글로벌 벤치마크 데이터 품질검증 방법 및 적용 사례

김아름[1], 이상복[1], 조연제[2], 유호진[2]
[1]한국정보통신기술협회 AI데이터품질팀
[2]한국지능정보사회진흥원 AI데이터사업팀
ark5139@tta.or.kr, jangpo@tta.or.kr, yjyjyjcho@nia.or.kr, yuhj@nia.or.kr

A Method and Application Case of Quality Verification of Global Benchmarking Data

A-Rum Kim[1], Sang-Bok Lee[1], Yeon-Je Cho[2], Ho-jin Yu[2]
[1]AI Data Quality Team, Telecommunications Technology Association
[2]AI Data Project Team, National Information Society Agency

요 약

본 논문은 글로벌 인공지능 모델 벤치마크에 사용되는 데이터의 품질검증 방법을 제안한다. 또한, 글로벌 규범 문화 평가를 위한 벤치마크 데이터 구축 사례를 소개하며, 주요 태스크별 품질검증 항목의 정량적, 정성적 검사 기준을 설명한다. 마지막으로, 신뢰성 있는 벤치마크 결과 도출을 위한 데이터 품질 확보의 중요성과 글로벌 데이터 품질 표준의 필요성을 강조한다.

1. 서론

최근 인공지능(AI, Artificial Intelligence) 기술의 급격한 발전과 함께 다양한 모델들이 등장하고 있으며 이들의 성능을 객관적으로 비교하고 평가하기 위한 벤치마크 데이터의 중요성이 증가하고 있다[1]. 벤치마크 데이터는 AI 모델 개발의 방향성을 제시하고 연구자 및 개발자들이 모델의 강점과 약점을 파악하여 성능 향상에 집중할 수 있도록 도움을 준다. 특히, 파라미터 수가 증가하면서 더욱 강력한 성능을 보이는 초거대 언어모델(LLM, Large Language Model)의 글로벌 활용이 확대됨에 따라, 특정 국가의 문화적 맥락과 사회적 규범에 대한 깊이 있는 이해를 기반으로 신뢰성 있는 답변을 제공하는 능력의 중요성이 부각되고 있다[2]. 또한 국가별 문화적 가치관, 행동 양식 등을 제대로 이해하지 못하는 LLM은 부적절한 답변을 생성할 위험이 있으므로 특정 국가의 문화와 규범을 정확하게 반영하는 벤치마크 데이터가 필요하다[3]. 이에 따라, 본 논문에서는 글로벌 인공지능 모델 벤치마크에 사용되는 데이터의 품질검증 방법을 살펴보고 실제 베트남어 벤치마크 데이터 구축 사례에 이를 적용하여 평가 방안의 유용성과 한계점을 고찰하고자 한다.

2. 본론

벤치마크 데이터 품질 검증의 핵심적인 역할은 모델의 변별력을 확보하고 평가 과정의 객관성 및 공정성을 보장하는 데 있다. 이는 마치 대학수학능력시험 문항 제작 시 수험생들의 학업 능력을 정확하게 변별하고, 문제에 대한 명확한 정답을 요구하는 것과 유사한 원리이다. 실제로 Open LLM Leaderboard와 같은 플랫폼에서는 LLM의 성능을 다각적으로 평가하기 위해 다양한 벤치마크 태스크를 활용하고 있다. 특히, 객관식 질문 답변(CSQA, HellaSwag, MMLU, BoolQ, Winogrande), 텍스트 생성(TruthfulQA, CommonGen, HHH), 그리고 독해(DROP)의 세 가지 주요 유형으로 분류되는 총 9개의 대표적인 태스크를 통해 LLM의 광범위한 능력을 종합적으로 측정한다. 각 태스크에 대한 상세한 설명은 표 1과 같다.

