MNA Maestría en Inteligencia Artificial Aplicada

Materia: TC5035.10 Proyecto Integrador
(Gpo 10)

Prof. Titulares:
Dra. Grettel Barceló Alonso
Dr. Luis Eduardo Falcón Morales

Avance 6: Conclusiones clave

Inteligencia Artificial en la Normativa Académica

EQUIPO 2
Matthias Sibrian Illescas - A01794249
José Ramiro Adán Charles - A00174646
Genaro Ramos Higuera - A00351269

Análisis del modelo

Análisis de los resultados obtenidos

Cuestionamientos de rendimiento

- **¿El rendimiento del modelo es lo suficientemente bueno para su implementación en producción?**

Se identificó que el rendimiento del modelo cuenta con dos aspectos a evaluar: la coherencia de las respuestas, así como la similitud que existe entre las respuestas generadas por el modelo y las respuestas generadas por el área funcional para cada pregunta. En las fases iniciales, empezando en la Fase 0, se definió que el rendimiento del modelo estaría fuertemente atado con la capacidad de responder adecuadamente a las interrogantes que generarían las áreas funcionales que solicitaron la solución.

Cabe recalcar que el equipo de desarrollo, Equipo 2, coincidió, junto con las áreas funcionales, que las respuestas eran coherentes y útiles. Fue recalcado en varias ocasiones por ellos que estaban muy contentos con las respuestas. El segundo aspecto, la similitud, es fácilmente medible por una métrica, por lo que se escogió la métrica F1 Score con el fin de permitir que se penalice tanto un bajo rendimiento en precisión, así como en recall.

Dicho esto, el valor promedio F1 Score entre las preguntas fue de 0.6429. Considerando que esta métrica va de 0.0 a 1.0, este valor es suficientemente bueno para pasar a producción. Asimismo, las respuestas generadas son coherentes y muestran los documentos que utilizó para generar la respuesta, una de las necesidades de las áreas funcionales. Considerando que la métrica se basa enteramente en la similitud en el espacio de embeddings entre la respuesta del modelo y la del área funcional, el resultado es bueno.

Debido a la alta cantidad de información que se debe manejar y que se puede incluir en las respuestas, por la manera en que se plantearon, el resultado es bueno. Asimismo, se debe considerar que las áreas funcionales plantearon las preguntas y fue después que intentaron responderlas. Con un corpus de documentos tan amplio, es difícil para un humano responder tantas preguntas revisando todos los documentos, a diferencia de la dificultad que esto presenta para un modelo LLM.

- **¿Existe margen para mejorar aún más el rendimiento?**

Técnicamente, sí existe un factor para mejorar el rendimiento, pues teóricamente, el valor promedio de F1 Score para el cuerpo de preguntas y respuestas puede tomar un valor de 1.0. Esto implicaría que la similitud de las respuestas generadas por el modelo y por las áreas funcionales sean iguales. Sin embargo, este margen de mejoría se debe dar en un ajuste fino entre el tipo de preguntas que se hace para evaluar el modelo más que solo enfocarse en la métrica.

Considerando que las preguntas son bastante amplias, y el requisito de negocio necesario es construir una solución que apoye a las áreas funcionales para agilizar este proceso, la solución en su estado actual ya alcanza estos requisitos, lo que lo hace un producto valioso.

Dicho esto, si solo se enfoca en la métrica en vez de otras acciones, se estaría solamente haciendo que las respuestas del modelo se parezcan textualmente a las de los humanos. Aunque puede ser beneficioso, una mejoría más adecuada sería lograr un seguimiento y ajuste más fino en el tipo de preguntas, así como un cuerpo de preguntas más fáciles de responder por un humano. Esto permitiría reducir la cantidad de documentos con los cuales se pueden responder las preguntas.

Recomendaciones para la implementación

- **¿Cuáles serían las recomendaciones clave para poder implementar la solución?**

1. Continuar con el seguimiento junto con Sistemas de TI del Tec de Monterrey para empezar a definir cómo se implementará la solución y se cómo se dará acceso a las áreas funcionales.

2. Definir con Sistemas de TI y las áreas funcionales la manera en que se suplirá de nuevos documentos al modelo. Es claro para el equipo funcional que, para que el modelo siga siendo útil, deberán estar constantemente actualizando el cuerpo de documentos accesibles por el mismo.

