인하우스 LLM 서비스를 위한 효율적 API 연동 방안 연구

우정주[1], 성 민[1], 조예영[1], 강지석[2], 송용성[3], 최영락[4]
[1]경북대학교 컴퓨터학부 학부생
[2]경북대학교 영어영문학부 학부생
[3]순순팩토리
[4]오픈인프라한국사용자모임
jayjay33@knu.ac.kr, pho0575@knu.ac.kr, betty56044@gmail.com,
,jiseok3530@gmail.com, soonsoon@soonsoons.com, ianyrchoi@gmail.com

A Study on Efficient API Integration Methods for In-house LLM Services

Jungju Woo[1], Min Sung[1], Yeyoung Cho[1], Jisuk Kang[2], Yong Seong Song[3], Yeongrak Choi[4],
[1]Dept. of Computer Science and Engineering, Kyungpook National University
[2]Dept. of English Language and Literature, Kyungpook National University
[3]Soonsoon Factory
[4]OpenInfra Korea Group

요 약

본 연구는 기업 내 데이터 보안 및 규정 준수 요구를 충족하는 인하우스 LLM 서비스 구축을 위해 오픈소스 프레임워크(Open WebUI, LiteLLM)를 활용한 효율적인 API 연동 아키텍처를 제안한다. 두 퍼블릭 클라우드에서 제공하는 LLM API와의 연동 실험 결과, 직접 API 호출 대비 소폭의 응답 시간 증가를 보인 반면, GUI 기반 설정과 다중 API 추상화에 따른 관리 및 유지보수 효율성에 대한 장점을 보였다. 이러한 결과는 제안 아키텍처가 기업 환경에서 인하우스 LLM 서비스 구축에 효과적으로 활용될 수 있음을 시사한다.

I. 서론

최근 LLM(Large Language Model)의 급속한 발전과 함께, 기업 내 데이터 보안 및 규정 준수에 대한 요구가 높아지면서 인하우스 AI 서비스에 대한 수요가 급증하고 있다 [1]. 하지만 자체 인프라 구축에는 비용과 기술적 부담이 수반되므로, 오픈소스 및 API 기반의 하이브리드 환경이 현실적인 대안으로 주목받고 있다 [2]. 본 연구에서는 오픈소스 프레임워크인 Open WebUI와 LiteLLM을 활용하여, 다양한 LLM API와 연동 가능한 인하우스 아키텍처를 제안한다. 또한 퍼블릭 클라우드 기반 LLM API와의 연동 성능을 실험적으로 분석하여, 효율적인 인하우스 LLM 서비스 구축 방안을 검증하였다.

II. 실험 설계

1. 실험 환경

본 연구에서는 Docker 기반 컨테이너 환경에서 Open WebUI와 LiteLLM 프레임워크를 활용하였다. 이를 통해 동일한 프론트엔드 환경에서 API 프로바이더(예: Google Vertex AI, OpenRouter)를 유연하게 변경하며 실험을 진행하였다. 실험에서 사용한 LLM 모델은 Gemini-2.0-flash로 고정하고, 각 API 프로바이더에서 해당 모델을 선택 후 제안하는 인하우스 프레임워크 아키텍처와 연동하였다. 대표 샘플 프롬프트 [3]를 주제별로 나누어 평가 지표에 대해 API 프로바이더와의 직접 API 호출 방식과 비교하여 실험을 진행하였다.

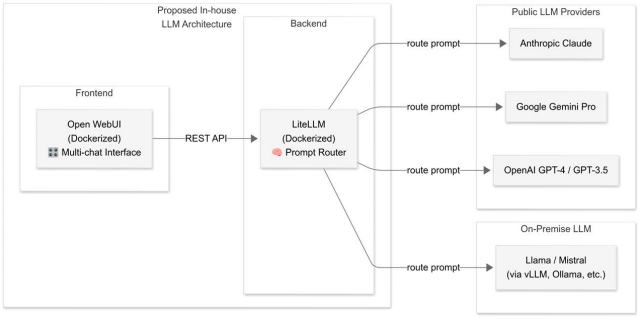

그림 1 제안 아키텍처 구조

2. 평가지표

본 실험에서 사용한 평가지표는 다음과 같다..

