)을 통해, 쌍별 차이는 본페로니 보정(Bonferroni Correction)을 적용한 맨-휘트니 U 검정(Mann-Whitney U test)으로 분석하였다. 각 분석의 효과 크기는 에타 제곱(eta-squared)과 코헨의 d(Cohen's d)를

통해 측정하였다.

응답 일관성(Response Consistency, RC)은 동일한 입력 조건 아래에서 모델이 얼마나 재현성 있게 응답을 생성하는지 측정하는 지표이다. 이는 해결 정확성(SA)이 높더라도 생성 결과가 일관되지 않으면 강건성이 떨어진다는 점에서 설정한 지표이다. RC 점수는 3회 반복 실험으로 얻은 3개의 응답 텍스트 간의 일치 여부를 1점(일치) 또는 0점(불일치)로 치환한 뒤 (2) 와 같은 방식으로 산출된다.

$$RC(m,f) = \frac{1}{3N}\sum_{i=11}^{N}\sum_{j<l\le 3} I(R_{m,i,f,j}, R_{m,i,f,l}) \quad (2)$$

$$I(R_a,R_b) = \begin{cases} 1 \text{ if } R_a = R_b \\ 0 \text{ if } R_a \neq R_b \end{cases}$$

성능 안정성(Performance Stability across Formats, PSF)은 프롬프트 형식 변화에 대한 모델 성능의 일관성을 평가하는 지표이다. PSF 점수는 SA 점수들의 표준편차()를 계산한 뒤 (3) 과 같은 공식을 통해 산출된다. PSF 점수는 1에 가까울수록 형식 간 성능이 균일하여 안정성이 높음을 의미한다.

$$PSF_m = 1 - \sigma_{SA_m} \quad (3)$$

실용 강건성 지수(Practical Robustness Index, PRI)는 세 가지 핵심 차원(SA, RC, PSF)을 종합적으로 고려하여 산출되는 값으로 모델의 절대적 서열을 보여주는 값이 아닌 특정 환경에서 최적의 모델을 탐색하고 선택하도록 돕는 위한 하나의 발견적 도구(heuristic tool)로서 기능하는 지표이다. PRI는 세 지표(, , PSF)를 기하평균으로 결합하여 산출한다. 이는 분산이 가장 큰 해결 정확성(SA)이 PRI 값에 과도한 영향을 미치는 것을 방지하기 위함으로 특정 지표가 낮을 때 총점이 함께 하락하는 기하평균의 특성을 활용하여 균형 잡힌 평가를 유도한다. 이러한 접근은 유엔(UN)이 인간개발지수(HDI)에서 산술평균의 한계를 지적하며 기하평균을 도입했던 것과 궤를 같이한다[20]. 본 연구의 모든 지표는 0과 1 사이의 값으로 정의되어 있으므로 별도의 스케일링 없이 (4) 와 같이 기하평균을 적용하였다.

$$PRI_m = \sqrt[3]{SA_{avg_m} \times RC_{avg_m} \times PSF_m} \quad (4)$$

Table 4. Statistical Summary of Solution Accuracy (SA) across Prompt Formats

Model	Korean				English			
	SA_{avg}	SA Range [Min, Max]	H-statistic	p-value	SA_{avg}	SA Range [Min, Max]	H-statistic	p-value
GPT-4.1	0.791	[0.778, 0.800]	0.249	.969	0.810	[0.786, 0.821]	0.168	.983
GPT-4.1-Mini	0.800	[0.785, 0.822]	0.324	.955	0.830	[0.821, 0.857]	0.023	.990
GPT-4.1-Nano	0.794	[0.778, 0.800]	0.224	.974	0.616	[0.583, 0.643]	0.252	.969
GPT-4o	0.800	[0.785, 0.807]	0.203	.977	0.786	[0.750, 0.857]	1.564	.668
GPT-4o-Mini	0.793	[0.770, 0.807]	0.263	.967	0.708	[0.679, 0.774]	0.599	.897
GPT-4-Turbo	0.794	[0.770, 0.807]	0.230	.973	0.717	[0.679, 0.750]	0.358	.949
GPT-3.5-Turbo	0.794	[0.770, 0.807]	0.321	.956	0.622	[0.571, 0.679]	0.842	.839
Gemini-2.5-Pro	0.891	[0.874, 0.919]	0.605	.895	0.946	[0.917, 0.964]	1.029	.794
Gemini-2.5-Flash	0.850	[0.815, 0.889]	1.480	.687	0.902	[0.893, 0.929]	0.300	.960
Gemini-2.0-Flash	0.835	[0.815, 0.852]	0.397	.941	0.789	[0.750, 0.810]	0.306	.959
Gemini-2.0-Flash-Lite	0.759	[0.667, 0.815]	2.748	.432	0.768	[0.726, 0.833]	0.924	.820
Gemini-1.5-Pro	0.872	[0.852, 0.889]	0.681	.878	0.839	[0.821, 0.857]	0.262	.967
Gemini-1.5-Flash	0.746	[0.719, 0.778]	0.773	.856	0.771	[0.714, 0.821]	1.062	.786
Gemini-1.5-Flash-8b	0.646	[0.585, 0.689]	1.164	.762	0.732	[0.679, 0.821]	1.985	.575
Claude-3-Opus	0.767	[0.689, 0.844]	3.160	.368	0.929	[0.893, 0.964]	1.067	.785
Claude-3.7-Sonnet	0.898	[0.881, 0.911]	0.540	.910	0.938	[0.893, 0.964]	1.661	.646
Claude-3.5-Sonnet	0.904	[0.881, 0.900]	0.672	.880	0.866	[0.821, 0.893]	0.839	.840
Claude-3-Haiku	0.587	[0.541, 0.607]	0.513	.916	0.786	[0.750, 0.821]	0.420	.936
HCX-003	0.659	[0.615, 0.689]	0.591	.898	0.616	[0.500, 0.786]	5.799	.122
HCX-DASH-001	0.287	[0.156, 0.348]	6.218	.101	0.527	[0.250, 0.750]	19.985	.000***

*p < .05, **p < .01, ***p < .001*

4. 연구결과

4.1. 해결 정확성(SA) 분석

모델별 해결 정확성을 계산하고 그 결과가 통계적으로 유

의한 지 검증한 결과는 Table 4 와 같다. 분석 결과 HCX-DASH-001 모델을 제외한 모든 모델의 형식 간 해결 정확성 차이는 통계적으로 유의하지 않았지만 계열별, 그룹별 편차는 존재하는 것으로 나타났다.

