Instituto Tecnológico y de Estudios Superiores de Monterrey

Maestría en Inteligencia Artificial:

Spin Voyager & Spin Compass: Avance 2. Ingeniería de características

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 1
   1.1. Situación Actual y Problemática Operativa (Pain Points) 1
2. Spin Compass y el Desafío de los Datos 2
   2.1. La Insuficiencia del «Chunking» Estático (Fixed-Size Splitting) 3
   2.2. Limitaciones de los Métodos Semánticos de Primera Generación (Context Drift) 3
   2.3. El Riesgo de Alucinaciones en Tareas Intensivas de Conocimiento 4
3. Estado del Arte: Nuevos Algoritmos de Segmentación Semántica (2025) 4
   3.1. Algoritmo de Segmentación *"Max-Min"* (Clustering Dinámico) 5
   3.2. Estrategia de Ventana Creciente (*Growing Window Strategy*) 6
   3.3. Comparativa de Rendimiento Esperado 7
4. Resultados de los notebooks de colab con Segmentación Semántica (Semantic Chunking) 8
   4.1. Análisis del algoritmo Max-Min Semantic Chunking 8
   4.2. Análisis del algoritmo Growing Window Semantic Chunking 9
   4.3. Comparativa de Estrategias de Chunking 10
5. Conclusiones 11
6. Referencias 12

1. Introducción

El presente documento define la estrategia de ingeniería de datos necesaria para maximizar la precisión de *Spin Compass*, el asistente conversacional diseñado para atender a los colaboradores («*Spinners*») en temas de gastos y políticas internas.

Actualmente, el área de **Finanzas de Spin** enfrenta una saturación operativa crítica debido a la gestión manual de validaciones y la dispersión de la información oficial en múltiples fuentes (*Google Drive*, *Canvas*, *La Órbita*). Para que la arquitectura RAG (*Retrieval-Augmented Generation*) propuesta funcione eficazmente, es imperativo que el sistema recupere respuestas precisas y completas, evitando alucinaciones que podrían derivar en incumplimientos normativos.

La calidad de las respuestas de *Spin Compass* depende intrínsecamente de cómo se procesan y fragmentan (*chunking*) los documentos oficiales. Si los datos de entrada son ruidosos o están mal segmentados, las respuestas del agente serán deficientes («*Garbage In, Garbage Out*»).

1.1. Situación Actual y Problemática Operativa (Pain Points)

La organización Spin enfrenta un desafío de escalabilidad en su área de Finanzas debido al crecimiento de su plantilla y volumen de operaciones. Actualmente, la gestión de viáticos y la resolución de dudas operan bajo un modelo manual insostenible:

Saturación del Canal de Soporte: El equipo de Finanzas recibe consultas constantes de los colaboradores («*Spinners*») a través de canales informales como **mensajes directos de Slack, correos y llamadas**. Esto genera interrupciones continuas, tiempos de respuesta variables y una alta dependencia del conocimiento *"tribal"* de ciertos empleados.

Dispersión de la Información: La ***"verdad oficial"*** sobre **políticas de viajes, montos de viáticos y lineamientos de hardware** no reside en un solo lugar. Está fragmentada en

múltiples repositorios como Google Drive, Canvas y La Órbita (*plataforma web interna*). Esta fragmentación provoca que los Spinners consuman información desactualizada o versiones contradictorias de los documentos.

Gestión de Gastos Manual: La validación de los reportes de gastos se realiza *"registro por registro"*, contrastando manualmente los recibos contra las políticas en PDF, lo que eleva el riesgo de error humano y la fricción operativa.

