Arquitectura multiagente auditable para la estimación del riesgo de fallo clínico en ensayos oncológicos: desarrollo y validación exploratoria de un pipeline secuencial de agentes LLM

Manuela Ibáñez Valenzuela
Grado en Estadística Aplicada, Facultad de Estudios Estadísticos, UCM

INFO

Palabras clave:
Sistemas multiagente; grandes modelos de lenguaje (LLM); inteligencia artificial en salud; ensayos clínicos; predicción de riesgo; auditabilidad; propagación de errores; evaluación de agentes.

ABSTRACT

Este trabajo presenta ORIÓN, un pipeline secuencial de cuatro agentes LLM especializados que estima el riesgo de fallo clínico de ensayos oncológicos de Fase II/III a partir del texto no estructurado de protocolo. Más allá del rendimiento predictivo, el trabajo propone un marco de evaluación transferible para sistemas multiagente que analiza la fidelidad de la comunicación entre agentes, la propagación de errores y la auditabilidad de una arquitectura descompuesta frente a un agente monolítico equivalente. En una evaluación piloto (n=20), ORIÓN superó al agente monolítico (Accuracy 75,0% vs. 65,0%; AUC 0,710 vs. 0,580), aunque esta ventaja no se reprodujo en la muestra ampliada (n=97), donde no se observaron diferencias significativas. La información se conserva con elevada fidelidad a lo largo de la cadena (100,0%→99,3%→97,9%), pero el sistema mostró una capacidad limitada para detectar corrupción introducida en su propio input. La etapa de enriquecimiento basada en una tasa base estimada empíricamente emerge como un componente crítico, mientras que distintas aproximaciones de modelado convergen en un techo predictivo moderado (AUC 0,53–0,61), sugiriendo una limitación intrínseca de la información disponible en el protocolo. El hallazgo más robusto fue un sesgo persistente asociado a la ausencia de comparador, que afecta de forma asimétrica a la clasificación de ensayos fallidos y exitosos. Los resultados deben interpretarse como evidencia exploratoria, dadas las muestras reducidas, la fiabilidad inter-anotador del *ground truth* aún pendiente de cálculo, el uso de un único modelo base por agente y un único dominio de aplicación. En conjunto, ORIÓN aporta evidencia metodológica sobre cómo diseñar, evaluar y auditar arquitecturas multiagente en tareas clínicas, sin plantearlas todavía como herramientas predictivas listas para uso clínico.

1. Introducción

1.1 Contexto: sistemas multiagente LLM y el problema de la auditabilidad

El despliegue de sistemas basados en modelos de lenguaje (LLM) organizados como cadenas de agentes especializados es una arquitectura cada vez más habitual en entornos de producción [1]. Su promesa es la descomposición: cada agente resuelve una subtarea acotada y comunica su salida al siguiente, de forma que el sistema completo resulta —en teoría— más controlable, más diagnosticable y fácil de mejorar por partes que una única llamada monolítica. Sin embargo, esa promesa rara vez se somete a evaluación empírica rigurosa: se reporta el rendimiento agregado del sistema, pero no la fidelidad con la que la información atraviesa cada punto de traspaso, ni la capacidad real de localizar el origen de un error cuando este ocurre, ni si la descomposición aporta algo medible frente a la alternativa monolítica equivalente. Este trabajo aborda exactamente esas preguntas.

El interés es doble. Por un lado, metodológico: cuantificar cómo se conserva —o se distorsiona— la información a lo largo de una cadena secuencial de agentes LLM, y qué mecanismos de detección de anomalías posee (o no posee) el sistema sobre su propio flujo interno de datos. Por otro lado, aplicado: demostrar el marco sobre un problema con relevancia real —la alta tasa de fallo de los ensayos clínicos oncológicos en fase de desarrollo [2, 3]— de forma que el trabajo tenga valor tanto como contribución metodológica como pieza de ingeniería reproducible. El dominio clínico se emplea como caso de estudio, no como objeto central: la contribución que se pretende comunicar es el framework de evaluación, transferible a cualquier pipeline de agentes en producción, y no un hallazgo clínico específico.

1.2 La cadena ORIÓN: cuatro agentes especializados

ORIÓN es un pipeline estrictamente secuencial de cuatro agentes LLM, cada uno con un rol cognitivo distinto y comunicándose exclusivamente a través de la salida estructurada del agente anterior: Alnitak (extracción de características clínicas del texto de protocolo), Alnilam (enriquecimiento con la tasa base empírica de fallo del estrato del ensayo), Mintaka

(estimación calibrada del riesgo de fallo clínico) y Rigel (generación de la explicación final en lenguaje natural para un usuario no técnico). El nombre del proyecto y de los agentes hace referencia a estrellas de la constelación de Orión. La arquitectura no nació con cuatro agentes. El diseño original contemplaba tres (extracción, estimación, explicación); el cuarto agente se incorporó a mitad del desarrollo experimental, al constatarse que la estimación de riesgo sin anclaje numérico empírico producía puntuaciones desconectadas de la probabilidad real de fallo y sesgos sistemáticos difíciles de corregir por prompt (sección 4.2). La reorganización resultante —introducir una etapa dedicada de enriquecimiento con tasa base empírica y un agente final de explicación— mejoró de forma medible la calibración del sistema y constituye, en sí misma, uno de los resultados de diseño del proyecto (secciones 4.4 y 6.4). Todo el proceso de iteración se documenta en un registro de cambios versionado (changelog de agentes) concebido como material de auditabilidad del propio experimento.