플래그십 모델 그룹은 대체로 프롬프트 형식 변화에도 불구하고 높은 해결 정확성을 보였다. Gemini 계열의 플래그십 모델인 Gemini-2.5-Pro은 국어와 영어 모두에서 가장 높은 해결 정확성을 보였다. 최고 해결 정확성과 최저 해결 정확성 간의 차이(SA)가 가장 작은 모델은 국어에서는 Claude-3.5-Sonnet(SA_국어=0.019)이었으며 영어에서는 GPT-4.1(SA_영어=0.035)이었다. 반면 Gemini-1.5-Flash(SA_국어=0.059, SA_영어=0.107), GPT-4o(SA_국어=0.022, SA_영어=0.107), GPT-4o-Mini(SA_국어=0.037, SA_영어=0.095) 등은 국어에서는 프롬프트 형식 간 해결 정확성의 차이가 작았지만, 영어에서는 상대적으로 프롬프트 형식 간 해결 정확성의 차이가 크게 벌어졌다.

전체 모델 가운데 HCX-DASH-001만이 통계적 유의성을 나타냈지만, H값의 경우 일부 경량화 모델들이 높은 값을 기록하며 프롬프트 형식에 민감한 것으로 나타났다. 구체적으로 Gemini-2.0-Flash-Lite, Claude-3-Opus 모델은 국어에서, HCX-003은 영어에서, HCX-DASH-001 모델은 국어와 영어 모두에서 높은 H값을 기록하며 프롬프트 형식에 민감하게 반응하는 것으로 나타났다. 이는 이들 모델이 특정 프롬프트 구조에 크게 의존하고 있으며 프롬프트 구조에 따라 성능이 급격하게 저하될 수 있음을 의미한다.

FIg. 2와 3은 국어와 영어 환경에서 각 모델의 프롬프트 형식 쌍별 효과 크기(Cohen's d)를 비교한 결과이다. 국어 환경(Fig. 2)의 평균 효과 크기는 0.085로 대부분 모델이 0점 기준선에 밀집되어 있는 것을 확인할 수 있다. 이는 국어 환경에서는 프롬프트 형식 변화가 모델의 해결 정확성에 미치는 영향이 적다는 것을 의미한다. 반면 영어 환경(Fig. 3)의 평균 효과 크기는 0.127로 국어에 비해 0점 기준선에서 좌우로 넓게 분포해 있는 것을 확인할 수 있다. 분산() 역시 국어(0.0147)에 비해 영어(0.0271)가 약 1.84배 높아 영어 환경에서 프롬프트 형식 효과의 변동성이 더 크게 나타났다. 특히 HCX 계열의 경우 JSON vs Markdown과 JSON vs XML 비교에서는 왼쪽(음수)으로 크게 치우쳐져 있으며 Markdown vs YAML과 XML vs YAML 비교에서는 오른쪽(양수)으로 치우친 양상을 보였다. 이는 HCX 계열 모델의 경우 프롬프트 형식 변화에 민감하게 반응함을 보여준다.

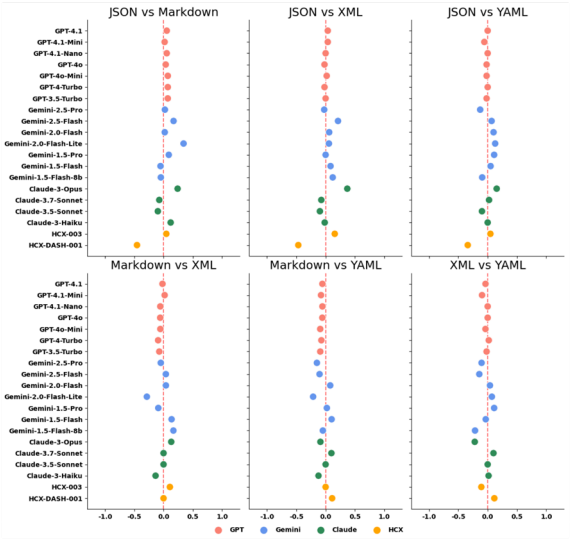

Figure 2. Effect Size Distribution of Prompt Format Comparisons in Korean

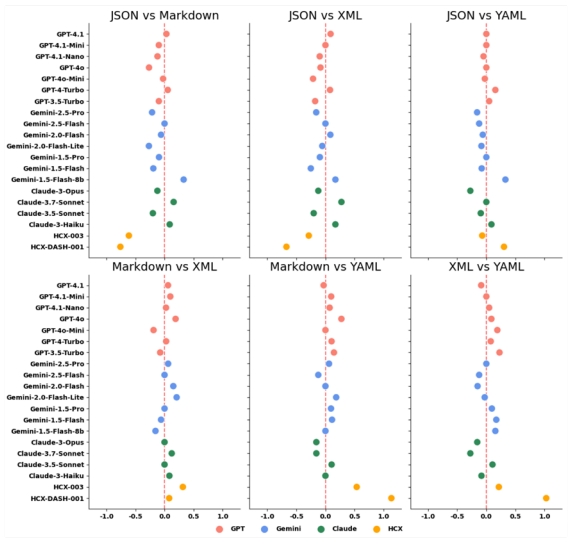

Figure 3. Effect Size Distribution of Prompt Format Comparisons in English

4.2. 응답 일관성(RC) 분석

네 가지 프롬프트 형식에 대한 20개 모델의 응답 일관성(RC)을 히트맵으로 표현한 결과는 Fig. 4, 5 와 같다. 응답 일관성 분석 결과 모델 계열에 따라 계층이 형성되는 모습이 보였으며 최적의 프롬프트 형식은 언어에 따라 다른 것으로 나타났다. 두 언어 모두 Claude 계열 모델들이 가장 높은 RC 점수를 기록하며 최상위 그룹을 차지했고 뒤이어 Gemini, GPT, HCX 계열 순으로 RC 수준이 계층적으로 구분되었다. 특히 국어에서는 Claude-3.5-Sonnet(M=0.996)이 가장 높은 응답 일관성을 보였으며 영어에서는 Claude 계열 전체와 Gemini-1.5-Flash-8b 모델이 완벽한 응답 일관성(M=1.000)을 보이며 결정론적(deterministic)으로 작동하는 특성을 나타냈다.

