

# Análisis de Arquitectura y Reestructuración Estratégica: Proyecto 'Caria'

De: Arquitecto de ML Cuantitativo y de Sistemas RAG

Para: El equipo del proyecto 'Caria'

Asunto: Análisis Crítico, Diagnóstico de Causa Raíz (Overfitting) y Plan de Reestructuración Estratégica

Este informe presenta un análisis crítico del proyecto 'Caria' en su estado actual. Se ha realizado una revisión exhaustiva de la arquitectura propuesta, los objetivos del modelo, el diccionario de datos y los artefactos de la interfaz de usuario. El análisis confirma la observación de *overfitting* y concluye que este no es un problema de ajuste de modelo, sino un síntoma de fallas fundamentales en el diseño de los datos y en la arquitectura del modelo.

El diagnóstico central es doble:

- Contaminación de Datos Fundacional:** El modelo predictivo está entrenado con etiquetas que introducen un severo **sesgo de retrospectiva (look-ahead bias)**, haciendo que el modelo sea inservible en un entorno de producción en tiempo real.
- Falla de Diseño Monolítico:** El proyecto intenta forzar a un solo "modelo" a realizar cuatro tareas conceptual y matemáticamente contradictorias (detección de régimen macro, selección de *outliers* de crecimiento, valuación fundamental y razonamiento cualitativo).

Este informe se estructura en cuatro partes. La Parte 1 proporciona un diagnóstico forense detallado de estas dos fallas. La Parte 2 presenta un plan de reestructuración completo, desacoplando 'Caria' en cuatro sistemas expertos independientes pero interconectados. La Parte 3 detalla el protocolo de MLOps correcto necesario para validar estos sistemas y cómo la interfaz de usuario existente se alinea perfectamente con la nueva arquitectura. La Parte 4 concluye con una hoja de ruta táctica para la implementación.

## Parte 1: Diagnóstico Crítico y Falla Fundamental del Modelo

El sentimiento de frustración y la incapacidad de "concretar nada" son el resultado directo de problemas no en la ejecución, sino en la concepción del sistema. El *overfitting* observado es el síntoma; a continuación, se detalla la enfermedad.

## 1.1 El Diagnóstico Forense del 'Overfitting': La Pistola Humeante en `data_dictionary.csv`

El problema principal no es que el modelo esté sobreajustado; es que está contaminado. La causa raíz es el diseño de la tabla `regime_labels`, como se detalla en el `data_dictionary.csv`.<sup>1</sup>

El diccionario de datos define la tabla `regime_labels` con las siguientes columnas y datos de ejemplo <sup>1</sup>:

- `start_date`: 2000-03-10
- `end_date`: 2002-10-09
- `regime`: crash
- `description`: Dot-com bubble burst

Este etiquetado es un ejercicio *ex-post-facto*. Se ha observado un período histórico (el estallido de la burbuja *dot-com*), se ha definido su inicio y fin con perfecta retrospectiva, y se ha etiquetado como "crash". Un modelo de aprendizaje automático entrenado con estos datos (presumiblemente usando *features* de la tabla `processed_features`<sup>1</sup> como `yield_curve_slope` y `vix` de ese mismo período) no está "aprendiendo a predecir *crashes*". Está **memorizando** la firma de un evento pasado específico.

Este es un caso de libro de texto de **sesgo de retrospectiva (look-ahead bias)**.<sup>2</sup> El modelo, durante su entrenamiento, tiene acceso a información que no estaría disponible en tiempo real. Sabe, por ejemplo, que el 11 de marzo de 2000 es el comienzo de un *crash*.<sup>2</sup> En un escenario de predicción real, esta etiqueta no existe.

El *overfitting* que se reporta en los "resultados del último entrenamiento" es la consecuencia inevitable.<sup>4</sup> El modelo probablemente muestra una precisión cercana al 100% en el conjunto de entrenamiento/validación porque simplemente ha aprendido a asociar "VIX alto en 2001" con "etiqueta de crash".<sup>6</sup> Este modelo fallará el 100% del tiempo en producción porque el futuro no viene pre-etiquetado.

## 1.2 La Trampa de la Validación: Por Qué las Métricas Mienten

La confianza en este modelo sobreajustado se ve probablemente exacerbada por el uso de métodos de validación incorrectos. Las técnicas de validación cruzada estándar (CV), como K-Fold, no son válidas para datos de series temporales financieras.<sup>7</sup>

En la CV estándar, los datos se dividen aleatoriamente.<sup>9</sup> Esto significa que un *fold* de entrenamiento podría contener datos del año 2005 para "predecir" un *fold* de prueba del año 2003. Esto contamina el entrenamiento con conocimiento futuro, un tipo de fuga de datos conocida como contaminación de entrenamiento-prueba (*train-test contamination*).<sup>10</sup> Para datos financieros, donde el orden temporal es la única característica que importa, este método es inválido.<sup>11</sup>

**Conclusión de la Sección 1:** El "modelo" cuantitativo actual, en lo que respecta a la detección de regímenes, no se puede "arreglar" o "ajustar". Es fundamentalmente insalvable porque se basa en datos contaminados. Debe ser descartado por completo. La *idea* de la detección de regímenes es correcta, pero la *implementación* (clasificación supervisada sobre etiquetas ex-post) es la antítesis de un sistema predictivo.

