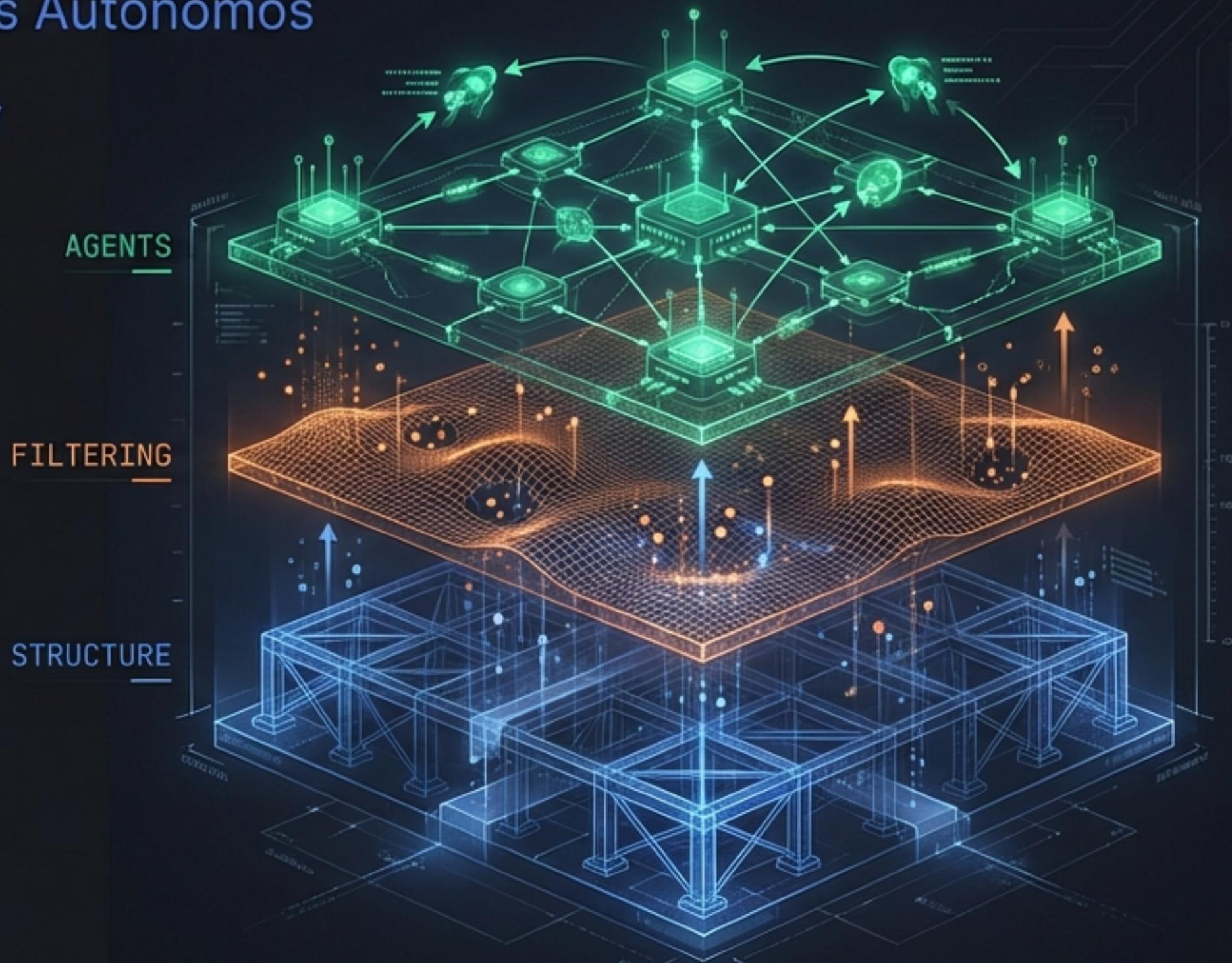


Arquitecturas Avanzadas de RAG

De la Búsqueda Híbrida a los Agentes Autónomos

Estrategias para maximizar relevancia, precisión y adaptabilidad con Google Gemini.



La evolución de un sistema RAG: Ver, Discernir y Actuar



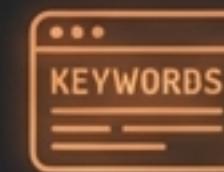
El Dilema de la Recuperación: Dense vs. Sparse



Dense / Embeddings

- ✓ Entiende Contexto
- ✓ Excelente con Sinónimos
- ✗ Falla con códigos exactos
- ✗ Ignora palabras clave específicas

“Entiende el **significado**.”



