The Transactions of the Korea Information Processing Society
TKIPS Vol.13, No.9, pp.421~428
eISSN: 3022-7011
https://doi.org/10.3745/TKIPS.2024.13.9.421

# A Study on the Evaluation Methods for Assessing the Understanding of Korean Culture by Generative AI Models

Son Ki Jun† · Kim Seung Hyun††

## ABSTRACT

Recently, services utilizing large-scale language models (LLMs) such as GPT-4 and LLaMA have been released, garnering significant attention. These models can respond fluently to various user queries, but their insufficient training on Korean data raises concerns about the potential to provide inaccurate information regarding Korean culture and language. In this study, we selected eight major publicly available models that have been trained on Korean data and evaluated their understanding of Korean culture using a dataset composed of five domains (Korean language comprehension and cultural aspects). The results showed that the commercial model HyperClovaX exhibited the best performance across all domains. Among the publicly available models, Bookworm demonstrated superior Korean language proficiency. Additionally, the LDCC-SOLAR model excelled in areas related to understanding Korean culture and language.

Keywords : LLM, Korean Culture, Culture Understanding, Evaluation Dataset

# 생성형 AI 모델의 한국문화 이해 능력 평가 방법에 관한 연구

손 기 준† · 김 승 현††

## 요 약

최근 GPT-4, LLaMA와 같은 초거대 언어모델을 활용한 서비스가 공개되며 많은 주목을 받고 있다. 이 모델들은 사용자들의 다양한 질문에 유창하게 응답할 수 있지만, 한국어 데이터의 학습량이 부족하여 한국 문화 및 한국어에 대한 잘못된 정보를 제공할 가능성이 있다. 본 논문에서는 한국어 데이터를 학습한 주요 공개 모델 8개를 선정하여, 5개 분야(한국어 이해 및 문화 영역으로 구성)에 대한 평가 데이터셋을 통해 한국 문화 이해 능력을 평가하였다. 그 결과, 상용 모델인 HyperClovaX가 전 분야에서 가장 뛰어난 성능을 보였으며, 공개용 모델 중에서는 Bookworm이 한국어 구사 능력에서 우수한 성과를 보였다. 또한, 한국어 이해 및 문화와 관련된 부문에서는 LDCC-SOLAR 모델이 뛰어난 성능을 확인할 수 있었다.

키워드 : 대규모 언어모델, 한국문화, 문화이해, 평가 데이터

## 1. 서 론

대규모 언어모델(Large Language Model, 이하 LLM)은 지난 5년간 지속적으로 성장해왔으며, OpenAI의 ChatGPT 서비스가 등장한 이후 LLM의 활용 방안에 대한 다양한 시도가 전 세계적으로 이루어지고 있다. LLM은 자연어를 이해하고 생성할 수 있는 대규모 딥러닝 모델로, 이를 개발하기 위해서는 방대한 학습 데이터와 모델링 작업이 필요하다. 이로 인해 충분한 데이터와 자본을 보유한 해외 유명 IT 기업들이 주도적으로 발전시키고 있으며, 국내 IT 기업들은 한국어 능력을 기반으로 한 맞춤형 서비스 제공에 한계가 있을 수 있다는 우려가 제기되고 있다.

또한, LLM 모델이 특정 국가의 시각으로 만들어져 각 나라의 관습, 역사, 지리 등 문화적 규범에 대해 충분히 학습 및 검증되지 않은 정보를 제공할 경우, 관련 분야에 대한 충분한 지식을 갖추지 못한 사람들이 이를 무비판적으로 사실로 받아들여 사회적 혼란을 초래할 수 있다는 우려가 커지고 있다.

따라서 이러한 우려를 해소하기 위해 LLM의 한국문화 이해도를 객관적으로 평가하여 문제점과 개선점을 찾아 수정하는 작업이 필요하다. 이를 위해 동시대 사람들의 생각과 행동 양식이 객관적으로 반영된 공인 인증 시험 문제를 활용하는 것은 새로운 문제를 만드는 수고를 줄이고 객관성을 확보하는 효과적인 방법이라고 할 수 있다.

이에 본 연구에서는 한국어를 학습한 LLM 모델들의 한국문

※ 이 논문은 2024년 ACK 2024의 일반논문으로 "대규모 언어모델의 한국어 이해 능력 평가 방법에 관한 연구"의 제목으로 발표된 논문을 확장한 것임.
† 비 회 원 : 포스트에이아이 AI 센터 이사
†† 정 회 원 : 한국지능정보사회진흥원 지능데이터본부 책임
Manuscript Received : July 11, 2024
Accepted : August 21, 2024
* Corresponding Author : Kim Seung Hyun(shkim@nia.or.kr)

화 이해 능력을 평가하기 위해 다양한 공인 시험과 자료를 활용하였다. ‘KBS 한국어 능력 시험’과 ‘TOPIK’을 통해 한국어 능력을 평가하였고, ‘한국사 능력 검정시험’, ‘국내여행안내서’, ‘공무원 9급 시험’을 바탕으로 한국문화 이해도를 평가할 수 있는 데이터셋을 구성하였다. 이러한 평가를 통해 LLM의 한국문화 이해 및 활용 능력을 객관적으로 평가하고, 서비스 제공을 위한 가장 적절한 LLM을 선정하는 데 필요한 기본 정보를 제공하고자 한다. 이 연구는 LLM이 한국문화와 언어에 대해 보다 정확하고 풍부한 이해를 갖추도록 도와줄 것이다.