## 1.3 La Falacia del Modelo Monolítico: Conflicto de Objetivos

La segunda falla fundamental es la ambición de crear un solo "modelo" para lograr objetivos múltiples y contradictorios. Los objetivos declarados incluyen:

1. **Detección de Regímenes Macro:** Identificar patrones en el "clima económico" (burbujas, crash).
2. **Identificación de Outliers:** Encontrar factores comunes en acciones *multibagger*.
3. **Valuación de Empresas:** Evaluar compañías en diferentes etapas (pre-ingresos vs. consolidadas).
4. **Razonamiento Psicológico:** Actuar como un *sparring partner* de RAG.

Estos objetivos están en conflicto directo, lo que hace imposible que un solo optimizador de modelo tenga éxito.

- **Conflicto 1: Reversión a la Media vs. Momentum de Cola.** La detección de regímenes macro (Objetivo 1) es, en gran medida, un ejercicio de identificación de reversión a la media (p.ej., un VIX extremadamente alto tiende a revertir a la baja).<sup>12</sup> Por el contrario, la búsqueda de *multibaggers* (Objetivo 2) es la búsqueda de eventos de cola extrema; se basa en la continuación de tendencias y la identificación de *outliers*.<sup>14</sup> Un modelo

entrenado para "reconocer patrones" de reversión a la media (descartando *outliers* como ruido) es matemáticamente opuesto a un modelo entrenado para "encontrar *outliers*" (buscando tendencias de cola).<sup>16</sup>

- **Conflict 2: Datos Estructurados vs. Cualitativos en Valuación.** La valuación de empresas consolidadas (Objetivo 3) se basa en datos financieros estructurados y predecibles (ingresos, FCF) de la tabla fundamentales.<sup>1</sup> La valuación de empresas pre-ingresos se basa en datos cualitativos, no estructurados (calidad del equipo, tamaño del mercado, propiedad intelectual) que ni siquiera existen en el data\_dictionary actual.<sup>19</sup> Un modelo DCF es inútil para una startup.<sup>22</sup>

Intentar construir un modelo monolítico para estas tareas es la fuente de la parálisis del proyecto. 'Caria' no debe ser *un* modelo. Debe ser una **plataforma de sistemas desacoplados** que se comunican.

**Tabla 1: Diagnóstico de Falla del Modelo Monolítico**

Objetivo Declarado (Usuario)	Implementación Implícita Actual	Falla Fundamental Identificada
"Detectar patrones tempranos de cambio en clima económico"	Modelo de clasificación supervisada usando regime_labels. <sup>1</sup>	<b>Contaminación por Sesgo de Retrospectiva (<i>Look-ahead bias</i>).<sup>2</sup></b> El modelo memoriza el pasado, no predice el futuro.
"Reconocer factores comunes en <i>outliers</i> que generan <i>multibaggers</i> "	Modelo monolítico único.	<b>Conflict de Objetivos (Reversión vs. Momentum).</b> <sup>15</sup> El modelo no puede optimizar simultáneamente para la estabilidad (regímenes) y los extremos de cola ( <i>outliers</i> ). <sup>16</sup>
"Enseñarle a realizar valuaciones de empresas	Un solo modelo de valuación.	<b>Heterogeneidad de Datos/Métodos.</b> Los

según el tipo o etapa"	<i>drivers</i> devaluación para una empresa pre-ingresos <sup>20</sup> (cualitativos) y una consolidada <sup>18</sup> (cuantitativos) no tienen superposición.
------------------------	--

## Parte 2: El Plan de Reestructuración: Desacoplamiento de 'Caria' en Cuatro Sistemas Especializados

La solución es abandonar el enfoque monolítico y re-arquitectar 'Caria' como un conjunto de cuatro microservicios de modelos especializados. Cada sistema tiene un propósito único y se comunica con los demás.

### 2.1 Sistema I: El Motor de Régimen Macroeconómico (El Enfoque Correcto)

Este sistema reemplaza la fallida clasificación supervisada. Su objetivo no es *predecir* un *crash*, sino *detectar* el estado latente actual del mercado.

- **Replanteamiento:** Se debe dejar de *etiquetar* regímenes (supervisado) y empezar a *descubrirlos* (no supervisado). El estado del mercado es una variable oculta, no una etiqueta observable.
- **Arquitectura Propuesta: Modelos Ocultos de Markov (HMMs)**
  - La metodología correcta para este problema es un **Modelo Oculto de Markov Gaussiano (Gaussian HMM)**.<sup>12</sup>
  - Los HMMs están diseñados para modelar sistemas donde se asume que las observaciones (los datos de mercado) son generadas por un estado subyacente (oculto) que sigue una cadena de Markov.<sup>24</sup> Esto coincide perfectamente con el concepto de "clima económico" o "régimen". El modelo puede modelar la probabilidad de transición entre estados (p.ej., la probabilidad de pasar de un régimen de 'Baja Volatilidad' a uno de 'Alto Estrés').<sup>12</sup>
- **Implementación Práctica:**
  1. **Eliminación:** Se debe eliminar permanentemente la tabla regime\_labels.<sup>1</sup>
  2. **Entrada (Features):** Se utilizarán los datos de la tabla processed\_features<sup>1</sup> (p.ej.,

- yield\_curve\_slope, vix, sentiment\_score) y se complementarán con métricas de volatilidad y autocorrelación calculadas a partir de prices.<sup>1</sup>
3. **Entrenamiento:** Se entrenará un HMM (con 3-5 estados latentes) sobre estas series temporales de *features*.
  4. **Salida (Output):** Para cualquier fecha, el HMM no generará una etiqueta de "crash". Generará:
    - El **régimen latente más probable** (p.ej., "Estado 2").
    - Un JSON de **probabilidades** de estar en cada estado: {"Estado 0": 0.1, "Estado 1": 0.3, "Estado 2": 0.6}.
  5. **Interpretación:** Posteriormente, se analizan las propiedades de estos estados descubiertos (p.ej., "El Estado 2 se caracteriza por un VIX alto y una curva de rendimiento invertida, lo llamaremos 'Régimen de Estrés'").
- **Implicación y Conexión (UI):** Esta salida (las probabilidades del régimen) es el *input* directo para el medidor MODEL OUTLOOK (Imagen 1), que actualmente muestra "Extreme Greed". El régimen se convierte en un *feature* de entrada para otros modelos, no en un *label* de salida.