3. Definir la cadencia del entrenamiento, ajuste y revisión de resultados del modelo, para poder garantizar su precisión en el tiempo. Debe definirse un modelo operativo para esta tarea. Esto podría realizarse por otras cohortes de estudiantes de la Maestría en Inteligencia Artificial Aplicada, así como por integrantes técnicos del área de Sistemas de TI del Tec de Monterrey.

Accionables: asignación hacia los stakeholders pertinentes

- **¿Qué tareas / procedimientos son accionables para las partes interesadas (stakeholders)?**

1. Para el área funcional:
    a. Una vez definido el proceso operativo por el que accederán al modelo, se debe empezar a utilizar en su día a día. Esto permitirá obtener más retroalimentación para el equipo de desarrollo correspondiente y podrá probar aún más su confiabilidad, desempeño y el grado de ayuda que les brinda.

    b. Se debe limitar su uso a preguntas que se pueden contestar con el cuerpo de documentos compartido con el equipo de desarrollo y con el que se entrenó el modelo. De momento, corresponde al cuerpo de documentos relacionados con el Reglamento Académico, que se alojaron en una carpeta de OneDrive. Si se deseara incrementar su alcance, se debe tomar en cuenta que debe existir un reentrenamiento y/o posible rediseño tras varios ciclos de experimentaciones.

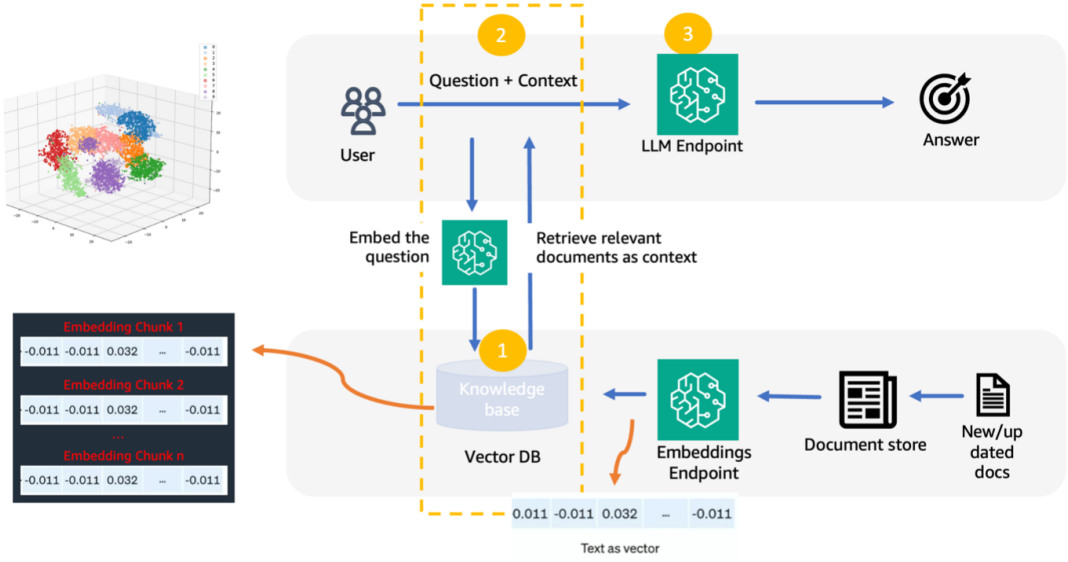
2. Para el área de TI:
    a. Se debe dar acceso a las áreas funcionales para que puedan hacer uso del modelo, tras una exitosa autenticación de su identidad. Aunque la solución maneja información que se ha hecho por el Tecnológico de Monterrey dominio público, las cargas computacionales que representan su utilización incurren en un costo, lo cual se debe considerar y limitar.

