메타버스 디지털트윈 챗봇 신뢰성 강화를 위한 한국어 LLM의 출력 민감도 분석

최연수, 제갈홍, 이현석*

세종대학교

c62041289@gmail.com, jagrhong@sju.ac.kr, *hyunsuk@sejong.ac.kr

Sensitivity Analysis of Korean LLMs for Enhancing Reliability of Metaverse Digital Twin Chatbots

Yeon-Su Choi, Hong Jegal, *Hyun-Suk Lee

Sejong Univ.

요 약

최근 컴퓨팅 리소스의 발전으로 인해 대규모 데이터를 처리하고 고도화된 모델을 학습시키는 것이 가능해지면서, 메타버스 디지털 트윈과 같은 분야에서도 챗봇의 형태로 고성능의 대규모 언어 모델(Large Language Model, LLM)이 널리 사용되고 있다. LLM은 확률적 생성 구조를 기반으로 하여 출력 문장을 생성하기 때문에 동일한 의미의 입력이라도 표현 방식의 작은 차이에 따라 출력 내용이나 방향이 달라질 수 있다. 이러한 출력의 민감성은 오용이나 정보 왜곡으로 이어질 가능성이 있다. 이에 따라, 입력 표현의 작은 변화에 대해 LLM이 출력에 얼마나 민감하게 반응하는지를 정량적으로 측정하는 작업은 LLM의 신뢰성과 안전성을 평가하는 데 필수적이다. 본 논문에서는 LLM의 신뢰도 및 안전성을 평가하기 위해 출력 민감도 정량화를 통하여 한국어 LLM에 원본 문장과 변화된 문장을 입력하고, 이에 따른 출력 문장들간 차이를 비교하는 실험을 진행한다.

Ⅰ. 서 론

최근 대규모 언어모델(Large Language Model, LLM)은 챗봇을 비롯해 의료, 법률, 고객 응대 등 다양한 분야에서 활발히 활용되고 있다[1]. LLM은 입력 문장에 따라 확률적으로 단어를 선택해 문장을 생성하는 구조를 가지므로, 동일한 입력에도 서로 다른 출력을 생성할 수 있다. 또한, 입력 문장의 표현이 조금만 달라져도 LLM은 그 차이에 민감하게 반응하여 출력의 의미나 방향이 달라질 수 있다. 이러한 특성은 단순한 질의응답을 넘어, 의료 상담, 법률 자문, 금융 조언 등 사용자의 판단과 의사결정에 실질적인 영향을 미치는 영역에서 특히 중요하게 작용한다. 동일한 의미를 전달하는 입력이라도, 표현 방식의 작은 차이에 따라 LLM의 응답이 달라질 수 있으며, 그로 인해 정보의 편차나 의미 왜곡이 발생할 수 있다.

따라서, LLM의 신뢰성과 일관성을 확보하기 위해서는, 입력 표현의 작은 변화에 대해 LLM이 출력에 얼마나 민감하게 반응하는지를 정량적으로 평가하고, 이를 바탕으로 LLM의 출력 결과를 이해하는 작업이 요구된다. 이를 위한 분석 기법으로 제안된 것이 분포 기반 분석 기법(Distribution-Based Perturbation Analysis, DBPA)이다[2]. 본 논문에서는 영문 기반으로 개발된 DBPA 구조를 한국어 LLM 환경에 맞게 구현하고 LLM의 출력 민감도 정량화 실험을 수행하였다.

Ⅱ. DBPA 구조

DBPA는 입력 표현의 작은 차이에 대해 LLM의 출력 의미가 얼마나 달라지는지를 정량적으로 평가하기 위해 설계된 분포 기반 분석 기법이다. LLM은 확률적 출력 생성 구조를 기반으로 동일한 입력에도 서로 다른 출력을 생성하며, 입력 문장의 표현이 조금만 달라져도 응답의 의미나 방향이 달라질 수 있다. 따라서 입력 문장의 일부를 변형하거나 제거해 출력 변화를 측정하는 입력 변형(perturbation) 기반 기법으로는 LLM의 출력 민감도를 정밀하게 평가하기 어렵다.

DBPA는 이러한 한계를 극복하기 위해, 기존 perturbation 기반 민감도 평가 방법에 빈도 가설(frequency hypothesis)을 적용하였다. LLM의 출력 문장 생성 유동성을 고려해 샘플의 개수를 증가시키고 통계적 검정 과정을 거쳐 안정화하는 방식으로 설계되었다.

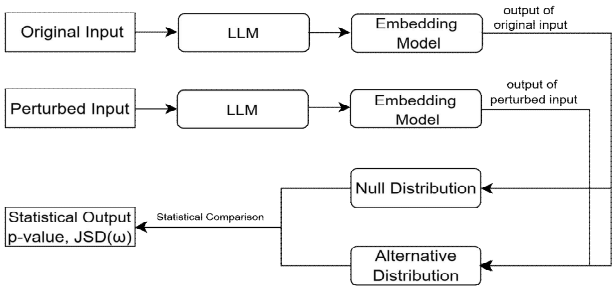

Fig. 1. DBPA 구조

Fig. 1은 본 논문에서 구현한 DBPA의 구조를 나타낸다. DBPA 분석 절차는 크게 네 단계로 구성된다. 첫째, 입력 변화가 LLM 출력 분포에 미치는 영향을 분석하기 위해, 분석 대상이 되는 원본 입력 문장과 원본 입력 문장에서 일부 속성을 변경한 변형 입력 문장들을 생성한다.