## 2.2 Sistema II: El "Socio Racional" (Arquitectura RAG y de Texto)

Este sistema es para el razonamiento cualitativo y está destinado a alimentar el "sparring partner" de "Challenge Your Thesis" (Imagen 4). La arquitectura RAG (Retrieval-Augmented Generation) propuesta (Ingesta → Embeddings → Vector DB → RAG) es correcta y estándar de la industria.<sup>26</sup> Sin embargo, las elecciones de tecnología deben ser refinadas.

- **Análisis Crítico: El Modelo de Embeddings**
  - El data\_dictionary<sup>1</sup> muestra el uso de text-embedding-ada-002 (1536 dimensiones).
  - Este modelo está obsoleto, es caro y tiene un rendimiento inferior en *benchmarks* semánticos en comparación con los modelos más recientes.<sup>28</sup>
  - **Recomendación:** Migrar a un modelo de embedding local, *state-of-the-art* y de código abierto. Modelos como mxbai-embed-large<sup>28</sup> o nomic-embed-text<sup>28</sup> ofrecen un rendimiento superior, se pueden ejecutar localmente (satisfaciendo el requisito de "Local model") y garantizan la privacidad de los datos.<sup>31</sup>
- **Análisis Crítico: La Base de Datos Vectorial (pgvector vs. Milvus vs. Pinecone)**
  - El data\_dictionary<sup>1</sup> ya está estructurado como una base de datos relacional PostgreSQL. La sintaxis de tipos de datos (VARCHAR(10), BIGINT, DATE, JSONB y, crucialmente, TEXT para themes en wisdom\_chunks) es inequívocamente PostgreSQL.
  - Su caso de uso requiere búsquedas híbridas: "encontrar textos (vector) sobre 'valuación' (SQL WHERE theme = 'valuation') de 'Warren Buffett' (SQL WHERE source = 'Warren Buffett')".

- **Decisión:** pgvector es la elección abrumadoramente superior para este caso.<sup>32</sup>
  - Milvus<sup>33</sup> y Pinecone<sup>33</sup> son bases de datos *externas* y especializadas. Integrarlas requiere un *pipeline* de sincronización de datos complejo para mantener los metadatos alineados con la base de datos de Postgres, aumentando el costo y la complejidad operativa.<sup>35</sup>
  - pgvector es una *extensión* de PostgreSQL.<sup>34</sup> Vive *dentro* de la base de datos existente. Permite ejecutar consultas híbridas que combinan filtros relacionales SQL (WHERE) y búsqueda de similitud vectorial (ANN) en una sola consulta, en una sola transacción.<sup>35</sup> Esta simplicidad operativa es decisiva.
- **Implementación del RAG (MCP Server)**
  - El *endpoint* RAG expuesto por el MCP (Microservices Compute Platform) Server ejecutará el siguiente flujo cuando un usuario ingrese una tesis (Imagen 4):
    1. **Enriquecimiento:** La consulta (p.ej., "Comprar NVDA") consulta las tablas fundamentals y prices para obtener el contexto financiero estructurado actual de NVDA.
    2. **Búsqueda Híbrida:** El MCP consulta pgvector (tabla wisdom\_chunks) para encontrar textos relevantes usando la consulta del usuario Y filtrando por metadatos (p.ej., themes como "burbuja", "disciplina").
    3. **Generación:** Un LLM local (p.ej., Llama 3) recibe un *prompt* que contiene: (a) la tesis del usuario, (b) el contexto financiero estructurado (de fundamentals), y (c) los *chunks* de sabiduría recuperados (de wisdom\_chunks).<sup>26</sup>
    4. **Respuesta:** El LLM genera una respuesta crítica ("desafío a la tesis") fundamentada en los datos recuperados.<sup>37</sup>

**Tabla 2: Comparativa de Bases de Datos Vectoriales para 'Caria'**

Criterio	pgvector (Recomendado)	Milvus	Pinecone
<b>Tipo de Despliegue</b>	Extensión de PostgreSQL <sup>32</sup>	Base de datos independiente <sup>33</sup>	Servicio gestionado en la nube <sup>33</sup>
<b>Sinergia con data_dictionary</b>	<b>Perfecta.</b> Reside en la misma BD que prices y fundamentals. <sup>1</sup>	Nula. Requiere sincronización de datos.	Nula. Requiere sincronización de datos.

<b>Búsqueda Híbrida (SQL + Vector)</b>	Nativa y en una sola consulta.	Compleja. Requiere dos consultas y lógica de unión en la aplicación.	Compleja. Filtrado de metadatos limitado vs. SQL completo.
<b>Complejidad Operativa</b>	<b>Muy Baja.</b> Es solo Postgres. <sup>36</sup>	Alta. Un segundo sistema distribuido para gestionar y escalar. <sup>33</sup>	Media. Gestionado, pero introduce <i>vendor lock-in</i> y costos. <sup>35</sup>
<b>Costo</b>	Mínimo (incluido en la instancia de Postgres). <sup>35</sup>	Alto (costos de infraestructura).	Alto (costos de suscripción). <sup>35</sup>

## 2.3 Sistema III: El Motor de Factores Cuantitativos (La Búsqueda de *Outliers*)

Este sistema aborda el objetivo de "reconocer factores comunes en *outliers* que generan *multibaggers*".