## 2. 관련 연구

### 2.1 LLM 개념

LLM(Large Language Model)은 대규모 데이터셋을 사용하여 훈련된 언어 모델을 의미한다. 이 모델은 사전에 방대한 양의 언어 데이터를 학습하여 문장 구조, 문법, 의미, 단어 내에 내재된 다양한 의미를 이해하고, 이를 바탕으로 사람이 말하는 것과 유사하게 자연어를 생성할 수 있다. 최근에는 국내 주요 기업들도 초거대 인공지능 시스템 구축에 주력하고 있으며, 이러한 LLM의 필요성은 더욱 커지고 있다. LLM은 자연어 처리(NLP) 분야에서 혁신적인 도구로 자리 잡았으며, 다양한 응용 분야에서 활용될 수 있다. 예를 들어, 고객 서비스 챗봇, 자동 번역 시스템, 콘텐츠 생성, 의료 상담 등에서 LLM은 중요한 역할을 하고 있다[1,2].

### 2.2 자연어 처리 초기 주요 기술

분산 표현(Word Embeddings)은 단어를 고정된 크기의 벡터로 변환하여 단어 간의 유사성을 측정하고 의미론적 정보를 보존하는 방법이다. 대표적인 예로는 워드 투 벡터(Word2Vec)와 글로브(GloVe)가 있다. 이 기술은 단어의 의미를 수치적으로 표현하는 데 큰 진전을 가져왔으며, 이후 자연어 처리 모델의 기반이 되었다[15].

장단기 메모리 네트워크(Long Short-Term Memory)은 순환 신경망(RNN)의 한계를 극복하기 위해 개발되었으며 기업 셀과 게이트 메커니즘을 도입하여 긴 시퀀스에서도 학습이 가능하게 하였다. LSTM은 긴 의존성을 처리하는 능력 덕분에 언어 모델링, 기계 번역 등 여러 자연어 처리 작업에서 널리 사용되었다[16].

어텐션 메커니즘은 입력 시퀀스의 모든 위치를 한 번에 참조하여 특정 단어에 대한 가중치를 계산하는 방법이다. 이는 RNN이나 CNN의 단점을 보완하며, 특히 기계 번역 작업에서 큰 성능 향상을 가져왔다. 어텐션 메커니즘은 이후 트랜스포머 모델의 핵심 요소로 발전하게 되었다[17].

이러한 기술적 발전들은 자연어 처리 모델이 점점 더 복잡한 언어 구조와 문맥을 이해할 수 있도록 하는 기반을 마련했으며, 이는 트랜스포머 모델의 개발로 이어져 현재의 LLM 발전에 중요한 역할을 하였다.

### 2.3 트랜스포머(Transformer) 모델

2017년 트랜스포머(Transformer) 아키텍처의 등장과 함께 거대 언어모델(LLM)이 본격적으로 발전하기 시작했다. 트랜스포머 모델은 RNN(순환 신경망)이나 CNN(합성곱 신경망)을 사용하지 않고, 포지셔널 인코딩(Positional Encoding)을 활용하며 다수의 인코더(Encoder)와 디코더(Decoder), 어텐션(Attention) 메커니즘을 사용한다. 포지셔널 인코딩을 통해 각 단어의 위치 정보를 인코딩하고, 어텐션 과정을 여러 레이어에서 반복 수행함으로써 문맥 간의 상호작용을 정교하게 모델링할 수 있다. 이러한 구조는 번역이나 요약과 같은 작업에서 어텐션과 정규화(Normalization) 과정을 통해 성능을 크게 향상시켰으며, 가장 중요한 정보를 강조하여 분석할 수 있게 되었다[3].

2018년, Google은 BERT(Bidirectional Encoder Representations from Transformers)를 공개하였으며, 이는 트랜스포머 모델의 인코더 아키텍처를 기반으로 양방향 해석을 통해 텍스트를 표현하는 학습 모델이다[4]. BERT는 입력된 텍스트의 좌우 문맥을 동시에 고려하여 더 풍부한 의미를 학습할 수 있게 한다. 대부분의 LLM은 트랜스포머 아키텍처에서 파생된 AI 모델로, 이를 통해 사람의 언어, 코드 등을 이해하고 생성할 수 있는 능력을 갖추게 되었다.

