Instituto Tecnológico y de Estudios Superiores de Monterrey

Maestría en Inteligencia Artificial:

Spin Voyager & Spin Compass: Avance 3. Baseline

Tania Alicia Caballero Saavedra - A01795957

Oscar Enrique García García - A01016093

Dante Rosas Fragoso – A01795850

Profesor Titular: Dra. Grettel Barceló Alonso

Asesor: Dr. Horacio Martínez Alfaro

Proyecto Integrador

Febrero, 2026

Índice

1. Introducción

A diferencia de los proyectos de **Machine Learning** tradicionales, donde el éxito se mide mediante la precisión predictiva, **Spin Compass** enfrenta un desafío de Integridad de la Información. El riesgo principal no es solo una respuesta incorrecta, sino la «alucinación» o la recuperación de una política obsoleta o fragmentada. Por tanto, se redefine el éxito del *baseline* alejándose de métricas genéricas para centrarse en la **Evaluación de Recuperación** (Retrieval Evaluation). No basta con que el bot responda con elocuencia; el sistema debe demostrar que es capaz de recuperar el «**fragmento de verdad**» (Gold Chunk) dentro del ecosistema documental de Spin.

2. Definición del Algoritmo (Baseline)

¿Qué algoritmo se puede utilizar como baseline para predecir las variables objetivo? Para Spin Compass, el algoritmo base no es un modelo predictivo tradicional, sino una arquitectura de recuperación estándar (Naive RAG):

1. Segmentación (Splitting): Se realizarán pruebas de comparación entre Max-Min Semantic Chunking y Growing Window Semantic Chunking.
2. Recuperación (Retrieval): Búsqueda vectorial simple (Dense Retrieval) utilizando similitud de coseno.
3. Generación: Debido a cuestiones internas de Spin, se realizarán las pruebas con modelos de LLM de Google **Gemini** (gemini-2.0-flash, gemini-2.5-flash, gemini-2.5-pro, y gemini-flash-lite-latest).

Justificación: Este enfoque permite medir si las políticas de Spin son recuperables «tal cual están» en los documentos oficiales después de aplicar limpieza a los documentos.

3. Importancia de las Características

¿Se puede determinar la importancia de las características para el modelo generado?

En el marco de arquitecturas **RAG (Retrieval-Augmented Generation)**, las «características» fundamentales son los fragmentos de texto (**chunks**) extraídos de los documentos oficiales. La inclusión de características irrelevantes afecta negativamente el rendimiento del modelo y aumenta la complejidad sin beneficios sustanciales, provocando alucinaciones.

- **Ruido en el Baseline:** El baseline probablemente recuperará «basura» (encabezados, pies de página, disclaimers legales) debido al corte fijo.
- **Integridad Semántica:** El análisis debe determinar si el chunk contiene la regla completa. Si el baseline corta una tabla de viáticos a la mitad, la característica pierde su valor predictivo.

Las pruebas realizadas dentro del notebook desarrollado demuestran que el límite del fragmento es la característica más crítica. El método **Max-Min** superó al método **Growing Window**, ya que mantiene la «regla de negocio» completa (e.g., una tabla de viáticos o una restricción) dentro del mismo bloque.

Por otro lado, al realizar una limpieza con `NoiseReducer`, se eliminaron características irrelevantes como números de página y encabezados. Con esto, se evita que el motor de búsqueda vectorial se distrayera y se logre centrar en el contenido normativo.

Característica	Impacto en el modelo	Evidencia en los resultados
Cohesión	Alta	El 73% de los chunks con el método Max-Min contiene la información completa, evitando «cortar» las políticas.
Concisión (chunks)	Crítica	Max-Min es más eficiente con un promedio de 84.2 tokens, facilitando al LLM procesar el contexto.
Relevancia (Hit Rate)	Determinante	La capacidad de capturar la intención del usuario subió al 93.3% con una mejor segmentación.

Tabla 1: Atributos de los chunks «ganadores»

4. Diagnóstico de Ajuste

¿El modelo está sub/sobreajustando los datos de entrenamiento?

- **Subajuste (Fallo de Recuperación):** Si el baseline tiene un Hit Rate bajo, indica que la estrategia de búsqueda no encuentra los documentos, posiblemente por la brecha entre el lenguaje natural del usuario y la terminología oficial.
- **Sobreajuste (Alucinación):** Si el modelo ignora el contexto recuperado y responde con conocimientos generales pre-entrenados, está «sobreajustando» a su entrenamiento original en lugar de adherirse a las políticas.

Se identificó un riesgo de subajuste principalmente en el método **Growing Window**. Con un Hit Rate de 50%, este método no logra adherirse a la base de conocimientos. Adicionalmente, las ventanas de texto demasiado grandes diluyen la semántica, haciendo que la búsqueda vectorial no identifique el fragmento exacto.

Por otro lado, respecto al sobreajuste, modelos «pesados» como Gemini 2.5 Pro muestran una latencia de 6.9 segundos. Aunque son potentes, este exceso de capacidad resulta en un desperdicio de recursos y una mala experiencia de usuario para tareas de consulta simple. Adicionalmente, si el MRR es bajo, el modelo puede recibir fragmentos irrelevantes en las primeras posiciones y derivar en alucinaciones.