- **Replanteamiento:** En lugar de intentar predecir el *próximo multibagger* (un evento raro y de cola, extremadamente propenso al *overfitting*), el sistema se reenfocará en la **Inversión por Factores (Factor Investing)**.<sup>39</sup> El objetivo es construir un screener que filtre empresas que exhiben las *características* de *multibaggers* pasados.<sup>41</sup>
- **Arquitectura del Modelo de Factores:**
  - Este es un modelo *cross-sectional* (a través de acciones en un punto en el tiempo), no un modelo de serie temporal.
  - **Entrada (Features):** Se utilizará la tabla fundamentals<sup>1</sup> para construir un universo de factores canónicos que la investigación ha demostrado estar asociados con altos retornos<sup>14</sup>:
    - **Valor:** FCF Yield (calculado de free\_cash\_flow / capitalización de mercado).
    - **Rentabilidad:** ROIC (existente en<sup>1</sup>), ROE.<sup>43</sup>
    - **Crecimiento:** Crecimiento de EPS<sup>14</sup>, Crecimiento de Ingresos.
    - **Solvencia:** Debt-to-Equity.<sup>43</sup>
    - **Momentum:** Calculado de la tabla prices<sup>1</sup> (p.ej., retorno a 12 meses).
- **La Integración Crítica (Conexión con Sistema I):**
  - La eficacia de los factores (Valor, Crecimiento, Momentum) es *dependiente del régimen*.<sup>46</sup> El factor "Valor" puede tener un rendimiento inferior durante años en un régimen de "Crecimiento/Manía", pero superar el rendimiento en una

"Recuperación".<sup>47</sup>

- **Implementación:** El motor de factores no será estático. La salida del **Sistema I (HMM)** (p.ej., "Probabilidad de Régimen de Estrés: 60%") actuará como un *feature* de contexto o un filtro.
- El sistema responderá a la pregunta: "Dado el régimen macroeconómico actual, ¿qué combinación de factores (p.ej., 'Alta Rentabilidad' + 'Bajo Apalancamiento') tiene la mayor probabilidad de superar el rendimiento?".
- **Conexión (UI):** La lista de acciones clasificadas generada por este sistema alimenta directamente el IDEAL CARIA PORTFOLIO y el widget TOP MOVERS (LIVE SIMULATION) (Imagen 1).

## 2.4 Sistema IV: El Motor de Valuación Híbrido y Consciente del Contexto

Este sistema aborda el objetivo de la valuación de empresas, reconociendo la heterogeneidad de las etapas de las compañías.

- **Arquitectura Propuesta: Un Modelo Condicional de Dos Vías**
  - El sistema primero clasifica la empresa: IF 'revenue' == 0 OR 'age' < 3 THEN GOTO Modelo B ELSE GOTO Modelo A.
- **Modelo A: Valuación de Empresas Consolidadas (Híbrido Quant+NLP)**
  - **Base:** Automatización de un modelo de Flujo de Caja Descontado (DCF).<sup>18</sup>
  - **Features:** Usar fundamentals<sup>1</sup> para proyectar Free Cash Flows (FCF).
  - **Integración del Sistema I (Macro):** Las proyecciones de FCF y la Tasa de Descuento (WACC) no deben ser estáticas. El WACC (específicamente la prima de riesgo del mercado) se ajustará *dinámicamente* basado en la salida del **Sistema I**.<sup>50</sup> Un "Régimen de Estrés" detectado por el HMM aumenta la prima de riesgo, disminuyendo el valor presente de la empresa.
  - **Integración del Sistema II (NLP):** Los modelos DCF son extremadamente sensibles a las suposiciones de crecimiento.<sup>49</sup> Se utilizará el **Sistema II (RAG/NLP)** para analizar datos no estructurados (transcripciones de *earnings calls*, informes 10-K) para extraer sentimiento y proyecciones cualitativas, ajustando así las proyecciones de FCF.<sup>51</sup>
- **Modelo B: Valuación de Empresas Pre-Ingresos/Etapa Temprana**
  - **Base:** Los modelos DCF son inútiles aquí.<sup>19</sup> Se debe cambiar a métodos cualitativos estructurados.<sup>22</sup>
  - **Metodología:** Implementar el **Método Berkus o Scorecard Valuation**.<sup>20</sup>
  - **Features (requerirán nuevas fuentes de datos):**
    1. Calidad del Equipo (0-30%)<sup>20</sup>

- 2. Tamaño de la Oportunidad (0-25%)<sup>20</sup>
- 3. Tecnología/Producto/Propiedad Intelectual (0-15%)<sup>20</sup>
- 4. Entorno Competitivo (0-10%)<sup>20</sup>
- El modelo de IA no predecirá el "valor", sino que asignará una *puntuación* en estos vectores, comparándola con una base de datos de valuaciones de rondas seed de la industria.<sup>56</sup>

**Tabla 3: Lógica del Motor de Valuación Híbrido (Sistema IV)**

Etapa de la Empresa	Metodología Primaria	Fuentes de Datos	Integración Sistema I (Macro)	Integración Sistema II (NLP)
<b>Consolidada</b>	DCF Asistido por IA <sup>18</sup>	fundamentals, prices	Ajuste dinámico de Tasa de Descuento (WACC). <sup>50</sup>	Ajuste de Proyección de FCF (Sentimiento de <i>Earnings Calls</i> ). <sup>52</sup>
<b>Pre-Ingresos</b>	Scorecard / VC Method <sup>19</sup>	(Requiere nuevos datos cualitativos)	Ajuste de Múltiplos Comparables de Mercado (Contexto).	N/A

## Parte 3: Implementación, Validación y La Interfaz de Usuario Unificada

La reestructuración del modelo requiere un protocolo de MLOps de nivel de producción para evitar futuras contaminaciones.

