

Informe sobre la evaluación de modelos de IA para soporte al diagnóstico clínico

Análisis metodológico de los pipelines v0-v4 y el fenómeno de la
saturación de la tarea

Autores: [Nombres de los Autores]

Institución: [Nombre de la Institución / Fundación]

Fecha: 16 de julio de 2025

Versión del Documento: 5.0

*Este documento presenta un análisis pormenorizado de los hallazgos y la evolución de los marcos de
evaluación para herramientas de diagnóstico asistido por IA, abarcando desde estudios clínicos iniciales
hasta el desarrollo de pipelines automatizados de alta fidelidad.*

Índice

1. Planteamiento del problema de evaluación	2
2. Composición del entorno de pruebas	2
2.1. Análisis exploratorio y diseño del scoring	3
2.2. Diversidad del dataset final	4
3. Evolución de los pipelines de evaluación	4
4. Análisis del fenómeno de la saturación de la tarea	5
4.1. La distorsión de la rigidez: el pipeline v2	6
4.2. La distorsión de la generosidad: el pipeline v3	6
4.3. La naturaleza de la tarea y la naturaleza de los LLM	8
5. Diseño y lógica del pipeline v4	8
6. Resultados y análisis del pipeline v4	9
6.1. Métricas cuantitativas y ranking final	9
6.2. Análisis de la precisión posicional	9
6.3. Análisis del método de resolución	10
7. Discusión y conclusión	11
7.1. Análisis crítico de los resultados del pipeline v4	11
7.2. Significancia estadística e incertidumbre	12
7.3. Consideraciones para futuras investigaciones	12

1. Planteamiento del problema de evaluación

La validación de sistemas de inteligencia artificial para el soporte al diagnóstico clínico es un desafío metodológico fundamental. No se trata meramente de verificar la corrección de una respuesta, sino de evaluar la calidad, robustez y relevancia de un proceso de razonamiento complejo. Como fundación dedicada al avance de estas tecnologías, nuestro objetivo es establecer un marco de evaluación que sea a la vez riguroso, escalable y transparente. Este marco debe ser capaz de discriminar con precisión el rendimiento entre diferentes modelos y arquitecturas, superando las métricas superficiales para capturar la esencia del juicio clínico.

El presente informe documenta el proceso iterativo que hemos seguido para construir dicho marco. Partimos de una premisa de escepticismo científico: toda metodología de evaluación introduce sus propios sesgos y artefactos. Nuestro trabajo, por tanto, no ha sido solo aplicar métricas, sino interrogar a las propias métricas. Este documento narra la evolución de nuestros pipelines de evaluación, desde los primeros intentos hasta el sistema actual, detallando cómo cada fase nos reveló tanto las capacidades de los modelos como las limitaciones de nuestras herramientas de medición.

El eje central de este informe es el análisis de un fenómeno recurrente y crítico: la “saturación de la tarea”, una aparente convergencia de rendimiento que amenazaba con enmascarar el progreso real. A través de este análisis, demostramos cómo un diseño metodológico cuidadoso puede desentrañar estas paradojas y proporcionar una visión clara y fiable del estado del arte en IA diagnóstica.

2. Composición del entorno de pruebas

Para llevar a cabo una evaluación rigurosa, es indispensable contar con un entorno de pruebas —un dataset— que sea representativo y desafiante. Nuestro análisis se basa en un conjunto de **450 casos clínicos**, extraídos de un benchmark más amplio mediante un proceso de selección estratificada.