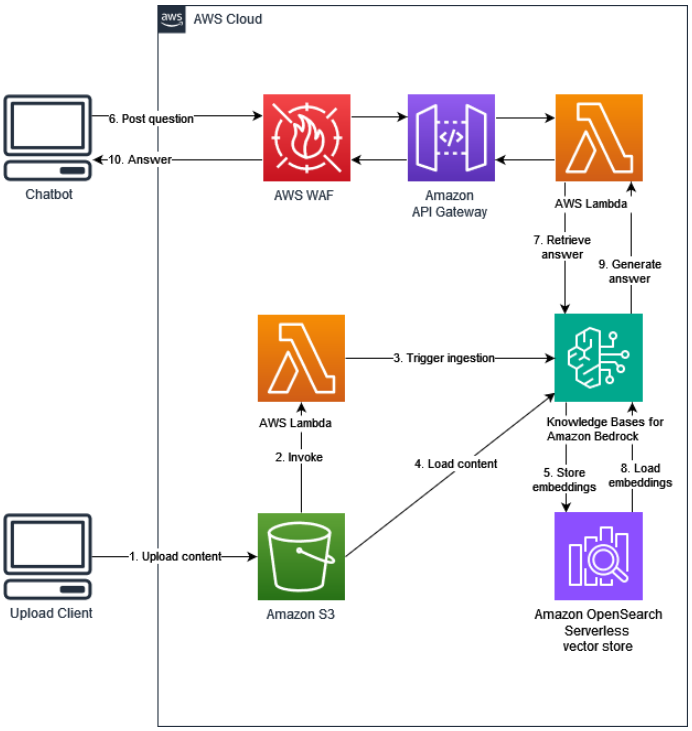
    b. Se debe dar acceso a los desarrolladores correspondientes para que se pueda reentrenar el modelo con una cadencia definida. A medida que esté disponible la retroalimentación futura proveniente de las áreas funcionales y peticiones para extender su alcance a más documentos, los equipos de desarrollo deben poder revisar la solución.

Análisis de proveedores para la implementación y selección

Proveedor 1 – Amazon Web Services (AWS)

Encontramos que Amazon AWS brinda las funciones necesarias para la arquitectura probada a lo largo del proyecto usando Haystack, muy similar a la siguiente imagen:





Ejemplo de interactividad.

Adicional a lo anterior, si queremos que los resultados sean muy similares a lo obtenido durante las diferentes etapas del proyecto, se buscarías usar el mismo LLM de Hugging Face, lo cual es viable usando las funciones descritas a continuación, de las cuales, presentamos la siguiente tabla con las características de los servicios requeridos, cabe destacar que por ser servicios de nube cumplen con las características de ser escalables y de que existe extensa documentación y ejemplos para su integración.

Servicio	Función	Costo
Amazon S3	Almacenamiento de Archivos PDF fuente.	0.023 por GB/mes
Amazon OpenSearch Service	Almacenamiento de los vectores indexados.	Instancia: t3.small.search (~$26/mes) Almacenamiento: 10 GB (~$3/mes)
Amazon Textract	Extracción e indexado de texto.	$0.003/página
Amazon SageMaker	Despliegue de Modelo LLM. Respuestas a preguntas en tiempo real.	ml.t3.medium (~$50/mes) $0.001 por inferencia
Amazon API Gateway	Enviar preguntas y recibir respuestas.	$1.08/mes para 6000 solicitudes (dependiendo de la región)

Basándonos en la arquitectura usada con Haystack como baseline el área funcional de Normatividad Académica nos proporcionó del tópico específico de Créditos Académicos:

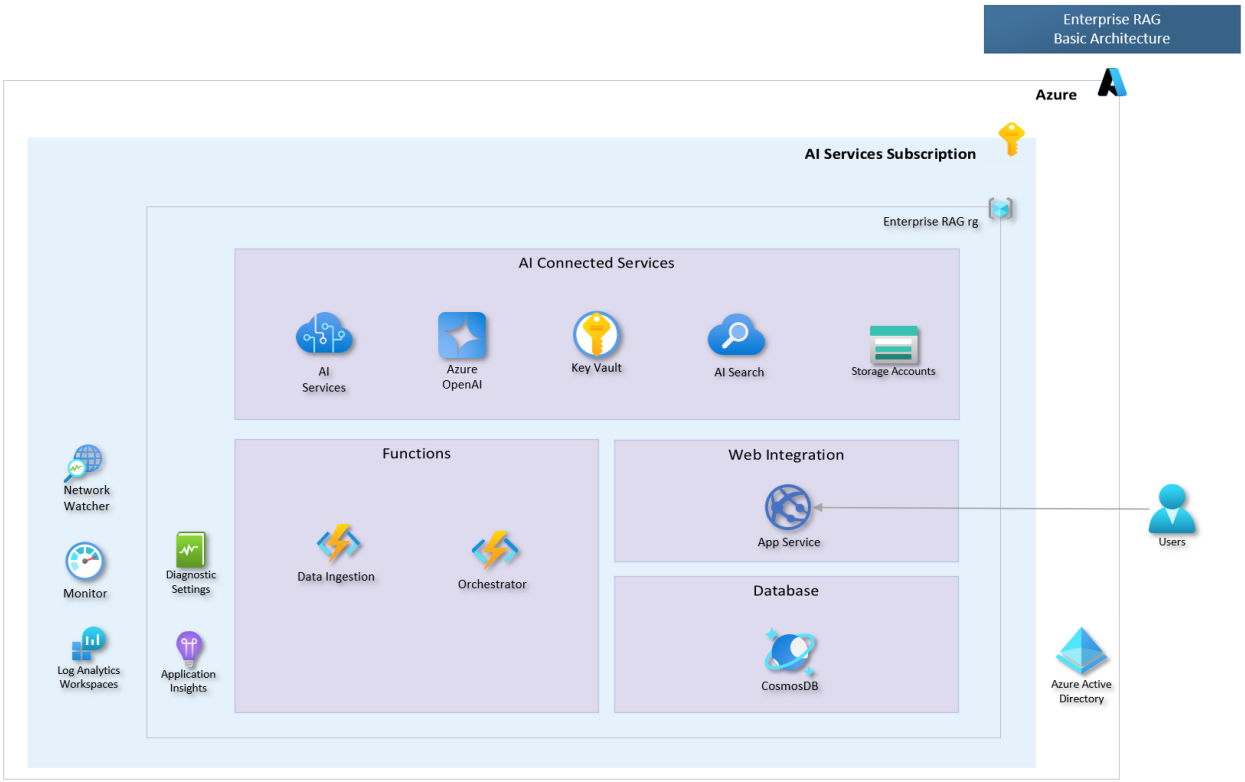
- 187 archivos en formato PDF ocupan 99 MB
- Un document store que contiene los embeddings de los 187 archivos usando transfomer Hugging Face genera 4765 documentos con un promedio de 0.013277 MB cada uno lo que da un total de 63 MB de tamaño del embedding.