### 3.1 Un Protocolo de MLOps a Prueba de Balas: La Cura Metodológica

Como se estableció en la Sección 1.2, la validación cruzada estándar es la causa de la falsa confianza en el modelo sobreajustado.<sup>8</sup>

- **La Solución: Validación Cruzada Purgada y Embargada**
  - La metodología correcta es la **Validación Cruzada Plegada (K-Fold) Purgada y Embargada** (*Purged and Embargoed K-Fold Cross-Validation*).<sup>58</sup>
  - Este método preserva el orden temporal de los datos<sup>7</sup> y previene dos tipos de fuga de datos:
    1. **Purgado (Purging):**<sup>58</sup> En finanzas, las etiquetas (p.ej., "retorno a 30 días") abarcan múltiples puntos de tiempo. El purgado elimina todas las observaciones del conjunto de *entrenamiento* cuyas etiquetas de tiempo se *superponen* con las observaciones en el conjunto de *prueba*.
    2. **Embargo (Embargo):**<sup>58</sup> Se elimina un pequeño conjunto de observaciones del conjunto de *entrenamiento* que *siguen inmediatamente* al conjunto de *prueba*. Esto previene que la información del *test* "regrese" al *train* debido a la autocorrelación de corto plazo.
- **Implicación:** Este es un protocolo de MLOps no negociable para los Sistemas I, III y IV. Las métricas de *backtest* resultantes (que probablemente serán mucho más bajas que las actuales) serán el primer reflejo verdadero del rendimiento esperado del modelo en producción.

### 3.2 Reinterpretación de la Interfaz de Usuario de 'Caria' (Imágenes 1-4)

La excelente noticia es que la interfaz de usuario existente (Imágenes 1-4) no necesita ser modificada. Ya está diseñada, quizás intuitivamente, para la arquitectura desacoplada propuesta. Simplemente necesita ser conectada a los *endpoints* correctos de los nuevos sistemas.

- **Imagen 1 MODEL OUTLOOK (Medidor "Extreme Greed"):**
  - **Conexión:** Es la visualización directa del **Sistema I (HMM)**. Muestra la probabilidad del régimen de mercado más probable (p.ej., Estado 2 = "Extreme Greed").
- **Imagen 1 PORTFOLIO SNAPSHOT / IDEAL CARIA PORTFOLIO:**
  - **Conexión:** Es la salida del **Sistema III (Factor Engine)** y **Sistema IV (Valuation)**. Crucialmente, esta cartera está *condicionada* por la salida del **Sistema I (Régimen)**.
- **Imagen 1 TOP MOVERS (LIVE SIMULATION):**
  - **Conexión:** Es una salida directa del screener del **Sistema III (Factor Engine)**.

- **Imagen 4 Challenge Your Thesis ("e.g., 'Buy NVDA because AI is the future...'""):**
  - **Conexión:** Esta es la interfaz de entrada principal para el **Sistema II (RAG)**. La consulta del usuario se envía al *pipeline RAG* (descrito en 2.2) para generar un desafío fundamentado.
- **Imágenes 2 y 3 (Landing Page):**
  - **Conexión:** El marketing ya refleja esta dualidad: "Enduring Principles" (Sistema II - RAG, wisdom\_chunks) y "Modern Insight" (Sistemas I, III, IV - Quant).

## Parte 4: Resumen Ejecutivo y Hoja de Ruta Táctica

El *overfitting* diagnosticado es el resultado de un sesgo de retrospectiva en el diseño de los datos (regime\_labels) y un intento de construir un modelo monolítico con objetivos contradictorios.

La solución es una **arquitectura desacoplada de cuatro sistemas especializados** (Régimen HMM, RAG de Texto, Motor de Factores, Valuación Híbrida) que se comunican, proporcionando contexto (Régimen → Factores) y enriquecimiento (NLP → DCF).

### 4.1 Hoja de Ruta de Implementación (Próximos Pasos)

1. **Día 1 (Fundacional):** Eliminar *inmediatamente* la tabla regime\_labels<sup>1</sup> de la base de datos y de todo el código de entrenamiento.
2. **Semana 1 (Sistema II - RAG):** Configurar la base de datos PostgreSQL y habilitar la extensión pgvector.<sup>32</sup> Migrar los datos de wisdom\_chunks<sup>1</sup> a esta nueva estructura. Reemplazar text-embedding-ada-002 por un modelo de embedding local como mxbai-embed-large.<sup>28</sup> Construir el endpoint RAG básico. Este es un *quick win* que produce valor tangible.
3. **Semana 2 (Sistema I - Régimen):** Construir el modelo HMM<sup>12</sup> sobre los datos de processed\_features.<sup>1</sup> Esta es ahora la máxima prioridad cuantitativa. El objetivo es tener un endpoint que devuelva las probabilidades de régimen actuales.
4. **Semana 3 (MLOps):** Implementar el backtester de series temporales usando validación cruzada purgada y embargada.<sup>58</sup> Este es el nuevo estándar. Ningún modelo cuantitativo se construye sin ser validado por este framework.
5. **Semana 4-5 (Sistema III y IV):** Construir el Motor de Factores (Sistema III) y el Motor de Valuación (Sistema IV) dentro del nuevo marco de MLOps. Asegurar que ambos sistemas consuman la salida del Sistema I como feature de entrada.

