Safe Generative Artificial Intelligence Conversational Prototype for Community Nutrition Education in Honduran Adolescents

Isaac Zablah[1]; Fiama Garcia[2]; Antonio Garcia Loureiro[3]; Maryori Pereira[4]; Sara Ney[5]; Juan Antonio Hernández[6]; Perla Simons[7]

[1,2,4,6,7]Universidad Nacional Autónoma de Honduras, Honduras, *jose.zablah@unah.edu.hn, fiama.garcia@unah.edu.hn, maryori.pereira@unah.edu.hn, antoniopavon@unah.edu.hn, perla.morales@unah.edu.hn*

[3]Universidad de Santiago de Compostela, España, *antonio.garcia.loureiro@usc.es*

[4]Universidade Federal de Viçosa, Brasil, *sara.ney@ufv.br*

Abstract*– Community nutrition education for adolescents is constrained by limited coverage, heterogeneous food environments, and the need for age appropriate messages that do not become individualized medical advice. This paper proposes NutriChat-HN, a safe conversational prototype for Honduran adolescents that combines a curated retrieval layer based on national food based dietary guidance, adolescent nutrition recommendations and school feeding context with a multilayer safety framework for generative artificial intelligence. The method follows a design science approach, using documentary sources, risk analysis and synthetic red teaming scenarios. The prototype separates educational dialogue from diagnosis, avoids restrictive diet plans, minimizes personal data from minors, and escalates high risk queries to human support. A desk evaluation with 240 synthetic interactions yielded 97.1 percent safe responses (233 of 240), 98.5 percent correct escalation in red flag cases (64 of the 65 prompts in the two domains where escalation applies), 94.6 percent citation coverage from the curated knowledge base, and 98.8 percent absence of critical hallucination (3 critical hallucinations in 240 interactions). All rates were recomputed from the per-domain counts and every interaction was rated by the authors, without an independent panel. The contribution is a replicable socio technical model for municipal nutrition education that treats generative AI as supervised educational infrastructure rather than an autonomous health adviser.

***Keywords*-- generative artificial intelligence, adolescent nutrition, Honduras, conversational agent, AI safety.**

Prototipo conversacional Inteligencia Artificial generativa segura para educación nutricional comunitaria en adolescentes hondureños

Isaac Zablah[1]; Fiama Garcia[2]; Antonio Garcia Loureiro[3]; Maryori Pereira[4]; Sara Ney[5]; Juan Antonio Hernández[6]; Perla Simons[7]
[1,2,4,6,7]Universidad Nacional Autónoma de Honduras, Honduras, *jose.zablah@unah.edu.hn, fiama.garcia@unah.edu.hn, maryori.pereira@unah.edu.hn, antoniopavon@unah.edu.hn, perla.morales@unah.edu.hn*
[3]Universidad de Santiago de Compostela, España, *antonio.garcia.loureiro@usc.es*
[4]Universidade Federal de Viçosa, Brasil, *sara.ney@ufv.br*

Resumen– ***La educación nutricional comunitaria dirigida a adolescentes enfrenta restricciones de cobertura, heterogeneidad territorial y necesidad de mensajes comprensibles, culturalmente pertinentes y seguros. Este artículo propone NutriChat-HN, un prototipo conversacional de IA generativa para adolescentes hondureños, diseñado como apoyo educativo no clínico y acompañado por un marco de seguridad multicapa. La metodología siguió un enfoque de ciencia del diseño, con revisión documental, base de conocimiento curada, análisis de riesgos y evaluación de escritorio mediante escenarios sintéticos. El sistema integra recuperación aumentada por generación, clasificación de intención y riesgo, reglas de privacidad para menores, restricciones explícitas contra diagnóstico o planes dietéticos individualizados, y derivación a personal de salud ante señales de alarma. La validación técnica con 240 interacciones simuladas mostró 97.1% de respuestas seguras (233 de 240), 98.5% de derivación adecuada en los casos de alto riesgo (64 de los 65 prompts de los dos dominios donde la derivación aplica), 94.6% de cobertura con fuente recuperada y 98.8% de ausencia de alucinación crítica (3 casos en 240). Todas las tasas se recalcularon a partir de los conteos por dominio y la calificación estuvo a cargo de los autores, sin panel independiente. El aporte principal es un modelo socio técnico replicable para municipios, centros educativos y promotores comunitarios, donde la IA generativa se usa como infraestructura pedagógica supervisada y no como sustituto de profesionales de salud.***