프롬프트 형식에 따른 응답 일관성은 언어별로 다른 양

상을 보였다. 국어에서는 YAML 형식(M=0.975, SD=0.027)이 가장 높은 응답 일관성을 보였고 JSON(M=0.970, SD=0.056), Markdown(M=0.959, SD=0.058), XML(M=0.947, SD=0.074)이 뒤를 이었다. 반면 영어에서는 XML(M=0.977, SD=0.034)이 가장 일관된 모습을 보였고 뒤이어 Markdown(M=0.973, SD=0.031), YAML(M=0.956, SD=0.073), JSON(M= 0.954, SD=0.077) 순으로 높은 응답 일관성을 보였다. 이러한 결과는 모든 모델과 다양한 언어에 보편적으로 우월한 특정 프롬프트 형식은 존재하지 않으며 최적 프롬프트 형식은 각 언어의 고유한 통사적 특성과 프롬프트의 구조적 표현 방식 간의 상호작용으로 결정됨을 의미한다. 따라서 LLM의 응답 일관성을 극대화하기 위해서는 대상 언어에 가장 적합한 프롬프트 형식을 선택해야 함을 시사한다.

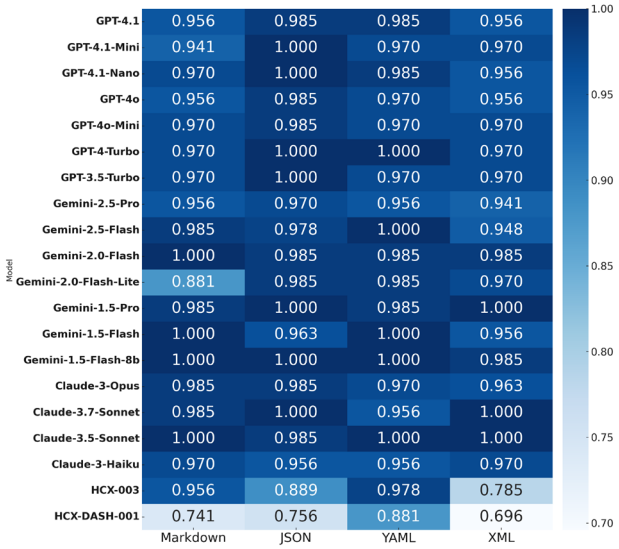

Figure 4. Analysis of Response Consistency (RC) by Model & Prompt Format in Korean

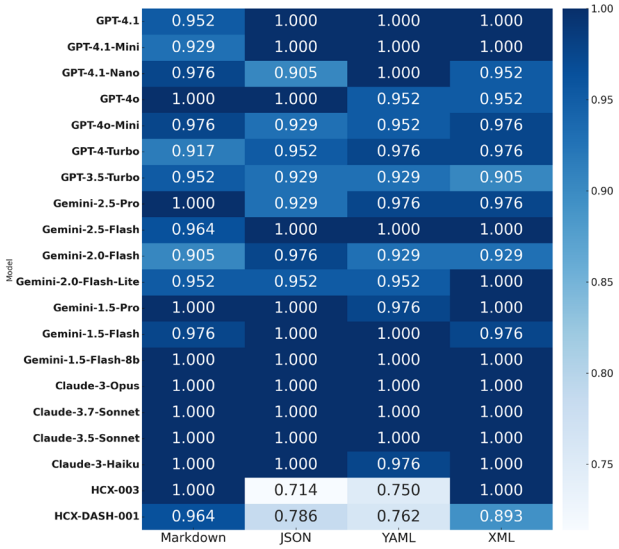

Figure 5. Analysis of Response Consistency (RC) by Model & Prompt Format in English

4.3. 성능 안정성(PSF) 및 실용적 강건성 지수(PRI) 분석

프롬프트 형식 변화에 따른 모델의 성능 안정성(PSF)과 실용적 강건성 지수(PRI)를 산출한 결과는 Table 5 와 같다. 각 모델의 성능 안정성(PSF)을 분석한 결과 Claude-3.7-Sonnet, Gemini-2.5-Pro, GPT-4.1 등 각 계열의 플래그십 모델은 국어와 영어 모두에서 높은 PSF 점수를 기록하여 프롬프트 형식 변화와 상관없이 안정적인 성능을 내는 것으로 나타났다. 반면 HCX 계열 모델과 일부 경량화 모델들은 낮은 PSF 점수를 기록하여 프롬프트 형식 변화에 굉장히 민감한 것으로 나타났다. 이는 고성능의 플래그십 모델일수록 프롬프트의 표면적 구조가 모델의 핵심 추론 능력에 크게 영향을 주지 않는다는 것을 의미한다. 반면 경량화된 모델은 추론 과정이 특정 프롬프트 구조와 강하게 결합되어 있거나 플래그십 모델 대비 기계적 파싱 능력이 상대적으로 떨어져 형식 변화에 더 민감한 결과가 나온 것으로 해석된다. 언어 간 PSF 점수를 비교 분석한 결과 Gemini 계열은 국어보다 영어에서 전반적으로 더 높은 성능 안정성을 나타냈다. HCX 계열은 국어에서 더 높은 성능 안정성을 나타냈다. 반면 GPT와 Claude 계열은 성능 안정성 측면에서 특정 언어에 대한 일관된 경향성을 보이지 않았다. 예를 들어, GPT-4.1은 영어에서 더 높은 성능 안정성을 보였지만 GPT-3.5-Turbo는 국어에서 더 높은 성능 안정성을 보였다. Claude 계열 역시 Sonnet 모델들은 국어에서 더 높은 성능 안정성을 보였지만 Opus 및 Haiku 모델은 영어에서 우세를 보이는 등 계열 내 패턴이 나타나지 않았다.

실용 강건성 지수(PRI) 분석 결과 Claude 계열과 Gemini 계열이 최상위 그룹을 형성하였으며 GPT 계열, HCX 계열이 뒤를 이었다. Claude 계열의 Claude-3.7-Sonnet과 Claude-3.5-Sonnet 모델은 국어와 영어 모두에서 최고점에 가까운 점수를 기록하며 언어와 관련 없이 가장 뛰어난 실용 강건성을 가진 모델임을 입증했다. Gemini 계열 모델들도 전반적으로 높은 PRI 점수를 나타냈다. Gemini 계열의 플래그십 모델인 Gemini-2.5-Pro는 Cluade-3.7-Sonnet과 더불어 가장 높은 PRI 점수를 나타냈다. 특히 주목할 점은, 이전 세대의 아키텍처를 사용한 Gemini-1.5-Pro와 속도와 비용의 측면에서 Gemini-2.5-Pro를 최적화한 Gemini-2.5-Flash 역시 매우 높은 PRI 점수를 기록했다는 점이다. GPT 계열 모델 중 GPT-4.1도 플래그십 모델답게 높은 PRI 점수를 기록하였다. 하지만 경량화 모델인 GPT-4.1-Mini의 PRI 점수가 GPT 계열 내에서 가장 높게 산출되어 경량화 또는 파생 모델이 과업을 수행하는 환경에 따라 플래그십 모델을 능가할 수 있음을 시사한다. 다만 일부 경량화 모델은 매우 낮은 PRI 값을 나타냈으며 이는 최적화 과정에서 성능 손실이 있었을 것이라고 해석할 수 있다. 마지막으로 HCX 계열의 두 모델은 실용 강건성의 핵심 지표인 해결 정확성(SA)에서 낮은 점수를 받아 20개의 모델 중 하위권에 머물렀다.