### 2.4 LLM 모델의 발전 과정

트랜스포머 모델의 등장 이후, LLM의 발전은 주로 모델의 크기 증가와 데이터 양의 확대를 통해 이루어졌다. 특히 OpenAI의 GPT 시리즈는 이러한 추세를 선도하였다. GPT-3와 GPT-4 같은 모델들은 수십억에서 수조 개의 파라미터를 사용하여, 더 깊고 넓은 문맥 이해를 가능하게 했다. 이러한 모델들은 문맥에 민감한 다양한 언어 작업을 수행할 수 있는 능력을 보여주며, 일반적인 언어 이해뿐만 아니라 특정 주제에 대한 지식 학습에서도 탁월한 성능을 보여주었다.

더욱 발전된 LLM은 언어를 넘어서 다양한 형태의 입력을 처리할 수 있는 능력을 가지고 있다. 예를 들어, OpenAI의 최신 모델인 GPT-4는 텍스트와 이미지를 동시에 이해하고 생성할 수 있는 멀티모달 능력을 갖추고 있다고 평가된다. 이는 AI가 단순한 텍스트 처리를 넘어서 시각적 데이터와 상호작용하며, 더욱 복잡한 인간의 언어와 커뮤니케이션 스타일을 모방할 수 있게 만들었다. 이러한 멀티모달 LLM은 사용자의 질문에 더욱 정확하고 상세한 답변을 제공할 수 있으며, 창의적인 콘텐츠 생성에서도 새로운 가능성을 열고 있다.

이와 같은 발전은 다양한 응용 분야에서 혁신을 가능하게 했다. 예를 들어, 고객 서비스, 교육, 의료, 법률 자문 등에서 LLM은 점점 더 중요한 역할을 하고 있으며, 이는 AI가 인간의 의사소통과 지식 전달 방식을 혁신적으로 변화시키고 있음을 의미한다.

## 3. LLM 모델 평가

### 3.1 평가 모델 및 항목 선정

최신 기술의 발전과 함께 다양한 LLM이 공개 및 개방되면서 사용자가 모델들의 성능을 객관적이고 세부적으로 판단할 수 있도록 다양한 평가 방안이 제시되고 있다. 가장 대중적으로 알려진 Hugging Face Open LLM Leaderboard는 Table 1과 같이 총 6가지 분야에서 모델의 성능을 평가하여 순위를 정하고 있다.

또한, 한국어 모델을 평가하기 위해 NIA-업스테이지에서 운영하는 Open Ko-LLM 리더보드도 있다. 이 리더보드는 역사 왜곡, 환각 오류, 형태소 오류, 불규칙 활용 오류, 혐오 표현 등을 고려한 상식 생성 기준을 바탕으로, 한국어 사용자가 가진 일반 상식에 부합하는지를 기준으로 모델의 성능을 평가한다. 이를 통해 한국어 LLM의 정확성과 신뢰성을 높이고, 사용자에게 적절한 언어모델을 제공할 수 있도록 돕고 있다.

공개된 다양한 모델 중에서 대표성이 있는 모델을 선정하여 평가하기 위해 다음과 같은 선정 방법을 사용하였다. 먼저, 공개용 한국어를 학습한 언어 모델을 내려받아 로컬 서버에 설치하여 활용할 수 있는 모델로 정의하였다. 평가 모델을 선정하기 위해 2024년 2월을 기준으로 Hugging Face Leaderboard의 상위 100위 이내 모델 중에서 Open Ko-LLM 리더보드의 상위 30위에 중복으로 포함된 모델을 우선 선정하였다.

선정된 모델 중 공공부문에서의 활용성을 고려하여 한국어 데이터를 학습한 공개 모델을 선택하였으며, 대다수 기업과 연구자들이 평가에 활용하는 인프라의 성능을 고려하여 사이즈가 12.8B 이하인 모델을 기준으로 검토하였다. 이와 같은 기준으로 최종적으로 Table 2에서 제시한 6개의 평가 모델을 선정하였다.

상용 모델의 경우, 로컬 PC에 직접 설치하여 사용하는 방식이 아니라 API나 웹 페이지를 통해 접근할 수 있는 모델을 의미한다. 이에 따라 네이버의 HyperClovaX와 OpenAI의 최신 모델인 GPT-4를 활용하여 분석을 수행하였다. 이러한 선정 과정을 통해 공공부문과 상용(민간)분야에서의 활용 모두를 고려한 포괄적인 평가를 진행할 수 있었다.