2.3 Construcción de la variable dependiente (ground truth)

El campo overall_status no distingue por sí solo entre un fallo clínico real (falta de eficacia, problemas de seguridad, futilidad en un análisis interino) y un cierre por motivos no relacionados con el desempeño clínico del tratamiento (financiación, reclutamiento insuficiente, decisiones de negocio, motivos regulatorios o de suministro). Confundir ambos tipos de cierre bajo una única etiqueta de "fallo" introduciría ruido sustancial en la variable dependiente. Por ello, los 3.869 ensayos Terminated se reclasifican en dos etapas a partir del texto libre del campo why_stopped.

Estado final (overall_status)	N
Completed	14.173
Terminated	3.869
Withdrawn	1.219
Suspended	115

Tabla 1. Distribución de la cohorte oncológica Fase II/III por estado final del ensayo, tras el filtrado inicial sobre AACT.

Etapa A — filtrado por reglas. Un diccionario de patrones léxicos (financiación, decisión del sponsor, reclutamiento, suministro del fármaco, motivos regulatorios) se aplica sobre el texto en minúsculas de why_stopped para preasignar categorías de alta confianza y reducir el volumen antes de la revisión más costosa. Dentro de la categoría administrativa, los motivos de reclutamiento no se colapsan en una etiqueta única: se distinguen explícitamente las subcategorías administrative_recruitment, administrative_enrollment y administrative_accrual, registradas por separado tanto en el filtro por reglas como en el prompt de clasificación de la Etapa B y en la plantilla de validación manual, dado que los tres términos, aunque relacionados, no son sinónimos en el lenguaje de protocolos. Tras varias iteraciones de refinamiento del diccionario —validadas manualmente sobre muestras sucesivas— la distribución resultante se muestra en la Tabla 2.

Categoría (Etapa A)	N	% s/ Terminated
Administrative	**2.192**	**56,7%**
Clinical candidate	1.248	32,3%
Unclear / missing	419	10,8%
Regulatory	10	0,3%

Tabla 2. Resultado del filtrado por reglas (Etapa A) sobre los ensayos Terminated, tras la última iteración del diccionario de patrones.

Etapa B — clasificación fina asistida por LLM. Los casos etiquetados como clinical_candidate y unclear_or_missing en la Etapa A (≈1.667 casos) se someten a clasificación individual con un LLM (Claude Sonnet, Anthropic), mediante un prompt de cinco categorías (clinical_failure, administrative, regulatory, design_feasibility, unclear) que incorpora ejemplos concretos extraídos del propio proceso de calibración. La categoría design_feasibility se incorporó tras observar en la validación manual casos en los que el ensayo se detenía no por fallo del tratamiento ni por decisión de negocio, sino porque los datos de cribado revelaban una prevalencia poblacional del biomarcador diana menor de la esperada, haciendo inviable completar el diseño original —una categoría de cierre distinta tanto del fallo clínico como del cierre administrativo. El proceso guarda cada clasificación de forma incremental (checkpointing línea a línea), permitiendo reanudarlo sin repetir llamadas ya completadas en caso de interrupción.

3. Arquitectura del pipeline ORIÓN

3.1 Roles y contrato de comunicación

Alnitak (extracción). Recibe el texto no estructurado del protocolo (resumen breve, descripción detallada, criterios de elegibilidad) y produce una representación estructurada en JSON de cinco características clínicamente relevantes: endpoint primario, población y biomarcadores, tipo de comparador, línea de tratamiento y criterios de inclusión/exclusión clave. Cada campo lleva una etiqueta de fuente (explicit, inferred, absent) que preserva la trazabilidad de qué se extrajo literalmente y qué se dedujo del contexto.

Alnilam (enriquecimiento). Recibe las características de Alnitak y las enriquece con la tasa base empírica de fallo clínico del estrato del ensayo, calculada sobre la propia cohorte mediante un procedimiento leave-one-out (el ensayo evaluado se excluye del cálculo de su propia tasa base), junto con un análisis contextual del estrato.

Mintaka (estimación de riesgo). Recibe exclusivamente la salida estructurada de Alnilam —no el texto original— y produce un risk_score definido como probabilidad derivada matemáticamente de la tasa base: el agente justifica un multiplicador acotado (rango

0,3x–3x) aplicado sobre la tasa base empírica, de forma que el score final es auditable aritméticamente (risk_score = tasa_base × multiplicador) y el campo multiplicador_aplicado queda expuesto en la salida junto con las variables más influyentes. Dado que la tasa base típica del estrato es baja (3–5%), el umbral de decisión se define de forma relativa: failure_probable si risk_score ≥ tasa_base × 1,5, con un suelo absoluto de 0,15.

Rigel (explicación). Recibe la puntuación y las variables influyentes de Mintaka y redacta una explicación en lenguaje natural del nivel de riesgo estimado, destinada a un usuario final no técnico (por ejemplo, un equipo de evaluación de oportunidades de licencia en una compañía farmacéutica).

3.2 Decisiones de configuración y registro

Los agentes se ejecutan sobre un mismo modelo base (Claude Haiku, Anthropic) con temperature=0 como configuración principal, decisión motivada por la detección temprana de varianza estocástica entre llamadas repetidas sobre el mismo caso, y validada posteriormente mediante un estudio dedicado de ruido de fondo (sección 6.5). Cada llamada a cada agente se registra de forma completa (entrada, salida, marca temporal, consumo de tokens), lo que permite el análisis de fidelidad de información y atribución de error de las secciones 5 y 6 sin coste de API adicional. Los errores de formato de salida (envoltura del JSON en bloques markdown) se tratan como un modo de fallo distinto de los errores de contenido y se resuelven en post-procesamiento determinista, no iterando el prompt.