**6. Semana 6 (Integración):** Conectar los *endpoints* de los 4 sistemas a la Interfaz de Usuario existente (Imagen 1).

Esta reestructuración alinea la arquitectura del proyecto con sus ambiciosos objetivos, curando la causa raíz del *overfitting* y sentando las bases para un sistema robusto, de nivel de producción y verdaderamente inteligente.

**Tabla 4: La Arquitectura Desacoplada de 'Caria' (El Futuro)**

Sistema	Objetivo Específico	Metodología de Modelo	Tablas de Datos	Salida (Output) / Consumidor
<b>Sistema I (Régimen)</b>	Detección de estado de mercado no supervisada	HMM / GMM <sup>12</sup>	processed_features, prices	current_regime_probabilities (JSONB) → Consumido por UI (Outlook) y Sistemas III, IV.
<b>Sistema II (RAG)</b>	Razonamiento cualitativo, desafío de tesis <sup>26</sup>	RAG sobre LLM (con pgvector) <sup>27</sup>	wisdom_chunks, fundamentals (para contexto)	generated_text_response (TEXT) → Consumido por UI (Chat).
<b>Sistema III (Factores)</b>	Screening de acciones basado en factores <sup>42</sup>	Regresión/Árbol de Decisión (Cross-sectional)	fundamentals, prices, processed_features	ranked_ticker_list (JSONB) → Consumido por UI (Top Movers).
<b>Sistema IV (Valuación)</b>	Valuación de empresas consciente del	Híbrido: (DCF-ML <sup>18</sup> + Scorecard <sup>20</sup> )	fundamentals, (Input cualitativo)	estimated_fair_value (FLOAT) → Consumido

	contexto <sup>52</sup>			por UI (Portfolio).
--	------------------------	--	--	------------------------

## Obras citadas

1. data\_dictionary.csv,  
<https://drive.google.com/open?id=1Wn7Kp3knOfMxiA79fCDEC1-NUh99wQS>
2. Look-Ahead Bias - Definition and Practical Example - Corporate Finance Institute, fecha de acceso: noviembre 11, 2025,  
<https://corporatefinanceinstitute.com/resources/career-map/sell-side/capital-markets/look-ahead-bias/>
3. A Survey of Large Language Models for Financial Applications: Progress, Prospects and Challenges - arXiv, fecha de acceso: noviembre 11, 2025,  
<https://arxiv.org/html/2406.11903v1>
4. Overfitting in finance: causes, detection & prevention strategies - OneMoneyWay, fecha de acceso: noviembre 11, 2025,  
<https://onemoneyway.com/en/dictionary/overfitting/>
5. What is Overfitting? - Overfitting in Machine Learning Explained - Amazon AWS, fecha de acceso: noviembre 11, 2025,  
<https://aws.amazon.com/what-is/overfitting/>
6. Overfitting In Financial Models - Meegle, fecha de acceso: noviembre 11, 2025,  
[https://www.meegle.com/en\\_us/topics/overfitting/overfitting-in-financial-models](https://www.meegle.com/en_us/topics/overfitting/overfitting-in-financial-models)
7. Time Series Cross-Validation - GeeksforGeeks, fecha de acceso: noviembre 11, 2025,  
<https://www.geeksforgeeks.org/machine-learning/time-series-cross-validation/>
8. k-fold CV of forecasting financial time series -- is performance on last fold more relevant?, fecha de acceso: noviembre 11, 2025,  
<https://stats.stackexchange.com/questions/14197/k-fold-cv-of-forecasting-financial-time-series-is-performance-on-last-fold-more-relevant>
9. 3.1. Cross-validation: evaluating estimator performance - Scikit-learn, fecha de acceso: noviembre 11, 2025,  
[https://scikit-learn.org/stable/modules/cross\\_validation.html](https://scikit-learn.org/stable/modules/cross_validation.html)
10. What is Data Leakage in Machine Learning? - IBM, fecha de acceso: noviembre 11, 2025, <https://www.ibm.com/think/topics/data-leakage-machine-learning>
11. Time Series Cross-Validation: Best Practices - Medium, fecha de acceso: noviembre 11, 2025,  
<https://medium.com/@pacosun/respect-the-order-cross-validation-in-time-series-7d12beab79a1>
12. Market Regime Detection Using Hidden Markov Models - QuestDB, fecha de acceso: noviembre 11, 2025,  
<https://questdb.com/glossary/market-regime-detection-using-hidden-markov-models/>
13. Trading the VIX with Hidden Markov Models? A Stochastic Modeling Case Study (MScFE @ WQU), fecha de acceso: noviembre 11, 2025,  
<https://medium.com/@eodenyire/trading-the-vix-with-hidden-markov-models-a->

[stochastic-modeling-case-study-mscfe-wqu-b597cabcb30](#)