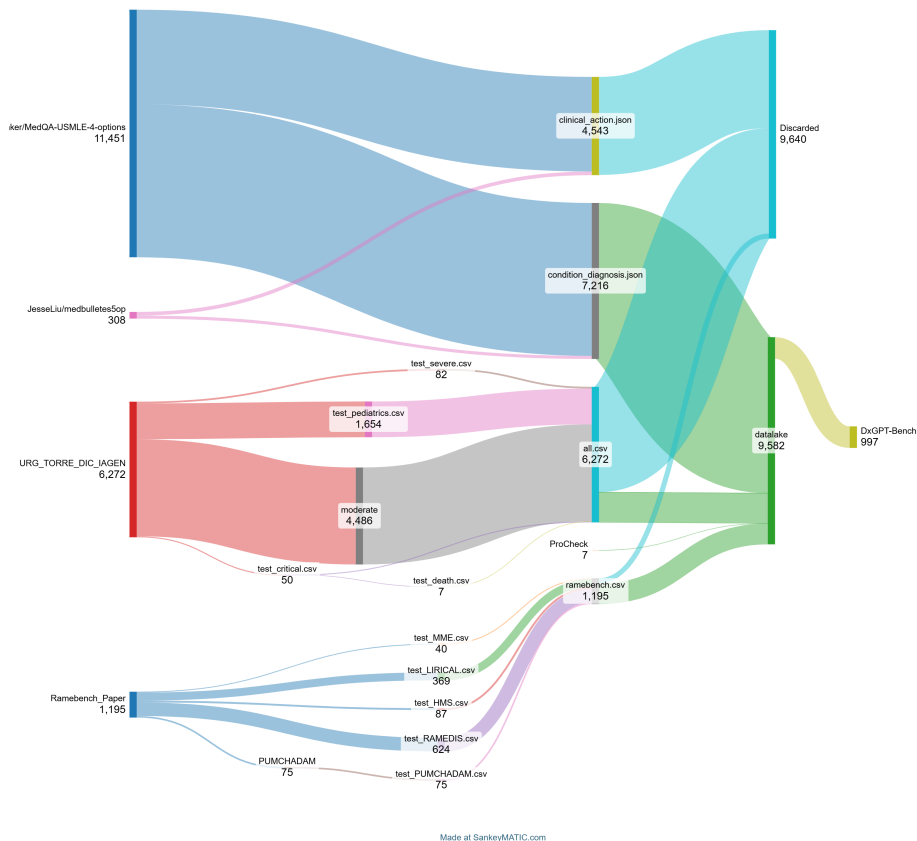


Figura 1. Diagrama de Sankey que visualiza el proceso de ETL para la composición del dataset de evaluación.

2.1 Análisis exploratorio y diseño del scoring

Antes de construir el pipeline de evaluación final (v4), se realizó un análisis exploratorio fundamental. Se generaron las respuestas de cinco modelos distintos para los 450 casos, creando un corpus de 2.250 diagnósticos diferenciales (DDX) para cada caso. Sobre una muestra aleatoria de este “dataset Frankenstein” se aplicaron técnicas de *text analytics* para entender la estructura de las respuestas y la viabilidad de usar vocabularios controlados.

Este análisis preliminar fue crucial y reveló tres hechos clave que informaron directamente el diseño del Pipeline v4:

1. **Viabilidad de SNOMED CT:** Se confirmó que SNOMED CT era el sistema de codificación con mayor cobertura y robustez, convirtiéndolo en el candidato ideal para ser el primer criterio de matching.
2. **Complejidad de ICD-10:** Se observó que las coincidencias exactas de ICD-10 eran poco frecuentes. Sin embargo, las relaciones jerárquicas (padre, hijo) eran comunes y clínicamente significativas. Esto llevó a la decisión de incluir estas relaciones en el scoring del Pipeline v4 para que fuera más justo y preciso.
3. **La inevitable brecha semántica:** El análisis confirmó la existencia de una gran cantidad de casos donde ningún código conectaba el diagnóstico principal con los diferenciales. Esto validó la necesidad de un mecanismo de evaluación semántica (como BERT o un Juez LLM) como último recurso.

Matriz de Transiciones GDX → DDX

GDX \ DDX	Range	Category	Block	Sub-block	Group	Subgroup	Total
Range	0	0	0	0	0	0	0
Category	1	22	53	6	1	0	83
Block	9	136	908	296	28	0	1377
Sub-block	4	45	227	122	15	0	413
Group	1	6	75	83	126	0	291
Subgroup	0	2	2	2	0	0	6
Total	15	216	1293	509	170	0	2170

Baja frecuencia
 Media frecuencia
 Alta frecuencia
 Coincidencias
 Hijos

Figura 2. Análisis de las relaciones jerárquicas ICD-10 en una muestra de los casos, justificando un scoring más sofisticado para el Pipeline v4.

2.2 Diversidad del dataset final

El dataset de 450 casos fue seleccionado para asegurar una amplia cobertura de patologías, garantizando que la evaluación no estuviera sesgada hacia una especialidad concreta.

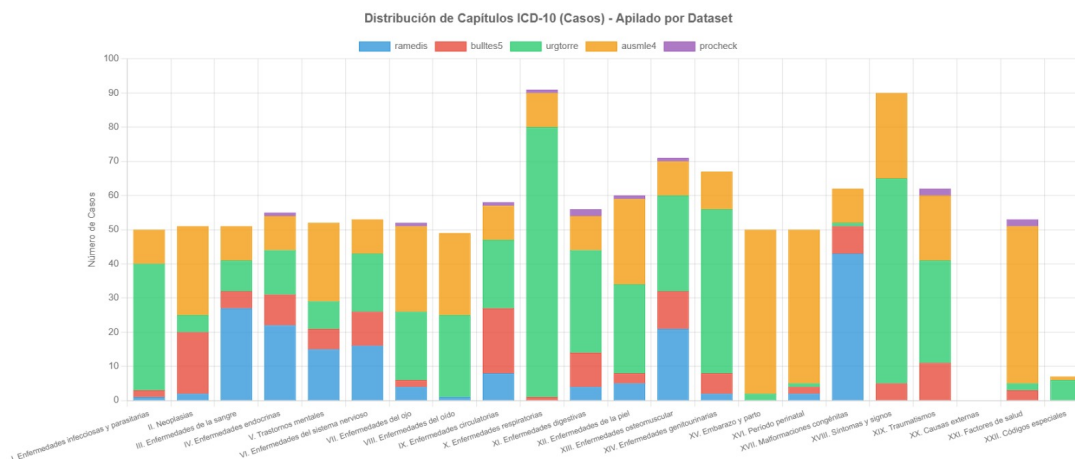


Figura 3. Distribución de los 450 casos clínicos por capítulos de la codificación ICD-10.