Si a esto establecemos que se estima un uso promedio de 200 preguntas por día por usuario, podríamos estimar un costo mensual de la siguiente manera, tomando en cuenta que actualmente el área de normatividad académica del Tec de Monterrey tiene almacenados 1.5GB de documentos PDF con todos los tópicos y temas de la Normatividad Académica y Políticas y estimando que pudiese crecer en el largo plazo a 10GB, pensando que lo que se tiene almacenado es de 12 años a la fecha.

Servicio		**Costo Estimado Mensual (USD)**	**Notas**
Amazon S3	**1.5 GB**	**$0.0345**	**Cubriendo el requerimiento completo de Normatividad Académica y no solo Créditos.**

Amazon OpenSearch Service	**10.0 GB**	**$29.00**	**Cubriendo el requerimiento completo de Normatividad Académica y no solo Créditos.**
Amazon Textract	**187 documentos * 20 páginas c/u**	**$112.00**	**Asumiendo 10 extracciones o actualización de documentos por mes**
Amazon SageMaker	**200 preguntas/día * 30 días/mes = 6000 solicitudes/mes**	**$50 (instancia) + $6 (inferencias) = $56.00**	**Para un usuario. Para múltiples usuarios se multiplicaría cantidad de usuarios por el costo estimado.**
Amazon API Gateway	**6000 solicitudes/mes**	**$1.08**	**Para un usuario. Para múltiples usuarios se multiplicaría cantidad de usuarios por el costo estimado.**
TOTAL		**$198.11 - $1,981.10**	**Para 10 actualizaciones de documentos por mes y 200 consultas diarias. 1 - 10 usuarios**

Proveedor 2. Microsoft Azure Cloud Computing Services.



En el diagrama de arquitectura (Azure, n.d.) para implementar un sistema RAG (Retrieval Augmented Generation) en Azure, explicamos a continuación cada componente, así como anexamos una estimación de los costos asociados a la solución para Information Retrieval para las Políticas y Normativa del Tec de Monterrey:

1. Azure AI Services Subscription:

Esta es la base de nuestro entorno en Azure, proporcionando acceso a los servicios de inteligencia artificial necesarios para construir el sistema RAG.

2. Enterprise RAG Resource Group o el grupo que designe al área de TI de Inteligencia Artificial del Tec de Monterrey:

Este grupo de recursos organiza y contiene todos los servicios de Azure que componen la solución RAG, facilitando la gestión y el control de costos.

