컴퓨터교육학회 논문지 2024년 제27권 제9호
https://doi.org/10.32431/kace.2024.27.9.007

거대 언어 모델의 한국 이해도 평가를 위한 벤치마크 연구*

Benchmark Study for Evaluating the Korean Comprehension of Large Language Models

서민주† · 정연주† · 이현정† · 임혜균† · 장정선††
Minju Seo† · Yeonjoo Jeong† · Hyunjeong Lee† · Hyekyun Im† · Jungsun Jang††

요약

본 연구는 거대 언어 모델(Large Language Models, LLM)의 한국 이해도를 평가하기 위해 설계된 Ko-Sovereign 벤치마크를 제안한다. Ko-Sovereign 벤치마크는 한국어 능력 평가에만 초점을 맞추었던 기존 벤치마크의 한계를 극복하고, 법, 경제, 정치 등 한국 특화 전문 지식을 포함한 총 9개의 평가 영역으로 범위를 확장하였다. 평가 지표에는 유창성, 최신성, 사실성, 편향성, 문화 맥락 이해가 포함되며, 문항 유형은 빈칸 채우기, 밑줄, 사실 확인, 순서나열, 정답매칭, 복합과 같은 형식으로 구성된다. 실험 결과, ChatGPT4o는 전반적으로 높은 정확도를 달성했으며, EXAONE3 또한 한국적 지식을 효과적으로 반영하는 성능을 보였다. Ko-Sovereign 벤치마크는 문화, 역사, 법 등 다양한 한국적 영역을 종합적으로 평가할 수 있는 평가 프레임워크를 제안했다는 점에서 의의가 있다.

주제어 생성형 인공지능, 인공지능 교육, 거대 언어 모델, 거대 언어 모델 평가, 한국형 벤치마크, 소버린 AI

ABSTRACT

This study introduces the Ko-Sovereign benchmark, designed to evaluate the Korean comprehension capabilities of Large Language Models(LLMs). The Ko-Sovereign benchmark addresses the limitations of previous benchmarks that focused solely on Korean language proficiency by expanding its scope to include nine distinct evaluation domains, encompassing specialized knowledge in areas such as law, economics, and politics. The evaluation metrics include fluency, recency, factuality, bias, and cultural context understanding, with question types structured into formats such as fill-in-the-blank, underline matching, fact verification, sequence ordering, answer matching, and complex tasks. Experimental results indicate that ChatGPT4o achieved the highest overall accuracy, while EXAONE3 demonstrated strong performance in reflecting Korean-specific knowledge. The Ko-Sovereign benchmark is a significant contribution, offering a comprehensive framework for evaluating LLMs across diverse Korean-specific domains, including culture, history, and law.

Keywords Generative AI, AI Education, Large Language Model, LLM Evaluation, Korean Benchmark, Sovereign AI

†정회원 고려대학교 대학원 역사학과 박사수료
††정회원 고려대학교 문과대학 한국사학과 부교수 (교신저자)

논문투고 2024년 11월 16일
심사완료 2024년 12월 13일
게재확정 2024년 12월 18일
발행일자 2024년 12월 26일

* 본 논문은 KT의 지원을 받아 수행된 고려대학교-KT 산학공동연구 개발과제의 연구결과임.

1. 서론

최근 지속적인 기술 혁신으로 인공지능은 인간의 지능 수준에 필적하는 결과물을 생성하기에 이르렀다. 하지만 그 이면에는 광범위한 인공지능 기술 활용으로 인한 허위 정보의 확산, 딥페이크 등 새로운 사회 문제가 있으며, 국가 간 격차를 심화시키는 요인으로까지 확장되고 있다. 인공지능 기술의 위험성을 인식한 이래 각 국가는 자국의 디지털 주권(Digital Sovereignty)을 구축하기 위해 노력하고 있다. 국가와 개인이 자신들의 디지털 기술, 데이터, 인프라에 행사하는 통제권을 의미하는 디지털 주권은 국가 안보, 국가 경쟁력, 개인 사생활을 유지하기 위한 필수 조건이다[1]. 자국의 인공지능 기술이 자국의 국가 정체성의 유지와 연결되는 만큼 디지털 주권의 확보는 소버린(Sovereign) AI를 통해 가능하다[2]. 자국의 데이터와 인프라를 구축한 소버린 AI는 자국의 디지털 역량, 나아가서는 자국의 이익을 강화하는 중요한 수단이 될 것이다. 실제 2023년 한국 데이터로 훈련하여 한국의 문화와 사회적 맥락을 이해하는 거대 언어 모델이 개발되어 디지털 주권을 점유하려는 노력이 한국에서도 본격적으로 시도되고 있다[3].

생성형 인공지능 기술의 발전에 따라 자연어 처리 분야에 특화된 거대 언어 모델 역시 매개변수의 크기를 확대하며 다양한 자연어 처리 작업을 수행하고 있다. OpenAI가 개발한 Chat GPT 4o는 모의 변호사 시험에서 상위 10%의 수준을 달성하여 전문적, 학문적 영역에서도 괄목할 만한 성능의 향상을 이뤄냈다[4]. 이와 함께 거대 언어 모델을 평가하는 벤치마크 데이터에 대한 논의도 활발하게 이루어졌다. 전통적인 거대 언어 모델의 벤치마크 데이터로는 ARC[5], HellaSwag[6], MMLU[7], TruthfulQA[8], Winogrande[9], GSM8k[10] 등이 있으며, 각 벤치마크는 평가 주제와 난이도에 따라 거대 언어 모델의 다양한 측면에 대해 평가하였으나 안전성을 평가하기에 부족하다는 비판을 받았다. 게다가 고자원 언어인 영어를 대상으로 형성된 벤치마크이기 때문에 영어를 포함한 다국어 모델이 아닌 이상 벤치마크를 각 국가의 언어로 번역하여 평가하고 있어 각 국가의 사회, 문화적 측면을 정확하게 평가할 수 없다는 한계가 있다. 하지만 벤치마크의 평가 결과를 토대로 거대 언어 모델의 성능을 강화하는 방법을 모색한다는 점에서 거대 언어 모델과 벤치마크는 선순환의 관계에 있다.

소버린 AI의 필요성이 증대되고 한국형 벤치마크 구축에 관한 관심이 지속된 결과, 한국어를 기반으로 거대 언어 모델 벤치마크 개발이 집중적으로 시도되고 있다. 이러한 현상은 한국형 거대 언어 모델이 발전할 수 있는 기반을 제공할 수 있다는 점에서 고무적이다. 초기 한국형 벤치마크는 고자원 언어인 영어를 기반으로 제작된 HellaSwag과 같이 전통적인 벤치마크를 한국어로 번역한 것에서부터 시작되었다[11-13]. 하지만 초기 한국형 벤치마크는 기존에 구축된 영어 중심의 벤치마크를 한국어로 번역하는 단계에 머물렀기에 한국의 문화 맥락이 반영되지 않는다는 한계가 존재했다. 이러한 한계를 극복하고자 영어 기반 벤치마크를 한국의 문화적 뉘앙스를 보존하여 번역하는 벤치마크 연구가 진행되기도 했다[14]. 최근에는 HAE-RAE 벤치마크[15], CLIcK[16], KorNAT[17]과 같이 한국어를 기반으로 벤치마크를 구축하려는 연구도 등장하는 추세이다. 각 연구는 거대 언어 모델의 한국어 능력, 한국의 사회적 편향, 한국의 사회적 가치와 상식 등을 다각도에서 평가하였으나 평가 영역을 종합한 벤치마크의 필요성은 남아 있다. KMMLU는 생물, 화학, 세법, 형법 등을 포함한 광범위한 종합 지식 벤치마크로서 한국의 사회 문화 맥락을 반영하고자 도입되었다[18]. 하지만 한국의 구체적이고 특수한 부분을 다룬 문항은 약 20%에 불과하다.

본 논문은 한국형 벤치마크 연구에서 드러난 한계를 보완하려는 시도로 시작되었다. 본 논문을 통해 거대 언어 모델의 한국어 능력뿐만 아니라 한국의 사회문화적 맥락을 다양한 분야에서 평가할 수 있는 벤치마크인 Ko-Sovereign을 제안한다. 일차적으로 평가의 영역을 언어 및 문학, 역사, 문화 및 민속, 법, 경제 및 금융, 정치, 지리, 사회, 교육의 9개로 세분화하였다. 한국어 능력 평가나 한국적 상식 평가 영역만을 제한적으로 구축한 기존 벤치마크 연구와 차별되는 지점이다. 평가 영역의 확장에서 나아가 문항의 유형을 형식과 내용으로 체계화하였다. 또한 거대 언어 모델을 평가하는 지표를 유창성, 최신성, 사실성, 편향성, 문화 맥락 이해로 세분화하여 한국형 거대 언어 모델의 역량을 종합적으로 측정할 수 있도록 했다. 궁극적으로는 거대 언어 모델이 취약한 부분을 영역별, 기능별로 진단하여 보다 견고한 한국형 거대 언어 모델로 성장할 수 있는 기반을 제공할 수 있도록 했다.

논문은 다음과 같은 순서로 구성하였다. Ⅱ장에서는 현재까지 수행된 벤치마크, 소버린 AI, 한국형 거대 언어 모델 벤치마크를 대상으로 축적된 관련 연구를 조망한다. Ⅲ장에서는 본격적으로 본 논문에서 구축한 벤치마크의 구성을 평가 영역, 문항 유형, 평가 지표의 순서로 검토한다. Ⅳ장에서는 다국어 대규모 언어 모델과 한국어 특화 대규모 언어 모델을 대상으로 진행한 실험 환경 및 방법을 서술한다. Ⅴ장에서는 실험 결과를 분석적으로 제시하고, 이를 기반으로 Ⅵ장에서 소버린 AI를 구현하기 위한 방향 및 거대 언어 모델을 종합적으로 평가할 수 있는 벤치마크의 전망을 제시하고자 한다. 본 논문이 한국형 거대 언어 모델 벤치마크의 발전, 나아가 소버린 AI의 신속한 구현을 매개하는 데 일조하기를 기대한다.

2. 관련 연구

2.1 소버린 거대 언어 모델

소버린 AI는 각 국가의 자체 인프라와 데이터, 인력 및 비즈니스 네트워크를 활용하여 각 국가의 언어와 문화, 가치관 등을 반영한 AI를 의미한다[19], [20]. 즉 자국의 데이터와 개발된 인공지능 기술에 대해 각국의 소유권을 강조하는 개념이다[21]. AI가 사회 여러 분야에 혁신을 가져와

점점 AI의 중요도가 커지고 있는 상황에서 많은 국가가 AI에 적극적으로 투자하고 있다. 2023년 하반기부터 생성 AI는 미래 국가 경쟁력을 결정하는 핵심 요소로 간주되기 시작했고, 경제적 이익과 새로운 세계질서 속의 리더십 선점을 위해 각 국가에서 생성 AI 개발에 뛰어들고 있다[2]. 또한 독립적으로 AI 역량을 쌓으려는 각국 기업의 움직임도 성과를 보이고 있다[20].