3. Evolución de los pipelines de evaluación

El desarrollo de nuestro framework de evaluación ha sido un proceso iterativo, donde cada pipeline representó una hipótesis sobre la mejor manera de medir el rendimiento. Esta evolución fue necesaria para confrontar y resolver el fenómeno de la saturación de la tarea.

- **Pipeline v0:** Fue el primer intento de automatización, utilizando exclusivamente un mode-

lo **BERT** para medir la similitud semántica. Se aplicó principalmente a modelos de Hugging Face. Su limitación era la incapacidad para entender el contexto clínico o las relaciones jerárquicas, reduciendo la evaluación a una simple comparación de proximidad textual.

- **Pipeline v1/v2 (ICD10+BERT):** Representó un salto en sofisticación al introducir los códigos **ICD-10** como primer criterio de evaluación. Si la coincidencia de código fallaba, se utilizaba BERT como red de seguridad. Este método, aunque más estructurado, introdujo la “paradoja del especialista castigado”, penalizando respuestas clínicamente superiores pero terminológicamente más específicas.
- **Pipeline v3 (Juez LLM):** Para corregir la rigidez anterior, este pipeline delegó la evaluación completa a un **Large Language Model** (GPT-4o) que actuaba como “juez”. Su capacidad de razonamiento contextual le permitió entender relaciones clínicas complejas, pero su excesiva “generosidad” llevó a la saturación de los resultados, como se discutirá en la siguiente sección.

La Figura 4 ilustra visualmente la diferencia en las distribuciones de resultados entre el enfoque del Pipeline v0 (aplicado a modelos Hugging Face) y el del Pipeline v3 (aplicado a modelos OpenAI), mostrando cómo diferentes metodologías producen realidades” de rendimiento muy distintas.



Figura 4. Comparativa de resultados entre el Pipeline v0 (BERT, izquierda) y el Pipeline v3 (Juez LLM, derecha).

4. Análisis del fenómeno de la saturación de la tarea

Durante el desarrollo de nuestros pipelines, nos enfrentamos a un fenómeno tan persistente como problemático: la “saturación de la tarea”. Con este término describimos la tendencia observada de que modelos de IA de diferentes generaciones y capacidades obtuvieran puntuaciones notablemente similares bajo ciertas métricas, creando una aparente meseta de rendimiento que contradecía el rápido avance teórico del campo. Este fenómeno no es una curiosidad, sino un obstáculo fundamental para la correcta valoración del progreso. Entenderlo es entender las trampas de la evaluación de la IA.

Este fenómeno se manifestó de formas distintas pero relacionadas en nuestros pipelines intermedios. Fue como observar un objeto distante a través de diferentes lentes: cada lente corregía una distorsión anterior, pero introducía una nueva, hasta que encontramos la combinación correcta que nos permitió ver con claridad.

4.1 La distorsión de la rigidez: el pipeline v2

Nuestro primer intento de automatización (Pipeline v2) buscaba la objetividad a través de la rigidez de los códigos médicos (ICD-10) y la sinonimia (BERT). El resultado fue un sistema que, si bien era objetivo, era ingenuo. Penalizaba la precisión clínica superior (la “paradoja del especialista castigado”) y era ciego a cualquier relación que no fuera una equivalencia terminológica. El ranking que producía era claro, pero estaba basado en una visión del mundo clínico excesivamente simplificada. La Figura 5 muestra la distribución de puntuaciones de este sistema: un paisaje de picos discretos, reflejo de su naturaleza binaria, incapaz de capturar los matices.