3. AI Connected Services:

- **AI Services:** Representa los modelos de lenguaje extenso (LLM) como es ChatGPT, que se utilizarán para generar respuestas coherentes y contextualmente relevantes.
- **Azure OpenAI:** Proporciona acceso a la API de OpenAI, permitiendo la integración con modelos como ChatGPT.
- **Key Vault:** Almacena de forma segura las claves de API y otros secretos necesarios para acceder a los servicios de Azure y OpenAI.
- **AI Search (Azure Cognitive Search):** Este servicio es fundamental para el RAG. Indexa y almacena los vectores (representaciones numéricas) de loos documentos PDF, lo que permite realizar búsquedas rápidas y eficientes basadas en similitud semántica.
- **Storage Accounts (Azure Blob Storage):** Aquí es donde se almacenarán los documentos PDF en su formato original.

4. Funciones (Azure Functions):

- **Data Ingestion:** Funciones serverless que se activan cuando se cargan nuevos documentos PDF en Blob Storage. Estas funciones procesan los documentos, extraen el texto y lo envían para su indexación.
- **Orchestrator:** Coordina el flujo de trabajo general, incluyendo la ingesta de datos, la indexación y la generación de respuestas.

5. Web Integration (Azure App Service):

- **App Service:** Alojará la aplicación web frontend, que interactuará con los usuarios, enviará consultas a ChatGPT y presentará las respuestas generadas.

6. Database (Azure Cosmos DB):

- **Cosmos DB:** Una base de datos NoSQL altamente escalable y flexible que puede almacenar datos estructurados y no estructurados, como registros de conversaciones, comentarios de usuarios y otros metadatos relevantes para la aplicación.

7. Monitor (Azure Monitor):

- **Diagnostic Settings:** Configura la recopilación de registros y métricas de los distintos servicios de Azure para su análisis y monitoreo.
- **Log Analytics Workspaces:** Almacena los registros recopilados por Azure Monitor, lo que permite realizar consultas y análisis para identificar problemas y optimizar el rendimiento.
- **Application Insights:** Proporciona información detallada sobre el rendimiento y el uso de la aplicación web, incluyendo tiempos de respuesta, tasas de error y seguimiento de dependencias.

8. Azure Active Directory:

Gestiona la autenticación y autorización de usuarios para acceder a la aplicación RAG.

9. Network Watcher:

Supervisa y diagnostica problemas de red en el entorno de Azure.

La implementación de Information Retrieval Augmented Generation (RAG) utilizando LLM con ChatGPT y vectorizacion Ada-002 con un corpus de documentos PDF, cuyo tamaño actual de corpus de PDFs es de 1.5GB y estamos suponiendo al momento un crecimiento de hasta 10GB e interacciones de usuario de 1Milllón de Tokens por mes nos lleva a estimar los siguientes costos:

Estimación de Costos (Mensuales):

(Utilizamos la Región Centro-Sur de los EE.UU., puede haber variaciones en otras regiones)

Servicio	Tamaño del Corpus (PDF)	Interacciones (Tokens/mes)	Costo Estimado Mensual (USD)	Notas
Azure Blob Storage	1.5 GB - 10 GB	N/A	$0.20 - $1.30	Depende del tipo de almacenamiento y la redundancia.
Azure AI Search	1.5 GB - 10 GB	1 millón	$100 - $600	Depende del nivel de servicio y la frecuencia de actualización del índice.

Azure OpenAI Service	N/A	1 millón	$10 - $75	Depende del modelo de ChatGPT utilizado y la cantidad de solicitudes.
Azure Functions	N/A	N/A	$1 - $10	Depende del plan de consumo y la cantidad de ejecuciones.
Azure Cosmos DB	N/A	N/A	$25 - $200	Depende del rendimiento y el almacenamiento utilizado.
Azure App Service	N/A	N/A	$55 - $220	Depende del plan de App Service y la cantidad de instancias.
Azure Monitor	N/A	N/A	$10 - $50	Depende de la cantidad de métricas recopiladas y alertas configuradas.
Log Analytics Workspace	N/A	N/A	$2 - $12	Depende del volumen de datos ingeridos.
Application Insights	N/A	N/A	$0 - $75	Depende del volumen de datos de telemetría.
Azure Key Vault	N/A	N/A	$0.93	Costo mínimo.
Total (Rango)			**$204.13 - $1,224.20**	**Importante:** Estos son rangos aproximados. Los costos reales pueden variar según la configuración y el uso.

Consideraciones Finales:

- **Azure Active Directory:** El costo puede variar según el plan de licenciamiento.
- **Network Watcher:** El costo suele ser mínimo aprox. $5-10 USD.

La optimización de costos es una prioridad para estos dos proveedores, con modelos de precios flexibles y herramientas de gestión de costos disponibles. AWS ofrece instancias reservadas y de pago por uso, junto con Cost Explorer y Budgets para monitoreo y administración. Azure proporciona instancias reservadas y de pago por uso, con administración de costos y facturación para controlar los costos, para ayudar a optimizar los costos.

