PUBLICACIONES DE LA
ACADEMIA NACIONAL DE
MEDICINA DE MÉXICO

ACTUALIDADES EN INTELIGENCIA ARTIFICIAL

Dr. Rodolfo Palencia Díaz
Dr. Rodolfo de J. Palencia Vizcarra
Dr. Raúl Carrillo Esper

Número 9

Reducción de la brecha de confianza: una hoja de ruta para el diagnóstico mediante IA

2 de mayo de 2026

Modelo de "backcasting" para cerrar la brecha de confianza entre los médicos y la inteligencia artificial diagnóstica hacia el año 2040. El autor sostiene que la baja adopción clínica no es un fallo técnico, sino un problema de diseño institucional que requiere intervenciones estructurales deliberadas. La hoja de ruta identifica tres hitos esenciales: la estandarización de arquitecturas verificables en 2030, la creación de la figura del Director de IA en 2035 y la integración de la alfabetización de futuros en la educación médica para 2040. En lugar de simplemente predecir el futuro, este enfoque define las condiciones necesarias para transformar la medicina en un ecosistema preparado para la IA. El objetivo final es alcanzar un estado de confianza estratificada por riesgo, donde las herramientas digitales y los profesionales colaboren bajo marcos de gobernanza transparentes y seguros.

Hoja de Ruta Estratégica para la Adopción Clínica de la Inteligencia Artificial en el Diagnóstico hacia 2040

Yu Y
Retrospectiva de la brecha de Confianza: Una hoja de ruta estratégica para la adopción de diagnóstico de IA por parte de los clínicos para 2040
J Med Internet Res 2026;28:e94234
doi: 10.2196/94234

Resumen Ejecutivo

La integración de la inteligencia artificial (IA) en la medicina clínica presenta una paradoja persistente: a pesar de que los modelos de diagnóstico demuestran una superioridad técnica en las pruebas de referencia, su adopción a pie de cama sigue siendo frágil y la confianza del clínico es baja. Este documento sintetiza un análisis de backcasting (proyección inversa), una metodología de futuros normativos, para identificar las intervenciones estructurales necesarias para alcanzar una confianza duradera de los médicos en la IA para el año 2040.

La brecha de confianza no es un problema técnico inevitable, sino un desafío de diseño institucional. Para resolverlo, se proponen tres hitos temporales críticos:

1. 2030 - IA Verificable: Estandarización de arquitecturas de proceso dual (LLM y SLM).
2. 2035 - Gobernanza Agéntica: Institucionalización del rol del Director de IA (CAIO) y transición hacia la IA como orquestadora de cuidados.
3. 2040 - Alfabetización de Futuros: Integración de competencias de colaboración humano-IA en el currículo médico estándar.

El éxito de esta hoja de ruta depende de transformar la pregunta central de *"¿cuándo estará lista la IA para la medicina?"* a "¿qué debemos construir para que la medicina esté lista para la IA?".

Análisis de la Brecha de Confianza y Metodología

El Fracaso del "Forecasting" Tradicional

Las proyecciones convencionales (forecasting) que se limitan a seguir líneas de tendencia optimistas sobre el rendimiento de los modelos son insuficientes. Estas no consideran las transiciones sociotécnicas no lineales necesarias para convertir la capacidad técnica en confianza institucional. En sistemas complejos como la atención médica, los futuros deseados deben construirse activamente mediante intervenciones deliberadas.

Justificación del Backcasting

Se identifica el backcasting como el método apropiado debido a cuatro condiciones críticas:

- Gravedad: La adopción fragmentada corre el riesgo de una "perpetuidad en fase piloto".
- Tendencias actuales problemáticas: El despliegue de modelos de lenguaje de gran tamaño (LLM) sin infraestructura de verificación aumenta los errores por alucinaciones, erosionando la confianza.
- Horizonte temporal largo: Los cambios institucionales (educación, regulación, gobernanza) operan en escalas de décadas.
- Intereses dominantes: Es necesario alinear a proveedores, sistemas de salud, pagadores y reguladores bajo una visión común.

El Estado de Visión para 2040

La meta final es un ecosistema de salud caracterizado por la transparencia semántica y una gobernanza integrada. Los componentes centrales de esta visión incluyen:

Matriz de Confianza Estratificada por Riesgo

La confianza no debe ser un valor único, sino una métrica basada en el nivel de autonomía de la tarea y sus posibles consecuencias.

- Tareas Autónomas (Triage simple, conciliación de medicamentos): Requieren un puntaje de confianza de ≥90%.
- Soporte de Decisión Asistida (Diagnóstico diferencial, interpretación de imágenes): Se acepta un puntaje de 70% a 85%, condicionado a la verificación humana ("human-in-the-loop").