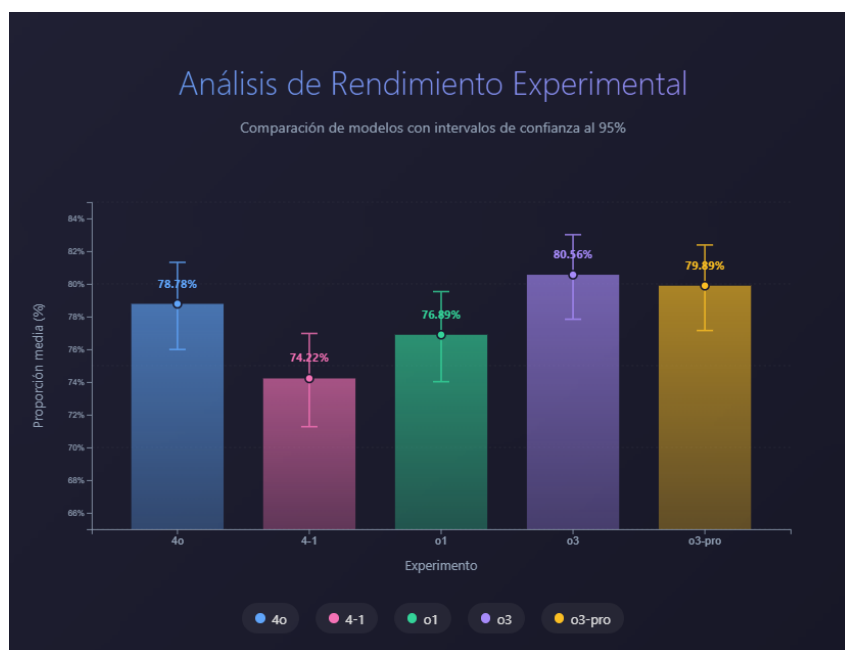


Figura 5. Histograma de resultados para el Pipeline v2 (ICD10+BERT).

4.2 La distorsión de la generosidad: el pipeline v3

Para corregir esta rigidez, el Pipeline v3 empleó un Juez LLM, esperando que su capacidad de razonamiento contextual proporcionara una evaluación más matizada. El resultado fue la manifestación más clara de la saturación. Como se observa en la Figura 7, las puntuaciones de todos los modelos se inflaron y se agruparon en una franja muy estrecha en el extremo superior de la escala. Un modelo de una generación anterior como ‘o1’ obtuvo una puntuación casi idéntica a los de vanguardia como ‘o3’.

¿Qué había ocurrido? El Juez LLM, al evaluar la “plausibilidad clínica”, se había vuelto un evaluador excesivamente generoso. Entendía las relaciones causa-efecto, las manifestaciones clínicas y las asociaciones diagnósticas, y premiaba todas estas conexiones. Al hacerlo, eliminó la distinción crucial entre una respuesta **correcta y precisa** y una respuesta meramente **relevante y plausible**. Esta generosidad actuó como un gran equalizador, borrando las diferencias de rendimiento y creando una falsa meseta. La tarea para los modelos ya no era ser preciso, sino sonar lo suficientemente convincente para otro LLM.

La Figura 6 es la evidencia visual más contundente de este efecto. Muestra la transición de un paisaje de puntuaciones dispersas (v2) a uno de convergencia casi total (v3).