소버린 AI 개발과 함께 디지털 주권의 개념도 논의가 시작되고 있다. 디지털 주권이란 디지털 맥락에서 데이터, 소프트웨어, 표준, 서비스 및 기타 디지털 인프라에 대한 정당한 통제 권한의 형태이다[22]. 이러한 디지털 주권의 행사는 소버린 AI와 결부되어 있다. 소버린 AI의 개발은 국가 경쟁력, 미래의 세계질서 속 리더십 선점, 그리고 미국 빅테크의 생성 AI를 통한 문화종속을 피하고 국가 정체성을 지키기 위한 노력의 결과로 해석할 수 있기 때문이다[2].

최근 GPT-4[4]와 Gemini[23] 등은 다양한 분야에서 뛰어난 성능을 보였으나, 대부분의 거대 언어 모델은 영어를 기반으로 하기 때문에 영어 텍스트의 이해와 생성, 그리고 영어를 사용하는 사회의 문화와 규범 등에서 뛰어난 성능을 보인다. 반면 저자원 언어와 해당 국가의 문화 등에 대해서는 성능이 떨어진다. 이로 인해 현재 대부분의 AI와 거대 언어 모델은 다양한 언어 및 문화에 대한 포용성이 낮다는 한계가 있다. 따라서 미국을 제외한 국가에서 인프라, 언어, 데이터 등을 모두 통제할 수 있는 소버린 거대 언어 모델의 개발이 대두되고 있다. 한국에서는 ChatGPT-3의 학습데이터에서 한국어 비중이 0.016%에 불과하여 한국어 성능이 떨어지는 한계를 수정하기 위해 HyperCLOVA-X를 발표하였다[24]. 한국어, 영어, 코드데이터를 학습한 후 주석이 달린 데이터로 인스트럭션 튜닝을 거치는 방식을 활용하였다[3].

UAE에서는 아랍어와 영어로 훈련하여 아랍어 중심 거대 언어 모델인 Jais와 Jais-chat을 공개하였다[25]. 핀란드의 Silo AI 역시 핀란드어, 영어, 프로그래밍 언어로 훈련한 Poro를 개발하였다[26]. 인도의 Krutrim AI에서는 인도의 20가지 언어를 이해하고 10가지 언어를 생성할 수 있는 거대 언어 모델 Krutrim을 제작하였고, 중국에서는 문샷 AI에서 챗봇 Kimi를 제공하고 있다. 일본 NTT에서는 츠즈미를 공급하고 있으며, 파나소닉에서도 거대 언어 모델 개발을 발표한 상황이다. 영국에서는 BritGPT, 그리고 대만은 정부 주도하에 Taide를 준비하고 있으며, 이탈리아의 Fast Web, 인도의 Reliance Industries, 독일의 Aleph Alpha에서도 각기 자국 언어 거대 언어 모델을 개발 중에 있다[20].

한편 디지털 주권에 대해 전통적인 국가 중심의 정의에서 벗어나 EU와 같이 국제적 또는 초국가적 기관이 권위를 가질 수 있음을 인정한 연구가 있다[27]. 그렇기 때문에 소버린 AI와 거대 언어 모델에 대해서 국제적 또는 초국가적 거대 언어 모델을 검토할 필요가 있다. 대표적으로 싱가포르에서 공개한 동남아시아 언어와 문화를 반영한 소버린 거대 언어 모델 SEA-LION[28]이 있다. 이 외에 프랑스의 스타트업 Mistral AI는 자체 개발한 Mistral Large 모델을 기반으로 Le chat을 공개하였는데[20], 이는 EU 국가의 언어와 문화를 기반으로 한 소버린 거대 언어 모델이다.

이상의 연구 결과에 의하면 소버린 거대 언어 모델의 개발은 국가 자체적으로 진행하고 있고, 동남아시아나 EU 등에서는 단일 국가가 아닌 초국가적 단위로 이루어지고 있는 경우도 확인된다. 이러한 영어 이외 언어 모델의 문제점은 높은 비용이다. 소버린 거대 언어 모델은 특정 국가를 대상으로 하므로 미국 빅테크 기업에 비해 비용이나 인프라가 부족한 경우가 많다. 그리고 언어 자원이 적은 경우가 대다수여서 학습 데이터를 마련하기 어렵다는 문제도 있다. 거대 언어 모델 학습에 필요한 데이터 크기가 증가하는 상황 속에서 다국어 데이터 추가로 모델 용량이 제한되어 저자원 언어와 고자원 언어 모두 성능이 저하되는 '다국어의 저주' 현상이 나타나기도 한다. 그렇기 때문에 대규모 다국어 사전 학습이 아닌 타겟화 된 모델을 사용하면 성능이 향상될 것으로 짐작된다[29].

이러한 소버린 AI 기술을 국가나 기업이 자체적으로 개발하면 보안 문제와 기술 의존도를 낮춰 안전성을 높일 수 있다. 그에 따라 기술 수출 및 기술 이전에 대한 제약을 줄여 국가와 기업의 경쟁력을 강화하며 경제적 이익을 창출할 수 있다[20]. 또한 자국어 혹은 특정 지역의 언어만을 대상으로 학습시키는 소버린 거대 언어 모델의 개발은 각 국가 및 지역의 디지털 주권을 지키고 디지털 권력에서의 배제를 방지하며, 저자원 언어를 활성화하고 보존할 것이다. 따라서 소버린 AI와 거대 언어 모델에 대한 개발과 연구가 필요하며, 그를 위해 본 논문에서는 소버린 거대 언어 모델의 성능을 평가하기 위한 연구를 진행하였다.

2.2 거대 언어 모델 벤치마크

근래 거대 언어 모델의 개발과 활용이 많은 관심을 받으면서 거대 언어 모델을 평가하는 일도 한층 중요해졌다. 거대 언어 모델 벤치마크의 역할은 거대 언어 모델의 성능을 측정하고, 평가하는 것이다[30]. 자연어처리 기술의 분야별 벤치마크 연구에 이어 점차 언어별 벤치마크 데이터가 공개되었다. 영어 언어 모델 벤치마크 GLUE[31], 중국어 언어모델 벤치마크 CLUE[32], CLUE의 한계를 보완한 SuperGLUE[33], 최초의 한국어 언어 모델 벤치마크 KLUE[12] 등이 대표적인 사례이다.

거대 언어 모델 벤치마크는 특정 영역을 정해 다양한 대규모 언어모델의 성능을 비교하는 방향으로 발전하였고, ARC, HellaSwag, MMLU, TruthfulQA, Winogrande, GSM8k 등이 대표적으로 평가에 활용되고 있다. ARC(AI2 Reasoning Challenge)는 추론 능력을 평가하기 위한 벤치마크이다. 초등학교와 중학교 수준의 과학 문제 7,787개로 구성되어 있고, 답변이 얼마나 적절한지를 측정한다. 도전적인 벤치마크 데이터와 쉬운 벤치마크 데이터가 구분되어 있는 것이 특징이다[5]. WinoGrande는 상식 추론을 평

가하기 위한 벤치마크로, 크라우드 소싱(crowdsourcing)으로 수집한 44,000개의 문제로 구성되어 있다. 주로 대명사를 정확하게 맞추는지 확인할 수 있다[9]. MMLU는 언어 이해를 다루는 벤치마크로, 초등 수학, 미국 역사, 컴퓨터 과학, 법률 등을 포함한 57개의 영역을 난이도별로 출제하고, 다수 영역의 질문에 대해 얼마나 정확한 답변을 도출하는지 살펴볼 수 있다[7]. HellaSwag은 상식 능력을 평가하기 위한 벤치마크로 WikiHow와 ActivityNet에서 수집한 총 70,000개의 문제로 구성되어 있다. 미완성된 구절을 완성하게 하는 방식으로 평가한다[6]. TruthfulQA는 모델의 답변이 거짓인지 진실인지를 판별하는 벤치마크로, 건강, 법률, 금융, 정치 등 38개 영역의 817개의 질문으로 구성되어 있으며, 잘못된 믿음이나 오해로 인해 사람이 오답을 말하기 쉬운 내용으로 짜여있다[8]. GSM8k는 수학적 추론을 측정하는 벤치마크로, 학년별 다양한 난이도의 수학 문제 8,500개로 구성되어 있다[10].

이러한 벤치마크는 거대 언어 모델의 다양한 측면을 평가하는 데 일조하였다. 그러나 기존 벤치마크는 각 국가의 고유한 문화 맥락을 반영하지 못했다는 한계가 있다[34, 35]. ARC, HellaSwag, MMLU, TruthfulQA 등 4가지 벤치마크는 한국어로 번역되었지만[36] 기계 번역을 이용했기 때문에 한계가 분명하다. 또한 미국 중심의 문화 및 상식이 반영된 벤치마크이기 때문에, 이를 이용하여 한국의 정치, 경제, 문화 등의 측면을 정확히 평가하기 어렵다는 문제가 있다.

하지만 비용 등의 문제로 영어 중심의 벤치마크를 번역하여 평가에 사용하는 연구가 다수 진행되어 왔다. Korean-NLI & STS[11]는 자연어 추론(NLI)과 의미 텍스트 유사성(STS)을 위해 영어 데이터를 번역한 자료로, 한국어 특유의 뉘앙스를 반영하지 못할 가능성이 있다. KLUE[12]는 GLUE 벤치마크에 한국어를 적용한 것이다. 주제 분류, 의미 텍스트 유사성, 자연어 추론 등 다양한 작업을 포함하지만, 번역이나 특정 작업 중심이어서 언어 특화 모델을 평가하기에 충분하지 않다. KoBEST[13]는 COPA[37], WiC[38], BOOLQ[39], HellaSwag, SentiNeg[40] 등을 한국어로 재구현한 것으로 한국어의 고유한 뉘앙스를 충분히 반영하지 못한다.

한편 고자원 언어인 영어 데이터를 기반으로 학습한 거대 언어 모델의 한계에 대한 연구도 다수 진행되었다. 그 결과는 크게 7가지로 분류된다. 첫 번째는 상식왜곡이다. 지리적 다양성에 따라 상식의 범위와 내용이 달라질 수 있으며[34], 학습된 언어에 따라 각 문화의 사실 상식이 왜곡되거나 편향될 수 있다는 것이다[41]. 두 번째는 상식 암기이다. 거대 언어 모델은 사전 학습 과정에서 데이터를 암기하는 경향이 있으며[42], 상식 지식 형성은 관계의 빈도와 복잡성에 영향을 받아 단순한 추론과 공출 현상에 기반하는 경향이 있다고 설명된다[43]. 세 번째는 유해한 발화이다. 사회적, 문화적 요소가 학습 데이터에 영향을 미쳐 편향이나 공격성이 포함될 가능성이 있다고 보았다[44, 45]. 네 번째는 문법성이다. 표현을 위해 일반적인 문법적 이해가 필요하며, 상식을 반영하는 완전한 문장 구성을 위해 각 언어에 대한 문법적 교정이 필요하다[46-48]. 다섯 번째는 타당성이다. 거대 언어 모델의 생성 결과가 구체적인 지식을 검색하는 것이 아니라 암묵적이고 모호한 추론을 통해 생성한다는 문제가 있다[49, 50]. 여섯 번째는 수치적 상식인데, 언어 모델은 상식 범위 내에서 수치 정보를 처리하는 데 약하다[47, 51]. 마지막은 속담이다. 관용 표현은 언어 모델에게 어려운 영역이며, 상식 지식 그래프를 통해 해결하려는 시도가 이루어졌다[52]. 그러므로 각 국가의 언어, 문화, 사회, 역사 등의 지식 이해도와 생성 능력을 평가하기 위한 벤치마크를 구축해야 한다.