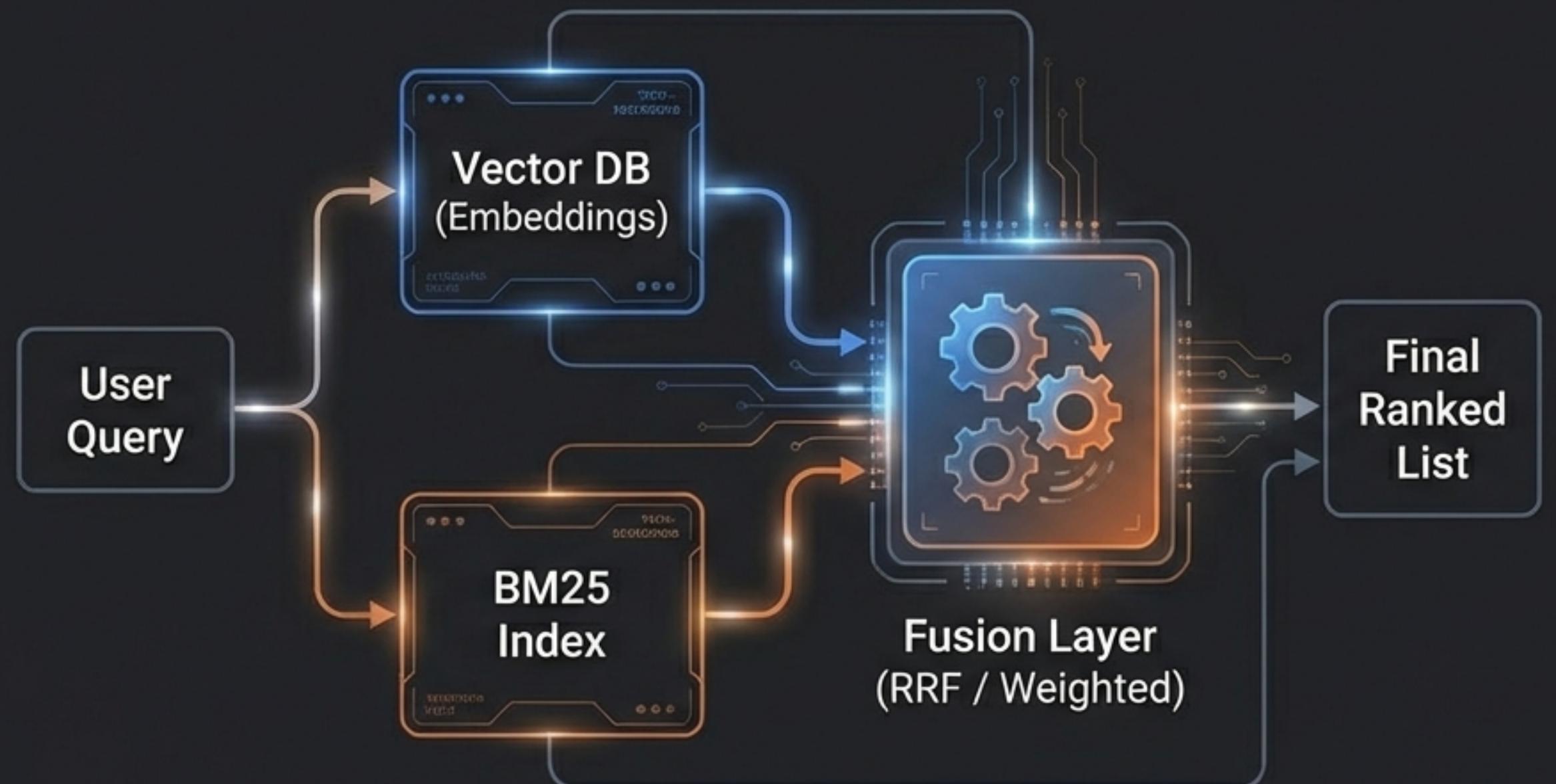
Sparse / BM25

- ✓ Excelente con Nombres/Códigos
- ✓ Muy Rápido (Inverted Index)
- ✗ Malo con sinónimos
- ✗ No entiende contexto semántico