Como se detalla en este documento, la combinación de Max-Min Semantic Chunking con el modelo Gemini 2.0 Flash resultó en el punto óptimo:

- **Alta Fidelidad:** Un MRR de 0.78 asegura que el contexto correcto sea lo primero que lee el modelo.
- **Eficiencia:** Una latencia de 0.53 segundos garantiza respuestas casi inmediatas en Slack.

5. Métricas de Desempeño

Para Spin Compass, el éxito del asistente no se define por la elocuencia de su lenguaje, sino por su capacidad para localizar la política exacta dentro del ecosistema documental de Spin (Google Drive, Slack, etc.). Una falla en la recuperación deriva inevitablemente en alucinaciones o respuestas negativas («No sé»), resultados inaceptables para el área de Finanzas. Por tanto, se adopta un marco de evaluación de dos niveles: **Evaluación de Recuperación (Retrieval)** y **Evaluación de Generación**.

5.1. Evaluación de Recuperación (Retrieval Evaluation)

La eficacia del sistema depende de la identificación precisa del documento fuente que contiene la «verdad» documental.

- **Hit Rate ($HR@K$)**:
  - *Definición*: Mide la frecuencia con la que el fragmento correcto aparece dentro de los primeros K resultados recuperados.
  - *Aplicación*: Permite determinar si documentos críticos, como *Política_Viajes_2025.pdf*, son identificados correctamente ante consultas específicas.
  - *Meta*: Se establece un objetivo de $HR@5 > 0.80$ para asegurar la disponibilidad del contexto necesario para el LLM.
- **MRR (Mean Reciprocal Rank)**:
  - *Definición*: Evalúa la posición relativa del primer resultado relevante, calculado como el promedio de los inversos de las posiciones ($\frac{1}{\text{rank}}$).
  - *Relevancia*: Un MRR elevado indica una comprensión precisa de la intención del usuario, reduciendo el riesgo de confusión por fragmentos irrelevantes previos.
  - *Impacto del Chunking*: El uso de estrategias semánticas como **Max-Min** mejora significativamente esta métrica al establecer límites lógicos claros en los fragmentos.
- **Métricas Semánticas (SRA y RCC)**:
  - **Same Relevant Article (SRA)**: Mide si los fragmentos recuperados pertenecen al mismo documento oficial que contiene la respuesta, evitando la mezcla de políticas incoherentes.
  - **Relevant Chunk Contained (RCC)**: Evalúa si el fragmento recuperado contiene la unidad de información completa. El **Semantic Chunking** busca mitigar el «corte» de oraciones presente en métodos de tamaño fijo, elevando la integridad del bloque informativo.

5.2. Evaluación de Generación (Generation Evaluation)

Una vez garantizada la recuperación, se procede a evaluar la respuesta final dirigida al colaborador.

- **Faithfulness (Fidelidad)**: Verifica que la respuesta se derive exclusivamente del contexto recuperado, empleando un LLM juez para detectar posibles alucinaciones.

- **Correctness (Exactitud) - Score@K**: Evaluación basada en un **Golden Dataset** de preguntas sintéticas, asignando puntajes según la presencia de información clave y la prontitud de la respuesta. Investigaciones indican que un chunking semántico superior puede incrementar este puntaje entre un 2% y 4%.

5.3. Estrategia de Implementación de Métricas

El proceso de evaluación se automatiza mediante la metodología de **Synthetic QA**:

1. **Generación de Datos Sintéticos**: Empleo de un LLM para extraer pares de Pregunta-Respuesta y fragmentos fuente de las políticas de Spin.
2. **Mapping de Recuperación**: Verificación de la coincidencia entre el chunk recuperado y la fuente original para el cálculo de HR y RCC.
3. **Benchmarking**: Comparación del desempeño entre el baseline y estrategias avanzadas para justificar técnicamente la reducción de ruido y mejora de integridad.

6. Desempeño Mínimo a Obtener

Para considerar el paso a producción o la iteración hacia arquitecturas de mayor complejidad, el sistema debe satisfacer los siguientes criterios técnicos fundamentales:

1. **Hit Rate > Azar**: El modelo debe demostrar una capacidad de recuperación que supere sustancialmente la probabilidad aleatoria. En un ecosistema de recuperación configurado para 10 chunks, el HR debe mantenerse consistente (ej. $> 70\%$ para consultas de preguntas frecuentes).
2. **RCC vs. Longitud**: El baseline debe exhibir un equilibrio óptimo entre la integridad de la respuesta y la eficiencia operativa. Si la obtención de un RCC elevado requiere la recuperación de fragmentos excesivamente extensos, el modelo incurrirá en costos computacionales y latencias inaceptables. Por tanto, el desempeño mínimo aceptable se define como la capacidad de lograr un RCC alto manteniendo chunks concisos (inferiores a 512 tokens).
3. **Latencia**: Con el objetivo de garantizar la fluidez de la experiencia del usuario en Slack, el tiempo total de procesamiento —incluyendo las fases de recuperación y generación— no debe exceder el rango de 5 a 8 segundos.