Para la ejecución a escala, las llamadas se paralelizaron con hasta 6 llamadas simultáneas en vuelo, tras validar sobre los mismos 20 casos del piloto que el modo concurrente reproduce el secuencial (diferencia máxima de risk_score por caso: 0,02, atribuible a variabilidad residual de la API incluso a temperature=0). La aceleración medida sobre timestamps reales fue de 5,6x, lo que además reduce la ventana de exposición a interrupciones externas (límites de uso de API) en corridas largas

4. Calibración iterativa de los agentes.

Todo el proceso de iteración de prompts se documentó como un registro versionado con estructura fija por entrada: qué fallaba, el caso concreto que lo evidenció, la causa raíz diagnosticada, el cambio exacto aplicado y el resultado tras el cambio. Esta sección resume los episodios con valor metodológico; el registro completo forma parte del material del proyecto.

4.1 Alnitak: de la extracción literal a la inferencia trazable

La primera versión del prompt de extracción prohibía cualquier inferencia ("no inventes ni infieras"), lo que produjo un 100% de valores nulos en endpoint_primario sobre los casos de prueba: en el lenguaje real de protocolos, el endpoint frecuentemente está implícito en el objetivo general del estudio sin terminología formal. El rediseño (v2) permitió inferencia razonable manteniendo auditabilidad mediante la etiqueta de fuente explicit/inferred/absent, con la restricción de que la inferencia debe apoyarse en el texto, nunca en conocimiento externo sobre el fármaco. Una tercera iteración amplió la definición de "comparador" para incluir diseños dosis-respuesta, corrigiendo tres casos discrepantes —dos de los cuales se habían diagnosticado inicialmente como "zona gris de interpretación irreducible" y resultaron ser, en realidad, una definición de prompt demasiado estrecha.

Las métricas finales de la versión estable, evaluadas frente a un gold-standard de 50 ensayos anotado manualmente, superan el 94% de coincidencia de presencia/ausencia en todos los campos (endpoint 96,0%; tamaño muestral 98,0%; población 100,0%; línea de tratamiento 94,0%; comparador 96,0%). Dos incidencias de este proceso merecen mención explícita como aprendizajes transferibles: (i) cambiar el criterio de un agente (permitir inferencia) invalida retroactivamente las comparaciones contra un gold-standard fijado bajo el criterio anterior, que debe re-versionarse igual que el agente; y (ii) errores en el propio código de evaluación pueden simular fallos del agente —un primer cálculo de coincidencia para "comparador" arrojó un 40% aparente porque la métrica no reconocía todas las formulaciones textuales de ausencia; corregida la métrica, la coincidencia real era del 90%.

Adicionalmente, el campo de tamaño muestral se eliminó del esquema de extracción al detectarse que su ausencia sistemática en el texto libre (el dato existe en AACT, pero en un campo estructurado aparte) estaba siendo tratada aguas abajo como señal de riesgo en el 100% de los casos —un artefacto de qué campos se pasan al pipeline, no una señal clínica—, comprimiendo el rango de risk_score y reduciendo la capacidad discriminativa.

4.2 El episodio del sesgo de "comparador ausente" (resultado negativo documentado)

Sobre una muestra piloto de 50 casos balanceados, la primera versión del agente de estimación obtuvo Accuracy 59,2% / AUC 0,590, con baja especificidad: "ausencia de comparador explícito" aparecía como factor de riesgo en 8 de 8 falsos positivos revisados —un sesgo sistemático, dado que los diseños de brazo único son comunes y metodológicamente normales en oncología temprana. Dos intervenciones de prompt sucesivas fracasaron en direcciones opuestas: instruir que la ausencia de comparador no es riesgo por defecto (v2) produjo sobrecorrección (el modelo pasó a tratarla como señal protectora, un salto lógico injustificado) y degradó el rendimiento a 49,0%/0,450; reformularla como estrictamente neutral (v3) lo degradó aún más

(42,0%/0,418). Se revirtió a v1 —la versión con mejor desempeño empírico pese a tener la justificación textual más cuestionable— y se extrajo el aprendizaje central del episodio: **el razonamiento textual explícito de un LLM no necesariamente correlaciona con la calidad de su predicción final**; una explicación plausible no garantiza ser la explicación causal real del comportamiento del modelo. Intentos posteriores (rúbrica de calibración con bandas numéricas explícitas; un campo adicional de intención de diseño en la extracción para desambiguar el brazo único deliberado) tampoco superaron a v1. A partir de la cuarta iteración se adoptó una partición dev/holdout (20/30 casos) para evitar el sobreajuste por consulta repetida del mismo conjunto de evaluación, y la cifra definitiva del estimador LLM zeroshot se declaró sobre el holdout nunca visto: **Accuracy 56,7%, AUC 0,527** —notablemente más modesta que las cifras de las iteraciones previas, caída consistente con la hipótesis de que parte del rendimiento aparente reflejaba ajuste involuntario al conjunto repetidamente consultado.

4.3 Techo epistémico de la tarea: triangulación con modelo de caja blanca

Para determinar si el techo observado era específico del razonamiento LLM o una propiedad de la información disponible, se entrenó un modelo alternativo completamente interpretable (regresión logística con regularización L1) sobre las mismas features extraídas por Alnitak, con evaluación por validación cruzada de 5 particiones (Tabla 3)

Método	Información utilizada	AUC
LLM zero-shot (holdout)	Features de Alnitak	0,527
Lasso, iter. 1	Solo texto de Alnitak	0,541
Lasso, iter. 2	+ sponsor (AACT)	0,595 ± 0,034
Lasso, iter. 3	+ semántica del texto	0,605 ± 0,018

Tabla 3. Triangulación del techo epistémico: cuatro aproximaciones de modelado independientes convergen en la banda AUC 0,53–0,61. El contenido semántico de la iteración 3 incluye biomarcador, enfermedad avanzada/metastásica, estado refractario y función orgánica, extraídos por palabras clave del texto de Alnitak.

