Desarrollo de un Chatbot Empresarial Funcional Basado en MCP

Marc Saez Rama
CE Inteligencia Artificial & Big Data
ITIC Barcelona
marc.saez

Abstract

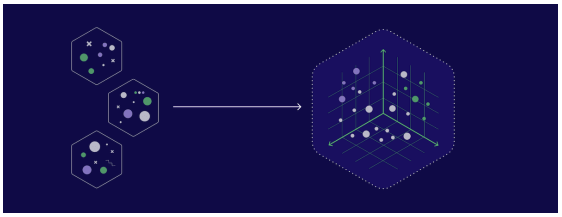
En esta investigación se presenta un enfoque innovador para la creación de chatbots empresariales inteligentes, centrado en el uso del protocolo MCP [1] (Model Context Protocol) como alternativa a metodologías tradicionales como RAG [2] (Retrieval-Augmented Generation). El objetivo es desarrollar un sistema genérico que permita a las empresas construir agentes conversacionales capaces de responder preguntas basadas en información privada, utilizando únicamente sus propios archivos como fuente de conocimiento. A través de la integración del servidor MCP filesystem, se automatiza la identificación contextual de los directorios relevantes y se proporciona al modelo solo la información necesaria para generar respuestas precisas y contextualizadas. Los resultados demuestran una mejora significativa en la fidelidad y utilidad de las respuestas frente a las obtenidas mediante RAG, estableciendo las bases para una solución de código abierto robusta y escalable que simplifica la implementación de agentes conversacionales empresariales basados en IA.

1. Introducción

En la actualidad existen diversas metodologías para crear chatbots basados en inteligencia artificial (IA), sin una estandarización clara sobre cuál es la más efectiva para el entorno empresarial. Esta investigación tiene como objetivo definir un sistema genérico que permita a las empresas crear agentes conversacionales capaces de responder preguntas basadas en su información privada, sin necesidad de programación avanzada ni configuraciones complejas. El enfoque parte de una necesidad real: simplificar el proceso de integración de IA en entornos corporativos mediante el uso de protocolos eficientes y abiertos.

2. Metodología

Inicialmente se utilizó un enfoque RAG, vectorizando archivos empresariales con Pinecone [4] y accediendo al contenido a través de embeddings. Sin embargo, este método generaba respuestas demasiado generales, imprecisas o incluso inventadas, lo que motivó el cambio hacia una metodología basada en MCP.

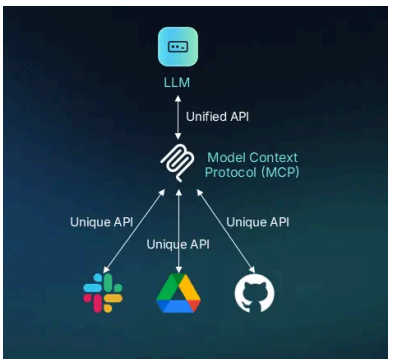


El nuevo sistema utiliza Python junto con el servidor MCP de tipo `filesystem`[3] (https://github.com/modelcontextprotocol/servers/tree/main/src/filesystem), que permite explorar el sistema de archivos de forma estructurada. Se ha consultado información y recursos clave disponibles en las plataformas oficiales de MCP (https://modelcontextprotocol.io/, https://mcp.so/) para comprender a fondo su funcionamiento y capacidades.

El proceso técnico es el siguiente:

1. El MCP escanea todos los directorios y genera un archivo JSON con descripciones breves de cada uno.
2. Cuando el usuario hace una consulta, el sistema compara la pregunta con las descripciones y selecciona el directorio más relevante.
3. Luego se leen los archivos de ese directorio y se pasan como contexto al modelo de lenguaje para generar la respuesta.

Esta arquitectura permite el uso modular de herramientas (tools), que permiten ampliar las funcionalidades del agente de forma flexible. Por ejemplo, se podría incorporar una tool para interpretar ficheros financieros, otra para analizar correos electrónicos o incluso para interactuar con APIs externas.



3. Resultados

La comparación entre RAG y MCP se realizó mediante pruebas con documentación empresarial propia. Aunque el análisis fue subjetivo, basado en la calidad percibida de las respuestas, se observó una clara mejora en la concreción, precisión y utilidad de las respuestas generadas con MCP. A diferencia de RAG, donde el modelo a menudo "alucinaba" respuestas o se desviaba del contenido real, con MCP el agente respondía de forma fidedigna al contenido proporcionado por los archivos.

Por ejemplo, al consultar sobre detalles técnicos de un producto documentado en PDFs, el agente con RAG ofrecía generalidades aprendidas del corpus de entrenamiento, mientras que el sistema con MCP extraía directamente los parámetros del archivo correspondiente, sin errores ni invenciones.

4. Discusión

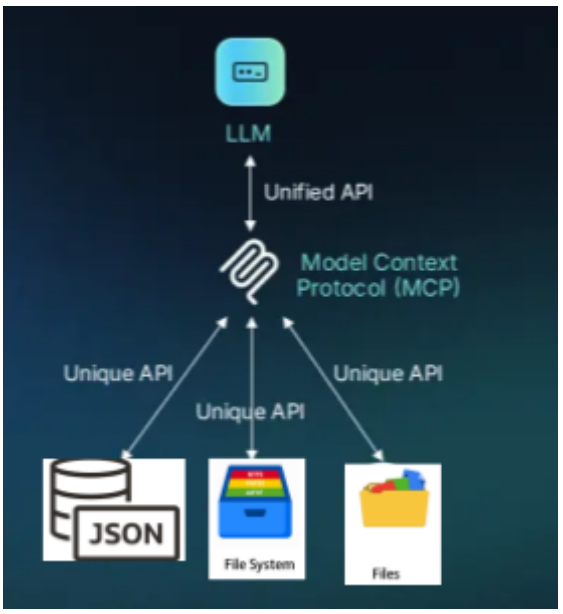
Este sistema permite a una empresa crear un agente que puede responder cualquier pregunta sobre su información interna simplemente subiendo sus archivos. La capacidad de identificar el contexto relevante antes de la generación de respuesta hace que MCP sea una herramienta poderosa. A diferencia de RAG, no depende de bases vectoriales externas, y su enfoque es mucho más determinista y controlado.

Además, el sistema MCP fomenta la trazabilidad y auditabilidad, permitiendo saber qué archivos han sido utilizados para cada respuesta generada. Esta propiedad es especialmente valiosa en entornos regulados como el financiero, médico o legal.

Aunque aún no se han realizado pruebas masivas ni con usuarios finales, los resultados preliminares demuestran que MCP es una tecnología con gran potencial, especialmente si se diseñan buenas "tools" que lo acompañen. Esta investigación plantea que el futuro de los chatbots empresariales puede pasar por arquitecturas como MCP, capaces de combinar el poder de los modelos de lenguaje con el acceso directo a conocimiento empresarial estructurado.

5. Conclusiones

Se concluye que la metodología basada en MCP proporciona mejores resultados en la creación de agentes empresariales inteligentes. Su enfoque estructurado y contextual permite generar respuestas mucho más útiles cuando se trata de información interna. MCP también abre la posibilidad de integrar otras herramientas avanzadas, como interfaces gráficas interactivas, dashboards, y sistemas de automatización.



6. Referencias

[1] MCP https://modelcontextprotocol.io/introduction

[2] RAG https://arxiv.org/pdf/2005.11401

[3] Filesystem https://github.com/modelcontextprotocol/servers/tree/main/src/filesystem

[4] Pinecone https://www.pinecone.io/

MCP Midudev