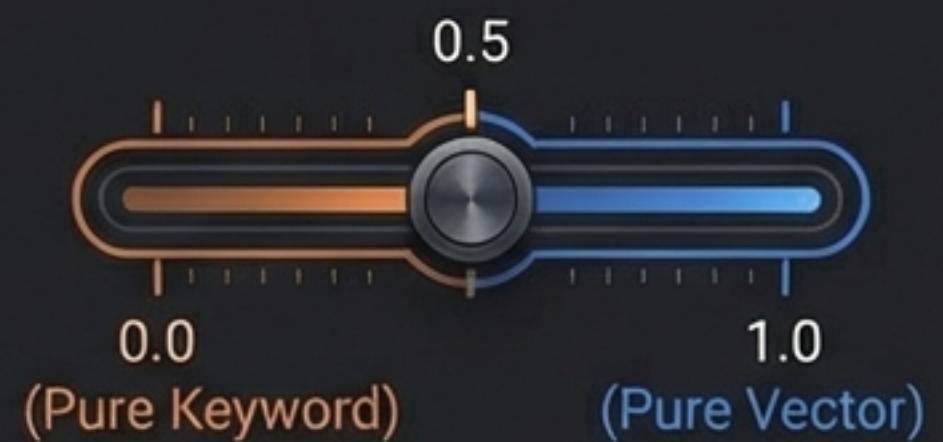
“Encuentra el **dato exacto**.”

Solución: Búsqueda Híbrida

Arquitectura Híbrida: Fusión de Resultados



Parámetro Alpha



Controla el balance entre coincidencia exacta y similitud semántica. Calibrar según datos de evaluación.

Cuándo implementar Búsqueda Híbrida

Criterio	Dense Only	Sparse Only	Híbrido (Winner)
Manejo de Sinónimos	Excelente	Malo	Excelente
Códigos Exactos (SKUs)	Medio/Bajo	Excelente	Excelente
Velocidad	Rápido	Muy Rápido	Medio
Complejidad de Infra	Media	Baja	Alta