Cuatro aproximaciones completamente distintas — un LLM sin entrenamiento alguno y tres iteraciones de un modelo lineal con selección de variables— convergen en la misma banda estrecha de AUC 0,53–0,61. Esta convergencia entre métodos heterogéneos es una evidencia considerablemente más sólida de la existencia de un techo genuino que cualquiera de los resultados individuales: se interpreta que la información de diseño de protocolo disponible en texto (y en las variables estructurales exploradas), sin datos de eficacia real, es insuficiente por sí sola para predecir con alta precisión el fallo clínico —una limitación de la tarea tal como está planteada, no de la implementación de ningún agente concreto. Como hallazgo bivariado destacado, el tipo de sponsor mostró la asociación marginal más fuerte del proyecto (58,3% de tasa de fallo en ensayos de industria frente a 35,2% en consorcios académicos/públicos grandes), pero su contribución incremental en el modelo multivariable fue modesta (AUC solosponsor: 0,546 ± 0,044) —ilustrando la diferencia entre asociación marginal y aporte predictivo real.

4.4 Reorganización a cuatro agentes: anclaje en tasa base empírica

El techo del estimador zero-shot y un fallo de calibración detectado en la primera prueba del pipeline ampliado —el agente mencionaba la tasa base en su razonamiento pero el risk_score final no guardaba relación aritmética alguna con ella— motivaron la redefinición del score como probabilidad derivada de la tasa base (sección 3.1) y la reorganización a la arquitectura de cuatro agentes. En la primera validación funcional, el modelo reportó risk_score=0,0618 con multiplicador 1,5 sobre tasa base 0,0412 —coherencia aritmética exacta—, confirmando el anclaje numérico. Este mismo periodo dejó registrado un episodio real de propagación de error con valor ilustrativo para el diseño experimental: un fallo mecánico en una etapa intermedia (respuesta truncada por límite de tokens, más envoltura markdown no depurada) produjo una aparente violación de fidelidad en la etapa siguiente, que "citaba" un número que la interfaz de datos limpia no debería haberle permitido ver. La fidelidad narrativa de un agente depende críticamente de la robustez mecánica de las etapas anteriores, no solo de sus propias instrucciones.

5. Hipótesis y diseño experimental

La hipótesis central es que un pipeline secuencial de agentes con puntos de traspaso explícitos y auditables mantiene paridad de rendimiento predictivo frente a un agente monolítico equivalente, a cambio de ganar capacidad de localizar y diagnosticar el origen de los errores. De ella se derivan cuatro sub-hipótesis falsables:

- **H1 (paridad de rendimiento):** el AUC del pipeline no difiere significativamente del de un agente monolítico realizando la tarea completa en una única llamada.
- **H2 (pérdida de información creciente con la profundidad):** la fidelidad de la información decae de forma medible y monotónica a lo largo de la cadena.
- **H3 (localización de error):** ante la inyección controlada de errores en la salida de Alnitak, es posible atribuir con precisión medible en qué etapa se originó la degradación.
- **H4 (necesidad del paso intermedio):** la eliminación de la etapa de enriquecimiento degrada

significativamente el rendimiento frente a la cadena completa.

Las condiciones comparadas son: (a) pipeline completo de cuatro agentes; (b) agente monolítico con el mismo input informativo (texto de protocolo + tasa base empírica del estrato, calculada con idéntico procedimiento leave-one-out) y la misma fórmula de score; (c) ablación de la etapa de enriquecimiento (conexión directa de la extracción a la estimación, con objeto de enriquecimiento vacío); (d) inyección de error controlada sobre la salida de Alnitak antes de su traspaso; y (e) condición opcional de contexto reducido. Las comparaciones de AUC se realizan mediante el test de DeLong [7, 8]; las de accuracy pareada mediante el test de McNemar [9]; las comparaciones pareadas de magnitud mediante el test de Wilcoxon [10]; y las proporciones con intervalos de Wilson [11].

6. Resultados preliminares

Los resultados se presentan en el orden en que se generó la evidencia —desde el hallazgo inicial en muestra piloto hasta su contraste en muestra ampliada— para reflejar fielmente el proceso de validación seguido.

Muestra	Acc. pipeline	Acc. monol.	AUC pipe./monol.
n=20, secuencial	75,0%	65,0%	0,710 / 0,580
n=20, concurrente (verif.)	75,0%	—	0,730 / —
n=97, intersección pareada	66,0%	66,0%	0,676 / 0,634
Significancia (n=97)	McNemar p=1,000	—	DeLong p=0,415

Tabla 4. Comparación H1 en las tres muestras evaluadas: piloto (n=20), piloto en modo concurrente de verificación (n=20) y muestra ampliada con intersección pareada (n=97).