- 응답 시간 (초): 인하우스 LLM 프레임워크(Open WebUI+LiteLLM) 또는 직접 API 호출 방식을 통해 요청~응답까지의 소요 시간
- 코드 및 설정 변경 횟수: 동일한 기능 구현을 위해 요구되는 소스 코드 및 설정 파일 변경 횟수

III. 실험

설계된 환경에서 각 연동 방식을 통해 7개 프롬프트 주제(Q1~Q7: 텍스트 요약, 정보 추출, 질의응답, 텍스트 분류, 대화, 코드 생성, 추론)별로 10회씩 반복 실험을 수행하고, 응답 시간을 측정하였다. 인하우스 LLM 방식(LLM)과 직접 API 호출 방식(Direct)별 평균 응답 시간을 측정하고, 기능 변경 시 필요한 코드 및 설정 변경 횟수를 집계하였다.

IV. 결과분석

API	Google AI		Open Router	
방식	LLM	Direct	LLM	Direct
Q1	1.486	0.988	1.813	1.009
Q2	1.262	0.698	1.686	0.942
Q3	1.342	0.76	1.527	0.874
Q4	1.227	0.662	1.891	0.853
Q5	2.173	1.77	2.859	1.915
Q6	1.273	0.773	1.727	1.013
Q7	1.961	1.35	2.279	1.369

(표1) API 호출 응답 속도(평균값, 단위: 초)

필요 수정횟수	Vertex AI	OpenRouter
LLM (Proposed)	13	11
Direct API 방식	12	2

(표2) API 호출을 위한 수정횟수 (라인 및 설정 수정횟수)

1. 응답 속도

제안 아키텍처를 사용할 때 직접 API를 호출할 때보다 평균 0.53-0.83초(최소 0.40-0.65초, 최대 0.61-1.04초) 더 소요되었다. 이는 프레임워크의 중간 처리(예: 요청 라우팅, 인증, 로깅 등)로 인한 오버헤드로 해석된다. 그럼에도 불구하고, 데이터 보안 및 규정 준수 측면에서 인하우스 방식이 제공하는 이점을 고려하면, 해당 수준의 latency 증가는 실무적으로 수용 가능한 범위로 판단된다.

2. 코드 및 설정 변경 횟수

제안하는 인하우스 LLM 방식의 경우 GUI 기반 설정이 다수 포함되어 있어 변경 횟수는 많았으나, 대부분 클릭이나 간단한 환경설정으로 이루어져 난이도는 높지 않았다. 특히, API 프로바이더가 달라져도 동일한 추상화 구조를 통해 일관된 연동이 가능했다. 반면, 직접 API 호출 방식은 코드 변경 횟수는 적었으나, 인증 등 복잡한 절차(예: Vertex AI - Google Auth 인증 반복)가 요구되어 실질적인 유지보수 부담이 컸다.

3. 논의

제안하는 인하우스 LLM 프레임워크는 퍼블릭 클라우드에서 제공하는 API 뿐만 아니라 로컬 및 온프레미스 환경에 구축한 LLM에 대한 API 연동 확장이 가능하다. 향후 연구에서는 이 부분을 고려하고, 인하우스 LLM이 가지는 데이터 보안 및 규정 준수를 평가하기 위한 객관적인 지표를 설정하여 기업 적용 가능성 및 성능을 실험해보고자 한다.

V. 결론

본 논문에서는 기업 환경에서 안전하게 LLM 서비스를 활용하기 위한 인하우스 LLM 연동 아키텍처를 제안하고, 퍼블릭 클라우드 API와의 연동 실험을 통해 응답 속도 및 코드 변경 효율성을 비교·분석하였다. 제안한 프레임워크는 오픈소스 기반으로 구축이 용이하며, 다양한 LLM API와의 연동을 지원한다. 실험 결과, 인하우스 방식은 약간의 latency 증가에도 불구하고, 보안성과 관리 편의성 측면에서 실질적인 이점을 제공함을 확인하였다. 향후에는 제안 아키텍처를 개선하고, 보안성·규정 준수·비용 효율성 등 다양한 실무 지표를 추가 분석을 진행해보고자 한다.

참고문헌

[1] Large Language Model Powered Tools Market Size, Share & Trends Analysis Report By Type (General-purpose Tools, Domain-specific Tools), By Deployment, By Application, By Region, And Segment Forecasts, 2024 - 2030.

[2] Wang and Xu, 16 Changes to the Way Enterprises Are Building and Buying Generative AI, Andereessen Horowitz, 2024.

[3] Prompt Engineering Guide, available at https://www.promptingguide.ai/introduction/examples (April 2024).

이 연구는 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업 지원을 통해 수행되었음 (2021-0-01082)