“Integración de LLMs en programación de APIs”

Sprint 3

Bernardo Corona Domínguez

NAO ID: 1085

Ciudad de México, México

23 de mayo de 2025

CueBot es un asistente virtual basado en LLMs que tiene como objetivo facilitar la interacción con el acervo digital de la Universidad de Cuévano. Este sistema incluirá una API desarrollada con FastAPI y una interfaz gráfica simple con Gradio para pruebas iniciales. Posteriormente, se planea su integración a la página web de la universidad.

1. Backlog.

A) Historias de usuario.

A continuación, las 5 (cinco) historias de usuario:

1. Como investigadora, quiero subir archivos en distintos formatos (.pdf, .txt, .html) para que CueBot pueda analizarlos.
2. Como docente, quiero realizar preguntas en lenguaje natural a CueBot para obtener respuestas rápidas y relevantes del acervo digital.
3. Como desarrolladora, quiero contar con una API documentada que me permita integrar CueBot a otras plataformas fácilmente.
4. Como estudiante, quiero acceder a una interfaz amigable para interactuar con CueBot sin necesidad de conocimientos técnicos.
5. Como administradora del sistema, quiero monitorear el desempeño del LLM para asegurar que funciona eficientemente y cumple con las expectativas.

B) Tablas de registro.

Lista de Requerimientos	
Historias de usuario	**Requerimientos**
Como investigadora, quiero subir archivos en distintos formatos (.pdf, .txt, .html) para que CueBot pueda analizarlos.	Soporte para carga y procesamiento de archivos .pdf, .txt, y .html Conversión de contenido a texto plano y limpieza de datos antes del análisis.

Como docente, quiero realizar preguntas en lenguaje natural a CueBot para obtener respuestas rápidas y relevantes del acervo digital.	Motor de preguntas y respuestas utilizando un LLM conectado al acervo digital. Capacidad para contextualizar respuestas en función del documento consultado. Capacidad para contextualizar respuestas en función del documento consultado.
Como desarrolladora, quiero contar con una API documentada que me permita integrar CueBot a otras plataformas fácilmente.	Autenticación básica en los endpoints para uso controlado. Soporte para consultas asincrónicas (async) para mayor rendimiento.
Como estudiante, quiero acceder a una interfaz amigable para interactuar con CueBot sin necesidad de conocimientos técnicos.	Interfaz web en Gradio que permita probar la funcionalidad básica de CueBot. Opción para cargar archivos directamente desde la interfaz. Visualización del texto procesado y respuesta del LLM en pantalla.
Como administradora del sistema, quiero monitorear el desempeño del LLM para asegurar que funciona eficientemente y cumple con las expectativas.	Registro de logs de consultas y métricas de desempeño del modelo. Almacenamiento de logs en un archivo o base de datos para auditoría. Panel básico de administración para revisión de logs y estadísticas de uso. Mecanismo de reinicio o actualización manual del modelo en producción.

Lista Priorizada			
Requerimientos	**Etapas**	**Estimación de tiempo**	**Entregables**
Soporte para carga y procesamiento de archivos (.pdf, .txt, .html).	**Fase 1 - Ingesta**	2 semanas	Módulo de ingestión de archivos y normalización de texto.
Motor de preguntas y respuestas con LLM.	**Fase 1 - Núcleo LLM**	4 semanas	Integración de LLM con funciones de búsqueda y respuesta.
Desarrollo de API con FastAPI	**Fase 1 - API**	2 semanas	API funcional con *endpoints* de carga, consulta y documentación.
Interfaz web en Gradio.	**Fase 2 - UI de prueba**	2 semanas	Prototipo funcional para pruebas de interacción con CueBot.
Registro de logs y métricas de desempeño.	**Fase 2 - Monitoreo**	2 semanas	Sistema básico de monitoreo (*logging*, tiempos de respuesta).

Tabla 2. Lista de Requerimientos Priorizados

A continuación, se presenta la actualización de la Tabla 2 de acuerdo con lo solicitado en el Sprint 3.

Requerimiento	**Etapa**	**Estimación de Tiempo**	**Entregable**
Soporte para carga y procesamiento de archivos (.pdf, .txt, .html)	Fase 1 – Ingesta	2 semanas	Módulo de ingestión y normalización de archivos
Conversión de contenido a texto plano y limpieza/preprocesamiento de datos	Fase 1 – Ingesta	1 semana	Función de limpieza y preprocesamiento
Motor de preguntas y respuestas con LLM conectado al acervo digital	Fase 1 – Núcleo LLM	4 semanas	Sistema de Q&A funcional
Capacidad de contextualizar respuestas según documento consultado	Fase 1 – Núcleo LLM	1 semana	Ajuste semántico de respuestas
Desarrollo de API RESTful con FastAPI y autenticación básica en endpoints	Fase 1 – API	2 semanas	API documentada y endpoints seguros (Swagger UI incluido)
Soporte para consultas asincrónicas (async) para mayor rendimiento	Fase 1 – API	1 semana	Endpoints optimizados para concurrencia
Interfaz web en Gradio para pruebas funcionales	Fase 2 – UI de prueba	2 semanas	Prototipo funcional accesible
Carga de archivos desde la interfaz	Fase 2 – UI de prueba	1 semana	Función de carga de archivos en interfaz
Visualización del texto procesado y respuestas generadas por LLM	Fase 2 – UI de prueba	1 semana	Vista previa de análisis y respuestas
Registro de logs de consultas y métricas de desempeño	Fase 2 – Monitoreo	2 semanas	Sistema básico de observabilidad

Almacenamiento de logs en base de datos o archivos persistentes para auditoría	Fase 2 – Monitoreo	1 semana	Solución de registro persistente
Panel básico de administración para visualización de logs y estadísticas	Fase 2 – Monitoreo	1 semana	Interfaz de administración
Mecanismo de reinicio o actualización manual del modelo en producción	Fase 2 – Monitoreo	1 semana	Función de mantenimiento para el modelo