Tecnologías

INTRODUCCIÓN

La Universidad de los Andes, reconocida como una de las instituciones educativas líderes en América Latina, enfrenta un desafío estratégico en la era de la aceleración tecnológica: integrar la inteligencia artificial generativa (GenAI) de manera transversal en sus procesos académicos y administrativos.

En un contexto en el que universidades globales de referencia, han comenzado a implementar soluciones de GenAI para potenciar la investigación, la docencia y la gestión institucional, la Universidad de los Andes busca posicionarse como pionera en la región. Su objetivo es lanzar **Inteligencia Aumentada**, una iniciativa y plataforma institucional que permita a profesores, investigadores, estudiantes y personal administrativo desarrollar, probar y desplegar agentes de GenAI adaptados a las necesidades específicas de la comunidad universitaria.

Además de facilitar la labor académica, administrativa y de investigación, esta plataforma busca promover la colaboración multidisciplinaria en el diseño y aplicación de soluciones de inteligencia artificial generativa.

La iniciativa se enmarca en un contexto global marcado por la competitividad académica y la rápida evolución tecnológica, donde las instituciones de educación superior buscan optimizar recursos, mejorar la experiencia estudiantil y potenciar la producción de conocimiento. No obstante, la adopción de GenAI conlleva retos que abarcan la infraestructura tecnológica, la gestión del talento humano, el cumplimiento normativo y la aceptación cultural de la innovación.

Por ello, la Universidad de los Andes se enfoca en analizar estos desafíos y construir una estrategia integral para el diseño, desarrollo y despliegue de soluciones de inteligencia artificial generativa que generen un impacto positivo y sostenible en la institución.

EL PROBLEMA

El desafío inicial no está en el diseño de la plataforma **Inteligencia Aumentada**, sino en la selección estratégica del **Large Language Model (LLM)** y su modelo de aprovisionamiento, debido a su impacto financiero y técnico.

La Universidad de los Andes se enfrenta al reto de elegir y aprovisionar un modelo de LLMs que sea financieramente viable y eficiente en términos de recursos institucionales. Por un lado, ecosistemas comerciales como el de OpenAI ofrecen soluciones sólidas. Sin embargo, su costo de USD 20 por usuario al mes escalaría

rápidamente, alcanzando aproximadamente USD 6 millones anuales (una comunidad de 25,000 usuarios activos), una cifra considerable para la institución.

No obstante, el acceso a estos servicios mediante APIs ofrece un modelo más flexible y potencialmente rentable, basado en pago por consumo de tokens. Aunque las APIs externas reducen la carga técnica inicial, también introducen dependencia de terceros, riesgos en la soberanía de los datos y costos variables por token, los cuales podrían incrementarse significativamente en aplicaciones con alto consumo, como la investigación.

En contraste, desplegar un LLM de código abierto (e.g., Llama 3, Falcon, DeepSeek) en infraestructura local (on-premise) implicaría mayores exigencias en infraestructura y soporte, aunque con la posibilidad de reducir costos recurrentes de suscripción.

Ante este escenario, la Universidad debe evaluar en detalle las implicaciones económicas y los requerimientos técnicos de ambas alternativas, con miras a garantizar un uso responsable y sostenible de la inteligencia artificial generativa.

La Universidad debe realizar un análisis detallado que contemple no solo los costos directos, sino también factores clave como la latencia, la capacidad de personalización del modelo, el cumplimiento de la Ley 1581 de protección de datos en Colombia y la escalabilidad futura. Esto permitirá definir si la autonomía tecnológica justifica una inversión inicial sustancial o si, en contraste, la flexibilidad del pago por uso - pese a sus riesgos a largo plazo - es la opción más viable dentro de sus capacidades actuales.

ACTIVIDADES

Paso 1 – Solución local

Determinar la capacidad y los recursos tecnológicos necesarios para implementar y operar el modelo de código abierto asignado a su equipo de trabajo, a nivel local (on-premise), evaluando en detalle los componentes de hardware, software y soporte necesarios para garantizar un desempeño eficiente.