6.1 H1 — Pipeline de cuatro agentes vs. agente monolítico

En una primera evaluación piloto sobre 20 casos balanceados, el pipeline alcanzó Accuracy 75,0% y AUC 0,710, frente a 65,0% y 0,580 del monolítico sobre los mismos casos, mismo input informativo y mismo umbral. El monolítico mostró sensibilidad perfecta pero especificidad muy baja (predijo "fallo" en 17 de 20 casos), un patrón de sobre-predicción similar al de las versiones tempranas no calibradas del estimador —sugiriendo que la descomposición ayuda a matizar cuándo un caso se aleja de la tasa base hacia el lado protector. La diferencia, sin embargo, no fue significativa (McNemar exacto p=0,625; solo 4 casos discordantes), y motivó la ampliación antes de considerarla concluyente. Cabe señalar que el umbral de esta evaluación piloto se calibró por índice de Youden [12] sobre los mismos 20 casos, una circularidad que probablemente infla el accuracy reportado y de la que se dejó constancia expresa para no reportarlo como cifra definitiva. Al escalar a 145 casos, una interrupción por límite de uso de la API dejó al monolítico con 98 casos completos; para mantener una comparación pareada válida, el análisis se restringió a la intersección exacta de casos con éxito en ambas condiciones (n=97). Sobre esta muestra, la Accuracy resultó idéntica (66,0% = 66,0%) y el AUC mostró una ventaja más modesta para el pipeline (0,676 frente a 0,634). Ninguna diferencia alcanzó significancia: McNemar p=1,000 (10 casos discordantes, repartidos exactamente 5/5) y DeLong p=0,415 (Tabla 4). De los 97 casos, ambas condiciones coinciden en su predicción en 87 (89,7%), lo que sugiere que la arquitectura determina el resultado solo en una fracción minoritaria de casos: la mayoría de aciertos y fallos parecen gobernados por la dificultad intrínseca del caso. La lectura conjunta: la señal inicial en n=20 fue más pronunciada de lo que la muestra ampliada sostuvo —un patrón de regresión hacia una diferencia más modesta consistente con lo observado en la calibración del estimador (dev vs. holdout). Con los datos disponibles, la evidencia es compatible con paridad de rendimiento entre arquitecturas en Accuracy y AUC agregados, si bien el pipeline mantuvo en todas las muestras un AUC igual o superior al monolítico. De existir una ventaja real de la descomposición, habría que buscarla en dimensiones distintas al rendimiento predictivo puro —auditabilidad, localización de error, trazabilidad— que abordan H2 y H3. Como dato de coste relevante para la comparación arquitectónica, el monolítico requirió más del doble de tokens de salida por llamada (1.400 frente a 500–650 por agente individual), al concentrar en una respuesta lo que el pipeline reparte en cuatro.

6.2 H2 — Fidelidad de información a lo largo de la cadena

Sobre los 144 casos exitosos del pipeline, se evaluó la fidelidad en cada traspaso mediante chequeos deterministas sobre los registros completos: (1) si la tasa base que reporta la etapa de enriquecimiento coincide con la efectivamente proporcionada; (2) si el risk_score es aritméticamente coherente con tasa_base × multiplicador; (3) si la cifra citada en la explicación final coincide con el risk_score real, en su formulación de riesgo o en su complementaria de éxito (Tabla 5). El patrón es monotónicamente decreciente, consistente con H2, con la mayor pérdida concentrada en el último traspaso. El propio proceso de medición dejó un hallazgo metodológico: una primera versión del chequeo del tercer traspaso, que solo aceptaba coincidencia con risk_score × 100, clasificó como "infiel" un 14,6% de los casos; la inspección manual reveló que la mayoría eran matemáticamente correctos —la explicación citaba la probabilidad de éxito complementaria (1−risk_score), una reformulación válida, no un error—, y un caso adicional resultó ser una cifra de contexto clínico del propio texto confundida por

la expresión regular con una cita del riesgo. Tras ambas correcciones, la fidelidad real fue del 97,9% (141 de 144). Los 3 casos genuinamente infieles comparten un mismo patrón mecánico: un risk_score de 0,0685 citado como "68,5%" —un error sistemático de escala (interpretar el score en base 0–100 en lugar de 0–1), no una alucinación de contenido ni una cifra inventada.

Traspaso	Fidelidad
Alnitak → Alnilam (tasa base transcrita)	100,0%
Alnilam → Mintaka (coherencia aritmética)	99,3%
Mintaka → Rigel (cifra citada, tras corrección)	97,9%

Tabla 5. Fidelidad por traspaso en la cadena de cuatro agentes (n=144)

6.3 H3 — Localización de error mediante inyección controlada

Sobre 25 casos con extracción ya validada se generaron dos condiciones de corrupción reutilizando la salida original de Alnitak: leve (criterios clave forzados a ausente) y severa (el campo comparador se invierte: se elimina si tenía valor real, o se fabrica un valor ficticio de "placebo control" si estaba ausente).

Resultado 1 — la magnitud no escala con la severidad. El delta medio de risk_score fue prácticamente idéntico entre corrupción leve (0,0123) y severa (0,0134); la severa superó a la leve en solo 11 de 24 casos (46%, compatible con azar). Wilcoxon pareado: p=0,616, no significativo. Tomado de forma aislada, este resultado no confirma H3 tal como se formuló.

Resultado 2 — el sistema no detecta la corrupción: la absorbe silenciosamente. Un primer chequeo automático (presencia de palabras clave como "comparador" o "ausente") reportó una aparente detección del 100% tras la corrupción severa. La inspección manual reveló un falso positivo del método de medición: en el caso examinado, el agente incorporó el comparador fabricado a su razonamiento con total normalidad —incluso citándolo como factor protector— sin ninguna señal de sospecha; la palabra "comparador" aparecía simplemente porque es una de las dimensiones que el agente evalúa siempre. Corregido el chequeo para buscar lenguaje explícito de inconsistencia ("inconsistente", "no coincide", "atípico", "discrepancia"), la tasa real de detección fue del 4,0% (1 de 25 casos).