둘째, LLM의 출력 분포를 수집하기 위해, 각 입력에 대해 LLM으로부터 다수의 출력을 생성한다. 생성된 문장들은 임베딩 모델을 통해 벡터화되며, 문장 길이 영향을 받지 않고 의미 유사성을 효과적으로 측정하기 위해 코사인 유사도를 사용하여 출력 간 유사도를 계산한다.

셋째, 입력 변화 전후의 출력 분포 차이를 분석하기 위해, 유사도 값을 기반으로 두 개의 분포를 구성한다. 원본 입력의 출력 간 유사도 분포는 Null 분포로 정의하고, 원본과 변형 입력의 출력 간 유사도 분포는 Alternative 분포로 정의한다. Null 분포는 하나의 입력에서 자연스럽게 발생할 수 있는 출력 간 변동성의 범위를 나타내고, Alternative 분포는 입력 변화가 출력에 유의미하게 영향을 미친 정도를 나타낸다.

넷째, 입력 변화가 실제로 출력에 유의미한 영향을 주었는지 판단하기 위해, Jensen-Shannon Divergence(JSD, ω)를 통해 두 분포 간의 차이를 정량화한다. JSD(ω)는 두 분포의 평균 분포를 기준으로, 각 분포와 평균 분포 간의 Kullback-Leibler Divergence(KLD)를 평균한 값으로, 두 분포 간의 유사성과 차이를 대칭적으로 정량화할 수 있는 지표이다. JSD(ω)값은 0에서 1 사이로 나타나며, 값이 클수록 두 분포의 차이가 크고 이는 입력 변화에 따른 출력 반응이 컸음을 의미한다. 이후, 관측된 JSD(ω)값이 통계적으로 유의미한지 검정하기 위해 순열 검정을 수행하여 p-value를 산출한다. 이때 두 분포를 무작위로 섞어 반복적으로 나눈 시뮬레이션 분포와 관측된 JSD(ω)를 비교해 입력 변화에 따른 것인지 아닌지를 판단한다. 계산된 p-value는 입력 변화 없이도 차이가 발생할 확률을 의미하며, 값이 작을수록 입력 변화가 출력에 유의미한 영향을 주었음을 나타낸다.

이러한 절차를 통해 DBPA는 입력 변화가 LLM의 출력 의미에 미치는 영향을 정량적으로 해석할 수 있다. 본 논문에서는 DBPA 분석 절차를 한국어 LLM 환경에 적용하여 실험을 구현하고, 그 결과를 분석하였다.

Ⅲ. 실험

본 논문에서는 DBPA 분석 절차를 한국어 기반 LLM 환경에서 구현하는 실험을 수행하였다. 실험에 사용된 한국어 LLM은 한국어 특화 모델인 Exaone-3.5와 다국어 모델인 Gemma-3이며, 각 모델에 대해 4종의 임베딩 모델(jhgan/ko-sbert-nli, multilingual-e5-large, bge-m3-korean, paraphrase-multilingual-mpnet-base-v2)을 적용하여 DBPA 결과를 비교 분석하였다. 임베딩 모델은 LLM의 출력 문장을 벡터로 변환해 의미적 유사도를 정량화하는 데 사용되며, 각 모델의 구조나 표현 방식에 따라 측정 결과가 달라질 수 있다. 이에 따라 다양한 임베딩 모델을 함께 적용해 DBPA 결과의 일반성과 신뢰도를 확보하고자 하였다.

LLM에 대한 입력 변화가 출력에 미치는 영향을 정량적으로 측정하기 위해, 다음과 같은 절차로 실험을 설계하였다. 먼저, 심혈관 질환 예방 및 치료에 대한 조언을 요청하는 내용을 기반으로 원본 입력 문장을 설정하고, 이 문장에서 나이, BMI, 혈압, 콜레스테롤 수치, HDL 수치, 흡연 상태, 당뇨병 유무, 가족력, 인종 등 일부 정보를 변경한 변형 입력 문장 10개를 구성하였다. 이러한 방식은 실제 활용 상황에서 입력 문장의 세부 속성이 변형되는 사례를 시뮬레이션하기 위한 목적으로 설계하였다. 입력 문장의 세부 변화에 LLM 출력이 얼마나 민감하게 반응하는지를 확인하기 위해, 원본 및 변형 입력 각각에 대해 LLM을 통해 10개의 출력을 생성하였으며, 해당 출력 문장들은 임베딩 모델을 통해 벡터로 변환되었다. 이후 코사인 유사도를 이용하여 출력 간 의미적 유사도를 산출하였고, 이를 기반으로 원본 입력에서 생성된 출력 간 유사도 분포는 Null 분포로, 원본과 변형 입력 간 출력 유사도는 Alternative 분포로 정의하였다. 두 분포 간 차이는 JSD(ω)를 통해 정량화하였으며, 입력 변화가 출력 분포에 유의미한 영향을 주었는지 확인하기 위해 순열 검정으로 p-value를 산출하였다.