Table 1. Hugging Face Open LLM Leaderboard Evaluation Criteria

| | Category | Content |
|---|---|---|
| 1 | ARC | Elementary-level science questions |
| 2 | HellaSwag | Evaluating language reasoning ability in specific situations |
| 3 | MMLU | Assessing the knowledge of pre-trained models |
| 4 | TruthfulQA | Ability to prevent hallucination phenomena |
| 5 | Winogrande | Evaluating common-sense reasoning |
| 6 | GSM8k | Multi-step mathematical reasoning evaluation |

Table 2. LLM Used for Evaluation

| | | Model |
|---|---|---|
| Public | 1 | LDCC-SOLAR-10.7B[6] |
| | 2 | Bookworm-10.7B[7] |
| | 3 | SOLARC-M-10.7B[8] |
| | 4 | LLaMA-2-13b-chat-hf[9] |
| | 5 | Kullm-solar[10] |
| | 6 | KoAlpaca-Polyglot-12.8B[11] |
| Commercial | 7 | HyperClovaX[12] |
| | 8 | GPT-4-turbo[13] |

* The public model was downloaded and analyzed from the March 20, 2024 version

공개용 모델로 최종 선정된 모델들의 특성을 간단히 살펴보면, 먼저 SOLAR 모델은 AI 스타트업인 업스테이지에서 자체 개발한 LLM 모델로 Hugging Face Leaderboard에서 1위에 오른 적이 있다. 이 모델은 효율적인 훈련과 빠른 성능 회복을 위해 'Depth Up-Scaling(이하 DUS)' 방법을 적용하여 깊이를 확장하고 지속적인 사전 훈련을 병행한다. 특히, 지시에 따른 질의응답 형식의 훈련과 인간 또는 고급 AI의 선호도에 맞춰 조정하는 'alignment tuning'을 통해 문맥에 적합하고 정확한 응답 생성 능력을 강화하였다[14].

LDCC-SOLAR 모델은 롯데 데이터 통신에서 개발한 모델로, SOLAR 모델을 바탕으로 자체 구축한 데이터를 추가 학습하여 만들어졌다. 이 모델 역시 DUS 기술을 사용하여 기존 언어 모델보다 효율적으로 성능을 향상시키며, 다양한 자연어 이해와 생성에서 뛰어난 성능을 보인다[6].

Bookworm 모델은 야놀자에서 업스테이지의 SOLAR 모델을 기반으로 자체 제작한 데이터셋을 학습한 후, 미세 조정을 통해 성능을 높인 모델이다[7]. SOLARC-M 모델은 앞선 모델들과 마찬가지로 DUS 기술을 적용하고, SOLAR -Instructions 모델에 Merge 기술을 사용하여 최적화된 모델로, 자연어 처리에서 우수한 성능을 보인다[8].

LLaMA-2 모델은 Meta가 개발한 LLM 모델로, 고급 감독 학습(Supervised Fine-Tuning)과 인간 피드백을 통한 강화 학습(Reinforcement Learning with Human Feedback, RLHF)을 통해 훈련되었으며, 사람의 선호도에 맞춘 대화를 할 수 있다[9]. Kullm-solar 모델은 고려대학교에서 제작한 구름 데이터셋을 학습한 SOLAR-Instruction 모델로, Open Ko-LLM 리더보드에서 상위권에 기록된 바 있다[10].

KoAlpaca-Polyglot 모델은 LLaMA나 Polyglot 모델에 한국어 instruction-following data를 학습시킨 모델로, 네이버 지식인 베스트 질문을 시드 데이터로 활용하여 추가 학습한 모델이다. 이 모델은 LoRA(Parameter Efficient Tuning) 방법을 적용해 파인튜닝한 모델 가중치를 제공하며, 베이스 모델로 많이 활용된다[11].

상용 모델로 최종 선정된 모델들의 특성도 간단히 살펴보면, HyperClovaX 모델은 네이버에서 개발한 LLM 모델로 복

잡한 자연어 처리 작업에 최적화되어 있다. 이 모델은 광범위한 사전 학습과 고급 감독 학습(SFT)을 통해 개발되었으며, 다양한 언어 데이터셋을 활용해 훈련되었다. HyperClovaX는 문맥을 정확히 파악하고 문법적으로 정확하며 의미론적으로 일관된 텍스트를 생성하는 능력이 뛰어난 것으로 평가된다. 또한, 인간의 피드백을 통한 강화 학습(RLHF)으로 사용자의 선호와 의도를 반영하여 높은 수준의 이해력과 응답 능력을 보여준다[12].

GPT-4는 OpenAI에서 개발한 LLM 모델로 GPT-3의 성능을 대폭 향상시킨 모델이다. 방대한 양의 인터넷 텍스트 데이터를 통해 사전 훈련되었으며, 새로운 토크나이징 기법과 향상된 어텐션 메커니즘을 적용하여 자연어 처리 작업에서 높은 정확도를 보인다. 자체적으로 진행된 강화 학습을 통해 인간의 피드백을 직접적으로 반영하여 모델 성능을 개선하였으며, 생성된 텍스트는 사람이 작성한 것과 구분하기 어려울 정도로 자연스럽다[13].