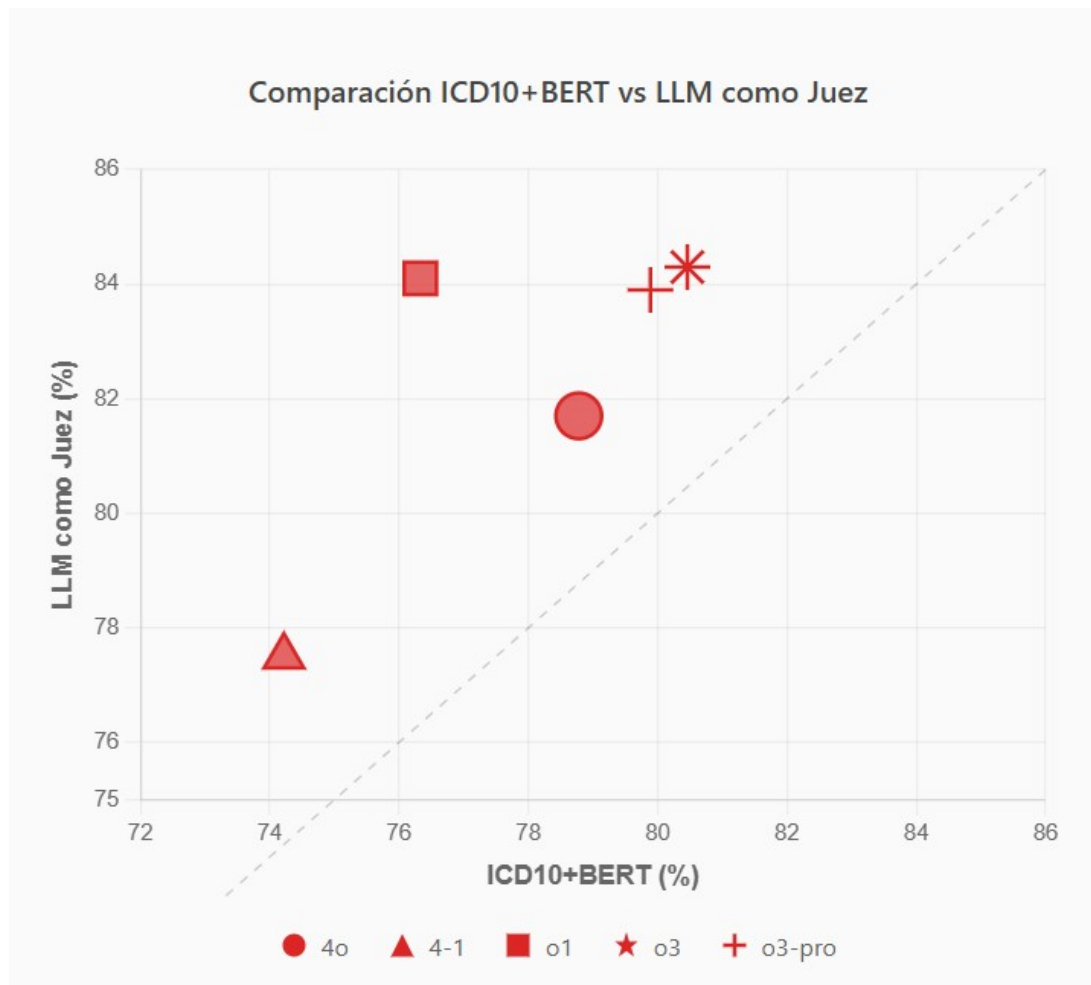


Figura 6. Comparativa directa de resultados entre el método ICD10+BERT (v2) y el Juez LLM (v3).

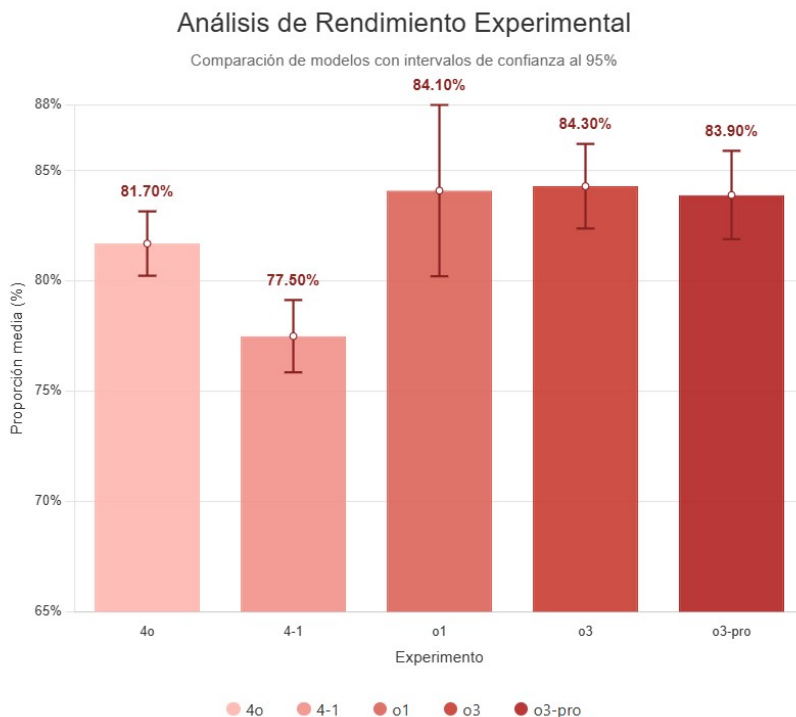


Figura 7. Histograma de rendimiento para el Pipeline v3 (Juez LLM).

4.3 La naturaleza de la tarea y la naturaleza de los LLM

La raíz de este fenómeno yace en la interacción de tres factores: la **definición de la tarea**, la **naturaleza probabilística de los LLM** y el **criterio de evaluación**. Nuestra tarea, generar una lista de 5 diagnósticos diferenciales, es fundamentalmente una tarea de recuperación y ranking de información, no de creación desde cero. Los LLM modernos, desde ‘o1’ hasta ‘o3’, poseen bases de conocimiento vastas. Con un prompt claro y restrictivo, todos son capaces de identificar un conjunto de diagnósticos plausibles.

La diferencia entre un modelo bueno y uno excelente no reside tanto en *si* puede encontrar el diagnóstico correcto, sino en *con qué prioridad y confianza* lo presenta. El Pipeline v3, al ser tan generoso, fallaba en medir esta dimensión de confianza. El Pipeline v4 fue diseñado específicamente para resolver este problema, reintroduciendo la rigidez de forma controlada y haciendo de la **precisión posicional** un criterio de desempate clave. Al hacerlo, finalmente logramos romper la ilusión de la convergencia y medir lo que realmente importa: no solo el acierto, sino la calidad y confianza de ese acierto.