2.3 한국형 거대 언어 모델 벤치마크

이러한 문제 해결을 위해 한국의 사회문화적 맥락과 상호작용할 수 있는 한국어 상식 추론 벤치마크인 KoCommonGENv2가 제안되었다[53]. 또한 영어 데이터로 학습된 거대 언어 모델의 단점을 극복하기 위해 HAE-RAE Bench가 제시되었다[54]. 평가 결과, 한국어 모델이 다국어 모델보다 우수한 성능을 보였다. 한국어 지식을 평가한 연구로 KoBBQ가 있다[14]. 한국 문화가 접목된 벤치마크를 제안하고, 한국의 상황과 사회적 편향을 반영하였다. KorFin-ASC[15]는 한국어 뉴스 데이터를 기반으로 하는데, 주로 금융 도메인에서의 감정 분류에 집중하여 좁은 범위를 대상으로 한 벤치마크를 제안하였다. KMMLU[18]는 한국어 시험에서 수집한 45개 영역 35,030개의 문제로 구성되어 있으나, 한국의 구체적이고 특수한 부분을 다룬 문항은 약 20%에 지나지 않는다.

기존 한국어 거대 언어 모델 벤치마크 연구는 연구 대상의 범위가 포괄적이지 못하다는 한계가 있다. 따라서 본 논문에서는 이러한 한계점을 보완하는 방향으로 한국형 벤치마크에 대한 연구를 진행하고자 한다.

3. 한국형 거대 언어 모델 역량 평가 벤치마크

3장에서는 한국형 거대 언어 모델의 역량을 평가하기 위해 벤치마크의 구성을 평가 영역, 문항 유형, 평가 지표의 순서로 검토한다. 난이도 변인을 분석한 연구에 따르면 전 영역에 걸쳐 나타나는 난이도의 공통적인 변인은 내용적 측면과 형식적 측면으로 구분된다[55]. 평가의 방식 또한 내용과 형식으로 분리할 수 있다고 판단하여 본 논문에서는 내용적 평가에 상응하는 평가 영역, 형식적 평가에 해당하는 문항 유형을 설정하였다. 평가 영역에서는 '한국적'인 것의 내용을 구체화하기 위해 영역을 9가지로 세분화하였다. 형식적 측면은 문제를 출제하는 방식으로 구현되며, 6가지의 문항 유형으로 확인할 수 있다. 마지막으로 평가하고자 하는 평가 주안점에 따라 평가 지표를 5가지로 정립하였다. 각 절에서는 본 논문이 제안하는 벤치마크 데이터셋의 사례를 제시하여 벤치마크의 구성을 시각적으로 파악할 수 있도록 하였다.

3.1 평가 영역

본 논문에서는 한국어 능력, 한국의 현행 정책, 한국의 실생활 경험 등을 포함한 종합적 평가를 진행한다. 한국 사회에 대한 이해도는 한국어뿐만 아니라 한국 사회 전반에 대한 이해 수준을 함께 고려해야 하기 때문이다. 한국 사회의 이해도를 시험과 같은 비교적 객관적 형태로 평가할 수 있는 방법에는 외국인의 한국 귀화 시험이 있다. 외국인이 한국 국적을 취득할 때 필수적으로 요구되는 한국 귀화 필기 시험은 한국어 능력 시험, 한국 사회 이해 능력 시험 2가지로 구성된다[56]. 외국인이 한국 귀화 시험에서 요구되는 능력을 갖추면 한국 국적을 취득할 수 있는 것과 마찬가지로 거대 언어 모델이 한국어와 한국 사회에 대한 이해도를 평가하는 시험에서 기준 이상의 점수를 획득하면 한국형 거대 언어 모델로서의 역량을 갖추었다고 전제한다.

본 논문은 현행 한국 귀화 시험의 영역 구분을 기반으로 세부 영역을 Table 1과 같이 재구성하였다. 수학이나 과학과 같이 보편적인 지식 분야는 국가별 특성이 약하므로[17], 소버린 AI의 평가와 거리가 멀다고 판단하여 평가 영역에서 제외하였다.

Table 1. Benchmark Evaluation Area

Evaluation Area	Detailed Area
Language &Literature	Non-fiction (expository writing, argumentative essays, personal essays, reports), Speech (spoken communication situations, dialects), Grammar, Literature (poetry, novels, prose, plays), Others
History	Prehistoric Era, Gojoseon and Various States, Three Kingdoms Period, Unified Silla and Balhae Period, Goryeo Dynasty, Joseon Dynasty, Japanese Occupation Period, Modern History, Historical Figures of Korea, Korean Cultural Heritage, Integrated Korean History, Others
Culture&Folklore	Traditional Values, Traditional Food, Clothing, and Housing, Rituals, Holidays, Religion, Popular Culture, Leisure Culture, Intangible Heritage, Commemorative Days, Others
Law	Constitution, Property Relations and Law, Family Relations and Law, Social Life and Law, Crime and Punishment, State and Institutions and Law, Everyday Laws, Others
Economy&Finance	Everyday Life, Economic Activities, Economic Policy, Financial Knowledge, Others
Politics	Korean Democratic Politics, Legislative Branch, Judicial Branch, Executive Branch, Elections and Local Autonomy, Civil Movements, Others
Geography	Climate, Topography, Population, Transportation, Capital Region (Seoul, Gyeonggi) and Provinces (Chungcheong, Jeolla, Gyeongsang, Gangwon, Jeju), Others
Society	Korean Symbols, Family, Workplace, Transportation and Communication, Housing, Urban and Rural Areas, Welfare, Healthcare and Safety, Others
Education	Early Childhood Education, Elementary School, Middle School, High School, University, Graduate School, Lifelong Education, Alternative Education, Others

각 영역에 따라 기대하는 거대 언어 모델의 역량은 다음과 같다.

(1) 언어 및 문학
문항의 내용을 이해하고, 적절한 어휘로 바꾸어 요약할 수 있다.
담화의 맥락을 파악하여 의사소통할 수 있다.
표준어와 지역의 방언을 이해한다.
한국어의 문법 구조를 이해하고 문법에 적합한 글쓰기가 가능하다.

(2) 역사
논란의 여지가 있거나 민감한 역사적 문제에 대해 편향된 시각을 갖지 않는다.
역사적 사건이나 인물에 대한 주관적 판단을 피하고 객관적 정보를 제공한다.
문화유산, 역사 해석 등 변화가 발생하는 영역에 대해 변화를 반영한 정보를 제공한다.

(3) 문화 및 민속
문화 및 민속의 형성 배경과 현재 시행 모습을 이해한다.
지역과 시기에 따른 다양한 문화 및 민속을 이해한다.
문화와 민속에 내재한 각각의 고유성을 인정하고 편향된 시각을 갖지 않는다.

(4) 법
여러 분야의 법에 대한 최신 정보 및 정확한 정보를 제공한다.
법률 지식을 전달할 때는 윤리적이고 중립적인 태도를 유지한다.
일상 용어와 법률 용어를 매끄럽게 변환한다.
제시된 상담 사례 및 판례에 대한 쟁점 및 요점을 전달한다.

(5) 경제 및 금융
한국의 경제정책 변화와 그에 따른 영향 및 경제 상황에 대한 최신 정보와 정확한 정보를 전달한다.
경제정책 및 경제 상황을 설명할 때 중립적인 태도를 유지한다.
일생에 필요한 경제 지식에 대해 정확한 정보를 전달한다.

(6) 정치
한국의 정치제도 변천과 현행 실태를 파악한다.
중앙과 지방의 정치 운영을 이해한다.
현재까지의 정치적 이슈와 지역 갈등을 중립적으로 다룬다.

(7) 지리
한국의 지리적 특성에 관한 정확한 정보를 제공한다.
한국 각 지역의 고유한 지리적, 문화적 특성을 이해한다.
현대 한국의 지리적 특성이 경제, 사회, 환경에 미치는 영향을 설명한다.

(8) 사회
현재 한국 사회에서 논의되고 있는 문제 또는 발생한 사건의 내용과 그에 대한 다양한 입장을 충분히 이해한다.
한국 사회 전반에 대한 정확한 사실을 제공한다.
한국 사회의 관습과 특징을 이해한다.

(9) 교육
한국 정부의 교육 정책에 따른 학교 제도, 교육 시스템 등에 대한 정확한 정보를 제공한다.
한국의 문화적 배경과 관련지어 교육 정책과 교육 환경에 대해 이해한다.
대학입시제도 등의 교육 정책과 교육 제도 내 특정 배경의 사용자에 편향되지 않는 전체적 정보를 제공한다.

3.2 문항 유형

본 논문에서 구축한 벤치마크는 4개의 선지에서 정답을 선택하는 객관식 사지선다의 형태이다. 문항의 형식은 '문제-지문-보기'의 세 부분으로 나뉜다. 문제는 답을 생성하기 위한 부가 설명 또는 해설과 문제 해결 방법을 지시하는 지시부(instruction)로 구성된다. 예를 들어 '빈칸에 들어갈 가장 적합한 단어를 보기에서 고르시오.'는 부가 설명 없이 지시부만 있는 문제이다. 지문은 문제를 풀기 위해 제시된 부연 설명 부분으로, 문제에 따라 선택적으로 제공된다. 보기는 1번부터 4번까지 정답과 오답이 혼재된 선지 전체를 의미한다. 실험 대상이 되는 거대 언어 모델은 문제를 확인하고, 지문을 읽으며, 보기에서 정답을 선택하는 과정을 거치게 된다. 또한 문제의 지시부에 '모두'를 표기하여 복수정답 여부를 명시하는 경우 복수정답도 가능하도록 구성하였다.

Table 2의 문제-보기 유형과 문제-지문-보기 유형 예시를 통해 문항의 형식을 확인할 수 있다. 문제-보기는 문항의 필수적 구성 요소이며 지문은 선택적으로 활용된다.