### 3.2 평가 데이터셋 구성 및 방법

앞서 설명한 Hugging Face Open LLM Leaderboard에서 선정한 6가지 분야의 성능 지표를 바탕으로, 공공부문과 민간부문의 한국문화와 한국어 사용성을 고려하여 평가를 진행하였다. 이를 위해 한국어 능력, 한국사, 한국지리, 문학, 문법 등을 폭넓게 평가할 수 있는 문제로 구성된 총 5가지 분야에서 431개의 문항을 작성하여 평가를 실시하였다. 이 평가 항목들은 다양한 측면에서 한국어와 한국문화에 대한 이해도를 측정하기 위해 선정되었다. KBS 한국어 능력시험은 한국어 문장 구조를 평가하며, TOPIK은 외국인을 대상으로 한국어 능력을 평가한다. 한국사 능력 검정시험은 한국의 역사와 문화를 평가하며, 국내여행안내사는 한국의 지리적 특성을 평가한다. 마지막으로, 공무원 9급 시험은 문학과 문법 등 한국어 전반에 걸친 종합 평가를 포함한다.

### 3.3 평가 결과

본 연구에서는 한국문화 이해 능력을 평가하기 위해 다양한 데이터셋(Table A1 참고)을 사용하여 객관식 문제 유형으로 평가를 실시하였다. 각 모델의 입력 구조와 일관된 답안을 제시할 수 있도록 프롬프트 엔지니어링을 통해 프롬프트를 구성하였으며, 평가를 위하여 구성한 문제의 정답과 모델이 제출한 답안을 비교하여 정답률을 확인하였다. 결과는 Table A2에서 확인할 수 있다.

Table 3. Test Name and Evaluation contents

| | Test(Number of Questions) | Content |
|---|---|---|
| 1 | KBS Korean Language Proficiency Test(88) | Evaluation of Korean Sentence Structure |
| 2 | TOPIK(203) | Korean Language Evaluation for Foreigners |
| 3 | Korean History Proficienct Test(40) | Evaluation of Korean History and Culture |
| 4 | Domestic Travel Guide Certificate(50) | Evalustion of Korea's Geographical Characteristics |
| 5 | Civil Service Exam, Grade 9(50) | Evaluation of Korean Literature, Grammar, etc. |

먼저, 'KBS 한국어 능력시험'은 한국어의 문법과 어휘를 중심으로 문장 구조 이해 능력을 평가한다. 상용 모델 중에서는 HyperClovaX가 49%의 정답률로 가장 높은 성능을 보였으며, GPT-4도 47%로 준수한 성능을 나타냈다. 공개용 모델 중에서는 Bookworm 모델이 24%로 가장 높은 정답률을 보였다.

'TOPIK'은 한국어를 모국어로 사용하지 않는 사람을 대상으로 한 시험으로, 언어 사용 및 이해 능력을 평가한다. HyperClovaX와 GPT-4 모두 89%의 높은 정답률을 기록하며 우수한 성능을 보였다. 공개용 모델 중에서는 Bookworm이 75%로 가장 높은 성능을 보였다.

'한국사 능력 검정시험'은 한국의 역사와 문화에 대한 지식을 평가하는 시험으로, HyperClovaX가 53%로 가장 높은 정답률을 보였으며, 공개용 모델 중에서는 LDCC-SOLAR가 45%로 우수한 성능을 나타냈다.

'국내 여행 안내사' 시험은 한국의 지리와 관광 명소에 대한 지식을 평가하는 시험으로, HyperClovaX 모델이 70%로 가장 높은 성능을 보였고, 공개용 모델 중에서는 LDCC-SOLAR 모델이 46%로 좋은 결과를 나타냈다.

마지막으로, '공무원 9급 시험'은 한국 문학, 문법, 한자를 포함한 폭넓은 한국어 능력을 평가하는 시험으로 GPT-4가 80%로 가장 높은 성능을 나타냈으며, 공개용 모델 중에서는 LDCC-SOLAR 모델이 56%의 높은 정답률을 기록하였다.

본 연구에서는 다양한 LLM 모델의 한국어 이해 능력을 종합적으로 평가하였으며, 각 모델은 동일한 문항 수인 431개의 문항으로 평가되었다. Table 4는 평가 결과를 종합적으로 확인할 수 있다.

상용 모델인 HyperClovaX와 GPT-4가 각각 73%와 71%의 정답률로 가장 높은 성능을 보였으며, 이는 상용 모델들이 정

Table 4. Summary of LLM Evaluation Results

| Rank | Model | Number of Questions | Number of Correct Answers | Ratio |
|---|---|---|---|---|
| 1 | HyperClovaX | 431 | 315 | 73 |
| 2 | GPT-4-turbo | 431 | 305 | 71 |
| 3 | LDCC-SOLAR-10.7B | 431 | 235 | 55 |
| 4 | Bookworm-10.7B | 431 | 230 | 53 |
| 5 | SOLARC-M-10.7B | 431 | 213 | 49 |
| 6 | LLaMA-2-13b-chat-hf | 431 | 121 | 28 |
| 7 | Kullm-solar | 431 | 113 | 26 |
| 8 | KoAlpaca-Polyglot-12.8B | 431 | 72 | 17 |

교한 데이터와 고급 최적화 기법을 활용함으로써 높은 성능을 달성할 수 있음을 시사한다. 특히 HyperClovaX 모델은 전반적인 평가에서 가장 높은 정답률을 기록하며 한국문화 이해도에서 우수한 성능을 입증하였다.