7. Análisis de Resultados

Tras la ejecución del baseline y la comparación de las estrategias de segmentación semántica, se presentan los hallazgos críticos divididos en tres dimensiones: eficacia de recuperación, integridad de la información y rendimiento generativo.

7.1. Comparativa de Estrategias de Recuperación (Retrieval)

Los experimentos demuestran una superioridad clara del algoritmo **Max-Min Semantic Chunking** sobre la estrategia de **Growing Window** en todas las métricas de recuperación evaluadas para el ecosistema documental de Spin:

- **Precisión de Búsqueda (Hit Rate@5):** El método **Max-Min** logró que el fragmento correcto apareciera en el Top-5 el 93.3% de las veces, superando drásticamente el 50.0% obtenido por **Growing Window**. Esto indica que la lógica de similitud mínima interna del Max-Min captura mejor la intención de las consultas de los usuarios.
- **Posicionamiento (MRR):** Con un MRR de 0.78, el sistema Max-Min tiende a colocar la respuesta correcta en la primera o segunda posición de los resultados. En contraste, el 0.50 de Growing Window sugiere que la respuesta aparece con menor relevancia, aumentando el riesgo de que el LLM procese ruido irrelevante.
- **Autoridad de la Fuente (SRA):** Ambos métodos demostraron una alta capacidad para identificar el documento correcto (100% para Max-Min y 90% para Growing Window), garantizando que la información proviene de la política oficial pertinente.

7.2. Integridad de la Información (RCC) y Eficiencia de Segmentación

La métrica **Relevant Chunk Contained (RCC)** es vital para el cumplimiento normativo en el área de Finanzas, ya que mide si la regla de negocio se recuperó íntegra.

- **Integridad Semántica:** El método **Max-Min** alcanzó un RCC de 0.73, lo que significa que en el 73% de los casos, la regla de negocio se mantuvo cohesiva dentro del fragmento. **Growing Window** solo

logró un 0.50, implicando que la mitad de sus respuestas podrían estar incompletas debido a cortes arbitrarios.

- **Control de Extensión (Tokens):** Mientras que el informe establece un límite deseable de 512 tokens por chunk , los datos muestran que **Growing Window** excede este límite con un promedio de 596.5 tokens y picos de hasta $4,336$. **Max-Min** es más eficiente con un promedio de 84.2 tokens, aunque registró un máximo de $1,301$ que requiere optimización.

7.3. Benchmark de Modelos Generativos

Se evaluaron cuatro variantes del modelo Gemini para determinar el equilibrio óptimo entre precisión y fluidez conversacional:

Modelo	Latencia (s)	Similitud Sem.	Longitud (tokens)
Gemini 2.0 Flash	0.53	0.648	19.8
Gemini 2.5 Flash	1.76	0.614	25.3
Gemini Flash Lite	0.40	0.550	20.3
Gemini 2.5 Pro	6.90	0.520	17.4

- **Rendimiento en Tiempo Real:** Para cumplir con el requisito de latencia menor a 8 segundos en Slack , los modelos **Flash** y **Flash Lite** son las únicas opciones viables. El modelo **Pro** consume casi la totalidad del tiempo permitido (6.9 s) solo en generación.
- **Calidad de Respuesta:** El modelo **Gemini 2.0 Flash** presenta el mejor balance, obteniendo la mayor similitud semántica (0.648) con una latencia extremadamente baja (0.53 s), minimizando el riesgo de alucinaciones.

7.4. Conclusión del Análisis

El baseline confirma que la segmentación semántica avanzada es necesaria para Spin Compass. Se recomienda adoptar la arquitectura basada en **Max-Min Semantic Chunking** y el motor **Gemini 2.0 Flash**, dado que esta combinación maximiza la recuperación de «fragmentos de verdad» ($HR = 93.3\%$) y garantiza la integridad de las políticas financieras.

8. Referencias

[1] L. Visengeriyeva, A. Kammer, I. Bär, A. Kniesz, y M. Plöd, «CRISP-ML(Q): The ML lifecycle process». Accedido: 22 de enero de 2026. [En línea]. Disponible en: https://ml-ops.org/content/crisp-ml

[2] R. Hernández-Sampieri y C. Mendoza, *Metodología de la investigación: Las rutas de la investigación cuantitativa, cualitativa y mixta*, 3.ª ed. McGraw-Hill, 2023.

[3] A. Moreno-Cediel, E. Garcia-Lopez, D. De-Fitero-Dominguez, y A. Garcia-Cabot, «Optimising retrieval performance in RAG systems: A new growing window semantic chunking strategy to address weak semantic boundaries». Alcalá de Henares, Madrid, Spain, 2025.

[4] C. Kiss, M. Nagy, y P. Szilagyi, «Max–Min semantic chunking of documents for RAG application». Budapest, Hungary, 2025.