Table 2. Examples of Question Type

	Q-O Type	Q-P-O Type
Question	Select all the incorrect statements regarding members of the National Assembly.	Choose the correct pair of words to fill in the blanks.
Passage		The 6 years of elementary school and ( ) years of middle school are provided as a ( ) period of free education.
Options	1. On December 28, 2007, an amendment changed the number of Supreme Court Justices to 13, excluding the Chief Justice. 2. On December 4, 1987, an amendment reinstated the title 'Supreme Court Justice' and changed the number of justices to 13, excluding the Chief Justice. 3. On December 15, 2005, an amendment changed the number of Supreme Court Justices to 12, excluding the Chief Justice. 4. On January 29, 1981, an amendment changed the number of Supreme Court judges to 12, excluding the Chief Justice.	1. 3, Compulsory Education 2. 6, Compulsory Education 3. 3, Paid Education 4. 3, Non-Compulsory Education

본 논문에서는 형식에 따른 문항 유형을 빈칸채우기, 밑줄, 사실확인, 순서나열, 정답매칭, 복합의 6가지로 정의하였다. 빈칸채우기는 빈칸에 들어갈 적절한 내용을 추론하여 문제를 해결하는 유형이다. 밑줄은 밑줄이 있는 단어나 문장의 의미를 고려하여 정답을 찾는 유형이다. 거대 언어 모델의 프롬프트에 글자와 함께 밑줄을 입력할 수 없으므로 밑줄이 있는 부분의 단어나 문장을 '해당 건물'과 같이 작은 따옴표를 붙여 문제를 정확하게 파악할 수 있도록 수정하였다. 다만 형식상의 문제 유형 구분에 가장 직관적이라고 판단하여 '밑줄 유형'이라는 명칭을 그대로 사용한다.

사실확인은 문제가 요구하는 지시 사항에 가장 적합한(또는 부적합한) 내용을 선택하는 유형이다. 예컨대 '가장 적절한 설명을 고르시오.' 또는 '가장 부적합한 서술을 고르시오.'가 있다. 순서나열은 문제에서 지정한 기준에 따라 지문의 각 항목을 순서대로 나열하여 정답을 찾는 유형이다. 시기순 나열, 맥락에 따른 문장 나열, 과정 순서의 나열 등의 문항이 출제될 수 있다. 정답매칭은 지문에 무작위로 배치된 항목을 관련 있는 것끼리 연결하는 형태의 문항 형식이다. 복합은 이상에서 서술한 5개 유형이 2개 이상 활용된 유형을 의미한다. 예컨대 빈칸이 있는 지문을 주고, '해당 빈칸에 들어갈 단어에 대한 설명으로 옳은 것을 고르시오.'와 같은 문항은 빈칸채우기와 사실확인이 결합한 유형이므로 복합유형이라고 지칭할 수 있다.

형식에 따른 문항 유형의 예시는 Table 3에 있다. 복합유형은 결합되는 유형에 따라 문항이 상이한데, Table 3의 복합 유형은 빈칸채우기와 사실확인이 결합한 사례이다.

Table 3. Examples by Question Type – Format

Type	Example
Fill-in-the-Blank	Choose the correct word to fill in the blank. To address the undervaluation of the domestic stock market and enhance shareholder value, the Korean government initiated the ( ) program in 2024. 1. K-Stock Activation **2. Corporate Value-Up-Correct** 3. Bubble 4. Giant-step
Underline	The following is part of a conversation between Yeongho, Minsu, and Jingu on their way home from school. Choose the one instance of 'our' that has a different meaning. Yeongho: Can we play at ㄱ.'our' house today? ㄴ.'Our' mom bought a delicious cake. Minsu: Really? I have time before going to my academy, so let's go together. Jingu, let ㄷ.'us' go together. Jingu: Great. Then we'll also get to see Sujin, whom Yeongho always calls ㄹ.'our' little sister and dotes on. 1. ㄱ 2. ㄴ **3. ㄷ-Correct** 4. ㄹ
Fact-Checking	Choose the incorrect statement regarding the differences in investigative authority between the prosecution and the police. 1. When an investigation begins through a complaint or accusation, the police make a decision on forwarding or not forwarding the case, while the prosecution decides whether to indict or not indict the forwarded case. **2. The prosecution can be involved in the investigation phase at the police level.-Correct** 3. If a warrant is required during the police investigation stage, the prosecutor is involved. 4. If the police determine that there is no suspicion of a crime after the investigation, they can decide not to forward the case and exercise primary authority to close the investigation.
Ordering	Choose the option that correctly arranges the events related to the opening of Joseon's ports in order. ㄱ. Japan sent ships to Joseon, modeling its diplomatic approach after the United States ㄴ. A treaty was signed that, for the first time, included provisions on tariffs and mediation. ㄷ. A treaty was concluded with a country where negotiations had been delayed due to issues related to the recognition of Catholicism. ㄹ. It was agreed to open not only Busan but also two other ports. **1. ㄱ-ㄹ-ㄴ-ㄷ-Correct** 2. ㄱ-ㄷ-ㄹ-ㄴ 3. ㄹ-ㄷ-ㄱ-ㄴ 4. ㄷ-ㄱ-ㄴ-ㄷ
Answer Matching	Choose the option that pairs A, B, and C with (가), (나), and (다) in the most natural way. A. Chungju B. Cheongju C. Gongju (가) The local government's YouTube channel recently became a major topic. (나) The Chungcheongnam-do History Museum is located there. (다) The King Sejong and Chojeong Mineral Water Festival takes place in October. 1. A-(나) 2. B-(나) 3. C-(다) **4. A-(가)-Correct**
Complex	Choose the correct word to fill in the blank and the correct description paired with the type of middle school. Middle schools are classified into general middle schools and ( ) middle schools. 1. National, admission is through a separate selection process. 2. Specialized, assigned to a location near home. 3. Private, offers career exploration education. 4. Specialized, educates students with talents in specific fields.-Correct

3.3 평가 지표

평가 지표는 모델 개발의 방향성을 안내하고, 거대 언어 모델의 발전 수준을 평가하는 정량적 도구이다[57]. 효과적인 평가 지표는 모델 출력에서 모델의 역량을 설명하고, 모델의 결함을 진단하며, 다양한 모델의 강점과 약점을 비교하는 데 유용하게 작용하기 때문에[58] 적절한 평가 지표의 설정이 중요하다. 자연어 생성 분야에서는 주요 평가 지표로 유창성, 사실성, 다양성을 제시한다[59]. 즉 생성된 텍스트가 맥락적으로 잘 구성되어야 하고, 의미를 잘 표현해야 하며, 최대한 다양한 단어를 활용해 생성해야 높은 성능을 갖는 모델로 간주할 수 있다.

본 논문에서는 기존 자연어 생성 분야의 평가 지표를 수용하여 완결성과 다양성의 측면에서 유창성으로, 의미 표현의 명시성 측면에서 사실성으로 재구성하였다. 최신성은 사실성과 연결되는 영역이지만 급변하는 정보에 거대 언어 모델이 얼마나 잘 적응하는지를 검사하기 위해 별도로 설정한 항목이다. 편향성은 거대 언어 모델에서 공정성을 보장하는 중요한 기준으로 작용하며[60], 역사적, 구조적으로 발생하기 때문에[61] 소버린 AI의 평가에 있어 필수적이다. 문화 맥락 이해는 한국적 상식 수준을 평가하기 위해 추가하였다. 나머지 4가지 항목에 포함되지 않으면서, 한국에서 생활하며 체화하는 사회문화적 습관을 거대 언어 모델이 얼마나 이해하고 있는지를 평가하기 위함이다.

이상의 내용을 기반으로 Ko-Sovereign 벤치마크 평가 지표인 유창성, 최신성, 사실성, 편향성, 문화 맥락 이해의 정의를 다음의 Table 4에 정리하였다.

Table 4. Evaluation Metrics of the Ko-Sovereign Benchmark

Evaluation Metric	Definition
Fluency	evaluates how well the content of the Korean language is understood from the perspectives of pragmatics and discourse.
Recency	assesses how accurately one responds to issues encompassing recent changes.
Factuality	evaluates cases where there is a clearly defined 'correct answer,' such as numbers or named entities.
Bias	assesses potential social biases that may occur within Korean society.
Cultural Context Comprehension	evaluate Korean culture by verifying culturally ingrained experiences of Koreans.

유창성은 문맥에 적합하게 대응하는지, 대명사 또는 대용어를 정확하게 해석하는지, 한국어 어휘를 잘 선택하여 답변을 생성하는지, 한국어의 방언을 표준어에 가깝게 인식하는지를 평가한다. 최신성은 최근 2년 내 발생한 변화에 한정한다. 학습 여부와 상관없이 거대 언어 모델이 최신 정보에 대해 어떻게 대응하는지 평가하는 부분이 필요하다고 판단하였다. 사실성은 정확한 정보를 요구하는 질문에 얼마나 정답에 가깝게 답변하는지를 측정한다. 편향성은 사회적으로 편향된 내용이 있는 질문과 선지를 함께 구성하여 평가한다. 이때 편향성은 최근 거대 언어 모델에서 밝힌 다양한 범주의 편향을 종합하여[61-64] '인종, 성별, 세대, 지역, 종교, 문화, 역사 등에 따라 행동이 정해진 것으로 보는 정도'로 정의한다. 문화 맥락 이해는 한국형 벤치마크에 필수적인 지표로, 다른 지표들에 해당하지 않으면서 한국인의 문화적 경험에 따라 정답을 고를 수 있는 문항에 활용된다.

이상에서 서술한 Ko-Sovereign 벤치마크의 5가지 평가 지표에 대한 문항 예시는 Table 5에서 확인할 수 있다.

Table 5. Examples According to the Ko-Sovereign Benchmark Evaluation Metrics

Metric	Example
Fluency	Choose the most appropriate expression to fill in the parentheses. a. The posters are ( ) stuck to the wall. b. The sand on the beach ( ) shines. 1. a – thickly b – brightly flashing 2. a – heavily b – brightly flashing **3. a – thickly b – sparkling –correct** 4. a – heavily b – sparkling
Recency	Choose all the incorrect statements regarding the Onnuri Gift Certificate. **1. It can be used at academies, hospitals, dental clinics, large supermarkets, etc., located within traditional markets nationwide.–Correct** 2. It is generally available for purchase at a 5–10% discounted price. 3. There are paper certificates, mobile certificates, and card certificates. **4. It can be purchased up to 2.5 million won per month.–Correct**
Factuality	Select all instances where the 22nd National Assembly member and their constituency match. 1. Kim Hyung-dong-Yangsan-si, Gyeongsangnam-do **2. Park Jeong-Paju-si,Gyeonggi-do –Correct** 3. Ahn Sang-hoon-Naju-si, Hwasun-gun, Jeollanam-do **4. Cha Ji-ho-Osan-si, Gyeonggi-do –Correct**
Bias	Choose the incorrect statement regarding childcare for children with disabilities. **1. Field trips are not conducted for children with disabilities because they may cause accidents outside the daycare center.–Correct** 2. Rehabilitation programs and individualized education programs are offered simultaneously at daycare centers for children with disabilities. 3. To train early childhood special education teachers, universities have established departments for early childhood special education. 4. The first childcare facility for children with disabilities was the St. John Paul II Specialized Childcare Center for Children with Disabilities in Daegu."
Cultural Context Comprehension	Choose the ingredient that is the least suitable as a component of kimchi stew. **1. Lamb–Correct** 2. Pork 3. Pacific Saury 4. Tuna

4. 실험 및 환경

4.1 평가 대상 모델

본 논문에서는 한국적 사회 맥락을 이해하는 능력과 한국 관련 지식의 서술 정확도를 비교 평가하기 위해, 총 8개의 언어 모델을 대상으로 실험을 수행하였다. 실험 대상 모델은 다국어 거대 언어 모델과 한국어 특화 언어 모델로 구성되었으며, 선정된 모델은 ChatGPT-3.5-turbo(OpenAI, 이하 ChatGPT3.5), ChatGPT-4o(OpenAI, 이하 ChatGPT4o), Claude-3-haiku(Anthropic, 이하 Claude3), Gemma-2-9b-it(Google, 이하 Gemma2), CLOVA X(Naver), KULLM-Uracle(고려대학교), EXAONE-3.0-7.8B-Instruct(LG AI Research, 이하

EXAONE3), Solar Mini-Chat(Upstage)이다.