5. Diseño y lógica del pipeline v4

El Pipeline v4 es el resultado de este proceso de aprendizaje. No es un método único, sino un sistema jerárquico que sintetiza las lecciones de sus predecesores, buscando un equilibrio entre la objetividad de los códigos y la inteligencia del análisis semántico.

Su lógica es una cascada de evaluación en tres niveles:

Nivel 1: Verificación por códigos (máxima prioridad): Intenta resolver el caso con la máxima objetividad, buscando coincidencias de código SNOMED CT y luego ICD-10 (exacta y jerárquica). Este nivel premia la disciplina y el rigor.

Nivel 2: Juicio semántico competitivo (IA vs. IA): Solo si los códigos fallan, se activa una competencia entre un análisis de similitud matemática (BERT) y un juicio clínico

simulado (Juez LLM).

Nivel 3: Criterio de desempate (precisión posicional): Se elige la coincidencia (BERT o LLM) que aparezca en la posición más alta de la lista. Este paso es crucial, pues mide la confianza del modelo en su propia respuesta.

6. Resultados y análisis del pipeline v4

La aplicación de este marco de alta fidelidad reveló una jerarquía de rendimiento clara y robusta.

6.1 Métricas cuantitativas y ranking final

Cuadro 1. Ranking y métricas clave de rendimiento por modelo (Pipeline v4).

Métrica	o3	o1	o3-pro	4o
Puntuación Final (%)	89.98 %	88.19 %	87.73 %	87.70 %
Posición Promedio	1.501	1.590	1.613	1.615
Total Casos Resueltos	433 (96.2 %)	437 (97.1 %)	437 (97.1 %)	434 (96.4 %)
Aciertos en Posición 1 (P1)	311	305	299	299
Aciertos en Posición 5 (P5)	9	17	24	20

Nota importante: A posteriori de la experimentación entre modelos, se ha probado a cambiar el prompt en el Pipeline v4 y se observó que una modificación sutil puede alterar la puntuación final de un modelo en torno al 1 %. Esto resulta especialmente llamativo si se considera que la diferencia total entre los modelos más avanzados apenas supera el 2.5 %. No se trata, por tanto, de un simple problema de estandarización, sino de una señal clara de saturación: en una tarea dominada por el reconocimiento de patrones, variaciones superficiales pueden generar efectos casi tan grandes como los que se atribuirían al cambio de modelo. Esto pone en entredicho la estabilidad del marco evaluativo y sugiere que estamos midiendo memoria más que razonamiento.

6.2 Análisis de la precisión posicional

La Figura 8 es la prueba visual de la superioridad de ‘o3’. No solo acierta más a menudo en la primera posición, sino que relega el acierto a la última posición con mucha menos frecuencia que sus competidores, un claro signo de mayor confianza y mejor ordenación de sus diferenciales.

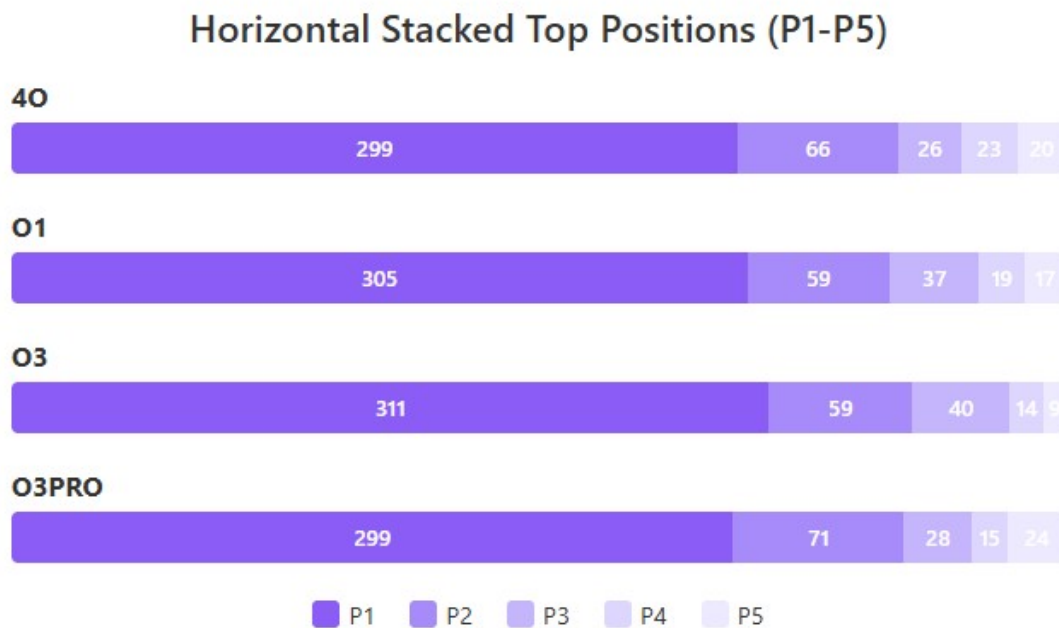


Figura 8. Distribución de los aciertos en las Top 5 posiciones por modelo.