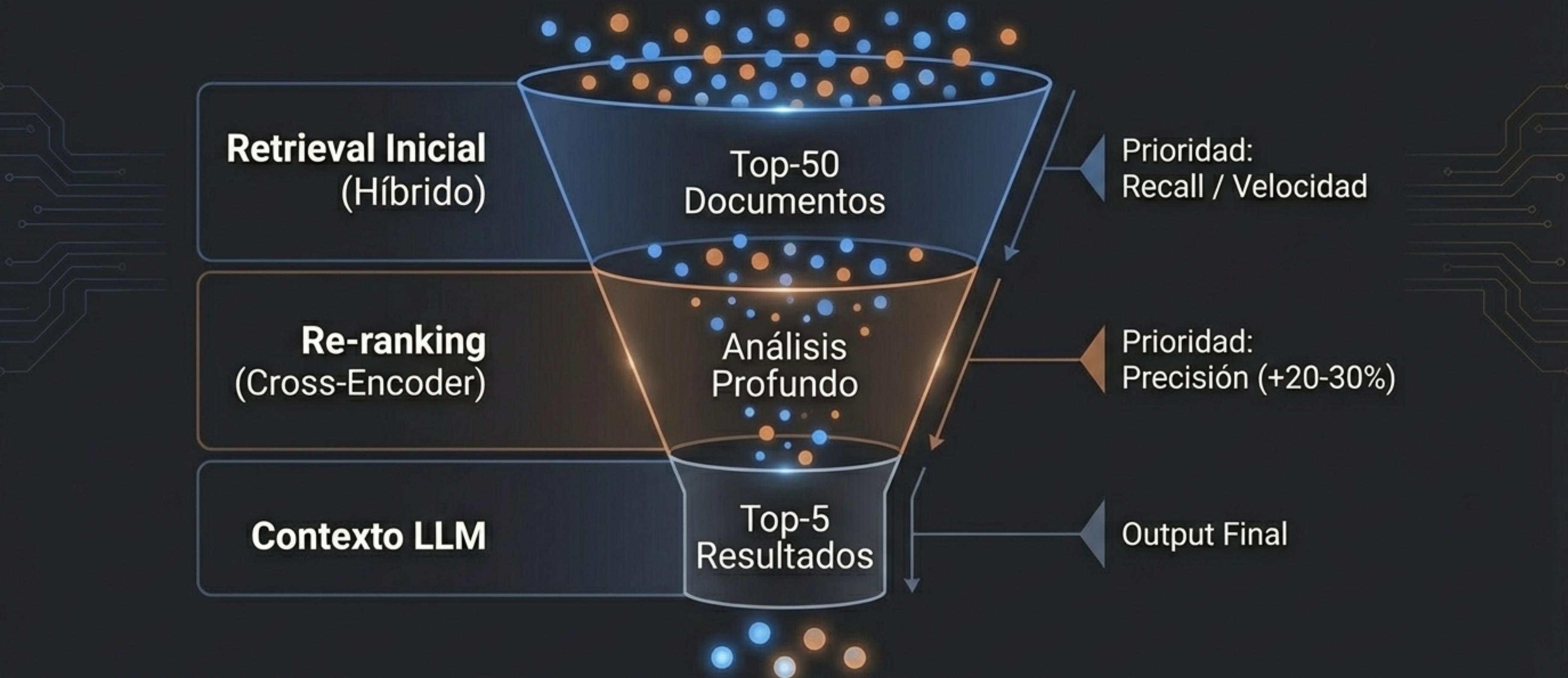
Checklist de Implementación

Implementar índices Dense y Sparse separados.

Elegir algoritmo de fusión (RRF o Weighted).

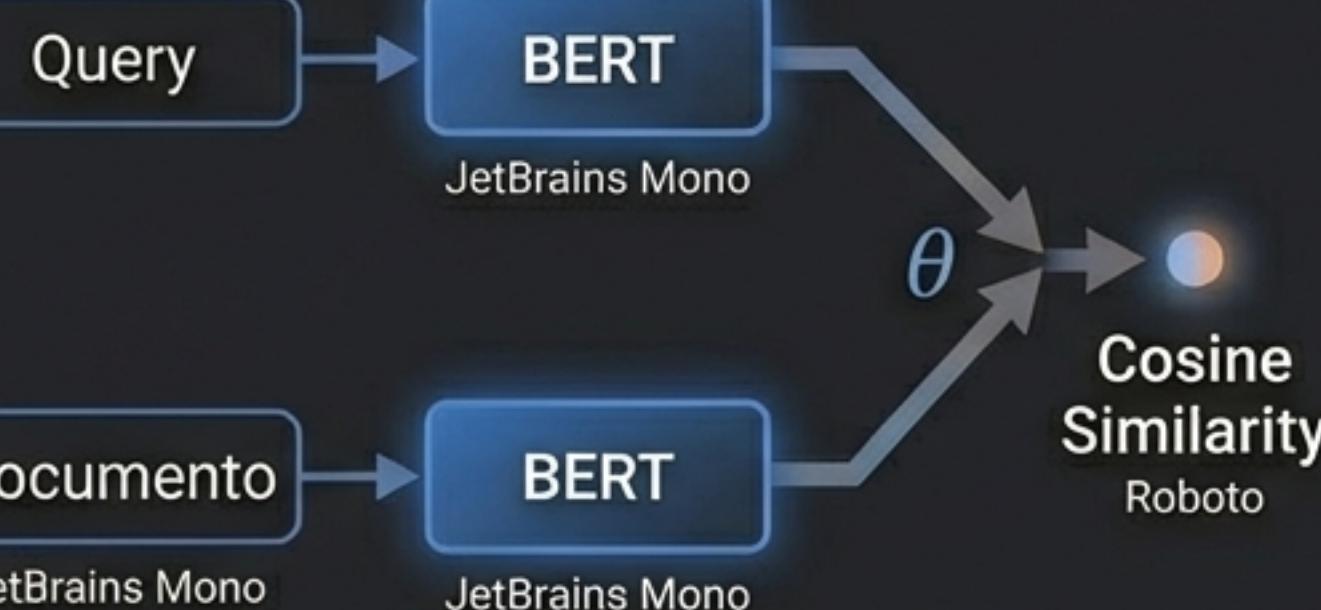
Calibrar alpha con set de validación.