모델 선정 이유는 다음과 같다. 첫째, 한국적 사회 맥락 이해와 한국 관련 지식 서술 능력을 평가하기 위한 실험이라는 점에서 한국어 특화 모델과 다국어 모델 간의 성능 차이를 분석하기 위해 대표적인 모델을 포함하였다. 둘째, ChatGPT3.5 모델은 베이스라인으로 설정하여 다른 모델과의 상대적 성능을 비교할 수 있도록 하였다. 셋째, 다국어 거대 언어 모델로는 세계적으로 널리 사용되며 비교적 언어 이해 능력이 높은 ChatGPT4o, Claude3, Gemma2를 선정하였다. 넷째, 한국어 특화 언어 모델로 국내 연구기관과 기업에서 개발한 CLOVA X, KULLM-Uracle, EXAONE3, Solar Mini-Chat을 포함하였다.

이들 모델은 토큰 정책, 응답 길이, 실험 환경 등을 종합적으로 고려하여 선정되었으며, 다양한 언어 모델이 한국적 맥락에서 언어 이해와 정보 서술 능력에서 어떻게 차별화되는지 정량적으로 평가한다.

Table 6. Overview of Experimental Models

	Model	Developer
Multilingual LLM	ChatGPT3.5	OpenAI
	ChatGPT4o	OpenAI
	Claude3	Anthropic
	Gemma2	Google
Korean-Specialized LLM	CLOVA X	Naver
	KULLM-Uracle	Korea Univ.
	EXAONE3	LG
	Solar Mini-Chat	Upstage

4.2 평가 데이터

벤치마크는 총 450개의 문항으로 구성된다. 내용에 해당하는 평가 영역과 형식에 해당하는 문항 유형은 Table 7과 같다. 영역별 50문항씩 총 450문항이며 가장 많은 비중을 차지하는 문항은 사실확인이다. 빈칸채우기-밑줄-정답매칭-순서나열-복합의 순서로 많이 출제되었다.

Table 7. Question Composition by Question Type

	Fill-in-the-Blank	Under-line	Fact-Checking	Ordering	Answer-matching	Complex
Language & Literature	8	9	26	0	5	2
Hitory	10	4	28	4	3	1
Culutre & Folklore	3	0	47	0	0	0
Law	7	3	35	2	3	0
Economy & Finance	12	3	33	0	2	0
Politics	10	6	25	3	6	0
Geo-graphy	11	5	27	2	2	3
Society	4	2	32	1	10	1
Education	7	7	23	3	6	4
Total	72 (16%)	39 (8.7%)	276 (61.3%)	15 (3.3%)	37 (8.2%)	11 (2.5%)

평가 지표별 문항 유형은 단일한 평가 지표를 가지는 문항이 총 346개이며 유창성 24문항, 사실성 307문항, 편향성 1문항, 문화 맥락 이해 14문항이다. 두 개의 복수 평가 지표를 가지는 문항은 총 100개로, 유창성-문화 맥락 이해 5문항, 유창성-사실성 24문항, 최신성-사실성 45문항, 최신성-문화 맥락 이해 2문항, 사실성-편향성 4문항, 사실성-문화 맥락 이해 20문항이다. 세 개의 복수 평가 지표를 가지는 문항은 총 3문항으로, 유창성-사실성-최신성 2문항, 최신성-사실성-편향성 1문항으로 구성된다.

4.3 평가 방법

Ko-Sovereign 벤치마크 데이터를 8개의 거대 언어 모델 대상으로 실험을 진행하기 위해 각 질문별로 제로샷을 수행하였다. 제로샷은 각 언어 모델에 별도의 추가 학습이나 훈련 없이 주어진 질문에 답변하도록 했다는 의미이다.

주어진 영역별 질문에 대해 각 언어 모델별로 응답 정확도를 측정했다. 그러나 선다형 질문임에도 불구하고, 각 언어 모델마다 서로 다른 형태로 답변이 생성될 수 있다. 본 논문에서는 선다형 질문임을 감안하여 생성 결과의 숫자를 기준으로 숫자와 설명이 모두 정확한 경우, 숫자는 정확하지만 설명이 틀린 경우를 정답으로 간주하여 각 모델의 정확도를 정량적으로 측정하였다.

5. 실험 결과

5.1 영역별 성능 평가

Table 8은 Ko-Sovereign의 평가 영역에 대한 모델별 정답률을 보여준다. 영역별로 가장 높은 정답률을 보인 모델의 정답률에 대해 음영 처리하였다. 역사 영역의 경우는 ChatGPT4o와 EXAONE3가 동일한 정답률을 보였고, 이외의 영역에 대해서는 ChatGPT4o가 가장 높은 정답률을 보였다. 정답률 평균에서 가장 높은 점수를 낸 것도 ChatGPT4o였다. ChatGPT4o가 한국 특화 거대 언어 모델이 아닌 다국어 언어 모델이라는 점, ChatGPT3.5와 비교하여 영역별 정답률 차이가 확연하게 드러난다는 점을 통해 ChatGPT4o의 성능이 이전 버전에 비해 주목할 정도로 향상된 것이 확인된다. ChatGPT3.5의 평균 정답률이 38.4%인데 반해 ChatGPT4o는 62.0%으로 성능이 23.6% 향상되었다. 게다가 ChatGPT4o의 정답률 1, 2, 3 순위에 문화 및 민속 영역, 지리 영역, 정치와 사회 영역(동

Table 8. Accuracy Rate (%) of Large Language Models by Ko-Sovereign Benchmark Area

Evaluation Area	ChatGPT 3.5	ChatGPT 4o	Claude3	Gemma2	CLOVA X	KULLM -Uracle	EXAONE3	Solar Mini-Chat
Language &Literature	42.0	64.0	58.0	42.0	42.0	40.0	42.0	34.0
History	34.0	46.0	36.0	38.0	34.0	28.0	46.0	44.0
Culutre&Folklore	46.0	72.0	48.0	50.0	44.0	52.0	64.0	58.0
Law	30.0	56.0	42.0	52.0	26.0	40.0	44.0	42.0
Economy&Finance	50.0	64.0	54.0	44.0	46.0	44.0	52.0	56.0
Politics	36.0	66.0	50.0	44.0	34.0	38.0	64.0	60.0
Geography	32.0	70.0	50.0	52.0	24.0	42.0	46.0	50.0
Society	38.0	66.0	52.0	52.0	38.0	32.0	54.0	54.0
Education	38.0	54.0	50.0	48.0	30.0	34.0	44.0	38.0
Average	38.4	62.0	48.9	46.9	35.0	38.9	50.7	48.4

률)이 포함된 것은 ChatGPT4o가 한국 사회만이 가지고 있는 문화적 특징과 사회적 경향성을 충분히 파악하고 있을 가능성을 시사한다.

한국어와 영어에 집중된 오픈 소스 모델 EXAONE3는 평균 50.7%의 정답률을 기록하며 ChatGPT4o 다음으로 높은 정답률을 기록했다. 역사 영역은 ChatGPT4o와 동일한 점수를, EXAONE3 자체 성적으로는 문화 및 민속, 정치 영역에서 가장 고점을 달성하였다. 한국형 거대 언어 모델에서 기대할 만한 언어 및 문학 영역이 다소 낮은 점수를 받은 것으로 보일 수도 있으나 공개된 모델의 크기가 7.8B인 것을 감안하면 비교적 높은 성능이다. KULLM-Uracle 모델과 Solar Mini-Chat 모델은 베이스라인인 ChatGPT3.5 모델에 비해서 평균 정답률은 높았으나 EXAONE3와 유사하게 언어 및 문학에서 낮은 점수를 기록하였다. 다국어 거대 언어 모델의 언어 및 문학 영역 평균 정답률이 51.5%, 한국형 거대 언어 모델의 그것이 38.6%임을 고려하면 후자의 자연어 이해 및 생성 영역에 대한 경쟁력은 추가적으로 확보되어야 할 것이다.

거대 언어 모델의 정답률 평균이 가장 낮은 영역은 역사였다. 전체적으로 성능이 가장 높다고 평가된 Chat GPT4o에서 조차도 역사는 50% 미만을 기록했다. 역사 영역의 최고 정답률이 Chat GPT4o와 EXAONE3의 46%인 것을 감안하면 여전히 거대 언어 모델의 한국 역사에 대한 학습이 부족하다고 평가할 수 있다.

5.2 문항 유형별 성능 평가

Table 9에 Ko-Sovereign의 450개 문항을 6개 유형으로 나눠서 유형별 정답률을 정리하였다. 각 문항 유형에서 최고점을 달성한 모델은 전 유형에서 ChatGPT4o였다. 그중 복합 유형의 정답률이 80% 이상으로 높은 편이었는데, 문항 유형이 결합한 복합 유형에서 높은 정답률을 보이는 것은 ChatGPT4o의 문제 이해도가 높다는 의미로 해석된다.

문항 유형별 성능을 파악하기 위해서는 각각의 모델 내에서 정답률이 높은 문항 유형이 무엇인지 살펴볼 필요도 있다. 먼저 ChatGPT3.5, ChatGPT4o, Gemma2, EXAONE3 총 4개 모델에서는 복합 유형이 가장 정답률이 높았다. ChatGPT4o를 제외한 모델 중 가장 평균 정답률이 높은 EXAONE3의 경우 복합을 제외한 다른 유형에서의 평균 정답률이 45.7%로, ChatGPT4o의 61.6%에 비해 낮아서 문제 자체에 대한 이해도가 높다고 담보할 수는 없다. 그리고 복합 유형은 여러 평가 모델에서 평균적으로 높은 정답률을 보이고 있으므로 복합 유형의 높은 정답률에 내용적 난이도가 영향을 미쳤을 가능성을 배제할 수 없다. 이 역시 벤치마크 추가 구축 등의 후속 연구가 필요한 지점이다.

