거대 언어 모델(LLM)의 비즈니스 활용을 위한 검증 미들웨어 시스템 구현

이세정, 조영준†, 이숙윤*
고려대학교 컴퓨터학과

seajo@korea.ac.kr, † whdudwns9909@korea.ac.kr, *uni7@korea.ac.kr

A Verification Middleware System for Business Utilization of Large Language Models(LLM)

Lee Se Jeong, Joh Yeong Jun† , Suk-Yun Lee*
Department of Computer Science and Engineering, Korea University

요 약

본 논문에서는 거대 언어 모델에서 입출력에 대한 신뢰성 확보가 비즈니스 활용의 중요 요소임을 전제하고, 모델과 데이터셋의 재구성 없이 작은 프로그램인 미들웨어들의 조합으로 입출력을 검증해 신뢰성을 향상시키는 방법을 제시한다. 제안된 방법으로 웹 애플리케이션 개발을 하고 미들웨어 예시를 통해 실제 사용 상황에서의 결과와 효과를 검증한다. 이를 통해 합리적인 비용으로 다양한 비즈니스 문제에 대응해 거대 언어 모델을 활용할 수 있음을 보인다.

Ⅰ. 서론

최근 많은 연구가 진행되고 있는 거대 언어 모델(Large Language Models : LLM)은 자연어, 프로그래밍 코드, 이미지 등 다양한 형태의 입력을 받아 유창하게 처리할 수 있는 능력을 지니고 있어 많은 분야에 활용이 되고 있다. 특히 비즈니스 분야에서는 생산성과 업무 효율성 증진에 큰 잠재력을 가지고 있다[1].

그러나 LLM 의 강력한 기능의 이면에는 한계점도 있다. 먼저 때때로 잘못된 정보를 제공하는 할루시네이션(hallucination) 출력[2] 발생으로 신뢰성에 문제가 생긴다. 다음으로는 입력된 기밀 데이터가 외부 서버로 무분별하게 전송되어 추출[3] 당할 수 있는 입력 보안 문제로 인해 기업에서 기술의 사용에 제약이 된다. LLM 의 오류를 사용자가 쉽게 감지하기 어려울 뿐만 아니라, 기업에서는 기밀성 유지에 대한 우려로 기술의 채택을 주저하게 만드는 요인이 되고 있다.

따라서 LLM 의 비즈니스 활용을 위해서는 LLM 의 입력과 출력 데이터에 대한 신뢰성이 확보되어야 한다. 본 연구에서는 합리적 비용으로 신뢰성을 확보하기 위하여 LLM 의 검증 미들웨어 시스템을 제시하고, 시스템과 데모 유저 인터페이스인 웹 어플리케이션을 구현하여 명시적이고 유연한 신뢰성 향상의 효과를 검증한다.

Ⅱ. LLM 의 할루시네이션(Hallucination) 문제

LLM 의 할루시네이션은 널리 알려진 문제로, 답변을 할 때 사실 관계를 왜곡한 허위 정보를 생성하거나 문맥과 전혀 맞지 않는 답변을 제공하는 것을 의미한다. 고도화된 LLM 일수록 잘못된 정보를 포함하여 유창한 답변을 만들어내므로 사용자는 이를 쉽게 감지하기 어렵고, 기업에서는 오류에 기반한 잘못된 정보를 재생산하므로 위험이 발생한다. 때문에 할루시네이션 문제를 개선하기 위한 방법들이 활발하게 연구되고 있다[2].