Pilares de la Visión 2040

- Transparencia Semántica: Los resultados de la IA están vinculados de forma nativa a evidencia clínica verificable y computable.
- Gobernanza Integrada: Cada sistema de salud cuenta con un Director de IA (CAIO) con responsabilidad equivalente al Director Médico (CMO).
- Alfabetización de Futuros: Los graduados en medicina están formados en equipos humano-IA (HAT) y competencias de prospectiva.

Puntos de Inflexión Temporales (Pivot Points)

2030: El Estándar de IA Verificable

Para mitigar el problema de las alucinaciones de los LLM, se propone una arquitectura de IA de proceso dual:

- Sistema 1 (LLM): Genera hipótesis diagnósticas rápidas a partir de datos no estructurados.
- Sistema 2 (SLM): Un Modelo de Lenguaje Pequeño (SLM) local que actúa como preceptor, verificando la salida del LLM frente a guías institucionales y literatura en tiempo real.
- Resultado: Un "puntaje de confianza calibrado". Las afirmaciones que no alcancen el umbral de verificación se marcan obligatoriamente para revisión humana.

2035: El Cambio a la Orquestación Agéntica

La IA evoluciona de ser un "consultor" pasivo a un "orquestador agéntico" que gestiona tareas longitudinales (monitoreo post-alta, manejo de enfermedades crónicas).

- Institucionalización del CAIO: Este rol cierra la brecha entre la autoridad clínica del CMO y la infraestructura técnica del CIO. El CAIO audita la calibración local y certifica la seguridad de los modelos.
- Reforma Política: Revisión de marcos de responsabilidad por mala práctica para reconocer las decisiones asistidas por IA como resultados clínicos colaborativos.

2040: Alfabetización de Futuros como Competencia Clínica

La integración educativa final implica:

- Sustituir parte del entrenamiento en diagnóstico diferencial lineal por razonamiento probabilístico de múltiples futuros.
- Talleres de prospectiva en programas de desarrollo de liderazgo médico.

Matriz de Hitos Temporales (Marco STEEP)

Dimensión	2026 (Línea Base)	2030 (Verificable)	2035 (Agéntico)	2040 (Visión)
Social	Escepticismo clínico y miedo a la automatización.	Explicabilidad por SLM genera confianza base.	La fluidez en IA entra en la identidad profesional.	Estándar de alfabetización de futuros.
Tecnológica	LLM generales sin capa de verificación.	Arquitectura dual y SLM en sitio (*on-premise*).	Orquestadores agénticos y monitoreo ambiental.	IA explicable (XAI) fluida e inteligencia ambiental.
Económico	Presupuestos de pilotos sin vías de facturación.	Retorno de inversión (ROI) de hardware	Introducción de códigos de facturación	Integración de costos de IA a
		SLM establecido.	específicos para IA.	nivel de sistema.
Política	Gobernanza fragmentada dirigida por TI.	Regulación hacia requisitos de verificación formal.	Mandato de CAIO y marcos de responsabilidad revisados.	Gobernanza nativa de IA como estándar.
Ambiental	Dependencia de la nube y alto costo energético.	Inferencia en sitio reduce la huella de carbono.	Arquitectura de computación de borde local.	Tejido de IA distribuido de baja latencia.

Riesgos y Consideraciones de Implementación

Sesgo de Automatización y Equidad Algorítmica

- Sesgo de Automatización: Existe el riesgo de que los médicos acepten los resultados de la IA sin evaluación crítica. La arquitectura de verificación debe ser "pedagógicamente visible", funcionando como una herramienta de enseñanza en tiempo real que refuerce el juicio clínico.
- Equidad: Los modelos SLM locales podrían optimizarse para datos demográficos específicos de una institución, sesgando los resultados para grupos minoritarios. Se proponen auditorías de equidad anuales obligatorias y el intercambio de datos de calibración entre instituciones.

Agencia médica colaborativa

El marco legal actual es incoherente con la agencia IA-humano distribuida. Para 2035, se requieren innovaciones legales:

1. Registro de decisiones: Un "registrador de datos de vuelo" para recomendaciones de IA.
2. Seguros de responsabilidad compartida: Que incluyan al CAIO y la calidad del historial de gobernanza institucional.
3. Protocolos de autonomía graduada: Especificando condiciones bajo las cuales la IA agéntica puede actuar sin confirmación humana en tiempo real.

Barreras Operativas

- Interoperabilidad: Necesidad de integración con sistemas de registros electrónicos de salud (EHR) heredados mediante APIs neutrales.
- Carga de trabajo: Una capa de verificación lenta podría provocar que los médicos busquen atajos que anulen la seguridad.
- Madurez digital desigual: Los sistemas de salud rurales podrían requerir un modelo de "CAIO-como-servicio" debido a la falta de infraestructura de cómputo local.