반면, 공개용 모델 중에서는 LDCC-SOLAR와 Bookworm 모델이 각각 55%와 53%로 비교적 높은 성능을 보였으나, 상용 모델에 비해 다소 뒤처지는 결과를 나타냈다. SOLARC-M은 중간 범위의 성능을 보였으며, 기타 모델들은 50% 미만의 정답률을 기록하여 상대적으로 낮은 성능을 보였다. LLaMA-2, Kullm-solar, KoAlpaca 모델은 각각 28%, 26%, 17%의 정답률을 보였으며, 한국문화 이해 능력에서 개선이 필요함을 보여주었다.

특히, 공개용 모델 중 상위권을 기록한 모델들의 공통점은 모두 업스테이지가 개발한 SOLAR 모델을 기반으로 하고 있다는 점이다. SOLAR 모델은 LLM의 효율적인 확장을 위해 DUS 기술과 지속적인 사전 훈련을 통해 개발되었으며, 롯데 데이터 통신과 야놀자는 자체 구축 데이터로 추가 학습을 지속하여 높은 정답률을 달성하였다. 이는 우수한 데이터를 지속적으로 학습하는 것이 LLM의 성능을 높이는 가장 효과적인 방법임을 확인시켜준다.

## 4. 결 론

본 연구에서는 다양한 대규모 언어 모델(LLM)들의 한국어 및 한국문화에 대한 이해도를 평가하여 각 모델의 성능을 종합적으로 분석하였다. 이 연구를 통해 얻은 결과는 모델 개발 및 튜닝 과정에서 언어와 문화적 맥락의 중요성을 강조하며, 특히 한국어 서비스를 목표로 하는 사용자에게 중요한 시사점을 제공한다.

공개용 모델과 상용 모델 간의 성능 비교에서, 상용 모델이 일반적으로 더 높은 정확도를 나타내는 것을 확인할 수 있었다. 특히 HyperClovaX와 GPT-4-turbo 모델은 최적화된 학습 방법과 광범위한 데이터 활용으로 높은 정답률을 기록하며, 한국어 이해 능력에 있어 상당한 우위를 확보하였다. 반면, 공개용 모델 중에서는 LDCC-SOLAR와 Bookworm 모델이 상대적으로 높은 성능을 보였으나, 상용 모델과의 차이는 여전히 존재했다.

이러한 결과는 LLM의 성능이 단순히 모델 크기나 파라미터 수에 의해 결정되지 않으며, 모델 훈련에 사용된 데이터의 품질과 다양성, 그리고 특정 언어 및 문화, 규범에 대한 깊이 있는 이해가 중요하다는 것을 시사한다. 특히 한국어와 같은 특정 언어를 대상으로 한 서비스 개발에 있어서는 지역적 특성과 문화적 요소를 반영한 데이터셋의 구축과 이를 효과적으로 활용하는 학습 전략이 필수적임을 보여준다.

또한, 본 연구는 LLM의 국제적 활용 가능성을 탐색하는 동시에, 특정 국가의 문화 및 언어적 특성을 간과하지 않는 방향으로 모델을 개발하고 최적화하는 방법론에 대한 중요한 통찰을 제공한다. 이는 궁극적으로 다양한 언어 및 문화권에서도 효과적으로 활용될 수 있는 보다 범용적이고 탄력적인 LLM 개발로 이어질 수 있을 것이다.

향후 연구에서는 추가적인 언어 및 문화적 맥락을 반영한 모델 튜닝 및 최적화 전략을 개발함으로써, 더욱 정교하고 신뢰할 수 있는 LLM의 구축을 목표로 해야 할 것이다. 이를 통해 LLM은 전 세계적으로 다양한 언어 사용자들에게 맞춤형 인공지능 기반 서비스를 제공하는 데 중요한 역할을 할 수 있을 것으로 기대된다.

이 연구는 특히 한국어 및 한국문화 이해도를 높이기 위한 LLM 개발의 중요성을 강조하며, 관련 데이터셋의 품질 향상과 학습 전략의 최적화가 성공적인 모델 개발의 핵심임을 시사한다. 따라서, 앞으로도 지역적 특성과 문화적 요소를 깊이 반영한 연구가 지속적으로 이루어져야 할 것이다.

## References


[1] S. Lim and S. Lee “Research trends in artificial intelligence language models,” *Information and Communication Magazine*, Vol.40, No.3, pp.42-50, 2023.

[2] M. Shanahan “Talking about large language models,” *Communication of the ACM*, Vol.67, No.2, pp.68-79, 2024.

[3] A. Vaswani et al., “Attention is All You Need,” *Advances in Neural Information Processing Systems*, pp.5998-6008, 2017.

[4] J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” *North America Chaper of the Association for Computational Liguistics*, pp.4171-4186, 2018.