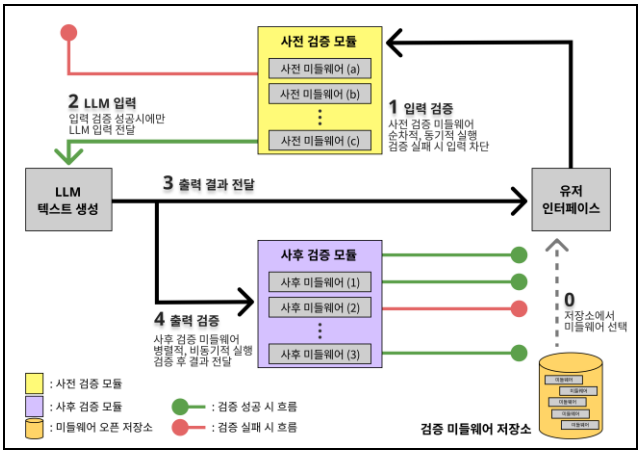
Ⅱ-1. 할루시네이션 제거

할루시네이션을 제거하거나 최소화하기 위해서는 모델을 새로 설계하거나 데이터셋의 수정 혹은 보완을 기반으로 이루어져야 한다. 작업 별 학습과 미세 조정이 필요하며 이 과정에서 방대한 컴퓨팅 파워와 시간이 소모된다. GPT-3[4] 모델이 학습하는 데에 1750 억개의 매개변수, 3000 억개의 데이터 수를 필요로 한 것처럼, 대부분의 LLM 은 많은 파라미터 수와 데이터 수를 가지고 있어 학습에 큰 비용이 발생한다.

Ⅱ-2. 할루시네이션 감지

할루시네이션의 감지는 오류의 종류에 따라 교차 검증으로 자동화할 수 있을 만큼 간단하며 모델의 구조를 변경하지 않아 새로운 학습을 필요로 하지 않는다. 검증 목표를 작은 도메인으로 특정한다면 간단한 자동 검증 프로그램으로도 결과를 확인할 수 있으므로 비교적 개발 방법이 수월하여 다양한 분야의 전문가들이 참여할 수 있다. 때문에 감지에는 유연한 개발로 물리적 비용과 시간을 절약할 수 있다는 장점이 있다.

할루시네이션 문제를 최소화해도 이를 완전히 해결할 수 없으므로, 그 사용에 대한 검증 책임은 사용자에게 달려있다. 따라서 본 연구에서는 비즈니스적 활용을 위해 LLM 의 할루시네이션 감지를 중심으로 신뢰성 검증 시스템을 설계하고자 한다. 이를 통하여 합리적인 비용으로 LLM 의 신뢰성을 확보하고, LLM 사용의 리스크를 예측 가능한 범위로 완화할 수 있다.



[그림 1] 제안된 시스템의 구조도 및 실행 흐름도

Ⅲ. 시스템 제안과 설계

본 연구에서는 LLM 사용 시 발생되는 오류를 입력문제와 출력문제로 구분하고, 잘못된 입력의 원천차단과 잘못된 출력의 선택적 대응을 수행하도록 시스템을 설계하였다. [그림 1]과 같이 시스템의 각 부분은 검증 미들웨어 저장소, 사전 검증 모듈, 사후 검증 모듈, 그리고 유저 인터페이스로 구분된다.

사용자의 질의 입력 값에 대해 1) 입력 검증을 순차적, 동기적 수행하여 검증 실패 시 차단, 2) 검증 성공 시 LLM 에 입력 텍스트 전달, 3) LLM 에서 출력된 텍스트를 사용자에게 1 차적 전달, 4) 출력 검증을 병렬적으로 수행하여 검증 결과를 사용자에게 전달하는 순으로 진행된다. 제안한 아이디어를 시스템으로 구현 및 설계하는데 있어 핵심 프로세싱은 다음과 같다.

① **(소규모 미들웨어의 조합)** 모든 오류를 감지하기 위해서는 복잡한 프로그램이 필요하지만, 작은 단위의 오류는 간단한 프로그램으로도 감지가 가능하다. 본 논문에서는 30 줄 이하의 미들웨어로 다양한 오류를 감지해 낼 수 있음을 보인다.

② **(사전 입력 검증과 차단)** 개인정보 유출, 기밀정보 입력과 같이 LLM 의 입력 단계에서 일어나는 문제는 데이터를 전송하고 나면 돌이킬 수 없으며, 치명적인 결과로 이어질 가능성이 높다. 따라서 입력 검증은 동기적으로 실행하여 엄격하게 검증하고 위배 시 전체 과정을 차단하는 흐름으로 이어진다. ([그림 1], 1 입력 검증)

③ **(사후 출력 검증과 병렬 실행)** 반면, LLM 의 출력 값에 대해서는 검증 과정이 상대적으로 복잡하고 실행 시간이 오래 걸리는 데다가, 경우에 따라서는 사용자가 직접 결과를 보고 판단하는 것이 더 빠르고 효율적일 수 있다. 따라서 출력 검증 단계는 사용자에게 출력을 전달하면서 동시에 병렬적으로 실행하여 전체 수행 시간을 줄이고 보조 역할을 수행하도록 설계되었다. ([그림 1], 4 출력 검증)