Table 5. Performance Stability across Formats (PSF) & Practical Robustness Index (PRI) Scores by Model & Language

Model	PSF		PRI		
	Kor	Eng	Kor	Eng	Avg
GPT-4.1	0.936	0.969	0.900	0.921	0.910
GPT-4.1-Mini	0.938	0.943	0.904	0.919	0.911
GPT-4.1-Nano	0.940	0.918	0.904	0.824	0.864
GPT-4o	0.943	0.927	0.905	0.896	0.901
GPT-4o-Mini	0.933	0.944	0.901	0.868	0.884
GPT-4-Turbo	0.936	0.952	0.904	0.873	0.889
GPT-3.5-Turbo	0.950	0.884	0.907	0.809	0.858
Gemini-2.5-Pro	0.944	0.980	0.937	0.970	0.954
Gemini-2.5-Flash	0.929	0.923	0.922	0.941	0.932
Gemini-2.0-Flash	0.940	0.949	0.921	0.898	0.909
Gemini-2.0-Flash-Lite	0.872	0.897	0.868	0.878	0.873
Gemini-1.5-Pro	0.933	0.969	0.932	0.932	0.932
Gemini-1.5-Flash	0.901	0.946	0.873	0.898	0.886
Gemini-1.5-Flash-8b	0.893	0.889	0.832	0.867	0.849
Claude-3-Opus	0.919	0.933	0.887	0.953	0.920
Claude-3.7-Sonnet	0.984	0.967	0.957	0.968	0.963
Claude-3.5-Sonnet	0.968	0.936	0.956	0.932	0.944
Claude-3-Haiku	0.884	0.951	0.798	0.907	0.853
HCX-003	0.855	0.833	0.824	0.801	0.812
HCX-DASH-001	0.769	0.736	0.596	0.729	0.662

Fig. 6 은 가로축을 국어 PRI 점수, 세로축을 영어 PRI 점수로 설정한 뒤 LLM의 언어별 실용 강건성 지수(PRI)를 나타낸 그래프이다. 우측 상단에 위치할수록 두 언어 모두에서 높은 실용 강건성을 나타내며 좌측 하단에 위치할수록 두 언어 모두에서 낮은 실용 강건성을 나타낸다. y=x는 언어 간 균형을 나타내는 기준선이다. Claude 계열(초록색)과 Gemini 계열(파란색)은 그래프의 우측 상단에 모여 있어 가장 뛰어난 실용 강건성을 나타내고 있다. GPT 계열(빨간색)은 위 두 그룹의 약간 아래쪽에 위치하고 있으며 세로로 길게 펼쳐져 있어 국어에서는 안정적인 실용 강건성을 나타내고 있으나 영어에서는 실용 강건성의 편차가 큰 것으로 나타났다. HCX 계열(노란색)은 좌측 하단에 위치하고 있어 국어와 영어 환경 모두에서 실용 강건성이 떨어지는 것으로 나타났다.

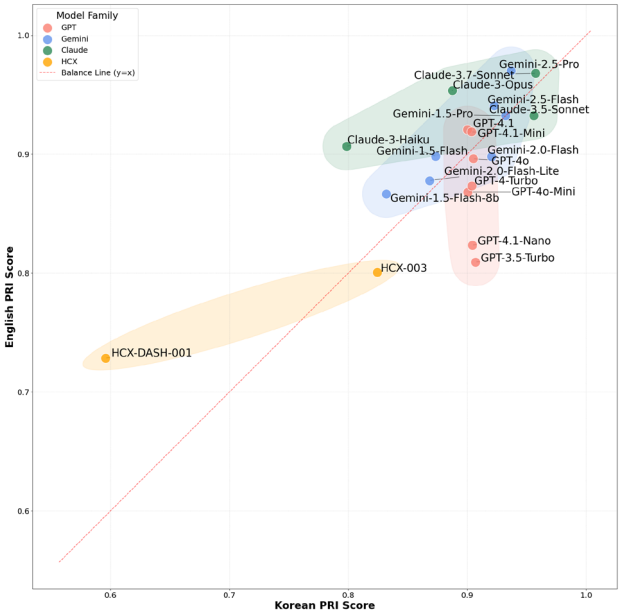

Figure 6. Distribution of LLMs by Korean vs English PRI Scores

5. 결론 및 제언

5.1. 결론

본 연구에서는 주요 대규모 언어 모델(LLM)을 대상으로 프롬프트 형식 변화가 모델의 강건성에 미치는 영향을 다각도로 분석하였다. 이를 위해 2025학년도 대학수학능력시험 국어 및 영어 문항 데이터셋을 활용하여 각 모델의 해결 정확성(SA), 응답 일관성(RC), 성능 안정성(PSF)을 측정하였다. 최종적으로 세 지표를 종합한 실용 강건성 지수(PRI)를 산출하여 모델의 성능을 입체적으로 평가하였으며 주요 분석 결과와 시사점은 다음과 같다. 첫째, 모델 성능에 따라 프롬프트 형식에 대한 민감도가 다르게 나타났다. 해결 정확성(SA)과 성능 안정성(PSF) 분석 결과 고성능 플래그십 모델은 프롬프트 형식 변화에 거의 영향을 받지 않았지만, 경량화 모델과 일부 모델은 프롬프트 형식에 따라 성능이 크게 좌우되는 형식 의존적 특성을 보였다. 이러한 결과는 모델 성능을 안정적으로 확보하기 위해서는 모델별 특성을 고려한 프롬프트 형식 최적화가 필요함을 시사한다. 둘째, 과업의 언어에 따라 최적의 프롬프트 형식이 다른 것으로 나타났다. 응답 일관성(RC) 측면에서 계열별 계층 구조가 형성되었으며 최적의 프롬프트 형식 또한 언어에 따라 다른 것으로 나타났다. 전반적으로 국어에서는 YAML 형식이, 영어에서는 XML 형식이 가장 안정적이었다. 본 연구에서 프롬프트 형식에 따른 지문, 지시문, 보기, 선택지의 구조적 차이는 없으며 각 프롬프트 형식이 이들 요소의 경계를 구분하는 방식에만 차이가 존재한다. 이러한 차이가 LLM이 구성 요소별 경계를 인식하고 파싱하는 능력에 영향을 미쳐 이와 같은 결과를 가져왔을 것이라고 판단된다. 따