Palabras clave-- ***IA generativa, nutrición adolescente, Honduras, agente conversacional, seguridad de IA.***

I. Introducción

La adolescencia es una etapa decisiva para consolidar patrones alimentarios, autonomía, identidad corporal y hábitos que condicionan la salud adulta. La Organización Mundial de la Salud (OMS) sostiene que las intervenciones nutricionales para adolescentes deben reconocer la heterogeneidad biológica, social y territorial de este grupo, así como incorporar su participación en el diseño, la implementación y la evaluación de los programas [1]. En paralelo, el Fondo de las Naciones Unidas para la Infancia (UNICEF) ha documentado que los sistemas alimentarios actuales exponen a niños y adolescentes a dietas de baja calidad, con coexistencia de desnutrición, deficiencias de micronutrientes, sobrepeso y obesidad [2].

En Honduras, la necesidad de educación nutricional comunitaria se observa en indicadores escolares y en el contexto alimentario nacional. La encuesta Global School based Student Health Survey de 2012, realizada en estudiantes de 13 a 15 años, reportó 17.8% de sobrepeso, 5.6% de obesidad y 74.3% de consumo de bebidas carbonatadas una o más veces al día durante los últimos 30 días [3]. En la Fig. 1, el dato más destacado es que tres de cada cuatro escolares de 13 a 15 años reportaron consumo diario de bebidas carbonatadas, lo cual contrasta con tasas combinadas de sobrepeso y obesidad de 23.4 por ciento. Estos datos, aunque de 2012, representan la fuente pública más reciente y específica para adolescentes escolarizados hondureños y justifican la orientación de NutriChat-HN hacia mensajes sobre hidratación, reducción de bebidas azucaradas y meriendas saludables.

Aunque Honduras no ha publicado una actualización del GSHS posterior a 2012, datos complementarios confirman la persistencia de la doble carga de malnutrición. El Índice Global de Hambre 2025 clasificó a Honduras con un puntaje de 12.5, correspondiente a un nivel moderado de hambre, ocupando el puesto 62 de 123 países evaluados [4]. El análisis del IPC confirma que 1.8 millones de personas, equivalentes a 18% de la población nacional, enfrentaron inseguridad alimentaria aguda entre diciembre de 2024 y marzo de 2025 [5]. Adicionalmente, Liu et al. documentaron que la doble carga de malnutrición, coexistencia de desnutrición y sobrepeso en adolescentes con hambre de 12 a 15 años es un fenómeno global identificado en datos GSHS de 41 países, incluyendo países de América Latina con perfiles similares a Honduras [6].

Las Guías Alimentarias para Honduras recomiendan consumo diario de frutas y verduras, reducción de frituras, embutidos, sal y productos de baja calidad nutricional, con mensajes adaptados a la cultura alimentaria nacional [7].

El Programa Nacional de Alimentación Escolar representa una plataforma pública de alto alcance para la niñez

hondureña. La Secretaría de Educación reporta que el programa beneficia a más de 1.3 millones de niños en 22,000 centros educativos gubernamentales, e incluye educación alimentaria, hábitos saludables y respeto a la cultura alimentaria [8]. Sin embargo, la cobertura de una ración o de una actividad escolar no garantiza por sí misma acompañamiento educativo continuo, personalizado por contexto, comprensible para adolescentes y disponible fuera del aula. La presión territorial aumenta cuando los hogares enfrentan inseguridad alimentaria, según las cifras del IPC ya señaladas. El Programa Mundial de Alimentos también describe a Honduras como un país expuesto a choques climáticos, pobreza e inequidades estructurales que afectan el acceso a alimentos y agua, especialmente en comunidades rurales y marginadas [9].

En este contexto, la justificación programática de una herramienta educativa digital no se limita a ampliar cobertura informativa. También responde a la necesidad de fortalecer prácticas preventivas frente a la inseguridad alimentaria, la doble carga nutricional y las limitaciones de recursos humanos especializados en nutrición. En establecimientos educativos, municipios y comunidades con escasa disponibilidad de profesionales de nutrición, un prototipo conversacional seguro podría funcionar como complemento de programas existentes, siempre que sus contenidos estén alineados con las Guías Alimentarias para Honduras, la disponibilidad local de alimentos y las rutas institucionales de supervisión.