El Embudo de Precisión: Del Recall a la Relevancia



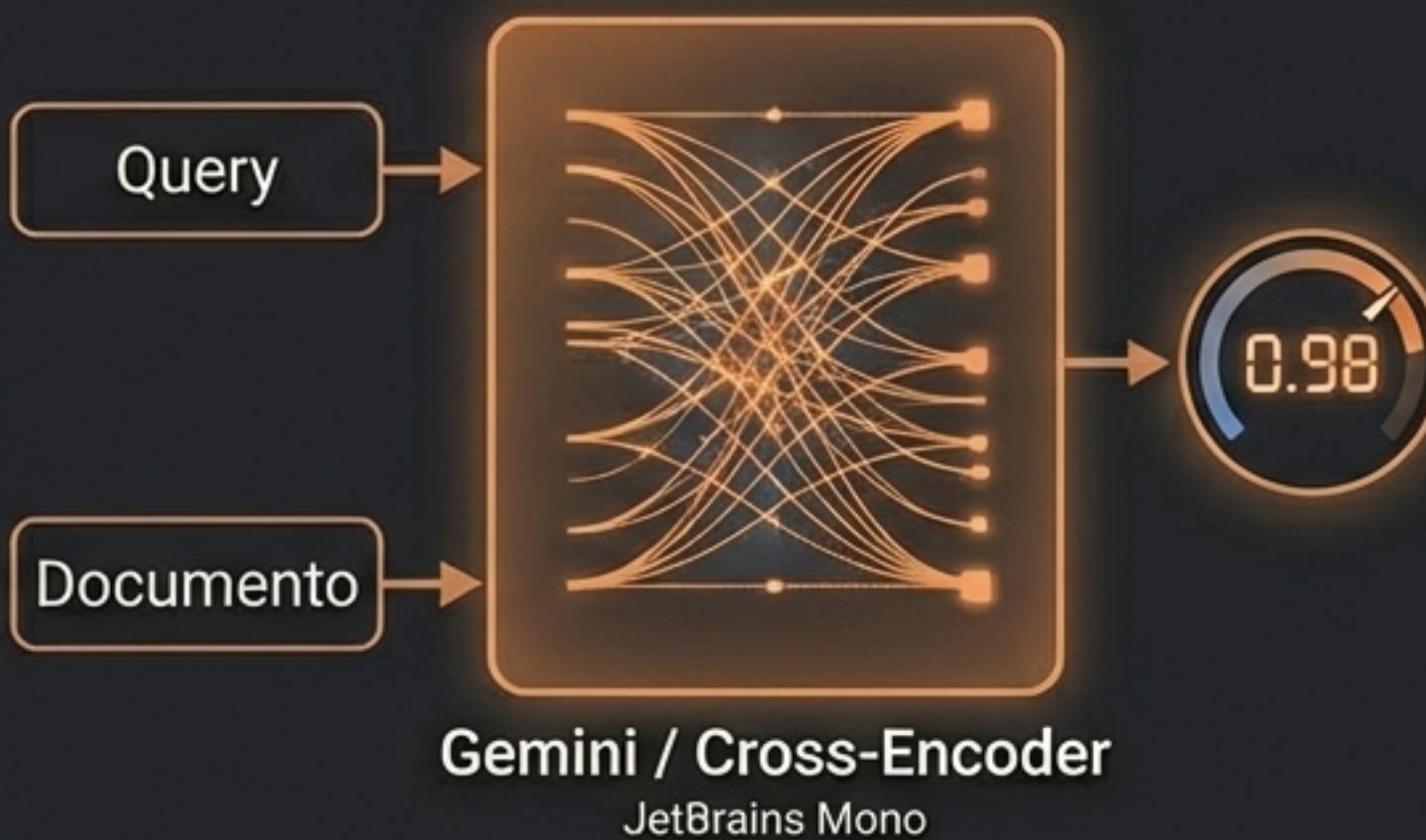
La Mecánica del Cross-Encoder con Gemini

Bi-Encoder / Standard



Rápido, pero pierde matices de interacción.

Cross-Encoder / Re-ranker



Lento, pero máxima comprensión semántica.

El Cross-Encoder 'lee' el par simultáneamente, entendiendo cómo la respuesta satisface específicamente la pregunta.

Estrategias de Filtrado Post-Retrieval

Filtros de Metadatos

Categoría, Fuente, Idioma.

JetBrains Mono

Filtros de Permisos (ACLs)

¿Tiene el usuario acceso a este documento?

JetBrains Mono

Filtros Temporales

Recencia y validez de datos.

JetBrains Mono



Impacto de Negocio

Evita alucinaciones sobre datos obsoletos y garantiza la seguridad de la información empresarial.

Impacto del Re-ranking en el Negocio

Re-ranking Básico

+30%

Mejora en Precisión



Multi-criterio

+25%

Relevancia en Dominio



Re-ranking Contextual

+40%

Satisfacción Usuario



Checklist de Re-ranking

- Retrieval inicial amplio (Top-50+).
- Implementar **Cross-Encoder** con Gemini.
- Aplicar filtros de seguridad (ACLs).
- Logging de pares query/doc para debug.