2. Spin Compass y el Desafío de los Datos

Para resolver esta problemática, se desarrollará Spin Compass, un asistente conversacional integrado en Slack. Su misión es democratizar el acceso a la información, permitiendo que un Spinner pregunte: «¿Cuánto puedo gastar en una cena con cliente?» y reciba una respuesta inmediata y normativa. Sin embargo, implementar un LLM (Modelo de Lenguaje Grande) por sí solo no es suficiente. Los modelos genéricos sufren de «alucinaciones» cuando no tienen acceso a datos privados y recientes. Por ello, se adopta una arquitectura RAG, que permite al agente recuperar información precisa de los manuales de Spin antes de responder. El Desafío Crítico: La eficacia de Spin Compass depende enteramente de la calidad de los datos que ingesta. Bajo el principio de «Garbage In, Garbage Out» (Basura entra, basura sale), si alimentamos al sistema con PDFs desordenados, tablas rotas o texto con «ruido» (encabezados repetitivos, disclaimers legales), el agente será incapaz de recuperar la respuesta correcta. Además, investigaciones recientes de 2025 señalan que las estrategias tradicionales de segmentación de texto (Standard Chunking) fallan en documentos financieros complejos, rompiendo la coherencia semántica entre una regla y sus excepciones. Por tanto, este proyecto requiere una estrategia de datos avanzada que trate a los documentos como activos críticos, optimizándolos para que sean legibles no solo por humanos, sino por la Inteligencia Artificial.

2.1. La Insuficiencia del «Chunking» Estático (Fixed-Size Splitting)

La práctica estándar en RAG es dividir el texto por un número fijo de caracteres (*ej. 500 tokens*). **Investigaciones recientes de 2025** demuestran que este método es ***"insensible a la estructura semántica"***, lo que provoca cortes arbitrarios que destruyen el significado.

Ruptura de Contexto: Al cortar estrictamente por tamaño, es común que el encabezado de una sección (*ej. «Gastos NO Deducibles»*) quede en un fragmento (chunk) y la lista de artículos prohibidos quede en el siguiente. Si un usuario pregunta *"¿Puedo comprar alcohol?"*, el sistema recuperará la **lista**, pero no el encabezado que prohíbe la compra, **induciendo al error**.

Pérdida de Jerarquía: Las tablas de límites de viáticos, si se procesan como texto plano y se cortan a la mitad, pierden la relación entre la **fila (*Concepto*)** y la **columna (*Monto*)**, generando respuestas incoherentes.

2.2. Limitaciones de los Métodos Semánticos de Primera Generación (Context Drift)

Aunque el **«Chunking Semántico»** (*agrupar oraciones por similitud*) es superior al estático, los métodos populares hasta 2024 (*como el enfoque de ventana deslizante de Kamradt*) presentan fallas críticas para manuales extensos como los de Spin.

El Problema de la *"Deriva Semántica"*: Según **Moreno-Cediel et al. (2025)**, estos métodos comparan una oración solo con sus vecinas inmediatas mediante una **ventana deslizante**. Esto provoca que el contexto inicial del párrafo se "olvide" a medida que la ventana avanza.

Caso Práctico en Spin: En una política larga sobre *"Uso de Tarjeta Corporativa"*, las oraciones finales pueden parecer semánticamente distantes del título inicial. El método de **Kamrad** podría agrupar incorrectamente las reglas finales de la tarjeta corporativa con

el inicio de la siguiente sección (*ej. "Reembolsos en Efectivo"*) simplemente porque usan vocabulario similar de gastos, creando ***"chunks híbridos"*** que confunden al modelo.

2.3. El Riesgo de Alucinaciones en Tareas Intensivas de Conocimiento

Los Modelos de Lenguaje Grandes (LLMs) son propensos a alucinar cuando se enfrentan a lagunas de información o contextos irrelevantes (*"noise"*).

Ruido en la Recuperación: Si los documentos de Spin no se limpian de encabezados repetitivos, disclaimers legales y pies de página antes de la segmentación, estos elementos actúan como distractores. El modelo puede recuperar un fragmento porque coincide con el texto del pie de página, ignorando que el contenido real no es relevante para la pregunta del usuario.

Consecuencia Operativa: Para **Spin Compass**, entregar una respuesta incorrecta sobre una política de gastos es más grave que no responder. Por tanto, la estrategia de segmentación debe garantizar la **Coherencia Semántica Total**: cada fragmento recuperado debe ser una unidad de verdad completa y autosuficiente, capaz de sostenerse por sí misma sin depender de fragmentos anteriores o posteriores.