14. How to Identify Multibagger Stocks Using Momentum Strategies | Wright Blogs, fecha de acceso: noviembre 11, 2025,  
<https://www.wrightresearch.in/blog/multibaggers-momentum/>
15. Outperforming the Stock Market Using Market Anomalies - ScholarWorks@UARK, fecha de acceso: noviembre 11, 2025,  
<https://scholarworks.uark.edu/cgi/viewcontent.cgi?article=1110&context=finnuht>
16. Identifying and Validating Outliers in Market Price Action | by Michael Harris | Medium, fecha de acceso: noviembre 11, 2025,  
<https://mikeharrisny.medium.com/identifying-and-validating-outliers-in-market-price-action-b28fb299b736>
17. Outlier Risk, Part I | Investing.com, fecha de acceso: noviembre 11, 2025,  
<https://www.investing.com/analysis/outlier-risk-part-i-200606354>
18. Build an AI-Powered DCF Valuation Explainer with FMP and Groq - Level Up Coding, fecha de acceso: noviembre 11, 2025,  
<https://levelup.gitconnected.com/build-an-ai-powered-dcf-valuation-explainer-with-fmp-and-groq-f6109cbd715b>
19. Six Valuation Models for Early-Stage Healthcare AI Companies in Europe: Methods to Calculate Enterprise Value by Nelson Advisors, fecha de acceso: noviembre 11, 2025,  
<https://nelsonadvisors.co.uk/blog/six-valuation-models-for-early-stage-healthcare-ai-companies-in-europe--methods-to-calculate-enterprise-value-by-nelson-advisors>
20. How to do a startup valuation using 8 different methods - Brex, fecha de acceso: noviembre 11, 2025, <https://www.brex.com/journal/startup-valuation>
21. Understanding AI Startup Valuations: Trends and Insights for Founders - DealMaker, fecha de acceso: noviembre 11, 2025,  
<https://www.dealmaker.tech/content/understanding-ai-startup-valuations-trends-and-insights-for-founders>
22. Machine Learning for Startup Valuation - Lucid.Now, fecha de acceso: noviembre 11, 2025, <https://www.lucid.now/blog/machine-learning-for-startup-valuation/>
23. Regime-Switching Factor Investing with Hidden Markov Models - MDPI, fecha de acceso: noviembre 11, 2025, <https://www.mdpi.com/1911-8074/13/12/311>
24. Market Regime Detection using Hidden Markov Models in QSTrader | QuantStart, fecha de acceso: noviembre 11, 2025,  
<https://www.quantstart.com/articles/market-regime-detection-using-hidden-markov-models-in-qstrader/>
25. [2007.14874] Detecting bearish and bullish markets in financial time series using hierarchical hidden Markov models - arXiv, fecha de acceso: noviembre 11, 2025, <https://arxiv.org/abs/2007.14874>
26. RAG for Finance: Automating Document Analysis with LLMs, fecha de acceso: noviembre 11, 2025,  
<https://rpc.cfainstitute.org/research/the-automation-ahead-content-series/retrieval-augmented-generation>
27. Intelligent Financial Data Analysis System Based on LLM-RAG - arXiv, fecha de

- acceso: noviembre 11, 2025, <https://arxiv.org/abs/2504.06279>
28. 13 Best Embedding Models in 2025: OpenAI vs Voyage AI vs Ollama | Complete Guide + Pricing & Performance - Elephas, fecha de acceso: noviembre 11, 2025, <https://elephas.app/blog/best-embedding-models>
29. Top Embedding Models in 2025 — The Complete Guide - Artsmart.ai, fecha de acceso: noviembre 11, 2025, <https://artsmart.ai/blog/top-embedding-models-in-2025/>
30. What are the best embedding models? : r/AI\_Agents - Reddit, fecha de acceso: noviembre 11, 2025, [https://www.reddit.com/r/AI\\_Agents/comments/1iqs84e/what\\_are\\_the\\_best\\_embedding\\_models/](https://www.reddit.com/r/AI_Agents/comments/1iqs84e/what_are_the_best_embedding_models/)
31. 5 Best Embedding Models for RAG: How to Choose the Right One - GreenNode, fecha de acceso: noviembre 11, 2025, <https://greennode.ai/blog/best-embedding-models-for-rag>
32. Vector database : pgvector vs milvus vs weaviate. : r/LocalLLaMA - Reddit, fecha de acceso: noviembre 11, 2025, [https://www.reddit.com/r/LocalLLaMA/comments/1e63m16/vector\\_database\\_pgvector\\_vs\\_milvus\\_vs\\_weaviate/](https://www.reddit.com/r/LocalLLaMA/comments/1e63m16/vector_database_pgvector_vs_milvus_vs_weaviate/)
33. We Tried and Tested 10 Best Vector Databases for RAG Pipelines - ZenML Blog, fecha de acceso: noviembre 11, 2025, <https://www.zenml.io/blog/vector-databases-for-rag>
34. The 7 Best Vector Databases in 2025 - DataCamp, fecha de acceso: noviembre 11, 2025, <https://www.datacamp.com/blog/the-top-5-vector-databases>
35. Pinecone? Milvus? PgVector Is 70% Faster and Cheaper and Open Source - Medium, fecha de acceso: noviembre 11, 2025, [https://medium.com/@Erik\\_Milosevic/pinecone-milvus-pgvector-is-70-faster-and-cheaper-and-open-source-3d051a9848ac](https://medium.com/@Erik_Milosevic/pinecone-milvus-pgvector-is-70-faster-and-cheaper-and-open-source-3d051a9848ac)
36. Who here has actually used vector DBs in production? : r/Database - Reddit, fecha de acceso: noviembre 11, 2025, [https://www.reddit.com/r/Database/comments/1myb6vc/who\\_here\\_has\\_actually\\_used\\_vector\\_dbs\\_in/](https://www.reddit.com/r/Database/comments/1myb6vc/who_here_has_actually_used_vector_dbs_in/)
37. Evaluating Retrieval-Augmented Generation Models for Financial Report Question and Answering - MDPI, fecha de acceso: noviembre 11, 2025, <https://www.mdpi.com/2076-3417/14/20/9318>
38. Evaluating Retrieval-Augmented Generation Models for Financial Report Question and Answering - ResearchGate, fecha de acceso: noviembre 11, 2025, [https://www.researchgate.net/publication/385034153\\_Evaluating\\_Retrieval-Augmented\\_Generation\\_Models\\_for\\_Financial\\_Report\\_Question\\_and\\_Answering](https://www.researchgate.net/publication/385034153_Evaluating_Retrieval-Augmented_Generation_Models_for_Financial_Report_Question_and_Answering)
39. Machine Learning for Factor Investing 9780367639747, 9780367639723, 9781003121596 - DOKUMEN.PUB, fecha de acceso: noviembre 11, 2025, <https://dokumen.pub/machine-learning-for-factor-investing-9780367639747-9780367639723-9781003121596.html>
40. Quant's Guide to Factor Investing: Theory, Practice, and Code | by Jakub Polec | Medium, fecha de acceso: noviembre 11, 2025, [https://medium.com/@jpolec\\_72972/quant-s-guide-to-factor-investing-theory-pr](https://medium.com/@jpolec_72972/quant-s-guide-to-factor-investing-theory-pr)