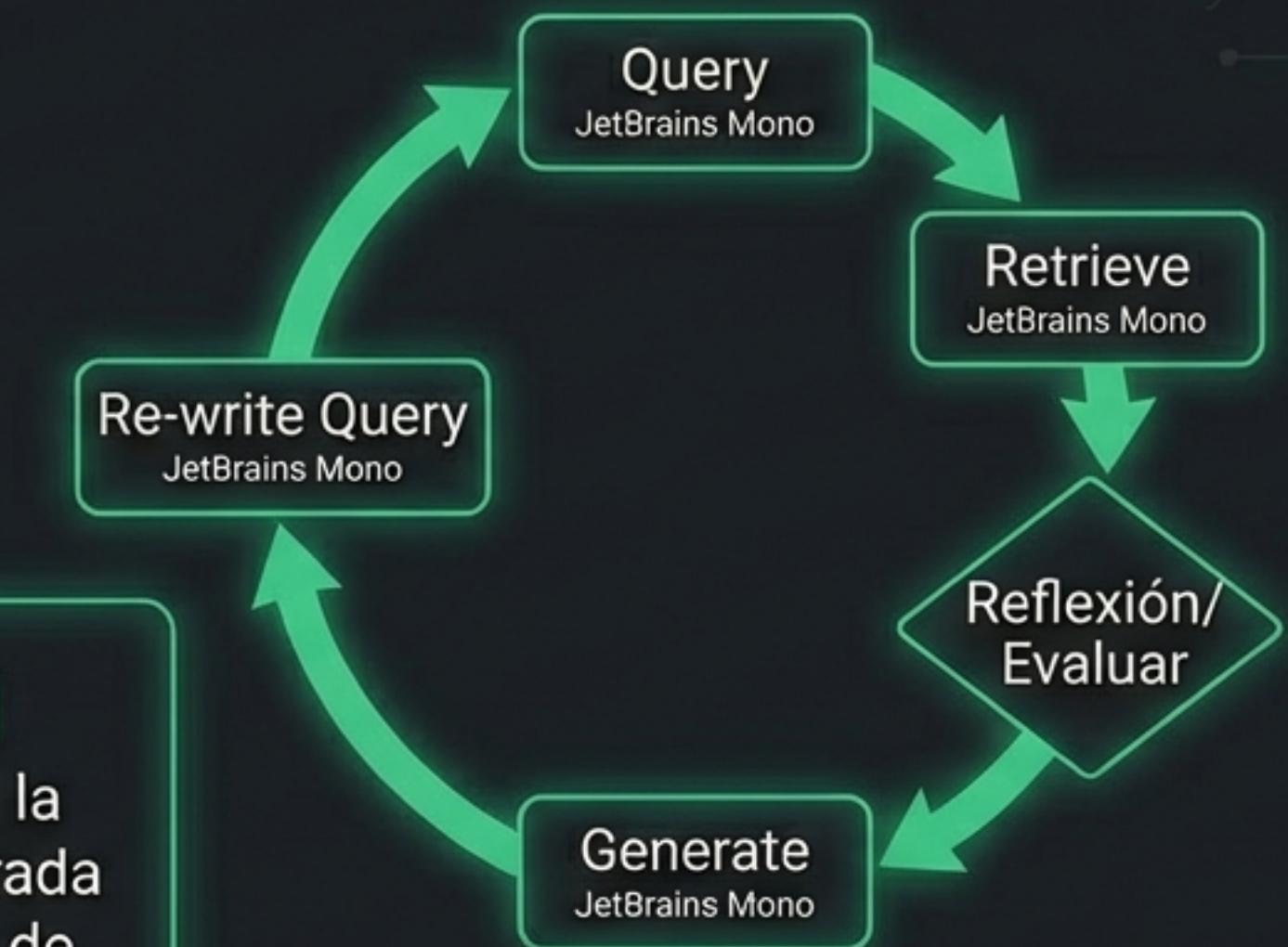
De Lineal a Cílico: El Salto a Agentic RAG

RAG Tradicional



One-Shot (Estático)