Conclusiones

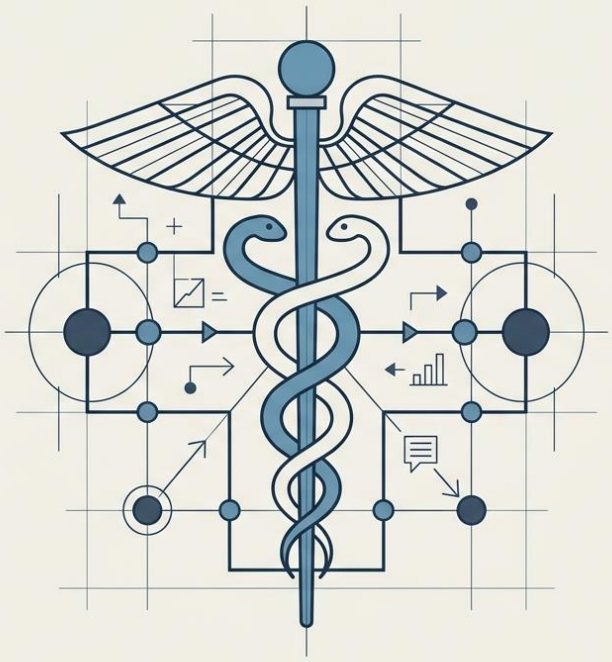
La brecha de confianza no es un destino inevitable, sino un problema de diseño institucional que requiere compromisos estructurales inmediatos. Los tres puntos de inflexión (IA verificable en 2030, gobernanza agéntica en 2035 y alfabetización de futuros en 2040) no son meras predicciones, sino los cimientos de un sistema de salud preparado para la IA. Se hace un llamado a la creación de consorcios multiinstitucionales para evaluar prospectivamente la arquitectura de proceso dual y validar los umbrales de confianza propuestos.

Reducción de la Brecha de Confianza: Una Hoja de Ruta Estratégica para el Diagnóstico mediante IA

Pasando de la predicción pasiva a la construcción activa de la medicina digital hacia 2040.

Presentado por los Dres. Rodolfo Palencia Díaz & Rodolfo de J. Palencia Vizcarra

Fundadores de TICC PalencIA



REDUCCIÓN DE LA BRECHA DE CONFIANZA: HOJA DE RUTA PARA LA IA MÉDICA (2040)

VISIÓN 2040: EL ECOSISTEMA DE CONFIANZA CALIBRADA

MATRIZ DE CONFIANZA ESTRATIFICADA
Niveles de confianza requeridos según el riesgo: >90% para tareas autónomas.

TRANSPARENCIA SEMÁNTICA OBLIGATORIA
Diagnósticos vinculados nativamente a evidencia clínica computable y guías institucionales.

GOBERNANZA INTEGRADA (CAIO)
Creación del Chief AI Officer con autoridad equivalente al Director Medico.

Arquitectura de IA de proceso dual

LLM (SISTEMA 1)
GENERACIÓN DE HIPOTESIS:
Rápido y Creativo

SLM (SISTEMA 2)
VERIFICACIÓN DE GUIAS:
Lento y Crítico

PUNTUACIÓN CALIBRADA
UMBRAL DE VISUALIZACIÓN:
Filtro de Seguridad

2026 — 2040

2040: ALFABETIZACIÓN EN FUTUROS
Integración de competencias de colaboración humano-IA en el currículo médico estándar.

2035: ORQUESTACIÓN AGÉNTICA
Cambio de IA "consultora" a agentes autónomos gestionados por gobernanza formal.

2030: ESTÁNDAR DE IA VERIFICABLE
Arquitecturas de proceso dual donde modelos locales (SLM) verifican a los LLM

INICIO DE LA HOJA DE RUTA ESTRATÉGICA
Identificación de intervenciones estructurales necesarias para que la medicina esté lista para la IA en 2040 mediante backcasting.

REFERENCIAS BIBLIOGRÁFICAS (FORMATO VANCOUVER)
1. Yu Y. Backcasting the Trust Gas: A Strategic Road Map for Clinician Adoption of AI Diagnostics by 2040. J Med Internet Res. 2026;28:e94234. 2. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;23(1).44,-96. 3. Meskil B, Kristof T, Dhunnoo P, Arvsi N, Katonal G. Exploring the need for medical futures studies: insights from a scoping review of health care foresight. J Med Internet Res. 2024;26:e57148. 4. Kahneman D. Thinking, Fast and Slow. New York: Farrar, Straus and Giroux; 2011.