Equipo de trabajo	Modelo	Tamaño
Equipo 1, 7	deepseek-r1	70B de parámetros
Equipo 2, 8	deepseek-r1	671B de parámetros

Equipo 3, 9	Llama 3.3	70B de parámetros
Equipo 4, 10	mistral-large	123B de parámetros
Equipo 5, 11	Qwen2.5-Max	de parámetros
Equipo 6, 12	deepseek-v3	671B de parámetros

En este ejercicio, se debe estimar el costo de la infraestructura de TI, incluyendo servidores de alto rendimiento con GPU. Asimismo, se debe proyectar el Capex anual de la inversión, considerando los gastos de adquisición, instalación, mantenimiento y escalabilidad de la solución, para obtener una visión integral del presupuesto necesario para la implementación exitosa del modelo en la Universidad de los Andes.

Tenga en cuenta las siguientes consideraciones:

- Incluir servidores y su almacenamiento de alta velocidad.
- Determinar el número de GPUs (ej.: NVIDIA A100/H100), memoria VRAM y ancho de banda de red, considerando su tamaño y la carga esperada (usuarios concurrentes, tokens generados por segundo).
- Calcular el consumo energético de esta infraestructura (ej.: 300W por GPU × 24/7 × tarifa industrial colombiana).

Nota: No es necesario que estime sistemas de almacenamiento adicionales, redes de alta velocidad y demás elementos de soporte técnico. Dado que la universidad cuenta con un centro de datos solido en estos aspectos.

PASO 2 – SOLUCIONES BASADAS EN APIS EN LA NUBE

Con el presupuesto estimado en el paso anterior, se debe proyectar la cantidad de tokens que podrían adquirirse a través de la API de OpenAI, tanto para el modelo GPT-4o como para un modelo de razonamiento avanzado, con el fin de comparar la inversión requerida frente a la implementación de infraestructura on-premise.

Estos tokens deben distribuirse a lo largo de un periodo de tres años, plazo definido para la amortización de los componentes de TI estimados en el paso anterior. De este modo, será posible realizar un análisis comparativo del costo por uso de la API frente a la inversión inicial y los gastos de mantenimiento asociados al despliegue local de la plataforma.

Nota: Debe comprender el modelo de costos del API, aspectos como el costo de los mensajes de entrada y el costo de los mensajes de salida, entre otros.

PASO 3 – COMPARACIÓN ENTRE LOS DOS ESCENARIOS

La Facultad de Ingeniería planea un piloto de seis meses con 120 agentes de inteligencia artificial generativa (GenAI), beneficiando a al menos 4000 estudiantes, con un crecimiento estimado del 15 % para el semestre siguiente. Según datos de pruebas preliminares, cada estudiante consume aproximadamente 40,000 tokens diarios en su interacción con un agente, y se proyecta que cada agente sea utilizado de manera intensiva durante un período de siete días en aproximadamente 4 momentos del semestre.

Adicional a esto se esperan desplegar aproximadamente 30 agentes para el personal administrativo de la facultad, alrededor de unas 100 personas. Estos agentes apoyaran a diario los procesos que realiza el personal administrativo y se estima que se utilicen de forma continua los días laborales del mes, en la jornada laboral. El equipo proyecta que cada usuaria utilizara alrededor de 3 agentes por día con un pico máximo de consumo de unos 50000 tokens.

Con base en esta información, y tomando como referencia las estimaciones de costos para el uso de APIs de OpenAI definidas en el paso 2, calcule el costo total de este ejercicio para un periodo de un año, considerando tanto el aumento en la cantidad de estudiantes como el número de agentes a desplegar.

PASO 4 – CONCLUSIONES Y CIERRE

Analizar los hallazgos obtenidos a partir del ejercicio de estimación de costos y comparar las diferentes alternativas de aprovisionamiento de Large Language Models (LLMs), con el fin de formular recomendaciones estratégicas para la Universidad de los Andes.

Instrucciones:

- Síntesis de hallazgos.
- Destaque las diferencias clave entre la opción on-premise y el uso de APIs de OpenAI en términos de costos, escalabilidad, soberanía de datos y mantenimiento.

Propuesta de recomendaciones:

- A partir de los hallazgos, elabore al menos tres recomendaciones para la Universidad de los Andes en relación con la estrategia de implementación de GenAI.
- Justifique cada recomendación con base en los datos obtenidos, considerando aspectos financieros, técnicos y de sostenibilidad a largo plazo.

Presentación:

- Comparta sus conclusiones y recomendaciones en una breve presentación

ENTREGA

Suba su documento con el análisis, conclusiones y recomendaciones para su evaluación y discusión.