Agentic RAG

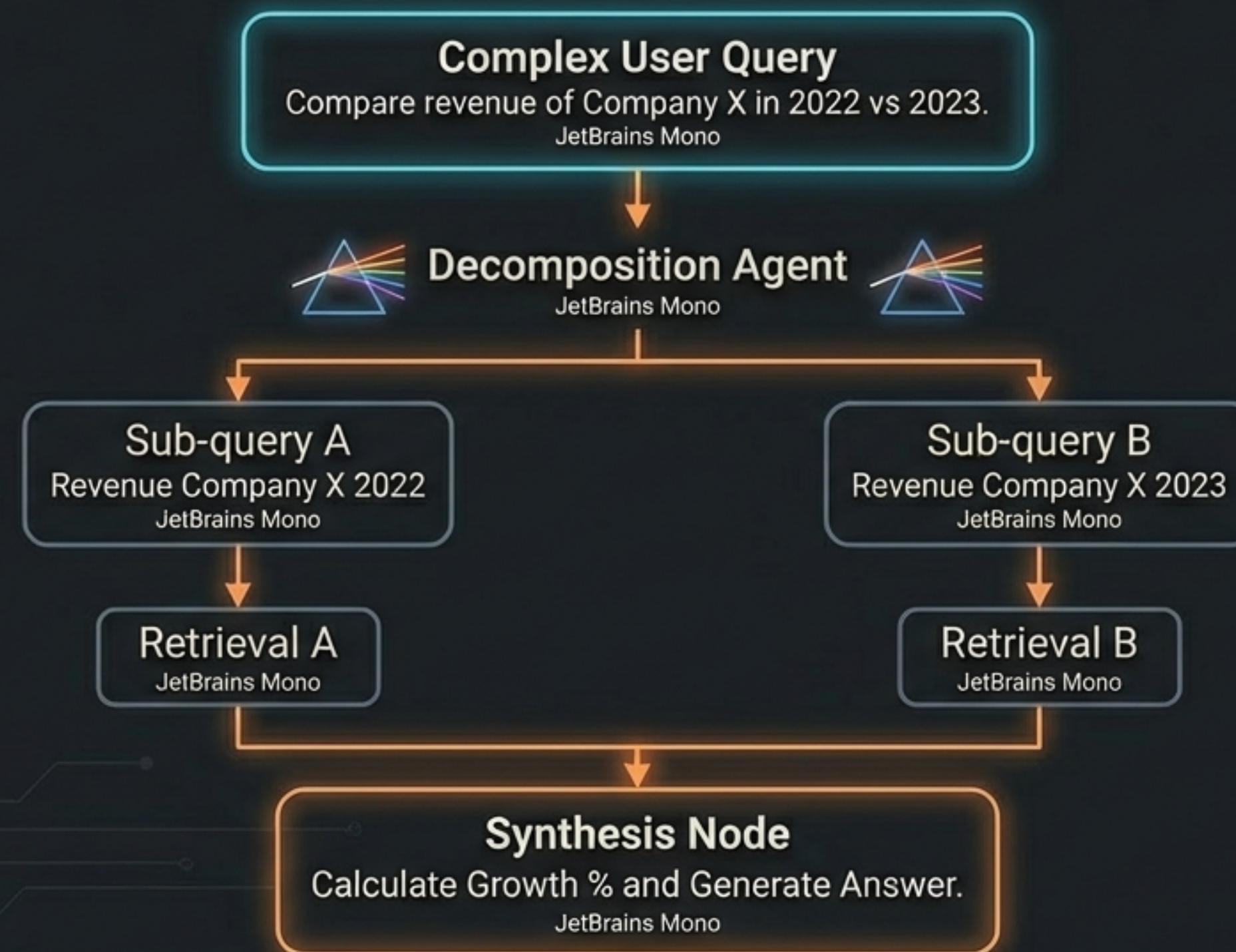


Iterativo (Autónomo)