La Paradoja de la IA: Precisión Técnica vs. Resistencia Clínica

Divergencia Sociotécnica

El Mito Tecnológico

La IA supera rutinariamente a los expertos humanos en entornos controlados y conjuntos de datos curados.

La Realidad Clínica

La adopción real en la cabecera del paciente sigue siendo frágil. Existe una resistencia a la automatización persistente.

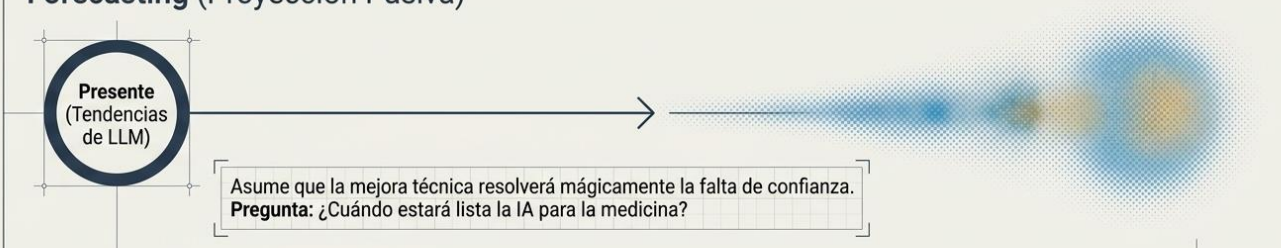
El Diagnóstico TICC PalencIA

Esto no es un déficit tecnológico. Es una falla de diseño institucional originada por la ausencia de infraestructura de verificación rigurosa.

NotebookLM

El Cambio Metodológico: De la Predicción a la Retrospectiva

Forecasting (Proyección Pasiva)



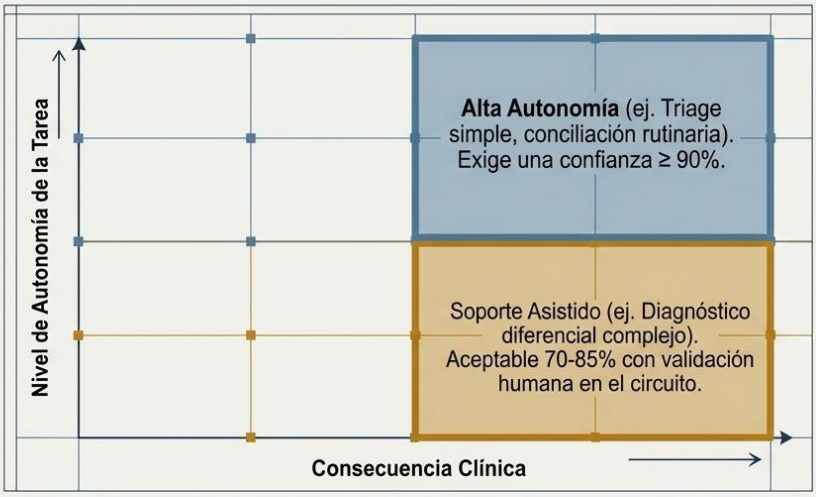
Asume que la mejora técnica resolverá mágicamente la falta de confianza.
Pregunta: ¿Cuándo estará lista la IA para la medicina?

Backcasting (Retrospectiva Activa)

Define un futuro normativo innegociable y planifica hacia atrás.
Pregunta: ¿Qué infraestructura debemos construir hoy para que la medicina esté lista para la IA?

NotebookLM

El Estado de Visión 2040: Confianza Calibrada, No Confianza Ciega



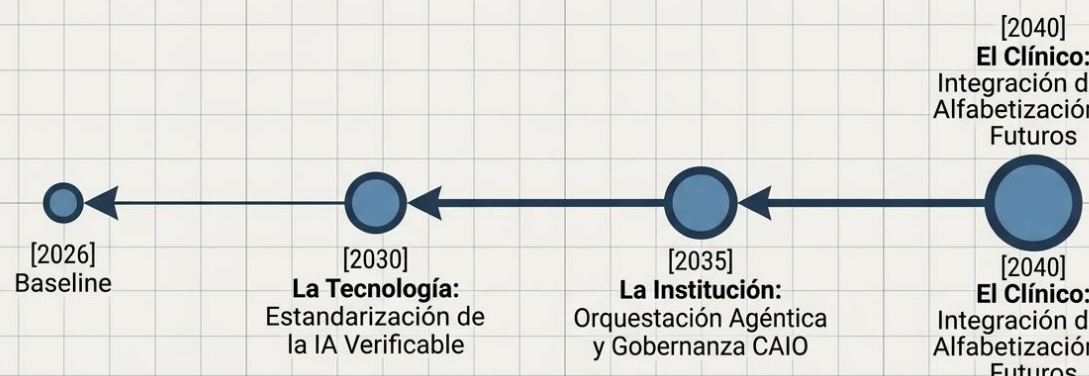
Los 3 Pilares del Ecosistema 2040

1. Transparencia Semántica Nativa.
2. Gobernanza Institucional Integrada.
3. Alfabetización en Futuros como Estándar Médico.