6.3 Análisis del método de resolución

El Pipeline v4 nos permite ver la “huella digital” de cada modelo: su preferencia por resolver casos mediante métodos objetivos o semánticos.

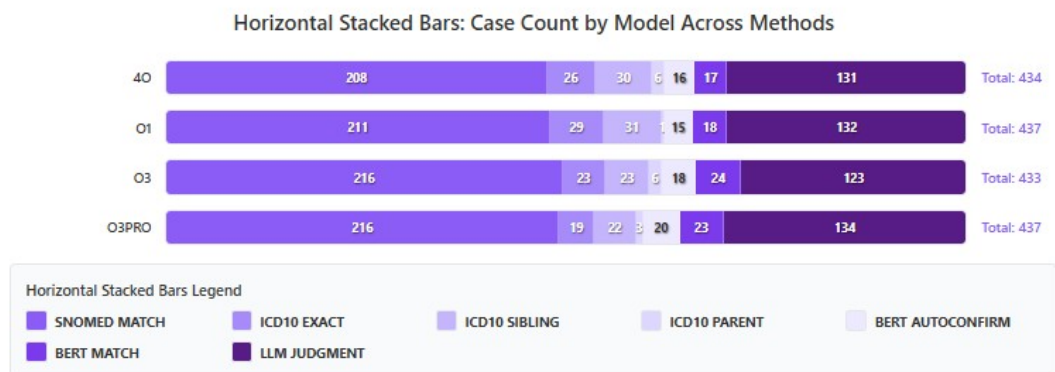


Figura 9. Desglose de los métodos de resolución por modelo.

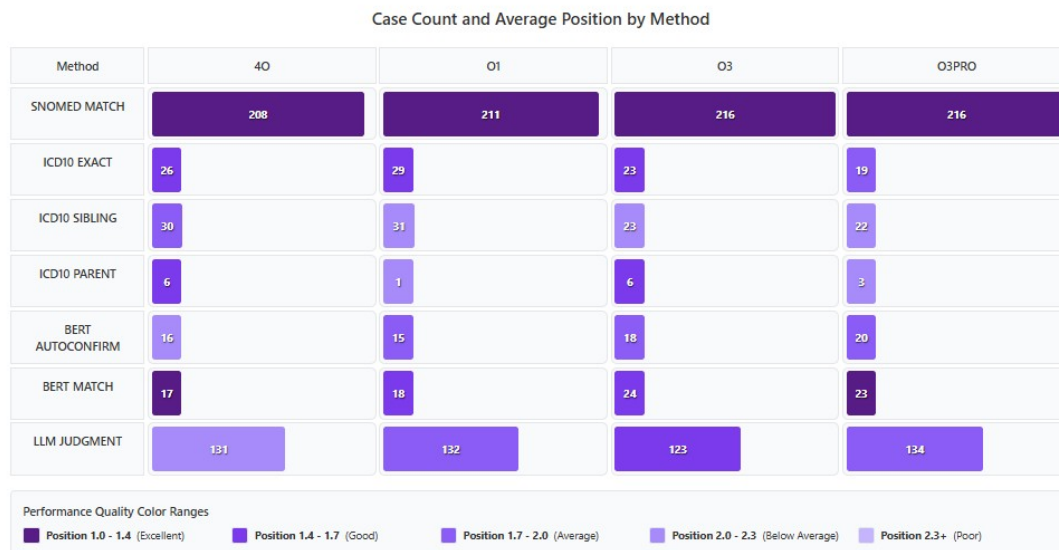


Figura 10. Número de casos resueltos y posición promedio por cada método y modelo.

Este análisis (Figuras 9 y 10) confirma que ‘o3’ basa su éxito en una mayor disciplina taxonómica (más aciertos por SNOMED e ICD-10), mientras que otros modelos como ‘o3-pro’ dependen más de la “red de seguridad” del análisis semántico para alcanzar una alta tasa de resolución, a menudo a expensas de la precisión posicional.

7. Discusión y conclusión

La metodología incremental documentada en este informe nos ha llevado a una conclusión compleja y matizada. Si bien el Pipeline v4 representa nuestro esfuerzo más sofisticado por medir el rendimiento de la IA diagnóstica, sus resultados, lejos de ofrecer una respuesta definitiva, nos enfrentan a una manifestación aún más sutil del fenómeno de la saturación de la tarea.