3. Estado del Arte: Nuevos Algoritmos de Segmentación Semántica (2025)

nte la **insuficiencia de los métodos de segmentación estática** para documentos con alta densidad normativa como los de **Spin**, es necesario adoptar estrategias de "Chunking Semántico" de segunda generación. Mientras que las estrategias de 2023–2024 (*como la ventana deslizante de Kamradt*) supusieron un avance, investigaciones publicadas en 2025 han identificado fallas críticas en su manejo del contexto a largo plazo.

Para Spin Compass, implementaremos y evaluaremos dos algoritmos de vanguardia diseñados para maximizar la coherencia y la recuperación de información (IR).

3.1. Algoritmo de Segmentación *"Max-Min"* (Clustering Dinámico)

Propuesto por Kiss et al. (2025), este método redefine el problema del **chunking** no como un corte de texto, sino como un problema de agrupamiento dinámico (*clustering*).

• **El Mecanismo**: A diferencia de cortar texto cuando se alcanza un límite de tokens, el algoritmo Max-Min procesa el documento oración por oración. Para decidir si una nueva oración (s_k) pertenece al fragmento actual (C), el algoritmo compara dos valores:

1. **Coherencia Interna (*min_sim*)**: La similitud mínima existente entre las oraciones que ya están dentro del fragmento actual.
2. **Afinidad de la Candidata (*max_sim*)**: La similitud máxima entre la nueva oración y cualquiera de las oraciones del fragmento.

• **La Lógica de Decisión**: Si la nueva oración tiene una afinidad mayor que la coherencia mínima del grupo, se agrega. Si no, se crea un corte. Esto asegura que no se añadan oraciones "intrusas" que diluyan el tema central.

• **Ventaja para Spin**: Este algoritmo ha demostrado superar estadísticamente a los métodos anteriores en preguntas de dificultad *"Alta" (Hard Questions)*. Esto es vital para las políticas de Spin, donde una regla compleja sobre excepciones de viáticos requiere que el fragmento mantenga una cohesión estricta y no mezcle temas adyacentes.

3.2. Estrategia de Ventana Creciente (*Growing Window Strategy*)

Desarrollada por **Moreno-Cediel et al. (2025)**, esta estrategia surge para corregir el defecto de la *"deriva semántica"* presente en los métodos de ventana deslizante tradicionales.

• **El Problema de la Ventana Deslizante (*Legacy*)**: Los métodos anteriores (*como el de Kamradt*) comparan la oración N con la N+1. A medida que la ventana avanza, se "olvida" el inicio del párrafo. Esto provoca que el final de una política larga se agrupe incorrectamente con el inicio de la siguiente sección simplemente porque usan palabras similares, perdiendo la referencia al título original.

• **La Solución *"Growing"***: En lugar de deslizar la ventana, este algoritmo hace crecer la ventana. Compara el embedding de las nuevas oraciones candidatas (*tamaño m*) contra el embedding acumulado de todo el fragmento actual (*tamaño n + acumulado*).

• **Mecanismo de Inclusión**:

1. Se inicia un **chunk** con un tamaño base (*n, ej. 8 oraciones*).
2. Se evalúa un grupo candidato (*m, ej. 4 oraciones*).
3. Se calcula la distancia coseno entre el vector del chunk entero y el vector del candidato.
4. Si la distancia es baja (*alta similitud*), el candidato se fusiona y el vector del chunk se recalcula para incluir la nueva información.

• **Ventaja para Spin**: Las pruebas realizadas en corpus de Wikipedia en español demostraron que esta estrategia **mejora un 4%** la entrega de respuestas correctas frente a los métodos de **Kamradt**. Dado que los manuales de Spin están en español y suelen tener una estructura jerárquica donde el contexto inicial es vital, este algoritmo previene que se pierda la relación entre el encabezado de la política y sus reglas finales.