<표 1> Open LLM Leaderboard 태스크 예시

태스크	설명
CSQA	상식적인 추론 능력을 평가하며, 일상적인 상황에 대한 질문에 상식 기반으로 답하는 태스크
HellaSwag	주어진 상황 다음에 이어질 가장 논리적인 문장을 선택하여 상식적인 상황 이해 능력을 측정하는 태스크
MMLU	다양한 학문 분야의 객관식 질문을 통해 모델의 광범위한 지식과 이해 능력을 평가하는 대규모 태스크

BoolQ	주어진 질문에 대해 "yes" 또는 "no"로 답변하여 모델의 사실 기반 이해 능력을 평가하는 태스크
Winogrande	문맥 속 대명사가 지칭하는 대상을 정확히 파악하는 대명사 해소 능력을 평가하는 태스크
TruthfulQA	모델이 사실에 기반한 답변을 생성하고, 오해를 일으킬 수 있는 잘못된 정보를 생성하지 않는지 평가하는 태스크
CommonGen	주어진 명사들을 사용하여 의미 있고 일관성 있는 문장을 생성하는 모델의 창의적 언어 생성 능력을 평가하는 태스크
HHH	모델이 인간에게 유용하고, 정직하며, 해롭지 않은 답변을 생성하는지를 평가하는 태스크
DROP	텍스트 단락을 이해하고, 그 안의 정보를 바탕으로 숫자 값이나 텍스트 범위를 추론하여 답하는 독해 및 추론 태스크

이러한 벤치마크 데이터의 품질검증은 크게 정량적 평가와 정성적 평가로 나뉘며, 각 평가 방법은 다음과 같은 측면을 고려하여 수행된다.

2.1 정량적 평가

정량적 평가는 데이터의 구조적 정확성, 다양성, 그리고 중복성 등을 객관적인 지표를 통해 측정하는 과정이다.

구문 정확성 검증에서는 각 태스크별로 정의된 데이터 구조와 형식의 정확성을 검증하는 것을 목표로 한다. 구체적으로 각 태스크는 고유한 데이터 구조를 가질 수 있으므로. 이에 대한 정의가 정확한지 확인하고, 데이터의 속성명(property)이 올바르게 정의되었는지, 그리고 데이터의 타입(string, number, array)이 각 속성에 맞게 지정되었는지 등을 검토한다. 이러한 평가는 주로 자동화된 스키마 검사나 데이터 타입 검사 등의 방법을 통해 이루어진다.

다양성 검증은 데이터의 균형성과 포괄성을 확보하기 위해 수행된다. 벤치마크 데이터를 구성하는 각 태스크의 구성비가 사전에 설정된 목표 비율에 부합하는지 확인하고, 데이터에 포함된 주제의 다양성을 평가하는 것이 주요 내용이다. 이러한 평가는 태스크별 데이터 수나 주제별 데이터 분포 분석 등의 정량적인 방법을 통해 이루어진다.

유사성 검증은 데이터 내의 중복을 최소화하기 위해 수행된다. 각 태스크의 특성을 고려하여 유사성 검증이 필요한 속성을 선정하고, 선정된 속성을 기준으로 데이터 간의 유사도를 측정한다. 이러한 평가는 텍스트 유사도 측정 알고리즘이나 데이터 중복률 계산 등의 방법을 통해 이루어진다.

2.2 정성적 평가

정성적 평가는 데이터의 수치화하기 어려운 품질을 평가하는 방법이다. 특히 의미 정확성이라는 품질 특성의 세부 항목을 협의하여 검증을 진행할 수 있다. 예를 들어, 어떤 텍스트 데이터에 대해 의미적 적절성을 평가한다면 해당 텍스트의 내용이 사실과 부합하는지, 용어의 사용이 적절한지 등을 검토할 수 있다. 또한, 문화적 적절성 측면에서 평가한다면 특정 문화권의 규범과 가치를 적절하게 반영하는지 등을 검토할 수 있다. 이러한 평가는 주로 해당 분야의 전문가나 실제 사용자의 검토를 통해 이루어진다.

2.3 적용 사례

2.2절에서 언급한 품질검증 방법을 '글로벌 규범·문화 평가 데이터'에 적용해 보았다. 이 데이터는 LLM이 베트남의 규범·문화를 고려한 답변을 제공할 수 있도록, 신뢰성, 지식 능력, 성능 등을 종합적으로 검증하는 데 활용될 목적으로 구축되었으며, 구축된 데이터의 품질은 다음의 네 가지 주요 품질 특성을 중심으로 평가를 진행했다.

첫째, 구문 정확성은 데이터의 구조적 정확성 및 형식적 정확성을 평가하는 데 중점을 두었다. 구체적으로, 각 태스크별 데이터의 구조와 형식이 다르므로 태스크별 구문 정확성을 검증하였으며, 품질 지표로는 정확도를 사용하였다.