라서 모델의 성능을 극대화하기 위해서는 과업의 언어적 특성을 고려한 프롬프트 형식을 선택하는 것이 중요하다는 시사점을 얻을 수 있다. 셋째, 본 연구에서 제안한 실용 강건성 지수(PRI)는 모델의 강건성을 실용적 관점에서 다각도로 진단하는 유용한 평가 도구가 될 수 있음을 확인하였다. 이는 단순히 개별 평가 지표에서 높은 순위를 차지한 모델이 강건한 모델이 아니라 여러 핵심 차원에서 균형을 이룬 모델이 실용적인 측면에서 강건하다는 새로운 관점을 제시한다. 특히, 과업 수행을 위한 모델 선택 시 고성능 플래그십 모델만을 고려하는 것이 아닌 비용 대비 가장 실용적 강건성이 높은 모델을 전략적으로 선택할 수 있는 합리적 의사 결정의 근거를 제공한다. 구체적으로, LLM을 활용한 교육 시스템 개발과 같은 과업에서 프롬프트 형식과 호출 비용, 과업의 언어 등 다양한 요소를 종합적으로 고려하여 어떤 LLM을 어떤 형식의 프롬프트로 설계하고 활용할 것인지 대한 최소한의 기준을 제시해 주었다는 점에서 그 의의가 있다. 넷째, 소버린 AI 개발의 필요성 및 전략적 과제를 확인하는 계기가 되었다. 글로벌 모델들은 프롬프트 형식 변화에도 비교적 안정적인 성능을 보였으나 HCX 계열 모델은 프롬프트 형식 변화에 민감하게 반응하며 세 가지 지표와 실용적 강건성의 측면에서 전반적으로 낮은 성능을 나타냈다. 이는 단순히 학습 데이터의 규모뿐만 아니라 다양한 프롬프트 구조와 문맥을 안정적으로 처리할 수 있는 역량 강화가 필요함을 시사한다. 따라서 소버린 AI 개발은 타 모델 대비 한국어 데이터의 비중이 높다는 점을 강점으로 하여 다양한 프롬프트 형식에 대응할 수 있는 실용적 강건성을 확보하는 방향으로 이루어져야 할 것이다.

본 연구는 한국어 중심 과업 수행에서 프롬프트 형식이 LLM의 강건성에 미치는 영향을 체계적이고 정량적인 분석의 영역으로 가져왔다는 점에서 의의를 지닌다. 또한, LLM 기반 교육용 AI 도구의 실용적 강건성 확보를 위한 실증적 평가 결과를 바탕으로 교육 현장에 적합한 체계적이고 효과적인 LLM 도입 전략 수립에 중요한 근거를 제공한다. 다만 본 연구는 다음과 같은 한계가 있다. 첫째, 본 연구 결과는 분석적, 추론적 과업에 대한 강건성의 측면만을 보여주고 있어 창의적인 글쓰기나 코드 생성과 같은 과업에서는 다른 결과가 나타날 수 있다. 둘째, 본 연구는 API 호출 비용과 시간이라는 현실적인 제약으로 인해 빠르게 발전하는 LLM 생태계의 모든 모델 및 최근 출시된 모델들을 평가에 포함하지 못했다.

5.2. 제언

프롬프트 형식에 따른 LLM의 실용적 강건성 확보를 위한 후속 연구의 제언은 다음과 같다. 첫째, 다양한 유형의 과업에서 프롬프트 형식과 실용적 강건성 간의 상관관계를 분석할 필요가 있다. 이를 통해 과업의 유형과 프롬프트 형식 간의 상관관계를 밝혀낼 수 있을 것이다. 둘째, 각 프롬프트 형식의 구조가 모델의 실용적 강건성에 미치는 영향에 대한 보다 체계적인 분석이 필요하다. 이를 위해서 형식별 토큰화 방식, LLM의 종류에 따른 사전 학습 과정에서의 프롬프트 형식 노출 빈도, 입력 구조에 따른 어텐션 분포의 차이 등과 같은 요인이 모델 응답에 미치는 영향을 규명할 수 있는 정교한 후속 실험이 필요하다. 셋째, 설명 가능한 AI(Explainable AI, XAI)를 활용하여 특정 언어에서 모델과 프롬프트 형식 간의 상호작용의 원리를 밝혀내고 이를 통해 모델의 강건성을 저해하는 측면을 찾고 보완하는 것에 관한 연구가 이루어질 필요가 있다. 넷째, 최신 모델 및 오픈 소스 모델의 강건성을 평가하고 그 결과를 체계적으로 정리할 필요가 있다. 다섯째, 개발 목적에 맞는 최적의 LLM과 그에 최적화된 프롬프트 형식을 선택하는 명확한 기준에 대한 연구가 필요하며 나아가 사용자가 입력한 내용을 모델별 특성에 맞게 최적의 프롬프트 형식으로 자동 변환해 주는 시스템 개발에 대한 논의가 필요하다. 이상의 후속 연구들은 사용자가 자신의 목적에 맞는 실용적 강건성을 갖춘 모델들을 선택하고 활용하는데 기여할 것으로 기대된다.

참고문헌

[1] Sahoo, P., Singh, A. K., Saha, S., Jain, V., Mondal, S., & Chadha, A. (2024). A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint* arXiv:2402.07927. https://doi.org/10.48550/arXiv.2402.07927

[2] He, J., Rungta, M., Koleczek, D., Sekhon, A., Wang, F. X., & Hasan, S. (2024). Does Prompt Formatting Have Any Impact on LLM Performance?. *arXiv preprint* arXiv:2411.10541. https://doi.org/10.48550/arXiv.2411.10541

[3] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *In Advances in Neural Information Processing Systems*, 30, 5998–6008. Curran Associates, Inc. https://doi.org/10.48550/arXiv.1706.03762

[4] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan , J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert- Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu , J., Winter, C., ... Amodei, D. (2020). Language models are few- shot learners. *Advances in neural information processing systems*, 33, 1877- 1901. https://doi.org/10.48550/ arXiv.2005.14165

[5] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2023). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35, 24824–24837. https://doi.org/10.48550/arXiv.2201.11903

[6] Sclar, M., Choi, Y., Tsvetkov, Y., & Suhr, A. (2023). Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting. *arXiv preprint* arXiv:2310.11324. https://doi.org/10.48550/arXiv.2310.11324