7.1 Análisis crítico de los resultados del pipeline v4

A primera vista, el Pipeline v4 establece una jerarquía: ‘o3’ (89.98 %) ¿‘o1’ (88.19 %) ¿‘o3-pro’ (87.73 %) ¿‘4o’ (87.70 %). Sin embargo, una mirada más crítica a estos números revela un panorama preocupante. La diferencia entre el primer y el cuarto modelo es de apenas 2.28 puntos porcentuales. Los modelos ‘4o’ y ‘o3-pro’ tienen un rendimiento prácticamente idéntico, y ‘o1’ se sitúa muy cerca. Esta compresión de los resultados es, de hecho, la evidencia más fuerte de saturación que hemos observado hasta ahora.

El argumento previo, de que la saturación en el v3 se debía a la “generosidad” del Juez LLM, se ve ahora desafiado. El Pipeline v4, diseñado para ser más riguroso y multifacético, debería haber ampliado estas diferencias, no mantenerlas tan estrechas. El hecho de que no lo haga sugiere que el problema raíz podría no estar (solo) en la metodología de evaluación, sino en una interacción más profunda entre la tarea y el dataset.

Planteamos la siguiente hipótesis: el dataset de 450 casos, a pesar de su diversidad taxonómica, podría estar concentrado en un espectro de dificultad que no exige un “razonamiento de primeros principios”, sino un “reconocimiento de patrones” altamente sofisticado. Si la mayoría de los casos, por complejos que parezcan, se resuelven identificando constelaciones de síntomas que los modelos ya han internalizado masivamente durante su entrenamiento, entonces es lógico que los modelos modernos converjan en su rendimiento. La tarea no estaría midiendo su capacidad de “pensar”, sino la exhaustividad de su “memoria” de patrones clínicos.

7.2 Significancia estadística e incertidumbre

Si bien la agrupación de los resultados es relativamente estrecha, la diferencia de aproximadamente un 2 % a favor de ‘o3’ frente a los demás parece lo suficientemente consistente como para ser considerada significativa en términos prácticos. Aunque en otro contexto podría justificarse un análisis estadístico más profundo —como un bootstrap para estimar la estabilidad de esta ventaja—, en este caso no se consideró necesario. La comparación se basa en 450 casos diversos, lo cual aporta una base empírica razonable para aceptar que la ventaja observada de ‘o3’ no es fruto del azar, sino una señal robusta dentro del marco evaluado. Por tanto, puede considerarse válida la jerarquía que se observa en los resultados.

En este contexto, el Pipeline v4, aunque metodológicamente es una agregación de criterios más completa, podría no estar aportando un poder discriminativo proporcional a su complejidad. Si la limitación fundamental reside en la naturaleza de la tarea que el dataset permite evaluar, añadir más capas de evaluación podría ser redundante.

7.3 Consideraciones para futuras investigaciones

Esta conclusión no es un punto final, sino una reorientación. Nos obliga a mover el foco de *cómo medimos* a *qué medimos*. Las futuras investigaciones deberían, por tanto, explorar vías para romper este ciclo de saturación, no mediante nuevos pipelines, sino alterando la naturaleza fundamental del desafío. Algunas direcciones posibles incluyen:

- **Diseño de tareas de razonamiento explícito:** Podemos evolucionar los prompts para que no solo pidan un diagnóstico final, sino que exijan explicaciones fisiopatológicas, la justificación del descarte de diagnósticos diferenciales, o incluso la elaboración de un plan diagnóstico escalonado. Esto transforma la tarea de ser principalmente de recuperación a una de explicación y síntesis, revelando mejor las capacidades latentes de modelos avanzados.
- **Curación de datasets de “frontera”:** Tiene sentido avanzar hacia la construcción de casos clínicos diseñados deliberadamente para ser ambiguos, contradictorios o multidominio. Estos escenarios más exigentes podrían servir como verdaderos diferenciadores entre modelos, revelando su capacidad para razonar bajo incertidumbre o resolver tensiones clínicas sutiles.
- **Análisis de la robustez ante la adversidad:** También se pueden explorar escenarios en los que los modelos reciban información engañosa o ruido clínico irrelevante. La forma en que responden ante este tipo de “pistas falsas” permite evaluar la profundidad y estabilidad de su razonamiento.
- **Integración de evaluadores como Deep Research:** Todas estas estrategias se vuelven más útiles si se combinan con evaluadores capaces de captar matices y profundidad contextual. Deep Research, al tener una arquitectura distinta y capacidades de análisis exhaustivo, permite valorar estas tareas con mayor sensibilidad y menor riesgo de sesgo por familiaridad. Su inclusión como juez complementario no reemplaza los sistemas actuales, pero sí los enriquece y valida desde otro ángulo.

En definitiva, nuestro viaje metodológico nos ha enseñado que la búsqueda de un “evaluador perfecto” es probablemente fútil si no va acompañada de una reflexión crítica y continua sobre la naturaleza de la tarea y los datos con los que trabajamos. La saturación no es un fracaso, sino un dato en sí mismo, que nos señala los límites de nuestro enfoque actual y nos guía hacia nuevos y más desafiantes horizontes de investigación.