둘째, 다양성은 구축된 베트남 벤치마크 데이터의 태스크와 주제를 다양하게 포괄하는지를 평가하는 항목이다. 태스크 분포 및 주제 분포의 균형성을 확보하기 위해 목표 구성비를 설정하고, 실제 구성비와 비교하여 중첩률을 측정하는 방식으로 데이터의 다양성을 정량적으로 평가하였다.

셋째, 유사성은 데이터 내 질의 및 답변의 중복성을 평가하는 데 초점을 맞추었다. 질의응답 데이터의 특성을 고려하여 각 태스크별로 적절한 속성을 선정하고, 중복률을 측정하여 데이터의 유사성을 정량적으로 평가하였다.

넷째, 의미 정확성에서 '질의응답 적정성' 항목은 크게 '베트남 규범·문화 연관성', '내용 및 구성 적절성', '문법적 오류'의 3가지 검사 기준을 중심으로 평가하였다. 각 주요 검사 기준에 대한 예시는 표 2와 같다.

<표 2> 글로벌 규범 문화 평가 데이터 의미 정확성 질의응답 적정성 검사 기준 예시

검사 기준	예시
베트남 규범·문화 연관성	- 질문 및 답변이 베트남의 공식 자료에 근거한 키워드를 포함하고, 전통 및 현대적 가치관과 문화적 요소를 반영하는지 - 베트남에서 금기시되는 민감한 주제나 부적절한 표현을 포함하는지 - 속담, 관용구, 특정 어휘 사용이 베트남 고유의 표현인지
내용 및 구성 적절성	- 질문에 대한 답변의 적절성, 질문 내 답 존재 여부, 불필요하거나 부족한 정보 포함 여부 - 객관식 선지의 형태 통일성 및 정답/오답 개수 준수 여부 - Vie-HellaSwag 및 Vie-Winogrande 태스크의 경우, 불완전한 질문 구성 여부
문법적 오류	- 맞춤법, 오탈자, 문장 잘림, 비문 표현, 시제 불일치 등 기본적인 문법 오류 - 조사, 전치사, 복수 접미사 등의 사용이 문맥에 적절한지, 문장이 일상적 사용 수준에서 자연스러운지 - 문맥상 의미 전달이 가능한 일상적인 표현인지

3. 결론

본 논문에서는 글로벌 벤치마크 데이터의 정량적 평가 방법인 구문 정확성, 다양성, 유사성과 정성적 평가 방법인 의미 정확성 검증 방법을 실제 베트남 벤치마크 데이터 구축 사례에 적용하여 다양한 검증 기준을 살펴보았다. 하지만 다양한 국가의 문화적 맥락을 포괄하는 데이터의 품질 표준이 아직 부재하여 이에 대한 연구가 요구된다[4]. 따라서 앞으로의 연구에서는 다양한 국가와 문화권의 특성을 반영하는 데이터 품질 표준을 정립하고, 이러한 표준을 기반으로 벤치마크 데이터의 품질을 검증하는 연구를 수행할 것이다[5]. 이는 데이터의 문화적 오류 및 편향성을 최소화하고, 글로벌 환경에서의 활용도를 높이는 데 중요한 역할을 할 것이다.

사사문구

본 연구는 과학기술정보통신부 초거대AI 확산 생태계 조성 사업(2600-2604-301, 2025년 초거대AI 확산 생태계 조성 사업)에 의해서 수행되었습니다.

참고문헌

[1] Eriksson, Maria, et al. "Can We Trust AI Benchmarks? An Interdisciplinary Review of Current Issues in AI Evaluation." arXiv preprint arXiv:2502.06559 (2025).

[2] Myung, Junho, et al. "Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages." Advances in Neural Information Processing Systems 37: 78104-78146 (2024).

[3] Belay, Tadesse Destaw, et al. "CULEMO: Cultural Lenses on Emotion--Benchmarking LLMs for Cross-Cultural Emotion Understanding." arXiv preprint arXiv:2503.10688 (2025).

[4] Pawar, Siddhesh, et al. "Survey of cultural awareness in language models: Text and beyond." arXiv preprint arXiv:2411.00860 (2024).

[5] Belay, Tadesse Destaw, et al. "CULEMO: Cultural Lenses on Emotion--Benchmarking LLMs for Cross-Cultural Emotion Understanding." arXiv preprint arXiv:2503.10688 (2025).