En 2040, el objetivo no es la adopción universal de la IA, sino umbrales de confianza estratificados por riesgo.

NotebookLM

La Hoja de Ruta Estructural: Tres Puntos de Pivote Interdependientes

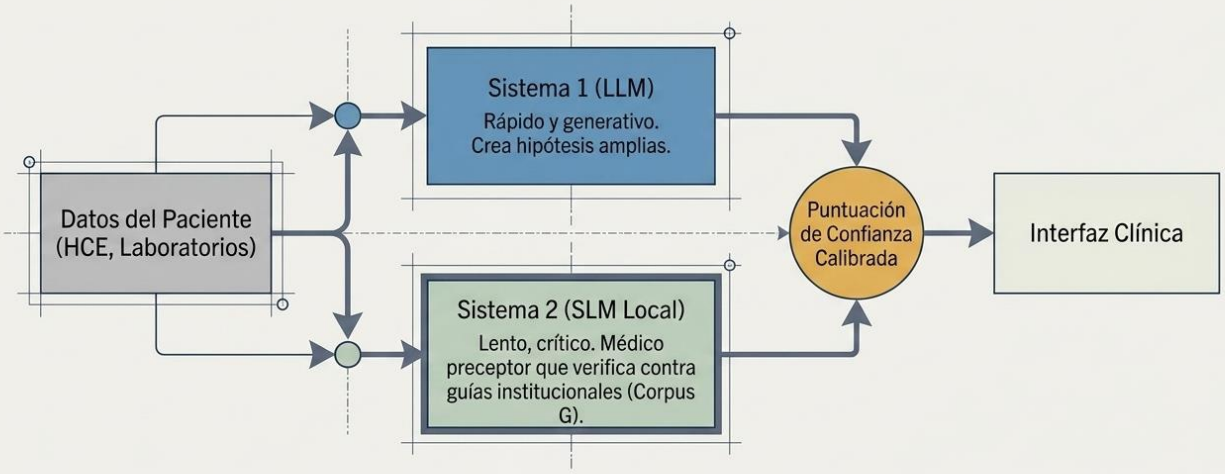


Condición Causal:
Cada punto de pivote es una precondición estructural del siguiente. Retrasar el estándar de verificación de 2030 imposibilita la gobernanza de 2035.

NotebookLM

Pivote 1 (2030): La Arquitectura de IA Verificable de Proceso Dual

El Problema: Las alucinaciones de los Modelos de Lenguaje (LLMs) destruyen la confianza clínica.
La Solución: Arquitectura de Proceso Dual.



NotebookLM

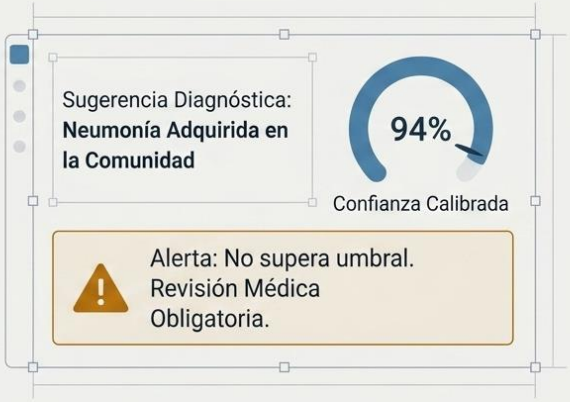
La Puntuación de Confianza Calibrada: Cuantificando la Certeza

TEORÍA

$$C(H \mid x, G) = P(H \mid x) \cdot 1\left[S_{\text{SLM}}(H, G) \geq \theta\right]$$

- La IA no da veredictos; cuantifica el respaldo de la evidencia.
- Si el respaldo no supera el umbral de verificación (θ), se exige revisión obligatoria.
- **Evidencia**: En estudios clínicos, esta calibración redujo las anulaciones médicas hasta 20 veces en predicciones de alta confianza.

APLICACIÓN CLÍNICA



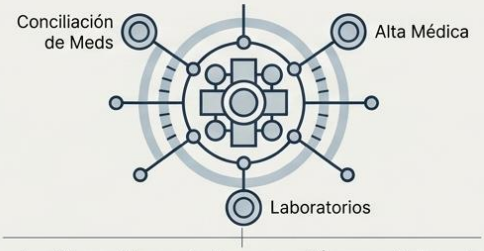
NotebookLM

Pivote 2 (2035): El Salto a la Orquestación Agéntica y el Vacío de Liderazgo

IA Consultiva (Hoy)

La IA responde preguntas pasivamente.