[7] Chander, B., John, C., Warrier, L., & Gopalakrishnan, K. (2025). Toward trustworthy artificial intelligence (TAI) in the context of explainability and robustness. *ACM Computing Surveys*, 57(6), 1-49. https://doi.org/10.1145/3675392
[8] Heo, Y., Lee, J., Lee, G., Kim, E., Jeong, H., & Cho, W. (2024). Understanding and Practice of AI Trustworthiness. *Chung Ram Publishing*.
[9] Li, B., Qi, P., Liu, B., Di, S., Liu, J., Pei, J., & Zhou, B. (2023). Trustworthy AI: From principles to practices. *ACM Computing Surveys*, 55(9), 1–46. https://doi.org/10.1145/3555803
[10] Kumar, P., Bade, &S. (2024). Robustness Testing for AI/ML Models: Strategies for Identifying and Mitigating Vulnerabilities. *International Journal of Science and Research (IJSR)*, 13(4), 923-930. https://doi.org/10.21275/SR24409085438
[11] Liu, S., Peng, F., & Tang, K. (2023). Reliable robustness evaluation via automatically constructed attack ensembles. *In Proceedings of the AAAI Conference on Artificial Intelligence*, 37(7), 8852-8860. https://doi.org/10.1609/aaai.v37i7.26064
[12] Chen, Z., Liu, Y., Shi, L., Wang, Z. J., Chen, X., Zhao, Y., & Ren, F. (2025). MDEval: Evaluating and Enhancing Markdown Awareness in Large Language Models. *In Proceedings of the ACM on Web Conference 2025*, 2981-2991. https://doi.org/10.1145/3696410.3714674
[13] Pezoa, F., Reutter, J. L., Suarez, F., Ugarte, M., & Vrgoč, D. (2016). Foundations of JSON schema. *In Proceedings of the 25th international conference on World Wide Web*, 263-273. https://doi.org/10.1145/2872427.2883029
[14] Ben-Kiki, O., Evans, C., & Ingerson, B. (2009). Yaml ain't markup language (YAML™) version 1.1. Retrieved from https://yaml.org/spec/1.1/
[15] Harold, E. R., & Means, W. S. (2004). XML in a nutshell: a desktop quick reference. *" O'Reilly Media, Inc.".*
[16] Yoo, K. M., Han, J., In, S., Jeon, H., Jeong, J., Kang, J., Kim, H., Kim, K.-M., Kim, M., Kim, S., Kwak, D., Kwak, H., Kwon, S. J., Lee, B., Lee, D., Lee, G., Lee, J., Park, B., Shin, S., ... Sung, N. (2024). HyperCLOVA X technical report. *arXiv*. https://doi.org/10.48550/arXiv.2404.01954
[17] Ackley, D. H., Hinton, G. E., & Sejnowski, T. J. (1985). A learning algorithm for Boltzmann machines. *Cognitive science*, 9(1), 147-169. https://doi.org/10.1207/s15516709cog0901_7
[18] Caccia, M., Caccia, L., Fedus, W., Larochelle, H., Pineau, J., & Charlin, L. (2018). Language gans falling short. *arXiv preprint* arXiv:1811.02549. https://doi.org/10.48550/arXiv.1811.02549
[19] Holtzman, A., Buys, J., Du, L., Forbes, M., & Choi, Y. (2019). The curious case of neural text degeneration. *arXiv preprint* arXiv:1904.09751. https://doi.org/10.48550/arXiv.1904.09751
[20] United Nations Development Programme. (2010). Human development report 2010. *Palgrave Macmillan.*

이승현

· 2024년 대구교육대학교 교육대학원 AI교육전공 (교육학석사)
· 2024년~현재 대구교육대학교 교육대학원 AI교육전공 교육학박사 과정
· 2022년~현재 대구매천초등학교 교사

관심분야 : AI교육, 데이터 과학, 교육데이터 분석
seunghyunlee@korea.kr

이영호

· 2018년 서울교육대학교 컴퓨터교육과(교육학박사)
· 2022년~현재 대구교육대학교 컴퓨터교육과 교수

관심분야 : 인공지능 교육, 디지털 리터러시, 데이터 리터러시, 교육 데이터 분석, 컴퓨터 교육 이론
yhlee@dnue.ac.kr

부 록

데이터 수집(Data Collection)

[1~3] 다음 글을 읽고 물음에 답하시오.

밑줄 긋기는 일상적으로 유용하게 활용할 수 있는 독서 전략이다. 밑줄 긋기는 정보를 머릿속에 저장하고 기억한 내용을 떠올리는 데 도움이 된다. 독자로 하여금 표시한 부분에 주의를 기울이도록 해 정보를 머릿속에 저장하도록 돕고, 표시한 부분이 독자에게 시각적 자극을 주어 기억한 내용을 떠올리는 데 단서가 되기 때문이다. 이러한 점에서 밑줄 긋기는 일반적인 독서 상황뿐 아니라 학습 상황에서도 유용하다. 또한 밑줄 긋기는 방대한 정보들 가운데 주요한 정보를 추리는 데에도 효과적이며, 표시한 부분이 일종의 색인과 같은 역할을 하여 독자가 내용을 다시 찾아보는 데에도 용이하다.

통상적으로 독자는 글을 읽는 중에 바로바로 밑줄 긋기를 한다. 그러다 보면 밑줄이 많아지고 복잡해져 밑줄 긋기의 효과가 줄어든다. 또한 밑줄 긋기를 신중하게 하지 않으면 잘못 표시한 밑줄을 삭제하기 위해 되돌아가느라 독서의 흐름이 방해받게 되므로 효과적으로 밑줄 긋기를 하는 것이 중요하다.

3. 윗글을 바탕으로 학생이 다음과 같이 밑줄 긋기를 했다고 할 때, 이에 대한 평가로 적절하지 않은 것은? [3점]

[독서 목적] 고래의 외형적 특징에 대한 정보 습득
[표시 기호] [ ], 1) , 2) , √ , ~~~

[독서 자료]
고래는 육지 [포유동물]에서 기원했지만, 수중 생활에 적응하여 새끼를 수중에서 낳는다. 1)암컷들은 새끼를 낳을 때 서로 도와주며, 2)어미들은 새끼들을 정성껏 보호한다.

고래의 생김새는 고래의 종류마다 다른데, √대체로 몸길이는 1.3m에서 30m에 이른다. √피부에는 털이 없거나 아주 짧게 나 있다. 지느러미는 배를 젓는 노와 같은 형태이고, 헤엄칠 때 수평을 유지하는 기능을 한다.

