Título: GUÍA DE DENSIDAD DE TOKENS: ARQUITECTURA DE LA ATENCIÓN

Subtítulo: Manual de optimización de contexto para ingenieros y Project Managers de IA.

Versión: 1.0 (Febrero 2026)

Autor: Antonio Gutiérrez – jefedeproyectos.com

EL CONCEPTO DE DENSIDAD SEMÁNTICA

1. El mito del contexto infinito

En la ingeniería de sistemas actual, la capacidad de una ventana de contexto (ej. 1M de tokens) no es equivalente a su utilidad. La "Arquitectura de la Atención" se basa en maximizar la **IUT (Información Útil por Token)**.

2. El fenómeno "Lost in the Middle"

Los modelos tienden a priorizar la información al principio y al final del contexto inyectado. Todo lo que reside en el "centro" de una ventana de contexto saturada tiene una probabilidad de recuperación (recall) hasta un 60% inferior.

MATRIZ DE UMBRALES DE EFICIENCIA

Tipo de Tarea	Modelo	Umbral Óptimo	Estrategia de Gestión
Resumen / Síntesis	GPT-4º	< 15k tokens	**Summarization Chains:** Condensar en bloques de 4k antes de la síntesis final.
Análisis de Código	Claude 3.5 Sonnet	< 40k tokens	**Expandir:** Priorizar firmas de métodos y definiciones de clases sobre cuerpos de función.
Extracción de Datos	Llama 3.1 (8B/70B)	< 8k tokens	**Pruning:** Eliminar metadatos y etiquetas HTML/JSON redundantes.
Razonamiento Lógico	o1-preview	< 10k tokens	**Precisión:** Priorizar la calidad de las instrucciones sobre la cantidad de ejemplos (Few-shot).

PROTOCOLO TÉCNICO DE PODA (PRUNING)

FASE 1: Filtrado de Metadatos

Eliminar cualquier información que no aporte carga semántica al problema (IDs internos, timestamps irrelevantes, cabeceras de log repetitivas).

FASE 2: Reranking Semántico

No inyectar todo lo recuperado por RAG. Utilizar un modelo de reranking (como Cohere o BGE) para seleccionar solo los 5-10 fragmentos con mayor relevancia vectorial.

FASE 3: Compresión de Historial

En chats de larga duración, aplicar técnicas de "sliding window" o resumir los turnos anteriores para mantener la atención del modelo en el objetivo actual.

CONCLUSIONES Y MÉTRICAS

Kpis de Gestión para el PM:

1. **Recall @ Context:** Capacidad del modelo para recuperar un dato específico en diferentes profundidades de la ventana.
2. **TTFT (Time To First Token):** Latencia inicial; aumenta linealmente con el tamaño del contexto.
3. **Efficiency Score:** Relación entre el acierto en la tarea y el coste de tokens consumidos.

Nota: "*Más contexto suele significar menos precisión. La excelencia en la gestión de proyectos de IA no reside en darle todo al modelo, sino en saber qué ocultarle para que su atención sea máxima.*"

Más recursos en: 🌐 https://jefedeproyectos.com