En última instancia, la elección entre AWS y Azure para la implementación de RAG depende de los requisitos específicos de una organización, la infraestructura existente y la familiaridad con cada plataforma. La amplia oferta y la flexibilidad de AWS lo convierten en un fuerte competidor para las empresas que buscan personalización e integración con una amplia gama de servicios.

El énfasis de Azure en la seguridad y la perfecta integración con los productos de Microsoft puede resultar atractivo para las organizaciones que ya han invertido en el ecosistema de Microsoft.

En conclusión, si bien estos dos proveedores de nube ofrecen soluciones sólidas para implementar RAG, cada uno tiene sus propias fortalezas y puntos de venta únicos. Las organizaciones deben evaluar cuidadosamente sus necesidades específicas, requisitos de seguridad y restricciones presupuestarias al seleccionar la plataforma más adecuada para su implementación RAG.

Al aprovechar las potentes herramientas y servicios proporcionados por AWS o Azure, las empresas pueden aprovechar el potencial de Retrieval Augmented Generation para ofrecer respuestas precisas, contextualmente relevantes y oportunas a las consultas de los usuarios, impulsando en última instancia la innovación y mejorando las experiencias de los clientes.

Para el caso que estamos abordando la Dirección de TI, a través de la Gerencia de Inteligencia Artificial nos ha enfatizado el hecho de que ellos desean que se implemente en el futuro la aplicación en la nube de Microsoft Azure, dado que ellos tienen la inmensa mayoría (sino es que todos) de sus productos con licencias de Microsoft y por supuesto cuenta empresarial en Microsoft Azure Cloud Services y ya utilizan servicios de Azure Open AI.

Referencias

Build a contextual chatbot for financial services using Amazon SageMaker JumpStart, Llama 2 and Amazon OpenSearch Serverless with Vector Engine. https://aws.amazon.com/blogs/machine-learning/build-a-contextual-chatbot-for-financial-services-using-amazon-sagemaker-jumpstart-llama-2-and-amazon-opensearch-serverless-with-vector-engine/

Concept | Retrieval Augmented Generation (RAG) approach and the Embed recipe. https://knowledge.dataiku.com/latest/ml-analytics/gen-ai/concept-rag.html

Amazon Web Services. (s/f). AWS Pricing - Amazon S3. Recuperado de https://aws.amazon.com/es/s3/pricing/.

Amazon Web Services. (s/f). AWS Pricing - Amazon EC2. Recuperado de https://aws.amazon.com/ec2/pricing/.

Amazon Web Services. (s/f). AWS Pricing - Amazon SageMaker. Recuperado de https://aws.amazon.com/es/sagemaker/pricing/.

Amazon Web Services. (s/f). AWS Pricing - Amazon RDS. Recuperado de https://aws.amazon.com/es/rds/pricing/.

Amazon Web Services. (s/f). AWS Pricing - Amazon CloudWatch. Recuperado de https://aws.amazon.com/es/cloudwatch/pricing/.

Amazon Web Services. (s/f). AWS Pricing - Amazon API Gateway. Recuperado de https://aws.amazon.com/es/api-gateway/pricing/.

Amazon Web Services. (s/f). AWS Pricing - AWS Lambda. Recuperado de https://aws.amazon.com/es/lambda/pricing/.

Azure. (n.d.). *GitHub - Azure/GPT-RAG: Sharing the learning along the way we been gathering to enable Azure OpenAI at enterprise scale in a secure manner. GPT-RAG core is a Retrieval-Augmented Generation pattern running in Azure, using Azure Cognitive Search for retrieval and Azure OpenAI large language models to power ChatGPT-style and Q&A experiences.* GitHub. https://github.com/Azure/GPT-RAG

Microsoft Azure. (n.d.). Microsoft Azure Pricing Calculator. https://azure.microsoft.com/en-us/pricing/calculator/

Richards, D., & Richards, D. (2024, March 12). RAG in the Cloud: Comparing AWS, Azure, and GCP for Deploying Retrieval Augmented Generation Solutions – News from generation RAG. *News from generation RAG – Dive deep into the transformative world of AI Retrieval Augmented Generation (RAG) technologies*. https://ragaboutit.com/rag-in-the-cloud-comparing-aws-azure-and-gcp-for-deploying-retrieval-augmented-generation-solutions/