Claude3와 CLOVA X, Solar Mini-Chat 총 3개 모델에서는 빈칸채우기 유형이 가장 정답률이 높았다. 이 유형은 빈칸에 들어갈 내용만을 추론하는 단순한 유형이기 때

Table 9. Accuracy Rate (%) of Large Language Models by Ko-Sovereign Benchmark Question Type

	ChatGPT 3.5	ChatGPT 4o	Claude3	Gemma2	CLOVA X	KULLM -Uracle	EXAONE3	Solar Mini-Chat
Fill-in-the-Blank	50.0	68.1	58.3	44.4	52.8	44.4	51.4	63.9
Underline	30.8	51.3	43.6	43.6	23.1	33.3	46.2	38.5
Fact-Checking	38.0	59.8	50.4	50.7	33.3	39.5	51.8	47.1
Ordering	26.7	53.3	26.7	26.7	26.7	46.7	33.3	40.0
Answer-Checking	27.0	75.7	32.4	29.7	37.8	29.7	45.9	37.8
Complex	54.5	81.8	54.5	63.6	18.2	27.3	72.7	63.6

문에 많은 모델에서 다른 유형에 비해 정답을 쉽게 맞출 수 있었던 것으로 보인다.

모든 모델에서 높은 정답률을 기록한 문항 유형은 빈칸 채우기와 사실 확인 유형이다. 이들은 단순한 유형의 문항이므로 비교적 정답률이 높은 것으로 보인다. 반면 밑줄, 순서나열, 정답매칭 유형은 두 단계 이상의 문제 풀이 과정을 거쳐야 한다. 예를 들어 밑줄 유형의 경우 밑줄이 가리키는 단어가 무엇인지 추론한 후 해당 개념에 대한 사실을 다시 검토하는 식이다. 따라서 대체적으로 빈칸채우기와 사실 확인 유형에 비해 정답률이 낮은 모습을 보여준다.

5.3 평가 지표별 성능 평가

유창성, 최신성, 사실성, 편향성, 문화 맥락 이해에 대한 모델 성능은 Table 10과 같다. Ko-Sovereign 문항의 평가 지표는 유창성, 최신성, 사실성, 편향성, 문화 맥락 이해 5가지로 구성된다. 다만 단일한 평가 지표가 아니라 2개 이상의 복수 평가 지표를 갖는 문항이 존재한다.

이 경우 해당 지표가 하나라도 있는 문항의 정답 여부를 포함하여 정답률을 계산하였다. 예를 들어 사실성과 최신성을 평가하는 문항에 대한 정답 여부는 사실성과 최신성의 정답률에 포함되는 방식이다. 이러한 계산식에 근거하여 평가 지표별 정답률을 Table 10으로 정리하였다. 그리고 각 평가 지표에서 가장 높은 정답률을 보인 모델을 음영 처리하였다.

각 평가 지표별로 높은 정답률을 보인 모델은 다음과 같다. 유창성은 Claude3, 최신성은 EXAONE3, 사실성은 ChatGPT4o, 문화 맥락 이해는 Solar Mini-Chat의 정답률이 높았고, 편향성은 ChatGPT4o, Claude3, EXAONE3, Solar Mini-Chat이 공동 순위를 기록했다. ChatGPT4o는 5개 지표 가운데 2개 지표에서 고득점을 차지하였고, 문화 맥락 이해와 유창성에서도 근소한 차이로 2위를 차지하였다. 이는 ChatGPT4o가 비교적 한국의 사회 문화 맥락에 적합하면서도 정확하게 정보를 제공하는 성능을 가지고 있다고 판단할 수 있는 근거가 된다. 최신성 분야에서는 EXAONE3가 ChatGPT4o에 약간 앞선 모습이 확인되는데, 2024년 8월에 공개되어 가장 최신 데이터로 학습되었기 때문으로 풀이된다(LG, 2024). 한편 문화 맥락 이해에서 Solar Mini-Chat이 보인 강세도 주목할 만하다. Solar Mini-Chat이 한국의 문화 현상을 맥락적으로 잘 파악하고 있음을 의미한다.

다음으로 각 모델별로 가장 높은 정답률을 보인 평가 지표를 검토한다. ChatGPT3.5는 유창성, CLOVA X는 문화 맥락 이해에서 최고점을 받았고, 이외 모델은 편향성의 평가지표에서 가장 높은 정답률을 기록했다. ChatGPT3.5, CLOVA X를 제외한 대상 모델이 한국 내에서 발생 가능한 편향의 정도를 판단하고, 편향성이 있는 선지를 적절하게 골라냈다는 의미이다. 다만 현재 벤치마크에서 편향성의 비율이 1.3%로 매우 적기 때문에 추후 편향성 지표를 보강하기 위하여 문항을 확대 구축하고, 추가 실험을 진행할 필요가 있다.

Table 10. Accuracy Rate (%) of Large Language Models by Ko-Sovereign Benchmark Evaluation Metric

	ChatGPT3.5	ChatGPT4o	Claude3	Gemma2	CLOVA X	KULLM-Uracle	EXAONE3	Solar Mini-Chat
Fluency	52.7	63.6	65.5	54.5	45.5	50.9	41.8	47.3
Recency	38.0	50.0	48.0	50.0	28.0	26.0	52.0	46.0
Factuality	37.2	60.5	47.1	46.2	32.8	37.2	50.1	46.7
Bias	33.3	83.3	83.3	66.7	50.0	66.7	83.3	83.3
Cultural Context Comprehension	43.9	68.3	53.7	53.7	53.7	53.7	63.4	70.7

6. 결론

본 논문에서는 소버린 AI 구현의 실질적인 방법으로 거대 언어 모델의 한국화 정도를 평가할 수 있는 Ko-Sovereign 벤치마크를 제안하였다. 기존에 제안된 한국어 벤치마크 연구가 한국어 능력이나 한국적 상식을 주로 평가하였던 것에서 나아가 한국의 법, 경제, 정치와 같은 한국에 특화된 전문 지식이 필요한 부분도 평가 영역에 포함하였다. 일차적으로 평가 영역을 언어 및 문학, 역사, 문화 및 민속, 법, 경제 및 금융, 정치, 지리, 사회, 교육의 9개로 세분화하고, 거대 언어 모델을 평가하는 지표를 유창성, 최신성, 사실성, 편향성, 문화 맥락 이해로, 문항의 유형을 빈칸채우기, 밑줄, 사실확인, 순서나열, 정답매칭, 복합으로 체계화하였다. 평가 항목을 여러 측면에서 세분화함으로써 한국형 거대 언어 모델의 역량을 종합적으로 측정할 수 있으며, 궁극적으로는 거대 언어 모델이 취약한 부분을 분야별, 기능별로 진단할 수 있다.

다국어 언어 모델과 한국형 거대 언어 모델 간의 성능 차이를 비교 분석하기 위하여 다국어 언어 모델인 ChatGPT3.5, ChatGPT4o, Claude3, Gemma2를, 한국형 거대 언어 모델인 KULLM-Uracle, EXAONE3, Solar Mini-Chat을 대상으로 실험을 진행하였다. 실험은 별도의 추가 학습이나 훈련 없이 주어진 질문에 답변하는 제로샷으로 수행하였다. 평가 영역, 문항 유형, 평가 지표를 종합했을 때 가장 높은 정확도를 보인 모델은 ChatGPT4o이다. 특히 내용에 해당하는 평가 영역 가운데 문화 및 민속, 지리, 정치, 사회 영역이 상위권이므로 ChatGPT4o가 한국 사회의 문화적 특징과 사회적 경향성을 충분히 파악하였다고 평가

할 수 있다. 한편 EXAONE3가 ChatGPT4o 다음으로 역사와 문화 및 민속 영역에서 높은 점수를 차지한 점은 고무적이다. 두 모델의 파라미터 크기를 감안하였을 때 EXAONE3는 한국의 문화적 맥락에 민감하게 반응하면서도 한국적 지식을 정확하게 파악하는 수준의 성능을 가졌다고 판단된다.

본 논문에서 제안한 Ko-Sovereign 벤치마크는 문법, 담화 등의 한국어 세부 영역뿐만 아니라 법, 정치, 경제 등 한국의 현행 정책, 관련 전문 지식, 한국의 역사와 문화를 평가 대상으로 포함하여 거대 언어 모델의 종합적 평가를 목표하였다. 다만 벤치마크 데이터셋 평가 내용의 다양화와 평가 지표의 체계화에 집중한 만큼 데이터셋의 크기를 확장하지 못하였다는 점은 한계로 지적될 만하다. 특히 편향성을 평가 지표로 하는 문항의 구축, 문항의 난이도를 고려한 복합 유형 문항의 추가 구축 등은 향후 검토가 필요한 부분이다. 본 논문이 제안한 Ko-Sovereign 벤치마크가 한국형 거대 언어 모델의 성능 평가에 활용됨으로써 모델의 성능 향상에 기여할 수 있기를 바라며, 궁극적으로는 한국의 소버린 AI 구축, 디지털 주권의 구현에 이바지할 수 있었으면 한다.

참고문헌

[1] Gillibrand, N., & Draper, C. (2023). *Informational Sovereignty: A New Framework For AI Regulation (Working Paper)*. University College Dublin. School of Law. http://hdl.handle.net/10197/24564

[2] Jung Woo Ha. (2024). The Era of Transformation Brought by Generative AI: Our Current Position and Strategies Needed for the Future(생성 AI가 불러온 대전환 시대 우리의 현재 위치와 미래를 위해 필요한 전략은?). *Digital Economy View*, 4-13.

[3] Yoo, K. M., Han, J., In, S., Jeon, H., Jeong, J., Kang, J., ... & Jung, J. (2024). HyperCLOVA X Technical Report. *arXiv preprint arXiv:2404.01954*. https://doi.org/10.48550/arXiv.2404.01954

[4] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... & McGrew, B. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*. https://doi.org/10.48550/arXiv.2303.08774

[5] Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., & Tafjord, O. (2018). Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*. https://doi.org/10.48550/arXiv.1803.05457

[6] Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., & Choi, Y. (2019). Hellaswag: Can a machine really finish your sentence?. *arXiv preprint arXiv:1905.07830*. https://doi.org/10.48550/arXiv.1905.07830

[7] Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2020). Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*. https://doi.org/10.48550/arXiv.2009.03300

[8] Lin, S., Hilton, J., & Evans, O. (2021). Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*. https://doi.org/10.48550/arXiv.2109.07958

[9] Sakaguchi, K., Bras, R. L., Bhagavatula, C., & Choi, Y. (2021). Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM, 64*(9). 99-106. https://doi.org/10.1145/3474381

[10] Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., ... & Schulman, J. (2021). Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*. https://doi.org/10.48550/arXiv.2110.14168

[11] Ham, J., Choe, Y. J., Park, K., Choi, I., & Soh, H. (2020). KorNLI and KorSTS: New benchmark datasets for Korean natural language understanding. *arXiv preprint arXiv:2004.03289*. https://doi.org/10.48550/arXiv.2004.03289

[12] Park, S., Moon, J., Kim, S., Cho, W. I., Han, J., Park, J., ... & Oh, T. (2021). Klue: Korean language understanding evaluation. *arXiv preprint arXiv:2105.09680*. https://doi.org/10.48550/arXiv.2105.09680

[13] Kim, D., Jang, M., Kwon, D. S., & Davis, E. (2022). Kobest: Korean balanced evaluation of significant tasks. *arXiv preprint arXiv:2204.04541*. https://doi.org/10.48550/arXiv.2204.04541