IA Agéntica (2035)



La IA gestiona de forma autónoma flujos de trabajo longitudinales.

El Riesgo Institucional: Vacío de Gobernanza

Director Médico (CMO)	Director de TI (CIO)	La Pregunta Clave
Tiene autoridad clínica, pero carece de alfabetización algorítmica profunda.	Posee la infraestructura, pero carece de licencia y autoridad clínica.	**¿Quién asume la responsabilidad cuando un agente de IA interviene longitudinalmente?**

NotebookLM

Institucionalizando el Liderazgo: El Director de IA (CAIO)

El Nuevo Mandato:

La seguridad clínica exige un **CAIO** que combine experiencia médica, competencia técnica y **autoridad de gobernanza**.

Funciones Críticas del CAIO:

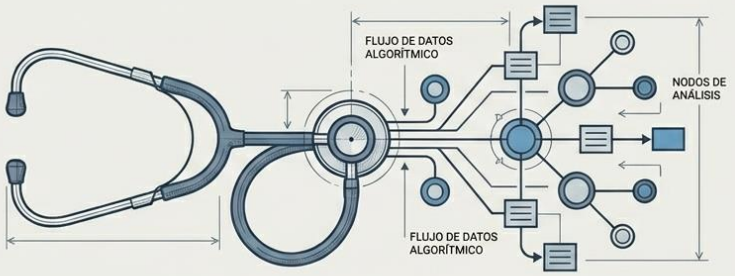
1. Certificación y re-certificación rigurosa de modelos de IA locales.
2. Auditorías de calibración y Auditorías de calibración y reporte estricto de equidad algorítmica.
3. Definición de la política institucional de responsabilidad IA.

Para redes de salud más pequeñas, se adoptará el modelo de 'CAIO-como-servicio' regional.

NotebookLM

Pivote 3 (2040): La Alfabetización en Futuros como Competencia Médica

El médico del futuro no es un receptor pasivo; es un orquestador algorítmico.

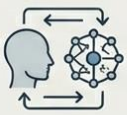


Integración Curricular Innegociable

1. Razonamiento Probabilístico

Transición del diagnóstico diferencial lineal al análisis de escenarios múltiples (*Multifuture reasoning*).

2. Módulos HAT



Entrenamiento obligatorio en "Human-AI Teaming" (Trabajo en equipo Humano-IA) durante el internado médico.

3. Talleres de Prospectiva

Cursos de liderazgo médico integrados con metodologías de previsión y diseño estratégico.

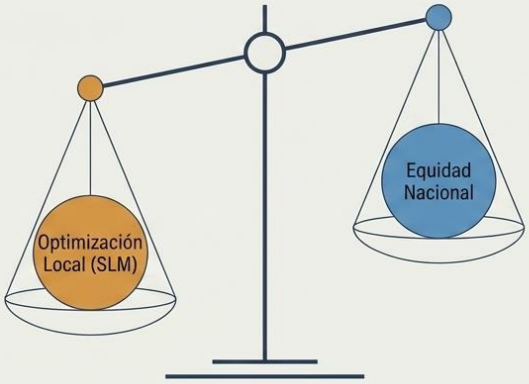
NotebookLM

Matriz de Madurez STEEP: Evolución Sistémica hacia 2040

	Baseline (2026)	Verificable (2030)	Agéntico (2035)	Visión (2040)
Social	Miedo a automatización	Confianza basal (SLM)	Identidad profesional	Estándar de Alfabetización
Tecnológico	LLMs sin verificación	Arquitectura dual	Orquestadores agénticos	Inteligencia ambiental
Económico	Presupuestos piloto	ROI de hardware local	Códigos de facturación IA	Integración de costos sistémica
Político	Gobernanza fragmentada	Requisitos de verificación	Mandato CAIO	Gobernanza nativa de IA

NotebookLM

Salvaguardas Críticas: Equidad y Mitigación de Sesgos Algorítmicos



El Riesgo Estructural: Los SLMs locales pueden calibrarse bien para una demografía y sesgarse severamente contra grupos subrepresentados.

Arquitectura de Protección TICC PalencIA

Auditoría de Equidad (CAIO)

Mandato de recertificación anual con desglose demográfico riguroso.

Aprendizaje Federado

Intercambio de calibración interinstitucional sin comprometer la soberanía de datos del paciente.

Diversidad desde el Origen

Requisitos innegociables de diversidad demográfica en los datos de entrenamiento.