고래는 폐로 호흡하므로 물속에서 숨을 쉴 수 없다. 고래의 머리 꼭대기에는 분수공이 있다. 물속에서 참았던 숨을 분수공으로 내뿜고 다시 숨을 들이마신 뒤 잠수한다. 작은 고래들은 몇 분밖에 숨을 참지 못하지만, 큰 고래들은 1시간 정도 물속에 머물 수 있다.

① 독서 목적을 고려하면, 1문단에서 '[ ]'로 표시한 부분은 적절하지 않게 밑줄 긋기를 하였군.
② 독서 목적을 고려하면, 1문단에서 '1) ', '2) '와 같이 순차적인 번호로 표시한 부분은 적절하지 않게 밑줄 긋기를 하였군.
③ 2문단에서 '[ ]'로 표시한 부분을 보니, 독서 목적에 관련된 주요 어구에 밑줄 긋기를 하였군.
④ 독서 목적을 고려하면, 2문단에서는 '지느러미는 배를 젓는 노와 같은 형태'에 '√ '를 누락하였군.
⑤ '~~~'로 표시한 부분을 보니, 독서 목적을 고려하여 3문단 내에서 정보 간의 상대적인 중요도를 판단해 주요한 문장에 밑줄 긋기를 하였군.

데이터 정형화(Data Structuring)

```
paragraph = """밑줄 긋기는 일상적으로 유용하게 활용할 수 있는 독서 전략 이다. ..."""
question3 = "Q3. 윗글을 바탕으로 학생이 다음과 같이 밑줄 긋기를 했다고 할 때, 이에 대한 평가로 적절하지 않은 것은?"
question_plus3 = """<보기> [독서 목적] 고래의 외형적 특징에 대한 정보 습득
[독서 자료] 고래는 육지 ㉠'포유동물'에서 기원했지만..."""
choices3 = [
  "① 독서 목적을 고려하면, 1문단에서 ㉠'포유동물'에 그은 밑줄은 적절하지 않게 밑줄 긋기를 하였군.",
  "② 독서 목적을 고려하면, 1문단에서 '1)', '2)'와 같이 순차적인 번호로 표시하고 밑줄 그은 부분은 적절하지 않게 밑줄 긋기를 하였군.",
  "③ 2문단에서 ㉡'고래의 생김새'에 밑줄 그은 것을 보니, 독서 목적에 관련된 주요 어구에 밑줄 긋기를 하였군.",
  "④ 독서 목적을 고려하면, 2문단에서는 '지느러미는 배를 젓는 노와 같은 형태'에 밑줄을 누락하였군.",
  "⑤ ㉢'피부에는 털이 없거나 아주 짧게 나 있다.'에 밑줄 그은 것을 보니 독서 목적을 고려하여 3문단 내에서 정보 간의 상대적인 중요도를 판단해 주요한 문장에 밑줄 긋기를 하였군."
```

데이터 포맷팅(Data Formatting)

Markdown

```
## paragraph
밑줄 긋기는 일상적으로 유용하게 활용할 수 있는 독서 전략 이다. ...

## problems
### problem 3
#### question
Q3. 윗글을 바탕으로 학생이 다음과 같이 밑줄 긋기를 했다고 할 때, 이에 대한 평가로 적절하지 않은 것은?
#### question_plus
<보기> [독서 목적] 고래의 외형적 특징에 대한 정보 습득
[독서 자료] 고래는 육지 ㉠'포유동물'에서 기원했지만...
#### choices
- ① 독서 목적을 고려하면, 1문단에서 ㉠'포유동물'에 그은 밑줄은 적절하지 않게 밑줄 긋기를 하였군.
- ② 독서 목적을 고려하면, 1문단에서 '1)', '2)'와 같이 순차적인 번호로 표시하고 밑줄 그은 부분은 적절하지 않게 밑줄 긋기를 하였군.
- ③ 2문단에서 ㉡'고래의 생김새'에 밑줄 그은 것을 보니, 독서 목적에 관련된 주요 어구에 밑줄 긋기를 하였군.
- ④ 독서 목적을 고려하면, 2문단에서는 '지느러미는 배를 젓는 노와 같은 형태'에 밑줄을 누락하였군.
- ⑤ ㉢'피부에는 털이 없거나 아주 짧게 나 있다.'에 밑줄 그은 것을 보니 독서 목적을 고려하여 3문단 내에서 정보 간의 상대적인 중요도를 판단해 주요한 문장에 밑줄 긋기를 하였군.
```

JSON

```
{
  "paragraph": "밑줄 긋기는 일상적으로 유용하게 활용할 수 있는 독서 전략 이다. ...",
  "problems": [
    {
      "question": "Q3. 윗글을 바탕으로 학생이 다음과 같이 밑줄 긋기를 했다고 할 때, 이에 대한 평가로 적절하지 않은 것은?",
      "question_plus": "<보기> [독서 목적] 고래의 외형적 특징에 대한 정보 습득\n[독서 자료] 고래는 육지 ㉠'포유동물'에서 기원했지만...",
      "choices": [
        "① 독서 목적을 고려하면, 1문단에서 ㉠'포유동물'에 그은 밑줄은 적절하지 않게 밑줄 긋기를 하였군.",
        "② 독서 목적을 고려하면, 1문단에서 '1)', '2)'와 같이 순차적인 번호로 표시하고 밑줄 그은 부분은 적절하지 않게 밑줄 긋기를 하였군.",
        "③ 2문단에서 ㉡'고래의 생김새'에 밑줄 그은 것을 보니, 독서 목적에 관련된 주요 어구에 밑줄 긋기를 하였군.",
        "④ 독서 목적을 고려하면, 2문단에서는 '지느러미는 배를 젓는 노와 같은 형태'에 밑줄을 누락하였군.",
        "⑤ ㉢'피부에는 털이 없거나 아주 짧게 나 있다.'에 밑줄 그은 것을 보니 독서 목적을 고려하여 3문단 내에서 정보 간의 상대적인 중요도를 판단해 주요한 문장에 밑줄 긋기를 하였군."
      ]
    }
  ]
}
```