3.3. Comparativa de Rendimiento Esperado

La adopción de estos algoritmos busca impactar directamente en las métricas de recuperación que definen el éxito de un sistema RAG corporativo:

Característica	**Chunking Estático** (*Baseline*)	**Max-Min** (Kiss et al.)	**Growing Window** (Moreno-Cediel et al.)
Enfoque	Mecánico (*Tokens / Caracteres*)	Semántico (*Coherencia Interna*)	Semántico (*Contexto Acumulado*)
Manejo de Contexto	Nulo. Corta oraciones a la mitad.	Alto. Mantiene grupos lógicos densos.	Muy Alto. Preserva el vínculo con el inicio del tema.
Riesgo Principal	Ruptura de frases y tablas.	Costo computacional medio.	Requiere ajuste de parámetros n y m.
Caso de Uso Spin	No recomendado.	Ideal para reglas densas y complejas.	Ideal para manuales extensos y narrativos.

Tabla 1: Comparación de estrategias de segmentación semántica para RAG.

La implementación en los notebooks de Colab comparará estos dos enfoques avanzados contra el baseline estático para determinar cuál ofrece la mayor precisión (Hit Rate y RCC) para el caso específico de Spin Voyager.

4. Resultados de los notebooks de colab con Segmentación Semántica (Semantic Chunking)

En esta sección se presentan y analizan los resultados obtenidos al aplicar dos estrategias de **segmentación semántica** (Semantic Chunking) sobre el corpus documental (políticas de gastos). El objetivo fue evaluar la **coherencia**, **la granularidad** y la **utilidad de los segmentos** (chunks) generados para su posterior indexación en un sistema RAG.

A continuación, se detalla el comportamiento de los algoritmos Max-Min y Growing Window.

4.1. Análisis del algoritmo Max-Min Semantic Chunking

El enfoque Max-Min demostró una capacidad superior para identificar **disrupciones semánticas claras**, generando segmentos de longitud dinámica pero con alta coherencia interna.

- **Alta Granularidad en Reglas de Negocio**: El algoritmo logró aislar reglas específicas. Por ejemplo, separó exitosamente la regla de tiempos de reservación (Art. 9.10) de la regla de aprobaciones (Art. 9.11). Esto es crítico para sistemas de recuperación de información, ya que permite devolver respuestas precisas sin contexto irrelevante.
- **Detección de Estructuras Administrativas**: Se observó que el algoritmo agrupó correctamente secciones de metadatos (Autorizaciones e Historial de Versiones) en un solo bloque lógico.
- **Sensibilidad al Ruido**: Aunque efectivo, el algoritmo identificó encabezados y pies de página repetitivos como bloques semánticamente fuertes independientes. Esto sugiere la necesidad de una etapa de pre-procesamiento para limpiar estos elementos antes de la segmentación.

4.2. Análisis del algoritmo Growing Window Semantic Chunking

El enfoque de Growing Window (Ventana Creciente) mostró una tendencia a la «**saturación**» de contexto en documentos con alta repetición estructural.

- **Fusión Excesiva (Over-chunking)**: Debido a su naturaleza acumulativa, el algoritmo fusionó múltiples páginas de metadatos, avisos legales y encabezados repetidos en un solo «Mega-Chunk» (superando los 2,000 caracteres).
- **Pérdida de Especificidad**: Reglas críticas quedaron «enterradas» dentro de bloques masivos de texto legal y administrativo. Esto dificulta que un LLM (Large Language Model) extraiga la respuesta correcta durante la generación, aumentando el costo computacional y el riesgo de alucinación.
- **Artefactos Visuales**: Se detectaron problemas de formato en los resultados (celdas aparentemente vacías en CSV o comillas no cerradas), indicativo de que el algoritmo absorbió saltos de línea y estructuras de formato sin discriminar el cambio de tema real.

4.3. Comparativa de Estrategias de Chunking

La siguiente tabla resume las diferencias clave observadas en la ejecución de ambos algoritmos sobre el documento de prueba `politica-gastos-viajes.pdf`.