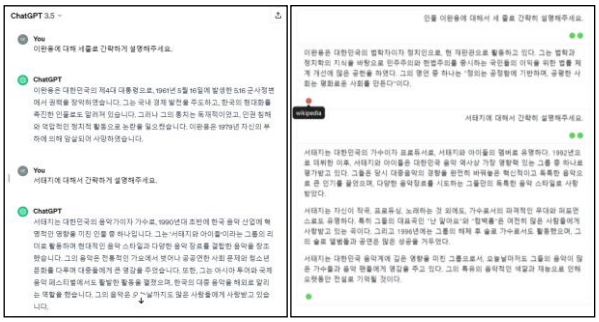
④ **(검증 미들웨어 및 LLM 맞춤화)** 최근의 LLM 은 상용 서비스, 사내 인프라, 오픈소스 등의 많은 형태로 제공되고 있어 사용자 환경에 따라 달라질 수 있으며, 검증 미들웨어 또한 시기, 상황 별로 달라질 수 있다. 제안된 시스템은 각 구성요소가 유연하게 결합하여 하나의 서비스에 종속되지 않고 선택이 가능하도록 설계되었다.

Ⅳ. 구현 결과 및 검증

제안한 아이디어의 개선 효과를 시각적으로 나타내고 실용성을 검증하기 위해, 유저 인터페이스와 코드 오류 검증, 출력결과 웹 교차 검증, 개인정보 검출 검증 등의 미들웨어 예시를 웹으로 구현하였다.

[그림 2]는 웹 교차 검증 미들웨어의 활용 예시로, '이완용'과 '서태지'라는 두 인물에 대해 질문했을 때 ChatGPT 서비스의 결과와 데모로 구현된 시스템의 결과를 비교하였다. '이완용'에 대한 사실 관계를 왜곡한 답변을 한 기존 ChatGPT 서비스와는 달리, 제안된 시스템에서는 해당 답변에 대해 병렬적인 검증을 수행함으로 오류가 발생했음을 시각적으로 표시하는 피드백을 제공한다. 이는 사용자가 답변에 별도의 조치를 취할 수 있도록 보조해주는 역할을 한다.

예시로 사용된 미들웨어 외에도, 출력된 코드에 오류가 있는지 검증하거나 입력에 대한 개인정보 검출 혹은 사내 시스템 코드 입력 감지 등의 미들웨어에도 사용될 수 있다. 이는 사용자 혹은 시스템 관리자의 요구나 기밀 내용에 따라 새로 작성되고 선택될 수 있다.



(a) ChatGPT 서비스 (b) 제안된 데모 웹 앱

[그림 2] 오류 비교 : 기존 LLM 서비스 *vs* 제안된 시스템

제안된 시스템은 구현한 웹 이외에도 API 를 통해 라이브러리나 데스크톱 앱과 같이 다양한 방법으로 유저 인터페이스를 구성할 수 있다. 또한 LLM 모듈과 미들웨어 저장소도 비즈니스 환경에서 여러 조건에 맞게 구현하여 사용할 수 있도록 유연하게 구성하였다.

Ⅴ. 결론 및 활용방안

본 논문에서는 거대 언어 모델(LLM)의 비즈니스적 활용을 위해 신뢰성을 확보할 수 있는 검증 미들웨어 시스템을 제안하고 미들웨어의 예시와 데모를 구현하여 효과를 확인하였다. 많은 비용을 들여 새로 모델을 학습시키고 튜닝하는 방법과 달리, 간단한 검증 미들웨어들을 통해서 민첩하게 비즈니스 문제에 대응할 수 있는 효과가 있으므로 엔터프라이즈 사례에서 의미를 가질 것으로 보인다.

ACKNOWLEDGMENT

본 논문은 2023 년도 SW 중심대학의 지원을 받아 수행되었음.

(No. 2023-0-00044).

참 고 문 헌

[1] Raj, R., Singh, A., Kumar, V., & Verma, P. (2023). Analyzing the potential benefits and use cases of ChatGPT as a tool for improving the efficiency and effectiveness of business operations. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations, 3*(3), 100140.

[2] Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., ... & Liu, T. (2023). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint* arXiv:2311.05232.

[3] Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., ... & Raffel, C. (2021). Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)* (pp. 2633-2650).

[4] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems,* 33, 1877-1901.