[5] A, Radfordm, J. Narasimhan, T. Salimans, and I. Sutskever, “Improving language understanding by generarive pre-training,” *OpenAI*, 2018.

[6] LDCC. (2024, Feb. 28). LDCC/LDCC -SOLAR-10.7B. Hugging Face. https://huggingface.co/LDCC/LDCC-SOLAR-10.7B

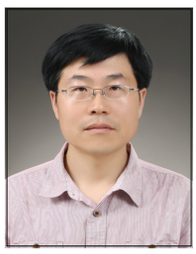
[7] Yanolja. (2024, Mar. 16). Yanolja/Bookworm- 10.7B-v0.4-DPO. Hugging Face. https://huggingface.co/yanolja/Bookworm-10.7B-v0.4-DPO

[8] Dopeornope. (2024, Jan. 15). DopeorNope /SOLARC -M-10.7B. Hugging Face. https://huggingface.co/DopeorNope/SOLARC-M-10.7B

[9] Meta. (2023, Nov. 13). Meta-Llama/L-ama -2-13b -Hf. Hugging Face. https://huggingface.co/meta-llama/Llama-2-13b-hf

[10] Heavytail, (2024, Jan. 28). Heavytail/Kullm-Solar. Hugging Face. https://huggingface.co/heavytail/kullm-solar

[11] Beomi.(2023, May. 3). Beomi/KoAlpaca-Polyglot-12.8B. Hugging Face. https://huggingface.co/beomi/KoAlpaca-Polyglot-12.8B
[12] Naver Cloud, (2024, Apr. 02), "HyperCLOVA X Technical Report," https://arxiv.org/html/2404.01954v1.
[13] OpenAI, (2024, Mar. 4),"GPT-4 Technical Report," https://arxiv.org/abs/2303.08774.
[14] Upstage AI, (2024, Apr. 04), "SOLAR 10.7B: Scaling Large Language Models with Simple yet EffectiveDepth Up-Scaling," https://arxiv.org/pdf/2312.15166.
[15] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," *arXiv preprint arXiv:1301.3781*, 2013. Retrieved from https://arxiv.org/abs/1301.3781
[16] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, Vol.9, No.8, pp.1735-1780, 1997.
[17] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2015.



**손 기 준**

https://orcid.org/0009-0005-1661-6453
e-mail : kijunson@post-ai.com
2005년 경북대학교 컴퓨터공학과(박사수료)
2013년 더아이엠씨 빅데이터센터 센터장
2019년 오피니언라이브 AI데이터센터 센터장
2024년 ~ 현 재 포스트에아아이 AI 센터 이사
관심분야 : 자연어처리, 기계학습, 정보검색, 시멘틱 웹

**김 승 현**

https://orcid.org/0009-0002-9170-9543
e-mail : shkim@nia.or.kr
2021년 서강대학교 경제학과(석사)
2016년 ~ 현 재 한국지능정보사회진흥원 지능데이터본부 책임
관심분야 : 통계분석, 자연어 처리, 기계학습