YAML

```
paragraph: |
  밑줄 긋기는 일상적으로 유용하게 활용할 수 있는 독서 전략 이다. ...

problems:
  - question: Q3. 윗글을 바탕으로 학생이 다음과 같이 밑줄 긋기를 했다고 할 때, 이에 대한 평가로 적절하지 않은 것은?
    question_plus: |
      <보기> [독서 목적] 고래의 외형적 특징에 대한 정보 습득
      [독서 자료] 고래는 육지 ㉠'포유동물'에서 기원했지만...
    choices:
      - ① 독서 목적을 고려하면, 1문단에서 ㉠'포유동물'에 그은 밑줄은 적절하지 않게 밑줄 긋기를 하였군.
      - ② 독서 목적을 고려하면, 1문단에서 '1)', '2)'와 같이 순차적인 번호로 표시하고 밑줄 그은 부분은 적절하지 않게 밑줄 긋기를 하였군.
      - ③ 2문단에서 ㉡'고래의 생김새'에 밑줄 그은 것을 보니, 독서 목적에 관련된 주요 어구에 밑줄 긋기를 하였군.
      - ④ 독서 목적을 고려하면, 2문단에서는 '지느러미는 배를 젓는 노와 같은 형태'에 밑줄을 누락하였군.
      - ⑤ ㉢'피부에는 털이 없거나 아주 짧게 나 있다.'에 밑줄 그은 것을 보니 독서 목적을 고려하여 3문단 내에서 정보 간의 상대적인 중요도를 판단해 주요한 문장에 밑줄 긋기를 하였군.
```

XML

```
<exam>
  <paragraph>밑줄 긋기는 일상적으로 유용하게 활용할 수 있는 독서 전략 이다. ...</paragraph>
  <problem id="Q3">
    <question>Q3. 윗글을 바탕으로 학생이 다음과 같이 밑줄 긋기를 했다고 할 때, 이에 대한 평가로 적절하지 않은 것은?</question>
    <question_plus><![CDATA[<보기> [독서 목적] 고래의 외형적 특징에 대한 정보 습득
[독서 자료] 고래는 육지 ㉠'포유동물'에서 기원했지만...]]></question_plus>
    <choices>
      <choice id="1">① 독서 목적을 고려하면, 1문단에서 ㉠'포유동물'에 그은 밑줄은 적절하지 않게 밑줄 긋기를 하였군.</choice>
      <choice id="2">② 독서 목적을 고려하면, 1문단에서 '1)', '2)'와 같이 순차적인 번호로 표시하고 밑줄 그은 부분은 적절하지 않게 밑줄 긋기를 하였군.</choice>
      <choice id="3">③ 2문단에서 ㉡'고래의 생김새'에 밑줄 그은 것을 보니, 독서 목적에 관련된 주요 어구에 밑줄 긋기를 하였군.</choice>
      <choice id="4">④ 독서 목적을 고려하면, 2문단에서는 '지느러미는 배를 젓는 노와 같은 형태'에 밑줄을 누락하였군.</choice>
      <choice id="5">⑤ ㉢'피부에는 털이 없거나 아주 짧게 나 있다.'에 밑줄 그은 것을 보니 독서 목적을 고려하여 3문단 내에서 정보 간의 상대적인 중요도를 판단해 주요한 문장에 밑줄 긋기를 하였군.</choice>
    </choices>
  </problem>
</exam>
```

〈그림 1〉 LLM 평가를 위한 데이터셋 구축 워크플로우

〈표 1〉 언어 환경에 따른 모델별 효과 크기(Cohen's d) 비교 분석

Model	JSON vs Markdown		JSON vs XML		JSON vs YAML		Markdown vs XML		Markdown vs YAML		XML vs YAML	
	Kor	Eng	Kor	Eng	Kor	Eng	Kor	Eng	Kor	Eng	Kor	Eng
GPT-4.1	0.06	0.03	0.04	0.09	0.00	0.00	-0.02	0.06	-0.06	-0.03	-0.04	-0.09
GPT-4.1-Mini	0.02	-0.10	0.04	0.00	-0.06	0.00	0.02	0.10	-0.08	0.10	-0.10	0.00
GPT-4.1-Nano	0.05	-0.12	0.00	-0.10	0.00	-0.05	-0.06	0.02	-0.05	0.07	0.00	0.05
GPT-4o	0.04	-0.27	-0.02	-0.09	-0.02	0.00	-0.06	0.19	-0.06	0.27	0.00	0.09
GPT-4o-Mini	0.07	-0.03	0.02	-0.21	-0.02	-0.03	-0.05	-0.19	-0.09	0.00	-0.04	0.19
GPT-4-Turbo	0.07	0.05	-0.02	0.08	0.00	0.16	-0.09	0.03	-0.07	0.11	0.02	0.08
GPT-3.5-Turbo	0.07	-0.10	0.00	-0.18	-0.02	0.05	-0.07	-0.08	-0.09	0.15	-0.02	0.23
Gemini-2.5-Pro	0.02	-0.21	-0.02	-0.16	-0.13	-0.16	-0.05	0.06	-0.15	0.06	-0.10	0.00
Gemini-2.5-Flash	0.17	0.00	0.21	0.00	0.07	-0.12	0.04	0.00	-0.10	-0.12	-0.14	-0.12
Gemini-2.0-Flash	0.02	-0.06	0.06	0.09	0.10	-0.06	0.04	0.15	0.08	0.00	0.04	-0.15
Gemini-2.0-Flash-Lite	0.34	-0.27	0.06	-0.05	0.13	-0.08	-0.29	0.21	-0.21	0.18	0.07	-0.03
Gemini-1.5-Pro	0.09	-0.10	0.00	-0.10	0.11	0.00	-0.09	0.00	0.02	0.10	0.11	0.10
Gemini-1.5-Flash	-0.05	-0.19	0.08	-0.25	0.05	-0.08	0.14	-0.06	0.10	0.11	-0.03	0.17
Gemini-1.5-Flash-8b	-0.05	0.33	0.12	0.17	-0.09	0.33	0.17	-0.16	-0.05	0.00	-0.22	0.16
Claude-3-Opus	0.24	-0.12	0.37	-0.12	0.15	-0.28	0.13	0.00	-0.09	-0.16	-0.22	-0.16
Claude-3.7-Sonnet	-0.07	0.16	-0.07	0.28	0.02	0.00	0.00	0.12	0.10	-0.16	0.10	-0.28
Claude-3.5-Sonnet	-0.10	-0.20	-0.10	-0.20	-0.10	-0.10	0.00	0.00	0.00	0.11	0.00	0.11
Claude-3-Haiku	0.12	0.09	-0.02	0.17	0.00	0.09	-0.14	0.08	-0.12	0.00	0.02	-0.08
HCX-003	0.05	-0.61	0.16	-0.29	0.05	-0.07	0.11	0.31	0.00	0.54	-0.11	0.22
HCX-DASH-001	-0.45	-0.76*	-0.47	-0.67	-0.34	0.30	0.00	0.08	0.11	1.13**	0.11	1.03**

*$p < .05$, **$p < .01$, ***$p < .001$*