UNIVERSIDAD EAFIT
SI7016 NLP Aplicado, 2025-2

Tarea 4 – Realizar el despliegue en simulación de producción de un modelo abierto LLM que implemente mínimo una interfaz tipo chatgpt. Realizar el analisis exploratorio del mecanismo de comunicación entre agentes Model Context Protocol.

Fecha de entrega: hasta el 21 de septiembre de 2025

Descripción de la tarea 4

Se realizara en los equipos de trabajo del proyecto final las siguientes dos actividades:

1. Despliegue de un modelo abierto LLM para mínimo la aplicación de chat (tipo chatgpt) y realizar alguna otra funcionalidad que soporte (ej: RAG, API, MCP, fine-tuning, etc)

Existen diferentes frameworks como ollama, vllm, lightning, lmstudio, etc., que pueden ser desplegados en diferentes ambientes de computo desde máquinas virtuales con gpu hasta servicios administrados en las nubes – aws,gcp,azure -, o plataformas como lightning.ai. tendiendo en cuenta que tambien puede se desplegado de forma local y/o en nube como lightning o lmstudio.

Por facilidad de este punto, se recomienda utilizar https://lightning.ai/ para desplegar modelos LLM, y realizar allí el despliegue de modelos abiertos y exponer interfaces gui tipo chat y APIs.

En el caso particular de lightning a través de la opción de “My Studios”, permite explorar ‘templates’, ‘models’ y ‘platform’, con especial énfasis en ‘platform’ porque es el ambiente donde permite desplegar los modelos.

Los ‘Studios’ que permite gestionar lightning permite tanto realizar el desarrollo (con Visual Studio Code y Notebooks python), y un ambiente de ejecución de computo con CPU, GPU, RAM y Disco. Además expone URLs publicas para ser accedidos a través de Internet.

Requerimientos especificos:

1.1 Desplegar modelos abiertos de ollama en un ambiente de plataforma lightning.
1.2 Desplegar modelos propios de lightning en plataformas lightning.
1.3 Explorar ventajas adicionales de la plataforma lightning como MCP, APIs, fine-tuning, etc.

Como entregable de este numeral, deberan realizar un pequeño informe a modo de reporte técnico y y un sustento a nivel de demostración (repositorio github, videosustentación y/o informe).

2. Realizar una exploración de posibilidades investigativas y practicas del protocolo y estándar de comunicación entre agentes MCP.

Realizar una exploración investigativa de la plataforma MCP: https://modelcontextprotocol.io/

Realizar alguna demostración funcional y básica de comunicación entre agentes basado en MCP. Explorar plataformas como LangChain y su ecosistema de frameworks y aplicación. Tambien hay soporte y demos en https://lightning.ai/. Tambien se puede explorar otras plataformas.

Como entregable de este numeral, deberán realizar una muy corta sintesis a nivel de ensayo de marco de referencia, y un sustento a nivel de demostración (repositorio github, videosustentación y/o informe).

Regla de ética y transparencia para todas las alternativas:

Si encuentran soluciones públicas o de reúso de código de alguna de las partes requeridas en este trabajo debe EXPLÍCITAMENTE DECLARAR:

- Declarar explícitamente: De que referencias en kaggle, médium, datacamp, toward data science, o de otro sitio, ud empleo parte del código y la solución para realizar su propio trabajo.
- Declarar explícitamente: cual fue el aporte específico que el grupo realizó en el trabajo.
- Uso de GenIA: deberá declarar explícitamente como fue el uso de chatgpt, copilot o similar empleo.

Entregables:

1. En la misma carpeta compartida para la tarea 1, con nombre: 'si7016-username-trabajos-252' en google drive compartida a: Edwin Montoya: edwin.montoya@gmail.com y allí contendrá:
    a. Directorio 'tarea4' y dentro:
        i. Datasets utilizado
        ii. Notebooks o programas en Python desarrollados
        iii. Toda la documentación completa debe estar en los mismos notebooks desarrollados
2. Enviar por BUZÓN DE ENTREGA de la plataforma de Interactiva Virtual declarando que ya está listo la tarea 4 y cuál es la URL en Google drive. (solo un miembro realiza la entrega especificando todos los integrantes del equipo, un solo directorio gdrive, no copiado en cada uno de los estudiantes, si trabajan en grupo)