# APPENDIX

Table A1. Example Composition of LLM Performance Evaluation Dataset

| Test set | Question | Answer | Type | classification |
|---|---|---|---|---|
| KBS Korean Languagee Proficiency Test | Q : 〈보 기〉에 제시된 단어의 발음이 표준 발음인 것끼리 묶인 것은?<br>Q : "Bojige jesin doen daeoeui bareumi pyojun bareumin geotggiri mookin geoseun?"<br>E : 〈보 기〉 ㄱ)의심[의심] ㄴ)본의[본이] ㄷ)닝큼[닝큼] ㄹ)무늬[무늬]<br>E : 〈Bo gi〉 ㄱ) uisim [uishim] ㄴ) bonui [boni] ㄷ) ningkeum [ningkeum] ㄹ) munui [munui]"<br>O : 1. ㄱ,ㄴ 2.ㄱ,ㄷ 3. ㄴ,ㄷ 4. ㄴ,ㄹ 5. ㄷ,ㄹ | 2 | (Single) Answer Type | Grammar |
| TOPIK | Q : ( )에 들어갈 말로 가장알맞은 것을 고르십시오.<br>Q : ( )-e deureogal malro gajang almajeun geoseul goreusipsio<br>다른 사람과 대화를 할 때는 적당한 거리를( ) 한다.<br>Dareun saramgwa daehwareul hal ttaeneun jeokdanghan georireul ( ) handa.<br>O : 1. 유지해야 2. 유지하는 3. 유지했고 4. 유지하니까<br>O : 1. yujihaya 2. yujihaneun 3. yujihago 4. yujihanikka | 1 | Incomplete Sentence Type | Reading |
| Korean History Proficienct Test | Q : (가)에 해당하는 인물로 옳은 것은?<br>Q : (Ga)-e haedanghaneun inmullo oreun geoseun?<br>E : 이곳 경복궁은 조선의 궁궐로 (가)이/가 이름 지었대. 국왕과 백성이 만년토록 태평하며 큰 복을 누리기를 바란다는 의미가 담겨 있어. 그는 새 왕조의 통치 방향을 제시한 조선경국전도 저술하였지.<br>E : Igot Gyeongbokgung-eun Joseon-ui gunggweollo (ga)-iga ireum jieotdae. Gukwanggwa baekseong-i mannyeontorok taepyeonghamyeo keun bog-eul nurigireul barandaneun uimiga damgyeo isseo. Geuneun sae wangjo-ui tongchi banghyang-eul jesihan Joseongyeonggukjeondo jeosulhaeotji.<br>O : 1.송시열 2.채제공 3.정몽주 4.정도전<br>O : 1. Song Si-yeol 2. Chae Je-gong 3. Jeong Mong-ju 4. Jeong Do-jeon | 4 | (Single) Answer Type | Basic |
| Domestic Travel Guide Certificate | Q : 소재지와 동굴의 연결이 옳은 것은?<br>Q : Sojaejiwa donggul-ui yeongyeol-i oreun geoseun?<br>O : 1. 경북 안동 - 성류굴 Gyeongbuk Andong - Seongnyugul<br>2. 강원 삼척 - 고씨굴 Gangwon Samcheok - Gossigul<br>3. 전북 익산 - 천호동굴 Jeonbuk Iksan - Cheonhodonggul<br>4. 충북 단양 - 초당굴 Chungbuk Danyang - Chodanggul | 3 | Matching Type | Tourist Resource Interpretation |
| Civil Service Exam, Grade 9 | Q : 관용표현 ㄱ~ㄹ의 의미를 풀이한 것으로 적절하지 않은 것은?<br>Q : Gwanyongpyo hyeon ㄱ~ㄹ-ui uimireul purihan geoseuro jeokjeolhaji anheun geoseun?<br>- 그의 회사는 작년에 노사 갈등으로 ㄱ홍역을 치렀다.<br>- Geuui hoesa-neun jagnyeone nosa galdeung-euro ㄱ hongyeog -eul chireotda.<br>- 우리 교장 선생님은 교육계에서 ㄴ잔뼈가 굵은 분이 십니다.<br>- Uri gyojang seonsaengnim-eun gyoyuk-gye-eseo ㄴ janppyeoga gulgeun bun-i simnida.<br>- 유원지로 이어지는 국도에는 차가 밀려 ㄷ입추의 여지가 없었다.<br>- Yuwonjiro ieojineun gukdoeneun chaga millyeo ㄷ ipchuui yeojiga eopseotda.<br>- 그분은 세계 유수의 연구자들과 ㄹ어깨를 나란히 하는 물리학자이다.<br>- Geubun-eun segye yusu-ui yeongujadeulgwa r eokkaereul naranhi haneun mullihagja-ida.<br>O : 1. ㄱ: 심한 어려움을 겪었다 2. ㄴ:오랫동안 일을 하여 그 일에 익숙한 3. ㄷ: 돌아서 갈 수 있는 방법이 없었다 4. ㄹ:비슷한 지위나 힘을 가지는<br>O : 1. ㄱ: simhan eoryeoum-eul gyeok-eotda 2. ㄴ: oraetdong-an il-eul hayeo geu il-e iksukan 3. ㄷ: dol-aseo gal su issneun bangbeob-i eopseotda 4. ㄹ: biseushan jiwi na him-eul gajineun | 3 | Incomplete Sentence Type | Vocabulary |

Table A2. LLM Evaluation Results

| | Model | KBS Korean Languagee Proficiency Test | | TOPIK | | Korean History Proficienct Test | | Domestic Travel Guide Certificate | | Civil Service Exam, Grade 9 | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Number of Correct Answers | Ratio | Number of Correct Answers | Ratio | Number of Correct Answers | Ratio | Number of Correct Answers | Ratio | Number of Correct Answers | Ratio |
| Public | LDCC-SOLAR-10.7B | 20 | 23 | 146 | 72 | 18 | 45 | 23 | 46 | 28 | 56 |
| | Bookworm-10.7B | 21 | 24 | 152 | 75 | 10 | 25 | 21 | 42 | 26 | 52 |
| | SOLARC-M-10.7B | 15 | 17 | 141 | 70 | 14 | 35 | 20 | 40 | 23 | 46 |
| | Llama-2-13b-chat-hf | 18 | 20 | 83 | 41 | 0 | 0 | 9 | 18 | 11 | 22 |
| | Kullm-solar | 8 | 9 | 87 | 43 | 2 | 5 | 6 | 12 | 10 | 20 |
| | KoAlpaca-Polyglot-12.8B | 13 | 15 | 33 | 16 | 8 | 20 | 12 | 24 | 6 | 12 |
| Commercial | HyperClovaX | 43 | 49 | 180 | 89 | 21 | 53 | 35 | 70 | 36 | 72 |
| | GPT-4-turbo | 41 | 47 | 180 | 89 | 17 | 43 | 27 | 54 | 40 | 80 |