NotebookLM

Soberanía de Datos y Agencia Médica Compartida

La incoherencia de responsabilizar al médico por decisiones de una IA agéntica

Aviación Civil

Medicina Digital (Nuestra Propuesta)

Innovación Legal Necesaria: Agencia Compartida

- **Caja Negra** (Flight Recorder)
- **Responsabilidad Compartida**
- **Autonomía por Fases**

- **Registro de Decisiones Auditable:** Captura inputs y puntajes de confianza.
- **Responsabilidad Solidaria:** El CAIO y la institución operan como partes co-aseguradas.
- **Autonomía Graduada:** Permisos de IA basados en rendimiento de calibración.

NotebookLM

Referencias Bibliográficas Clave

1. Yu Y. Backcasting the Trust Gap: A Strategic Road Map for Clinician Adoption of AI Diagnostics by 2040. J Med Internet Res. 2026;28:e94234.
2. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25(1):44-56.
3. Meskó B, Kristóf T, Dhunnoo P, Árvai N, Katonai G. Exploring the need for medical futures studies. J Med Internet Res. 2024;26:e57148.
4. Kahneman D. Thinking, Fast and Slow. Farrar, Straus and Giroux; 2011.
5. Yu Y, Gomez-Cabello CA, Haider SA, et al. Enhancing clinician trust in AI diagnostics: a dynamic framework for confidence calibration and transparency. Diagnostics (Basel). 2025;15(17):2204.

NotebookLM

Agentes de IA en el Cuidado de la Salud: Aplicaciones, Evaluación y Direcciones Futuras

Zaho, L., Liu, S., T. et al Agentes de la IA en la atención médica: aplicaciones evaluaciones y direcciones futuras. npj Artif. Intell. 2, 31 (2026)
https://doi.org/10.1038/s44387-026-00076-4

Resumen

El vertiginoso avance de los Modelos de Lenguaje de Gran Escala (LLM) ha propiciado la emergencia de los "Agentes de IA" en el sector salud. A diferencia de los modelos de lenguaje convencionales, estos agentes actúan como sistemas autónomos capaces de percibir su entorno, razonar basándose en objetivos y ejecutar tareas complejas mediante el uso de herramientas externas. Este documento sintetiza un análisis exhaustivo sobre su evolución, capacidades actuales, aplicaciones clínicas y el marco necesario para su evaluación y gobernanza. Los agentes de IA prometen optimizar el diagnóstico, el apoyo a decisiones clínicas y la gestión hospitalaria, aunque su implementación enfrenta desafíos críticos como las alucinaciones, la falta de interpretabilidad y dilemas éticos sobre la responsabilidad legal.

1. Fundamentos y Evolución Conceptual

El concepto de "Agente" ha evolucionado desde la contemplación filosófica (el telos de Aristóteles) hasta sistemas tecnológicos prácticos. La arquitectura moderna de un agente de IA, basada en la definición de Weng, se fundamenta en un LLM como controlador central complementado por cuatro módulos críticos:

- Planificación: Descomposición de tareas complejas en pasos manejables.
- Memoria: Retención de información a corto y largo plazo para contextualizar acciones.
- Uso de herramientas: Capacidad de invocar APIs externas para obtener datos o realizar operaciones.
- Autorreflexión: Capacidad de corregir errores y mejorar el desempeño de forma autónoma.

Hitos en la Evolución

- 1950s: Propuesta de la Prueba de Turing.
- 1970s: Surgimiento de sistemas expertos basados en reglas.
- Siglo XXI: Avances en aprendizaje profundo y aprendizaje por refuerzo multi-agente (MARL).
- Post-2022: Proliferación de agentes basados en LLM (GPT-4, Gemini, LLaMA) que ofrecen bases de conocimiento más ricas e interacciones humanas naturales.

2. Características Distintivas de los Agentes de IA

Los agentes poseen cinco capacidades centrales que los diferencian de las tecnologías de IA tradicionales:

1. Comprensión y Generación de Texto: Procesamiento profundo de información contextual para interacciones personalizadas.
2. Uso de Herramientas e Interactividad: Capacidad de autoaprendizaje para seleccionar e invocar herramientas externas (EHR, PACS, LIS) mediante APIs.
3. Procesamiento de Tareas y Generalización: Integración fluida con diversos sistemas de información para la colaboración interdisciplinaria.
4. Razonamiento Lógico y Descomposición de Tareas: Uso de estrategias de prompting para resolver problemas complejos mediante lógica deductiva.
5. Capacidad de Aprendizaje y Adaptación: Optimización continua basada en grandes volúmenes de datos con mínima intervención manual.

3. Panorama de Aplicaciones en el Cuidado de la Salud

Los agentes de IA se están desplegando en múltiples dominios médicos, buscando aliviar la carga de trabajo de los profesionales y mejorar la precisión clínica.