Ambos resultados son consistentes entre sí: si el sistema no detecta la corrupción como anómala, no existe mecanismo por el cual debiera reaccionar con mayor fuerza ante una corrupción más severa —ambas magnitudes se procesan igual porque, desde la perspectiva del agente, ambas son simplemente "el input recibido". Esto redefine H3: la pregunta relevante no es cuánto se propaga el error según su severidad, sino si el pipeline posee algún mecanismo de detección de anomalías en su propio input —y la respuesta, con esta evidencia, es que prácticamente no lo posee. Un eslabón corrompido (por fallo de extracción, ataque adversarial o simple ruido) se propaga aguas abajo sin que el sistema emita ninguna señal de alerta interna: una limitación de auditabilidad genuina y directamente relevante para la discusión de riesgos de despliegue en producción.

6.4 H4 — Necesidad del paso intermedio: ablación de Alnilam

Sobre los mismos 25 casos de H3, se conectó la extracción directamente a la estimación, sustituyendo el enriquecimiento por un objeto vacío —sin tasa base ni análisis contextual (Tabla 6).

El delta medio de risk_score entre condiciones fue 0,0744 (máximo 0,295) —sustancialmente mayor que los deltas de H3 (0,012–0,013). McNemar p=0,070; DeLong p=0,075: ninguno alcanza el umbral convencional de significancia, pero ambos se sitúan notablemente cerca, a diferencia del patrón de H1 (p=1,000 y p=0,415). Dado el tamaño de muestra (n=25, heredado de H3 por diseño para mantener comparabilidad), se interpreta este resultado como limitado por potencia estadística más que como ausencia de efecto: la magnitud práctica (AUC cruzando por debajo de 0,5 sin la etapa de enriquecimiento) es considerable y consistente en ambas métricas, y el hallazgo es coherente con la evidencia cualitativa de la calibración —la etapa se diseñó precisamente porque la estimación sin anclaje empírico mostraba sesgos sistemáticos. Se requiere replicación sobre muestra mayor antes de reportarlo como resultado definitivo

Métrica	Con Alnilam	Sin Alnilam
Accuracy	60,0%	36,0%
AUC	0,667	0,438 (< azar)

Tabla 6. Comparación con y sin el paso de enriquecimiento de Alnilam (n=25).

6.5 Cuantificación del ruido de fondo del modelo

Como validación del compromiso de fijar temperature=0, se ejecutaron 12 casos, cada uno 5 veces, a temperatura por defecto y sin manipulación experimental alguna. La desviación estándar media del risk_score dentro de un mismo caso fue 0,0109 (rango máximo observado en un caso: 0,0577). La predicción binaria cambió entre repeticiones del mismo caso en 3 de 12 casos (25%), puramente por azar de generación (Tabla 7).

Esta cuantificación reinterpreta H3 y refuerza H4: la ausencia de gradiente entre severidades en H3 se explica, al menos parcialmente, porque ambos efectos son del mismo orden que la variabilidad normal de generación —no hay margen para que se distingan de ella—, mientras que el efecto de la ablación es casi 7 veces mayor que el ruido, apoyando su interpretación

como real. El hallazgo de que una cuarta parte de los casos cambia de categoría de decisión sin cambio alguno en el input es, en sí mismo, relevante para la fiabilidad de despliegue: sugiere que, en un entorno real, las decisiones cercanas al umbral deberían tratarse con cautela adicional (por ejemplo, mediante múltiples muestreos y agregación en lugar de una única llamada).

6.6 Fidelidad narrativa completa y confianza narrativa como predictor

Más allá de la cifra numérica, se evaluó si el contenido de la explicación final es consistente con las variables realmente influyentes generadas por la etapa de estimación, mediante solapamiento léxico sobre los 144 casos. Los cinco casos con menor solapamiento (9,2%–23,3%) fueron revisados manualmente: en ninguno se encontró contenido fabricado —todos eran paráfrasis legítimas con sinónimos accesibles para un lector no técnico, tal como instruye el propio prompt del agente, penalizadas artificialmente por el método de medición léxica (en un caso, además, por identificadores en snake_case frente a lenguaje natural: una diferencia de formato, no de contenido). Este es el tercer episodio del proyecto en que un chequeo automatizado por coincidencia textual requirió corrección mediante inspección manual antes de poder reportarse con confianza (sección 7.2).

Adicionalmente, se evaluó si la "confianza narrativa" (longitud del análisis contextual, longitud media del razonamiento por variable, balance de léxico de certeza frente a duda) predice el acierto de la predicción final. Ninguno de los tres proxies mostró correlación significativa con el acierto ($r=0,005$ a $r=0,030$; $p=0,72$ a $p=0,95$), y el balance de certeza resultó cercano a cero en ambos grupos: los agentes tienden a un registro consistentemente asertivo con independencia de la incertidumbre real de la estimación subyacente. La evidencia cuantitativa confirma así la observación cualitativa del episodio de calibración: **la elocuencia o el tono asertivo de una explicación generada por un agente LLM no es un indicador fiable de la corrección de su predicción**, y no debe utilizarse como señal de confianza sin validación externa adicional.