La educación alimentaria convencional ha mostrado beneficios cuando se implementa de forma escolar, participativa y sostenida. Revisiones sistemáticas han encontrado efectos favorables de intervenciones educativas escolares sobre consumo de frutas y verduras, reducción de conductas de riesgo alimentario y prevención de ganancia excesiva de peso, aunque la magnitud depende de intensidad, participación familiar y entorno escolar [10], [11]. Además, una síntesis de encuestas GSHS en 72 países evidenció baja frecuencia de consumo de frutas y verduras y alta exposición a bebidas carbonatadas y comida rápida en adolescentes escolarizados, con amplia variabilidad por país y subpoblación [12].

Los agentes conversacionales ofrecen una vía complementaria para ampliar cobertura, reforzar mensajes y responder dudas frecuentes en lenguaje natural. Una revisión mixta reciente sobre agentes conversacionales para conductas dietarias encontró potencial para mejorar conocimiento nutricional, consumo de frutas y verduras, adherencia a patrones saludables y experiencia de usuario, aunque todavía existen limitaciones metodológicas, variabilidad en resultados y necesidad de evaluación rigurosa [13]. La inteligencia artificial (IA) generativa permite adaptar ejemplos, lenguaje y secuencias de aprendizaje, pero también introduce riesgos de alucinación, sobreconfianza, sesgo, privacidad y emisión de consejos de salud fuera de competencia. Por ello, la Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura (UNESCO) propone que la IA generativa en educación se use bajo un enfoque humano, apropiado por edad y regulado, con protección de privacidad y validación ética [16]. La Organización Mundial de la Salud (OMS) advierte que los modelos multimodales grandes pueden generar afirmaciones falsas, inexactas, sesgadas o incompletas, con potencial de daño cuando se usan para decisiones de salud [17].

Este trabajo plantea la siguiente pregunta: ¿Cómo diseñar una IA generativa segura para educación nutricional comunitaria de adolescentes hondureños, ampliando cobertura sin convertir el sistema en un asesor clínico autónomo? El objetivo fue elaborar y validar conceptualmente un prototipo conversacional, denominado NutriChat-HN, junto con un marco operativo de seguridad, gobernanza y evaluación para uso municipal, escolar y comunitario.

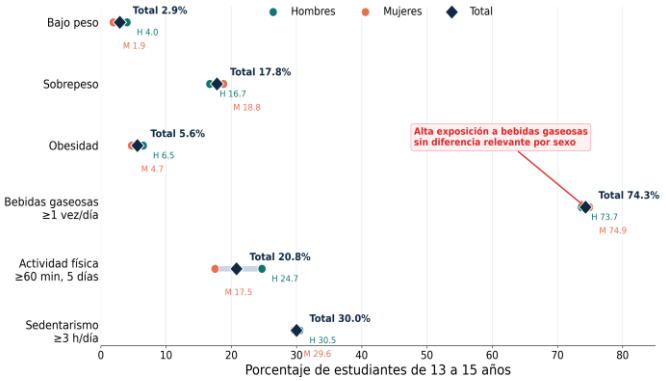


Fig. 1 Indicadores nutricionales y conductuales en escolares hondureños de 13 a 15 años según GSHS 2012 [3].

II. Trabajos Relacionados

El uso de agentes conversacionales en salud ha evolucionado desde sistemas basados en reglas hasta chatbots impulsados por modelos de lenguaje grande (LLM). Van Kolfschooten et al. analizaron S.A.R.A.H., el chatbot de la OMS para promoción de salud, identificando tensiones entre personalización, autonomía del usuario y gobernanza ética bajo los seis principios de la OMS para IA en salud [18]. En salud mental adolescente, una revisión sistemática reciente encontró que las intervenciones basadas en IA dirigidas a adolescentes producen reducciones modestas de síntomas depresivos y ansiosos, aunque la evidencia de largo plazo es limitada y la adaptación cultural insuficiente **¡Error! No se encuentra el origen de la referencia.**.

La arquitectura de recuperación aumentada por generación (RAG) se ha consolidado como el enfoque dominante para reducir alucinaciones en chatbots de dominio específico. Documentación técnica de la industria reporta reducciones de la tasa de alucinación de entre 60 y 80 por ciento frente a modelos sin recuperación, así como mejoras de adherencia

dietaria cercanas a 25 por ciento en prototipos nutricionales con bases curadas [19]. Estas cifras proceden de material no arbitrado y se citan como referencia de orden de magnitud, no como evidencia consolidada; su verificación en estudios revisados por pares sigue pendiente, y la magnitud del efecto depende de la cobertura y la actualización de la base de conocimiento.