Áreas de Aplicación Clínica y Administrativa

Dominio	Aplicaciones Específicas	Ejemplos de Sistemas/Modelos
Diagnóstico Asistido	Detección y clasificación de imágenes multimodales; simulaciones de interacciones médico-paciente.	Agent Hospital, ClinicalAgent, ZODIAC (cardiología).
Apoyo a la Decisión	Colaboración multidisciplinaria (MDT); análisis de errores y corrección de decisiones médicas.	MedAgents, MDAgents, MEDAIDE.
Generación de Informes	Interpretación de rayos X de tórax; creación de informes legibles para el paciente.	CheXagent, CXR-agent, MGA.
Gestión de Salud	Chatbots para salud mental (depresión, estrés); asesoría en pérdida de peso y manejo de piel.	Agent Mental Clinic (AMC), MISHA, Polaris.
Educación Médica	Simulación de pacientes y casos clínicos para entrenamiento de estudiantes.	AI Patient, MEDCO, ChatCoach.
Gestión de Medicación	Validación de prescripciones; farmacovigilancia y predicción de eficacia de fármacos.	Rx Strategist, MALADE, ClinicalAgent.
Gestión Hospitalaria	Automatización de registros electrónicos (EHR); codificación ICD; simplificación de autorizaciones previas.	EHRAgent, Almanac Copilot, ColaCare.

4. Desafíos Críticos para la Implementación

A pesar de su potencial, existen barreras significativas que dificultan la adopción a gran escala en entornos reales:

- Alucinaciones: Generación de conclusiones incorrectas pero convincentes en casos de enfermedades raras o presentaciones clínicas ambiguas.
- Falta de Interpretabilidad: Procesos de "caja negra" que impiden a los clínicos rastrear el razonamiento detrás de una recomendación.
- Ambigüedad en la Responsabilidad: Incertidumbre legal y ética sobre quién asume la responsabilidad ante un error diagnóstico o terapéutico generado por un agente.
- Problemas de Datos: Sesgos en los conjuntos de entrenamiento (género, etnia, geografía) y riesgos de violaciones a la privacidad de datos sensibles.

5. Marco de Evaluación Multidimensional

Para asegurar la confiabilidad y seguridad, se propone un marco de evaluación dividido en indicadores básicos y de desarrollo.

Dimensiones e Indicadores de Evaluación

Dimensión	Indicadores Primarios	Propósito
Indicadores Básicos	Exactitud, Precisión, Sensibilidad, F1-score, BLEU, ROUGE, Tasa de éxito en tareas.	Garantizar la corrección objetiva y semántica, así como el cumplimiento de tareas específicas.
Eficiencia	Tiempo de respuesta, número de rondas de interacción.	Evaluar la velocidad operativa y la facilidad de interacción en escenarios críticos.
Calidad de Contenido	Riqueza, utilidad, seguridad, cumplimiento ético, legibilidad.	Asegurar que el contenido sea clínicamente significativo y comprensible.
Cuidado Humanístico	Empatía, confianza del usuario, satisfacción, cumplimiento del paciente.	Evaluar la atención a las necesidades psicológicas y la aceptabilidad del sistema.

6. Direcciones Estratégicas para el Futuro

El documento identifica siete direcciones críticas para el desarrollo futuro de los agentes de IA:

1. Integración con Robots Empoderados: Transición de respuestas mecánicas a interacciones humanizadas en robots quirúrgicos y de asistencia.
2. Modelos de Expertos Híbridos (MoE): Uso de arquitecturas de "Mezcla de Expertos" para activar sub-modelos especializados, mejorando la precisión en tareas clínicas específicas.
3. Expansión de Métricas de Evaluación: Incorporación de indicadores económicos (costo-beneficio), seguridad clínica (eventos adversos) y métricas subjetivas de satisfacción.
4. Gestión de Riesgos y Seguridad: Creación de directrices estandarizadas y marcos de supervisión perpetua para responder a emergencias.
5. Revisión Moral y Ética: Establecimiento de comités de ética independientes y protocolos claros de privacidad y transparencia algorítmica.
6. Confianza del Usuario y Retroalimentación: Integración directa de la retroalimentación de pacientes y médicos en el ciclo de desarrollo de la IA.
7. Desarrollo Profesional del Personal Médico: Fomento de la co-adaptación entre la tecnología y la práctica humana, redefiniendo el papel del médico hacia la colaboración con sistemas inteligentes.

Conclusión

Los agentes de IA basados en LLM representan una frontera tecnológica con la capacidad de transformar radicalmente el cuidado de la salud. Sin embargo, su éxito no depende exclusivamente de los avances técnicos, sino de la creación de marcos de gobernanza robustos, evaluaciones científicas exhaustivas y una integración ética que priorice la seguridad del paciente y la colaboración con los profesionales de la salud.