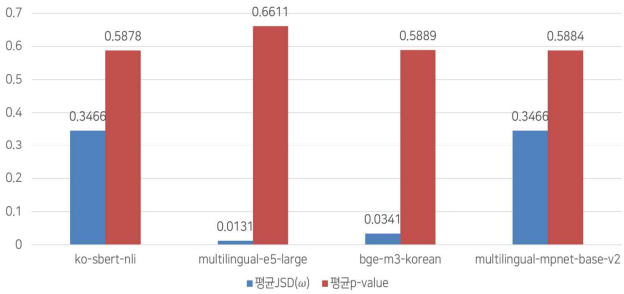

Fig.. 2. Exaone-3.5 실험결과

Exaone-3.5 실험 결과는 Fig. 2와 같다. 평균 p-value가 전반적으로 높아 입력 변화가 출력에 통계적으로 유의한 영향을 주지 않은 경우가 많았다. 그러나 ko-sbert-nli와 paraphrase-multilingual-mpnet 임베딩 모델에서 JSD(ω)값이 0.3466으로 상대적으로 높게 나타났고 multilingual-e5-large는 JSD(ω)값이 0.0131로 낮았다. 이처럼 임베딩 모델에 따라 평균 JSD 값의 편차가 커, 민감도 분석 결과가 달라질 수 있음을 보여준다.

Fig. 3. Gemma-3 실험 결과

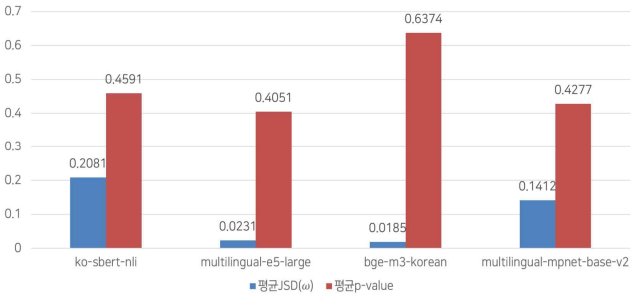

Fig. 3은 Gemma-3 실험 결과이다. 대부분의 임베딩 모델에서 p-value가 높고 JSD(ω)값이 낮게 측정되었다. 그러나 multilingual-e5-large와 bge-m3-korean에서는 상대적으로 JSD(ω)값이 높게 나타났다. 전반적으로는 Exaone에 비해 Gemma가 입력 변화에 덜 민감하게 반응한 것으로 해석할 수 있다.

실험 결과를 통해 DBPA 구조를 한국어 환경에 맞게 구현함으로써, 입력 변화가 출력 의미에 미치는 영향을 정량적으로 측정하고 비교할 수 있는 분석이 가능함을 확인하였다. 또한, DBPA 분석 결과가 사용하는 임베딩 모델의 특성에 따라 달라질 수 있음을 보여준다.

Ⅳ. 결론

본 논문에서는 DBPA를 한국어 LLM 환경에서 구현하고, 이를 통해 LLM의 출력 변화를 정량적으로 분석할 수 있음을 확인하였다. 또한, 다양한 임베딩 모델을 적용한 결과, 임베딩 모델의 종류에 따라 민감도 분석 결과가 달라질 수 있음을 관찰하였다. 이는 임베딩 모델마다 문장의 의미를 벡터로 표현하는 방식, 학습 데이터의 범위, 언어적 특성에 대한 민감도가 다르기 때문이며, 동일한 출력이라도 의미 유사도 인식에 차이가 발생할 수 있다. 이러한 결과는 DBPA의 분석 결과가 사용하는 임베딩 모델에 따라 달라질 수 있으며, 결과 해석 시 임베딩 모델의 선택이 중요한 요소가 될 수 있음을 보여준다. 향후에는 임베딩 분석 결과의 편차를 줄이기 위한 보완이 이루어진다면, DBPA의 해석 안정성과 다양한 언어 환경에서의 일반화 가능성이 더욱 강화될 것으로 기대된다.

ACKNOWLEDGMENT

이 논문은 2025년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(No. RS-2022-II220545, 지능형 디지털 트윈 연합 객체 구성 및 데이터 프로세싱 기술 개발, 50%)과 정보통신기획평가원-대학ICT연구센터(ITRC)의 지원(IITP-2025-RS-2021-II211816, 50%)을 받아 수행된 연구임.

참 고 문 헌

[1]HADI, Muhammad Usman, et al. Large language models: a comprehensive survey of its applications, challenges, limitations, and future prospects. Authorea Preprints, 2023, 1: 1-26.

[2]RAUBA, Paulius; WEI, Qiyao; VAN DER SCHAAR, Mihaela. Quantifying perturbation impacts for large language models. arXiv preprint arXiv:2412.00868, 2024.