6.7 El determinante real del acierto a nivel de caso individual

Descartados los proxies narrativos, se analizó si el error se concentra en subgrupos estructurales (Tabla 8). El hallazgo más fuerte del proyecto surgió del desglose por resultado real: accuracy del 84,9% en ensayos que efectivamente fallaron frente a solo 45,1% en ensayos que se completaron con éxito —una asimetría de sensibilidad/especificidad de casi 40 puntos. La revisión manual de 8 casos fallidos y su confirmación sobre el conjunto completo revelaron la causa: el 59,0% de los falsos positivos carecen de comparador, frente al 25,0% de los verdaderos negativos —más del doble de frecuencia. El determinante más fuerte del acierto identificado en este proyecto es, por tanto, la persistencia del sesgo de "comparador ausente = riesgo" documentado extensamente durante la calibración (sección 4.2): a pesar de las mejoras sustanciales del pipeline final (tasa base empírica, agente de explicación dedicado), este sesgo específico nunca fue corregido del todo —solo diluido parcialmente por el anclaje empírico— y sigue explicando una parte sustancial y cuantificable del error residual, concentrada en ensayos de brazo único que en la práctica se completan con éxito. Un intento adicional de corrección (reintroducción de un campo de clasificación de intención de diseño, integrado con la arquitectura de tasa base) fue diseñado e implementado, pero no se completó su validación en esta fase, dado el historial de dos intentos previos de corrección de este mismo sesgo que empeoraron el rendimiento; queda documentado como línea de trabajo futura prioritaria, con la advertencia expresa de que cualquier resultado positivo en muestra de desarrollo deberá validarse en un holdout independiente antes de reportarse como mejora confirmada.

Variable	Categoría	Accuracy
Fase	Fase II	61,7% (n=107)
Fase	Fase III	75,0% (n=32)
Sponsor	Industria	58,1% (n=43)
Sponsor	Académico/público grande	78,6% (n=14)
Label real	Completado (0)	45,1% (n=71)
Label real	Fallo clínico (1)	84,9% (n=73)

Tabla 8: Accuracy del pipeline por subgrupos estructurales (n=144). La asimetría por resultado real del ensayo (casi 40 puntos) es muy superior a cualquier otro efecto medido en el proyecto.

7. Discusión

7.1 Paridad predictiva, ventaja de auditabilidad

La evidencia disponible es compatible con la formulación más conservadora de la hipótesis central: el pipeline de agentes especializados no supera al monolítico en rendimiento predictivo agregado, pero tampoco lo empeora, y a cambio ofrece algo que el monolítico no puede ofrecer por construcción —cada uno de los hallazgos de las secciones 6.2 a 6.7 (fidelidad por traspaso, error de escala sistemático, tasa real de detección de anomalías, contribución de cada etapa, concentración estructural del error) solo es medible porque los puntos de traspaso son explícitos y están registrados. La auditabilidad no es un subproducto del pipeline: es su producto. Al mismo tiempo, H3 delimita con precisión el alcance de esa ventaja: la arquitectura permite al evaluador externo localizar y atribuir errores, pero no dota al sistema de capacidad alguna de detectarlos por sí mismo —auditabilidad no es autovigilancia.

7.2 Los chequeos automáticos requieren verificación manual: un patrón recurrente

En tres ocasiones independientes (probabilidad complementaria no reconocida en H2; falso positivo de detección por palabras clave en H3; paráfrasis legítima penalizada por solapamiento léxico en la fidelidad narrativa), un chequeo automatizado basado en coincidencia de superficie textual produjo conclusiones erróneas que solo la inspección manual de casos concretos permitió corregir —en las tres, además, en la dirección de subestimar la fidelidad real del sistema. El patrón es consistente: los métodos de detección basados en superficie (regex, palabras clave, solapamiento léxico) son propensos tanto a falsos positivos como a falsos negativos cuando evalúan contenido generado por LLM, que varía legítimamente en formulación mientras preserva el significado. Se recomienda, como práctica general para la evaluación de fidelidad en sistemas multiagente, complementar los chequeos automáticos de superficie con métodos de similitud semántica o con muestreo de verificación manual sistemático, y no reportar métricas de coincidencia textual sin validación cualitativa.

7.3 Los límites de la interpretabilidad narrativa

Dos resultados independientes convergen en la misma conclusión: una heurística con justificación textual cuestionable superó empíricamente a dos versiones con razonamiento más matizado y defendible (sección 4.2), y ningún proxy de confianza narrativa correlacionó con el acierto (sección 6.6). Es posible que "ausencia de comparador" actúe como proxy indirecto de otras características del diseño correlacionadas con el resultado real (fase, tipo de patrocinador), no como mecanismo causal directo —el modelo podría estar acertando por razones distintas a las que articula. La explicación que un agente LLM da de su propia predicción debe tratarse como una hipótesis sobre su comportamiento, no como su descripción.

7.4 Implicaciones para el despliegue en producción

Tres hallazgos tienen traducción operativa directa. Primero, la ausencia de detección de anomalías en el input (4,0%) implica que cualquier despliegue serio de una cadena de agentes necesita mecanismos de validación externos a los propios agentes —chequeos deterministas en los traspasos, como los aquí empleados, o agentes verificadores dedicados. Segundo, la inestabilidad de clasificación cerca del umbral (25% de flips a temperatura por defecto) desaconseja decisiones binarias basadas en una única llamada. Tercero, el patrón de fallo dominante de fidelidad (error de escala 0–1 vs. 0–100) es mecánico y detectable de forma determinista, lo que sugiere que una parte relevante de la infidelidad observable en estos sistemas es prevenible con contratos de interfaz más estrictos, no con mejores prompts.