## actice-and-code-09ce1c06c3e8

41. How to implement a multi factor quantitative investment strategy - Quant-Investing, fecha de acceso: noviembre 11, 2025,  
<https://www.quant-investing.com/blog/how-to-implement-a-multi-factor-quantitative-investment-strategy>
42. How to build a factor model? - Quantitative Finance Stack Exchange, fecha de acceso: noviembre 11, 2025,  
<https://quant.stackexchange.com/questions/17125/how-to-build-a-factor-model>
43. How to find multibagger stocks: A guide for long-term value investors, fecha de acceso: noviembre 11, 2025,  
<https://www.valueresearchonline.com/stories/222387/how-to-find-multibagger-stocks-guide-long-term-value-investors/>
44. Find Next 10 Bagger Using Data Driven Screening | Quant Investing, fecha de acceso: noviembre 11, 2025,  
<https://www.quant-investing.com/blog/find-next-10-bagger-using-data-driven-screening>
45. Building a multiple regression model to beat the benchmark : r/quant - Reddit, fecha de acceso: noviembre 11, 2025,  
[https://www.reddit.com/r/quant/comments/1jdal59/building\\_a\\_multiple\\_regression\\_model\\_to\\_beat\\_the/](https://www.reddit.com/r/quant/comments/1jdal59/building_a_multiple_regression_model_to_beat_the/)
46. Can Machine Learning Improve Factor Returns? Not Really - Alpha Architect, fecha de acceso: noviembre 11, 2025,  
<https://alphaarchitect.com/can-machine-learning-improve-factor-returns-not-really/>
47. A Machine Learning Approach to Regime Modeling - Two Sigma, fecha de acceso: noviembre 11, 2025,  
<https://www.twosigma.com/articles/a-machine-learning-approach-to-regime-modeling/>
48. Chapter 18 Python notebooks | Machine Learning for Factor Investing, fecha de acceso: noviembre 11, 2025, <http://www.mlfactor.com/python.html>
49. Automate your Discounted Cash Flow model in Python | by Gianluca Baglini | Medium, fecha de acceso: noviembre 11, 2025,  
<https://medium.com/@gianlucabaglini/automate-your-discounted-cash-flow-model-in-python-cdf98eb0924d>
50. How AI Enhances DCF Valuation Accuracy - Lucid.Now, fecha de acceso: noviembre 11, 2025,  
<https://www.lucid.now/blog/how-ai-enhances-dcf-valuation-accuracy/>
51. Unstructured Data and AI: Fine-Tuning LLMs to Enhance the Investment Processes - CFA Institute Research and Policy Center, fecha de acceso: noviembre 11, 2025,  
<https://rpc.cfainstitute.org/sites/default/files/-/media/documents/article/industry-research/unstructured-data-and-ai.pdf>
52. Journal of Artificial Intelligence, Machine Learning and Data Science - urfjournals.org — Virtualmin, fecha de acceso: noviembre 11, 2025,  
<https://urfjournals.org/open-access/quantitative-ai-models-for-company-valuation>

ns.pdf

53. Discovering the sentiment in finance's unstructured data | LSEG, fecha de acceso: noviembre 11, 2025,  
[https://www.lseg.com/content/dam/lseg/en\\_us/documents/white-papers/discovering-sentiment-in-finances-unstructured-data.pdf](https://www.lseg.com/content/dam/lseg/en_us/documents/white-papers/discovering-sentiment-in-finances-unstructured-data.pdf)
54. Building a Pre-Revenue Startup Valuation in 2025 | Finro Financial Consulting, fecha de acceso: noviembre 11, 2025,  
<https://www.finrofca.com/news/pre-revenue-valuation-2025>
55. How to value a pre revenue startup company? - Eqvista, fecha de acceso: noviembre 11, 2025,  
<https://eqvista.com/company-valuation/value-pre-revenue-startup-company/>
56. Five charts showing how AI is dominating the venture fundraising market - Carta, fecha de acceso: noviembre 11, 2025,  
<https://carta.com/data/ai-fundraising-trends-2024/>
57. Valuing Pre-Revenue Startups: It's All About the Future - Texas based venture capital investor in AI, fecha de acceso: noviembre 11, 2025,  
<https://www.sentiero.vc/2025/05/01/valuing-pre-revenue-startups-its-all-about-the-future/>
58. Cross-Validation in Finance, Challenges and Solutions - RiskLab AI, fecha de acceso: noviembre 11, 2025,  
[https://www.risklab.ai/research/financial-modeling/cross\\_validation](https://www.risklab.ai/research/financial-modeling/cross_validation)