REFLEXIÓN

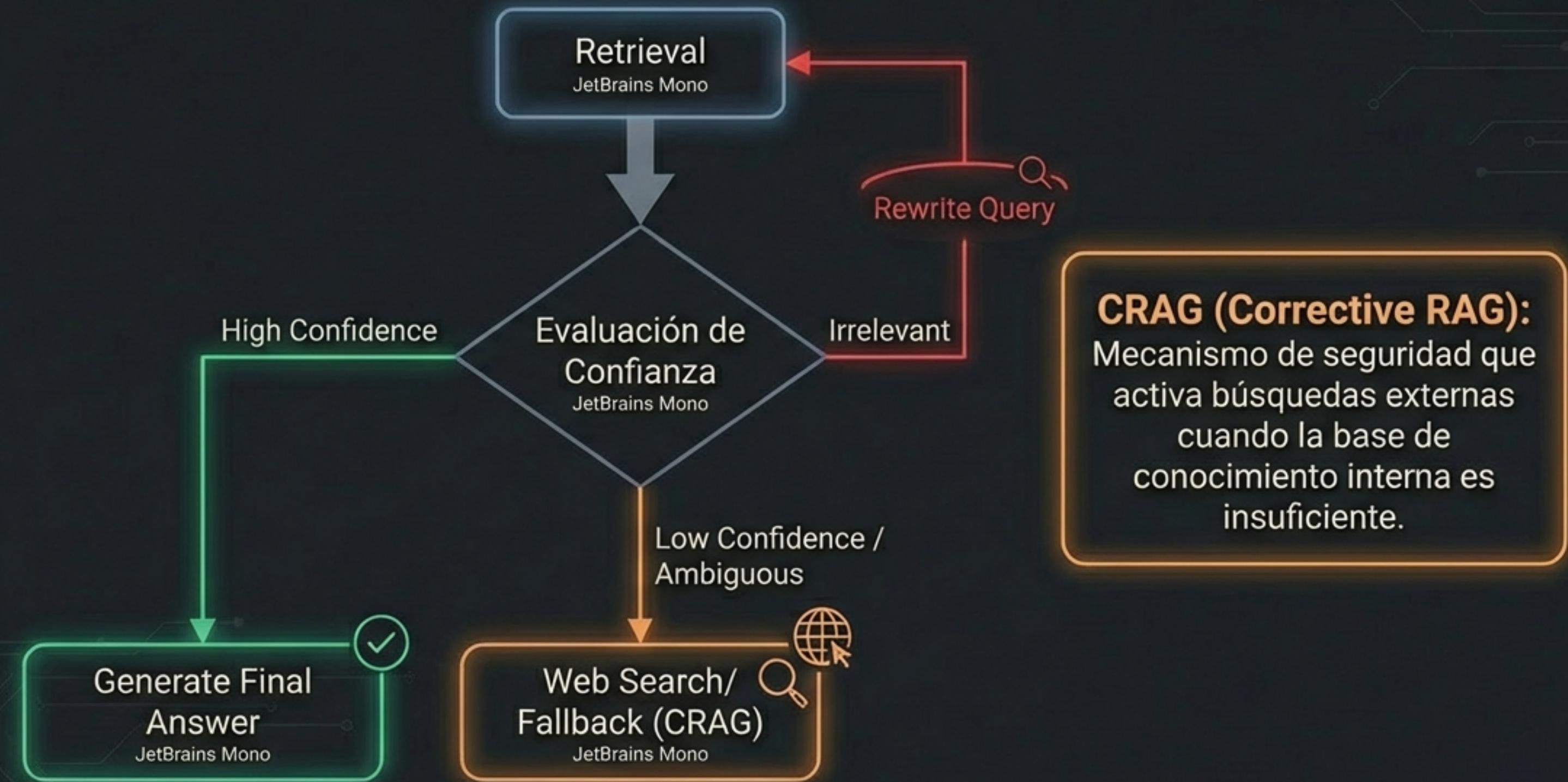
El agente evalúa si la información recuperada es suficiente antes de responder.

Estrategia Agéntica I: Descomposición de Queries

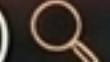


Ideal para preguntas multi-aspecto donde una búsqueda semántica simple fallaría.

Estrategia Agéntica II: Self-RAG y CRAG



Matriz de Selección de Estrategias Agénticas

Estrategia	Caso de Uso Ideal	Complejidad/Costo
RAG Simple JetBrains Mono	Queries directas y factuales	Baja (Low Latency)
Iterativo JetBrains Mono	Información profunda/difícil	Media
Self-RAG JetBrains Mono	Decisión automática de calidad	Media
Descomposición  JetBrains Mono	Queries comparativas/multi-aspecto	Alta
Corrective (CRAG)  JetBrains Mono	Alta precisión requerida	Alta (Web Search Costs)

Trade-off: Mayor autonomía implica mayor latencia y consumo de tokens.



Master Checklist: RAG en Producción

1. Hybrid Search

- Combinar Dense + Sparse.
- Calibrar Alpha.
- Métricas de Recall.

2. Re-ranking

- Initial Retrieval Top-50.
- Cross-Encoder (Gemini).
- Filtros de Metadatos.

3. Agentes

- Evaluación de Contexto.
- Límite de iteraciones.
- Monitoreo de Tokens/Costo.

Meta Final: Un sistema equilibrado en Recall, Precisión y Autonomía.

Conclusión y Siguientes Pasos



No perder nada.

Confiar en el dato.

Responder la intención.

Próximo Módulo

RAG en Producción

- Estrategias de Chunking
- Evaluación de Calidad (RAGAS)
- Actualización de Índices Vectoriales