[14] Jin, J., Kim, J., Lee, N., Yoo, H., Oh, A., & Lee, H. (2023). KoBBQ: Korean bias benchmark for question answering. *arXiv preprint arXiv:2307.16778*. https://doi.org/10.48550/arXiv.2307.16778

[15] Son, G., Lee, H., Kang, N., & Hahm, M. (2023). Removing non-stationary knowledge from pre-trained language models for entity-level sentiment classification in finance. *arXiv preprint arXiv:2301.03136*. https://doi.org/10.48550/arXiv.2301.03136

[16] Kim, E., Suk, J., Oh, P., Yoo, H., Thorne, J., & Oh, A. (2024). CLIcK: A Benchmark Dataset of Cultural and Linguistic Intelligence in Korean. *arXiv preprint arXiv:2403.06412*. https://doi.org/10.48550/arXiv.2403.06412

[17] Lee, J., Kim, M., Kim, S., Kim, J., Won, S., Lee, H., & Choi, E. (2024). KorNAT: LLM Alignment Benchmark for Korean Social Values and Common Knowledge. *arXiv preprint arXiv:2402.13605*. https://doi.org/10.48550/arXiv.2402.13605

[18] Son, G., Lee, H., Kim, S., Kim, S., Muennighoff, N., Choi, T., ... & Biderman, S. (2024). Kmmlu: Measuring massive multitask language understanding in korean. *arXiv preprint arXiv:2402.11548*. https://doi.org/10.48550/arXiv.2402.11548

[19] Angie Lee. (2024, February 28). What Is Sovereign AI?. NVIDIA Blog. Retrieved from https://blogs.nvidia.com/blog/what-is-sovereign-ai/

[20] Institute of Information & communications Technology Planning & Evaluation(IITP)(2024). ICT Brief 2024-25.

[21] Brian Caulfield. (2024, February 12). *NVIDIA CEO: Every Country Needs Sovereign AI*. NVIDIA Blog. Retrieved fromhttps://blogs.nvidia.com/blog/world-

governments-summit/

[22] Floridi, L. (2020). The fight for digital sovereignty: What it is, and why it matters, especially for the EU. *Philosophy & technology, 33*. 369-378. https://doi.org/10.1007/s13347-020-00423-6

[23] Team, G., Anil, R., Borgeaud, S., Alayrac, J. B., Yu, J., Soricut, R., ... & Blanco, L. (2023). Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*. https://doi.org/10.48550/arXiv.2312.11805

[24] Jung Woo Ha(2021). Naver's AI Platform CLOVA and Hyper-Scale AI HyperCLOVA(네이버 AI플랫폼 CLOVA 그리고 초대규모 AI HyperCLOVA). *The Transactions of the Korea Information Processing Society, 28*(3). 55-66.

[25] Sengupta, N., Sahu, S. K., Jia, B., Katipomu, S., Li, H., Koto, F., ... & Xing, E. (2023). Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *arXiv preprint arXiv:2308.16149*. https://doi.org/10.48550/arXiv.2308.16149

[26] Luukkonen, R., Burdge, J., Zosa, E., Talman, A., Komulainen, V., Hatanpää, V., ... & Pyysalo, S. (2024). Poro 34B and the Blessing of Multilinguality. *arXiv preprint arXiv:2404.01856*. https://doi.org/10.48550/arXiv.2404.01856

[27] Roberts, H., Cowls, J., Casolari, F., Morley, J., Taddeo, M., & Floridi, L. (2021). Safeguarding European values with digital sovereignty: an analysis of statements and policies. *Internet Policy Review, Forthcoming*. http://dx.doi.org/10.2139/ssrn.3937345

[28] Ong, D., & Limkonchotiwat, P. (2023, December). SEA-LION (Southeast Asian Languages In One Network): A Family of Southeast Asian Language Models. *In Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023). 245-245.* https://doi.org/10.18653/v1/2023.nlposs-1.26

[29] Chang, T. A., Arnett, C., Tu, Z., & Bergen, B. K. (2023). When is multilinguality a curse? language modeling for 250 high-and low-resource languages. *arXiv preprint arXiv:2311.09205*. https://doi.org/10.48550/arXiv.2311.09205

[30] Kang Yejee, Kim Hansaem, Park Seoyoon, Kang Joeun, Kim Yujin, Lee Jaewon, Jung Gayeon, Choi Gyuri, Choi Changsu, Won Inho, Kim Minjun, Lim Hyeonseok, Lim GyeongTae, Ham Yeonggyun. (2024). Comprehensive Evaluation of LLMs' Korean Language Proficiency(인공지능의 한국어 능력 종합 평가를 위한 벤치마크). *Hanguel, 85*(3). 679-715. https://doi.org/10.22557/HG.2024.9.85.3.679

[31] Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. (2018). GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In T. Linzen, G. Chrupała, & A. Alishahi (Eds.), *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP* (pp. 353-355). Brussels, Belgium: Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-5446

[32] Xu, L., Hu, H., Zhang, X., Li, L., Cao, C., Li, Y., ... & Lan, Z. (2020). CLUE: A Chinese language understanding evaluation benchmark. *arXiv preprint arXiv:2004.05986*. https://doi.org/10.48550/arXiv.2004.05986

[33] Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., ... & Bowman, S. (2019). Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems, 32*. https://doi.org/10.48550/arXiv.1905.00537

[34] Yin, D., Bansal, H., Monajatipoor, M., Li, L. H., & Chang, K. W. (2022). Geomlama: Geo-diverse commonsense probing on multilingual pre-trained language models. *arXiv preprint arXiv:2205.12247*. https://doi.org/10.48550/arXiv.2205.12247

[35] Nguyen, T. P., Razniewski, S., Varde, A., & Weikum, G. (2023, April). Extracting cultural commonsense knowledge at scale. *In Proceedings of the ACM Web Conference 2023*. 1907-1917. https://doi.org/10.1145/3543507.3583535

[36] Park, C., Kim, H., Kim, D., Cho, S., Kim, S., Lee, S., ... & Lee, H. (2024). Open Ko-LLM Leaderboard: Evaluating Large Language Models in Korean with Ko-H5 Benchmark. *arXiv preprint arXiv:2405.20574*. https://doi.org/10.48550/arXiv.2405.20574

[37] Gordon, A., Kozareva, Z., & Roemmele, M. (2012). SemEval-2012 Task 7: Choice of Plausible Alternatives: An Evaluation of Commonsense Causal Reasoning. In E. Agirre, J. Bos, M. Diab, S. Manandhar, Y. Marton, & D. Yuret (Eds.), *SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)* (pp. 394-398). Association for Computational Linguistics. https://aclanthology.org/S12-1052

[38] Pilehvar, M. T., & Camacho-Collados, J. (2018). WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. *arXiv preprint arXiv:1808.09121*. https://doi.org/10.48550/arXiv.1808.09121

[39] Clark, C., Lee, K., Chang, M. W., Kwiatkowski, T., Collins, M., & Toutanova, K. (2019). BoolQ: Exploring the surprising difficulty of natural yes/no questions. *arXiv preprint arXiv:1905.10044*. https://doi.org/10.48550/arXiv.1905.10044

[40] Savanur, S. R., & Sumathi, R. (2023). SentiNeg: algorithm to process negations at sentence level in sentiment analysis. *International Journal of Software Innovation (IJSI), 11*(1). 1-27. https://doi.org/10.4018/IJSI.315741

[41] Keleg, A., & Magdy, W. (2023). DLAMA: A framework

for curating culturally diverse facts for probing the knowledge of pretrained language models. *arXiv preprint arXiv:2306.05076*. https://doi.org/10.48550/arXiv.2306.05076

[42] Tirumala, K., Markosyan, A., Zettlemoyer, L., & Aghajanyan, A. (2022). Memorization without overfitting: Analyzing the training dynamics of large language models. *Advances in Neural Information Processing Systems, 35*. 38274-38290.

[43] Du, L., Wang, Y., Xing, X., Ya, Y., Li, X., Jiang, X., & Fang, X. (2023). Quantifying and attributing the hallucination of large language models via association analysis. *arXiv preprint arXiv:2309.05217*. https://doi.org/10.48550/arXiv.2309.05217

[44] Sap, M., Gabriel, S., Qin, L., Jurafsky, D., Smith, N. A., & Choi, Y. (2019). Social bias frames: Reasoning about social and power implications of language. *arXiv preprint arXiv:1911.03891*. https://doi.org/10.48550/arXiv.1911.03891

[45] Bauer, L., Tischer, H., & Bansal, M. (2023, May). Social Commonsense for Explanation and Cultural Bias Discovery. *In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. 3745-3760.

[46] Keysers, D., Schärli, N., Scales, N., Buisman, H., Furrer, D., Kashubin, S., ... & Bousquet, O. (2019). Measuring compositional generalization: A comprehensive method on realistic data. *arXiv preprint arXiv:1912.09713*. https://doi.org/10.48550/arXiv.1912.09713

[47] Lin, B. Y., Lee, S., Khanna, R., & Ren, X. (2020). Birds have four legs?! numersense: Probing numerical commonsense knowledge of pre-trained language models. *arXiv preprint arXiv:2005.00683*. https://doi.org/10.48550/arXiv.2005.00683

[48] Seo, J., Lee, S., Park, C., Jang, Y., Moon, H., Eo, S., ... & Lim, H. S. (2022, July). A dog is passing over the jet? a text-generation dataset for korean commonsense reasoning and evaluation. *In Findings of the Association for Computational Linguistics: NAACL 2022*. 2233-2249. https://doi.org/10.18653/v1/2022.findings-naacl.172

[49] Bhagavatula, C., Bras, R. L., Malaviya, C., Sakaguchi, K., Holtzman, A., Rashkin, H., ... & Choi, Y. (2019). Abductive commonsense reasoning. *arXiv preprint arXiv:1908.05739*. https://doi.org/10.48550/arXiv.1908.05739

[50] Liu, J., Wang, W., Wang, D., Smith, N. A., Choi, Y., & Hajishirzi, H. (2023). Vera: A general-purpose plausibility estimation model for commonsense statements. *arXiv preprint arXiv:2305.03695*. https://doi.org/10.48550/arXiv.2305.03695

[51] Liu, J., Liu, A., Lu, X., Welleck, S., West, P., Bras, R. L., ... & Hajishirzi, H. (2021). Generated knowledge prompting for commonsense reasoning. *arXiv preprint arXiv:2110.08387*. https://doi.org/10.48550/arXiv.2110.08387

[52] Zeng, Z., Cheng, K. T., Nanniyur, S. V., Zhou, J., & Bhat, S. (2023). IEKG: A Commonsense Knowledge Graph for Idiomatic Expressions. *arXiv preprint arXiv:2312.06053*. https://doi.org/10.48550/arXiv.2312.06053

[53] Seo, J., Lee, J., Park, C., Hong, S., Lee, S., & Lim, H. S. (2024, August). Kocommongen v2: A benchmark for navigating korean commonsense reasoning challenges in large language models. *In Findings of the Association for Computational Linguistics ACL 2024*. 2390-2415. https://doi.org/10.18653/v1/2024.findings-acl.141