7.5 Limitaciones

- Tamaños muestrales reducidos en varios experimentos (n=20–145; H3/H4 sobre n=25), con potencia estadística limitada declarada explícitamente donde corresponde.
- Fiabilidad inter-anotador del ground truth (kappa de Cohen sobre 50–80 casos, con anotación ciega e independiente) pendiente de ejecución.
- El umbral de decisión de la evaluación piloto de H1 se calibró sobre los mismos casos evaluados (circularidad reconocida); la cifra pareada de n=97 no está afectada por comparación entre condiciones, pero el nivel absoluto de accuracy debe leerse con cautela.
- Un único modelo base por agente y una única familia de modelos: los resultados de fidelidad y sesgo podrían no transferirse a otros modelos sin reevaluación.
- Un único dominio de aplicación (oncología, Fase II/III); la transferibilidad del framework es una afirmación de diseño, no un resultado demostrado empíricamente en otros dominios.
- El intento de corrección v6 del sesgo de compara dor está implementado pero no validado; el análisis formal de coste y latencia como dimensión arquitectónica está pendiente de formalización

8. Conclusiones

- Un pipeline secuencial de cuatro agentes LLM con traspasos explícitos mantiene paridad de rendimiento predictivo frente a un agente monolítico equivalente (n=97; McNemar p=1,000; DeLong p=0,415), con un AUC igual o superior en todas las muestras evaluadas.
- La fidelidad de la información decae de forma monotónica a lo largo de la cadena (100,0% → 99,3% → 97,9%), y el fallo dominante es mecánico (error de escala), no una alucinación de contenido.
- El pipeline no detecta corrupción inyectada en su propio input (4,0% de detección real): la auditabilidad de la arquitectura beneficia al evaluador externo, no al sistema.
- La eliminación de la etapa de enriquecimiento con tasa base empírica degrada el rendimiento de forma sustancial (AUC 0,667 → 0,438; efecto 6,8 veces mayor que el ruido de fondo),

aunque la significancia formal queda limitada por el tamaño de muestra.

- Cuatro aproximaciones de modelado independientes convergen en un techo epistémico de AUC 0,53–0,61: la información de diseño de protocolo, sin datos de eficacia, es insuficiente para predecir con alta precisión el fallo clínico —un límite de la tarea, no de la implementación.

- El determinante más fuerte del error residual es la persistencia del sesgo de "comparador ausente = riesgo" (59,0% de falsos positivos sin comparador vs. 25,0% de verdaderos negativos), nunca corregido del todo por ninguna de las intervenciones probadas.

- La elocuencia y el tono asertivo de las explicaciones generadas no correlacionan con el acierto; los chequeos automáticos de fidelidad por superficie textual requieren verificación manual sistemática antes de reportarse.

9. Líneas de desarrollo futuro

Los próximos pasos del proyecto, en orden de prioridad, son los siguientes. En el plano del ground truth, la ejecución de la validación inter-anotador ciega e independiente (kappa de Cohen sobre 50–80 casos) es el requisito previo a cualquier lectura confirmatoria de los resultados. En el plano experimental, la replicación de H4 sobre una muestra ampliada con partición dev/holdout —siguiendo el mismo procedimiento aplicado a H1— permitiría resolver la cuestión de potencia que hoy limita ese resultado, y la validación en holdout independiente del diseño v6 (campo de intención de diseño integrado con el multiplicador sobre tasa base) es la línea prioritaria para atacar el sesgo de comparador, con la salvaguarda metodológica ya declarada frente al historial de sobrecorrección. En el plano del framework, se plantea sustituir los chequeos de fidelidad por superficie textual por métodos de similitud semántica, incorporar mecanismos explícitos de detección de anomalías en los traspasos (chequeos deterministas de rango y consistencia, o un agente verificador dedicado), explorar el muestreo múltiple con agregación para decisiones cercanas al umbral, y formalizar el análisis de coste y latencia como dimensión de comparación arquitectónica. Finalmente, la incorporación del tamaño muestral estructurado de AACT como covariable complementaria —fuera del texto libre— y la demostración del framework sobre un segundo dominio no clínico completarían la validación de su transferibilidad, que en esta fase es una afirmación de diseño.

Referencias

[1] Guo T, Chen X, Wang Y, et al. Large language model based multi-agents: a survey of progress and challenges. Proceedings of IJCAI. 2024.

[2] Wong CH, Siah KW, Lo AW. Estimation of clinical trial success rates and related parameters. Biostatistics. 2019;20(2):273–286.

[3] Hay M, Thomas DW, Craighead JL, Economides C, Rosenthal J. Clinical development success rates for investigational drugs. Nature Biotechnology. 2014;32:40–51.

[4] Tasneem A, Aberle L, Ananth H, et al. The database for aggregate analysis of ClinicalTrials.gov (AACT) and subsequent regrouping by clinical specialty. PLoS ONE. 2012;7(3):e33677.

[5] Clinical Trials Transformation Initiative (CTTI). AACT Database. aact.ctti-clinicaltrials.org. Snapshot de 19 de agosto de 2026.

[6] Cohen J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement. 1960;20(1):37–46.

[7] DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988;44(3):837–845.

[8] Sun X, Xu W. Fast implementation of DeLong's algorithm for comparing the areas under correlated receiver operating characteristic curves. IEEE Signal Processing Letters. 2014;21(11):1389–1393.

[9] McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika. 1947;12(2):153–157.

[10] Wilcoxon F. Individual comparisons by ranking methods. Biometrics Bulletin. 1945;1(6):80–83.

[11] Wilson EB. Probable inference, the law of succession, and statistical inference. Journal of the American Statistical Association. 1927;22(158):209–212.

[12] Youden WJ. Index for rating diagnostic tests. Cancer. 1950;3(1):3