Característica	Max–Min Semantic Chunking	Growing Window Semantic Chunking
Lógica de agrupación	Separa el texto cuando detecta una **caída brusca** en la coherencia semántica.	Acumula texto continuamente mientras la **similitud semántica** se mantenga estable.
Manejo de reglas específicas	**Alta precisión**: segrega reglas individuales (p. ej., **9.10** vs. **9.11**) en **chunks** distintos.	**Baja precisión**: mezcla reglas específicas dentro de contextos amplios o metadatos.
Sensibilidad a metadatos	Identifica encabezados repetitivos como **bloques independientes** (ruido aislado).	Fusiona encabezados y pies de página de múltiples páginas en un solo bloque (ruido acumulado).
Longitud del chunk	Variable y controlada; tiende a ser **breve** y **temática**.	Variable, con tendencia a ser excesiva (**"mega-chunks"**) en documentos estructurados.
Idoneidad para RAG	**Ideal**: facilita la recuperación de respuestas exactas (**"aguja en el pajar"**).	**Limitada**: útil para resúmenes generales, pero ineficiente para búsquedas puntuales.
Hallazgo principal	Detectó correctamente **cambios de tema** entre párrafos normativos.	Generó una **"ilusión de continuidad"** al encontrar encabezados idénticos en páginas consecutivas.

Tabla 2: Comparación entre estrategias de **chunking semántico** y su impacto en sistemas de **Retrieval-Augmented Generation (RAG)**.

5. Conclusiones

A partir de todo lo presentado en el documento, se puede concluir que la implementación de estrategias de chunking semántico representa un avance frente a los métodos tradicionales para el desarrollo de nuestro chatbot: Spin Compass.

A continuación, se deteallas los puntos clave de la conclusión, basada en los resultados mostrados en la sección anterior:

1. El valor del chunking semántico vs. estático

La implementación de la segmentación semántico es fundamental para evitar el ya conocido fenómeno de «Garbage In, Garbage Out». A diferencia del chunking estático (basado en x número de caracteres), que fragmenta de manera fija el texto y destruye la relación entre reglas y sus excepciones, el enfoque semántico garantiza la coherencia semántica total; es decir, la relación entre un fragmento y otro, con el contexto necesario. Esto permite que cada fragmento recuperado sea una unidad de información autosuficiente y precisa.

2. Análisis de resultados y ventajas

Tras evaluar los nuevos algoritmos propuestos para el 2025, se determinaron la siguientes ventajas y comportamientos:

- **Algoritmo Max-Min (Clusterindg Dinámico):** Resultó ser el más idóneo para el sistema RAG de Spin Compass. Su principal ventaja es la alta granularidad, permitiendo separar reglas de negocio específicas con gran precisión.
- **Estrategia de ventaja creciente (Growing Window):** Aunque preserva mejor el vínculo con el inicio del tema en textos narrativos, presentó una tendencia de «fusión excesiva» en documentos con mucha estructura, creando bloques demasiado grandes que dificultan la extracción de respuestas puntuales.

3. Posibles Desventjas y Desafíos

- **Sensibilidad al ruido:** Ambos métodos identificaron de buena forma los encabezados, pies de página y *disclaimers* como bloques semánticos, lo que puede actuar como distractor para el modelo.
- **Costo computacional:** El algoritmo Max-Min conlleva un costo de procesamiento medio en comparación con el método estático. Aunque esto puede significar un problema, el equipo no lo prioriza tanto, ya que el volumen de documentos y contenido de los mismos, no es excesivo y puede manejarse de buena forma.
- **Riesgo de «Mega-Chunks»:** En el caso de *Growing Window*, existe el riesgo de generar bloques de más de 2,000 caracteres, que aumentan el costo del LLM y la posibilidad de alucinaciones, cosa que no sería ideal para un chatbot que responde dudas de políticas internas. Un error en las respuestas, podría incluso hasta generar un problema mayor.

Dicho lo anterior, el equipo concluye que para maximizar la precisión de Spin Compass y reducir el riesgo de incumplimientos normativos derivados de respuestas incorrectas, se recomienda priorizar el algoritmo Max-Min. No obstante, es importante añadir una etapa de pre-procesamiento para *parsear* el documento, limpiar metadatos, disminuir el ruido, etc., antes de la segmentación para optimizar totalmente la calidad de los datos ingeridos.

6. 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.