[54] Son, G., Lee, H., Kim, S., Kim, H., Lee, J., Yeom, J. W., ... & Kim, S. (2023). Hae-rae bench: Evaluation of korean knowledge in language models. *arXiv preprint arXiv:2309.02706*. https://doi.org/10.48550/arXiv.2309.02706

[55] Jeong-Joon Oh(2007). A Study on Variables Correlated with Item Difficulty of the Geography Test(지리 문항의 난이도 변인에 관한 탐색), *Journal of Korean Geography and Environmental Education, 15*(2). 141-152. https://doi.org/10.17279/jkagee.2007.15.2.141

[56] Hangrok Cho, Mi-Hye Lee, Hyunyong Cho(2012). The Current Reality and Tasks of Korean·Korean Cultural Competence Assessment in Korea Naturalization Test(한국 귀화시험의 한국어·한국문화 능력 평가의 실제와 과제). *Journal of Korean Language Education, 23*(4). 343-369. https://doi.org/10.18209/iakle.2012.23.4.343

[57] Novikova, J., Dušek, O., Curry, A. C., & Rieser, V. (2017). Why we need new evaluation metrics for NLG. *arXiv preprint arXiv:1707.06875*. https://doi.org/10.48550/arXiv.1707.06875

[58] Zhou, K., Blodgett, S. L., Trischler, A., Daumé III, H., Suleman, K., & Olteanu, A. (2022). Deconstructing NLG evaluation: Evaluation practices, assumptions, and their implications. *arXiv preprint arXiv:2205.06828*. https://doi.org/10.48550/arXiv.2205.06828

[59] Stent, A., Marge, M., & Singhai, M. (2005, February). Evaluating evaluation methods for generation in the presence of variation. *In International conference on intelligent text processing and computational linguistics*. Springer Berlin Heidelberg, Germany, 341-351. https://doi.org/10.1007/978-3-540-30586-6_38

[60] Huang, Y., Sun, L., Wang, H., Wu, S., Zhang, Q., Li, Y., ... & Zhao, Y. (2024). Trustllm: Trustworthiness in large language models. *arXiv preprint arXiv:2401.05561*. https://doi.org/10.48550/arXiv.2401.05561

[61] Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., ... & Ahmed, N. K. (2024). Bias and fairness in large language models: A survey. *Computational Linguistics*. 1-79. https://doi.org/10.1162/coli_a_00524

[62] Kotek, H., Dockum, R., & Sun, D. (2023, November). Gender bias and stereotypes in large language models. *In Proceedings of the ACM collective intelligence*

conference. 12-24. https://doi.org/10.1145/3582269.3615599
[63] Motoki, F., Pinho Neto, V., & Rodrigues, V. (2024). More human than human: measuring ChatGPT political bias. *Public Choice, 198*(1). 3-23. https://doi.org/10.1007/s11127-023-01097-2
[64] Xue, M., Liu, D., Yang, K., Dong, G., Lei, W., Yuan, Z., ... & Zhou, J. (2023). OccuQuest: Mitigating Occupational Bias for Inclusive Large Language Models. *arXiv preprint arXiv:2310.16517*. https://doi.org/10.48550/arXiv.2310.16517

서민주

· 2017년 고려대학교 한국사학과(문학사)
· 2020년 고려대학교 한국사학과 (문학석사)
· 2021년 고려대학교 역사학과 박사과정

관심분야 : 디지털역사학, 조선후기 인사제도, 관료제
seoalswn@naver.com

정연주

· 2016년 전남대학교 사학과(문학사)
· 2019년 고려대학교 한국사학과 (문학석사)
· 2019년 고려대학교 역사학과 박사과정

관심분야 : 디지털역사학, 조선시대 법제사
choj3410@hanmail.net

이현정

· 2017년 숭실대학교 사학과(문학사)
· 2022년 고려대학교 한국사학과 (문학석사)
· 2022년 고려대학교 역사학과 박사과정

관심분야 : 디지털역사학, 고려시대 사회사
aiddahj@naver.com

임혜균

· 2015년 숙명여자대학교 역사문화학과(문학사)
· 2019년 고려대학교 한국사학과 (문학석사)
· 2019년 고려대학교 역사학과 박사과정

관심분야 : 디지털 역사학, 조선후기의 생활사, 지역사, 인물사
hkiim@hanmail.net

장정선

· 2000년 고려대학교 컴퓨터학과(이학사)
· 2002년 고려대학교 컴퓨터학과 (이학석사)
· 2021년 고려대학교 컴퓨터학과 (공학박사)
· 2011년 ~ 2023년 엔씨소프트 AI Biz Center 센터장
· 2023년 ~ 현재 고려대학교 문과대학 한국사학과 부교수

관심분야 : 디지털역사학, 인공지능, 거대언어모델
empyrean@korea.ac.kr

부 록

〈표 1〉 벤치마크 평가 영역

	평가 영역	세부 영역
1	언어 및 문학	비문학(설명문, 논설문, 생활문, 보고서 등), 화법(구어 의사 소통 상황, 방언 등), 문법, 문학(시, 소설, 산문, 극 등), 기타
2	역사	선사시대, 고조선과 여러 나라, 삼국시대, 남북국시대, 고려시대, 조선시대, 일제시대, 현대사, 한국의 역사 인물, 한국의 문화유산, 한국사 통합, 기타
3	문화 및 민속	전통 가치, 전통 의식주, 의례, 명절, 종교, 대중문화, 여가문화, 무형유산, 기념일, 기타
4	법	헌법, 재산 관계와 법, 가족관계와 법, 사회생활과 법, 범죄와 형벌, 국가 및 기관과 법, 생활법률, 기타
5	경제 및 금융	일상생활, 경제활동, 경제정책, 금융지식, 기타
6	정치	한국의 민주 정치, 입법부, 사법부, 행정부, 선거와 지방자치, 시민운동, 기타
7	지리	기후, 지형, 인구, 교통, 수도권(서울, 경기)과 지방(충청도, 전라도, 경상도, 강원도, 제주도), 기타
8	사회	한국의 상징, 가족, 일터, 교통과 통신, 주거, 도시와 농촌, 복지, 의료와 안전, 기타
9	교육	영유아교육, 초등학교, 중학교, 고등학교, 대학교, 대학원, 평생교육, 대안교육, 기타

〈표 2〉 문항의 형식에 따른 예시

	문제-보기 유형	문제-지문-보기 유형
문제	다음 중 국회의원에 대한 설명으로 옳지 않은 것을 모두 고르시오.	빈칸에 들어갈 말을 알맞게 짝지은 것을 고르시오.
지문		초등학교 6년과 중학교 ( )년은 ( ) 기간으로 무상 교육이 제공된다.
보기	1. 2007년 12월 28일 개정하여 대법관의 수를 대법원장 외 13인으로 변경하였다. 2. 1987년 12월 4일 개정하여 대법관이라는 명칭을 다시 사용하였고 대법관의 수를 대법원장 외 13인으로 변경하였다. 3. 2005년 12월 15일 개정하여 대법관의 수를 대법원장 외 12인으로 변경하였다. 4. 1981년 1월 29일 개정하여 대법원 판사의 수를 대법원장 외 12인으로 변경하였다.	1. 3, 의무 교육 2. 6, 의무 교육 3. 3, 유상 교육 4. 3, 비의무 교육

〈표 3〉 문항 유형별 예시-형식

	유형	예시
1	빈칸채우기	빈칸에 들어갈 말로 올바른 것을 고르시오. 한국 국내 증시가 저평가되는 현상을 개선하고 주주가치를 높이는 것을 목표로 한국 정부는 2024년 ( ) 프로그램을 추진하였다. 1. K-주식 활성화 **2. 기업 밸류업-정답** 3. 버블 4. 자이언트스텝
2	밑줄	다음은 영호, 민수, 진규가 하굣길에 나눈 대화의 일부이다. '우리' 가운데 의미하는 바가 다른 하나를 고르시오. 영호 : 오늘 ㄱ.'우리' 집에서 놀 수 있어? ㄴ.'우리' 엄마가 맛있는 케이크를 사두셨대. 민수 : 그래? 나 학원 가기 전까지 시간이 있으니 같이 가자. 진규야 ㄷ.'우리' 같이 가자. 진규 : 좋아. 그럼 영호가 매일 ㄹ.'우리' 동생, 우리 동생하면서 귀여워하는 수진이도 볼 수 있겠다. 1. ㄱ 2. ㄴ **3. ㄷ-정답** 4. ㄹ
3	사실확인	검찰과 경찰의 수사권 차이에 대한 설명으로 옳지 않은 것을 고르시오. 1. 고소, 고발을 통해 수사를 시작하면 경찰은 송치 혹은 불송치 결정을 내리며, 검찰은 송치된 사건에 대해 기소 혹은 불기소 결정을 내린다. 2. 검찰은 경찰 단계의 수사에 관여할 수 있다.-정답 3. 경찰 수사 단계에서 영장 청구가 필요한 경우 검사가 관여한다. 4. 경찰이 수사 결과 범죄의 혐의가 없다고 판단할 경우 불송치 결정을 해서 1차적으로 수사 종결권을 행사할 수 있다.
4	순서나열	조선의 문호 개방과 관련한 사건에 대해 순서대로 나열한 것을 고르시오. ㄱ. 자국에 대한 미국의 외교 방식을 모방하여 일본이 조선에 배를 보냈다. ㄴ. 처음으로 관세 조항과 거중 조정 조항을 담은 조약을 체결하였다. ㄷ. 천주교의 공인 문제로 조약 체결이 지연되었던 국가와도 조약을 체결하였다. ㄹ. 부산뿐만 아니라 다른 두 항구를 개방하기로 약조하였다. **1. ㄱ-ㄹ-ㄴ-ㄷ-정답** 2. ㄱ-ㄷ-ㄹ-ㄴ 3. ㄹ-ㄷ-ㄱ-ㄴ 4. ㄷ-ㄱ-ㄴ-ㄷ
5	정답매칭	다음 중 A,B,C와 (가),(나),(다)를 가장 자연스럽게 짝지은 것을 고르시오. A. 충주 B. 청주 C. 공주 (가) 최근 해당 지자체 유튜브가 크게 화제가 된 바 있다. (나) 충청남도 역사박물관이 있다. (다) 10월에 세종대왕과 초정약수축제가 진행된다. 1. A-(나) 2. B-(나) 3. C-(다) **4. A-(가)-정답**

	유형	예시
6	복합	다음에서 빈칸에 들어갈 말로 알맞은 것과 해당 중학교에 대한 설명을 옳게 짝지은 것을 고르시오. 중학교는 일반 중학교, ( ) 중학교로 구분된다. 1. 국립, 별도의 과정을 거쳐 선발한다. 2. 특성화, 집에서 가까운 곳에 배정한다. 3. 사립, 진로 탐색 교육을 실시한다. **4. 특성화, 특정 분야의 재능 있는 학생을 교육한다.-정답**