El marco de seguridad para chatbots en salud enfrenta desafíos específicos. Un protocolo de "red teaming" evaluado en un chatbot RAG para necesidades sociales relacionadas con salud identificó que las pruebas de un solo turno ocultan vulnerabilidades que solo emergen en interacciones multi-turno, donde las tasas de error pueden alcanzar 50 por ciento en consultas de asesoramiento y 40 por ciento en situaciones de angustia del usuario [14]. Esta evidencia respalda la necesidad de evaluaciones multi-turno como las propuestas en el presente estudio.

En el contexto de regulación, la Comisión Federal de Comercio de Estados Unidos inició en 2025 una investigación formal sobre medidas adoptadas por desarrolladores de IA generativa para mitigar daños a menores derivados de interacciones con chatbots [15]. California aprobó la ley SB 243 que exige protocolos de prevención de suicidio, divulgación de que la interacción es con IA, y restricciones de contenido sexual para usuarios menores [15]. Estos desarrollos regulatorios confirman la pertinencia del marco de seguridad propuesto en NutriChat-HN.

No se identificaron estudios previos que combinen RAG, marco de seguridad multicapa y adaptación cultural para educación nutricional de adolescentes en países de ingreso medio-bajo de América Latina, lo cual constituye la brecha que este trabajo busca abordar.

III. Materiales y métodos

Se desarrolló un estudio de ciencia del diseño orientado a prototipo. La unidad de análisis fue el sistema socio técnico de educación nutricional comunitaria, no individuos ni historias clínicas. El producto esperado fue un prototipo conversacional y un marco de seguridad aplicable a pilotos municipales. Las cifras de contexto proceden de fuentes públicas agregadas, documentos técnicos y literatura científica. Tabla I resume las fuentes utilizadas y su función dentro del prototipo.

La base de conocimiento se estructuró con tres capas. La primera capa contiene mensajes normativos de nutrición adolescente, guías alimentarias y principios de promoción de salud. La segunda capa contiene contexto hondureño, incluyendo Guías Alimentarias para Honduras, Programa Nacional de Alimentación Escolar, disponibilidad de alimentos locales y restricciones económicas frecuentes. La tercera capa contiene políticas de seguridad y derivación, con reglas para no diagnosticar, no indicar dietas restrictivas, no calcular planes individualizados para pérdida rápida de peso y no responder consultas clínicas complejas.

El prototipo fue definido como un agente conversacional supervisado. La arquitectura incluye interfaz web o mensajería, aviso de uso apropiado por edad, minimización de datos, clasificador de intención y riesgo, recuperación aumentada por generación (RAG), modelo generativo, validador posterior y registro técnico anónimo. La recuperación aumentada por generación se utiliza para recuperar fragmentos de fuentes curadas antes de generar la respuesta, de modo que el modelo no dependa únicamente de memoria paramétrica. Cuando el clasificador detecta síntomas, embarazo, alergias, trastornos alimentarios, dietas extremas, autolesión o solicitud de diagnóstico, el sistema activa una respuesta de apoyo y derivación.

La implementación de RAG se configuró con los siguientes parámetros de diseño. La base de conocimiento fue segmentada en fragmentos de 300 a 500 tokens con solapamiento de 50 tokens, utilizando separación semántica por secciones temáticas. La recuperación se configuró para devolver los tres fragmentos más relevantes por consulta, con un umbral mínimo de similitud coseno de 0.75. Cuando ningún fragmento superó el umbral, el sistema generó una respuesta genérica de orientación sin afirmaciones factuales específicas y sugirió consultar fuentes oficiales o personal de salud. El modelo generativo recibió instrucciones de sistema que incluían restricciones explícitas: no diagnosticar, no prescribir, no recomendar suplementos, no calcular índice de masa corporal, no interpretar resultados de laboratorio y no sugerir pérdida de peso como objetivo.

La gobernanza de riesgo se alineó con el marco NIST AI RMF 1.0, que organiza la gestión de IA en funciones de gobernar, mapear, medir y gestionar [20]. La dimensión de ciberseguridad incorporó riesgos de aplicaciones con modelos de lenguaje descritos por OWASP (*Open Worldwide Application Security Project*), como *prompt injection*, manejo inseguro de salidas, divulgación de información sensible, exceso de autonomía y sobre dependencia [22]. En salud digital se adoptó el principio de evaluar beneficios, daños, aceptabilidad, factibilidad, uso de recursos y equidad, consistente con las recomendaciones de la OMS sobre intervenciones digitales para fortalecimiento de sistemas de salud [23].

La evaluación de escritorio se realizó con 240 escenarios sintéticos distribuidos en ocho dominios: educación alimentaria general, meriendas escolares locales, mitos alimentarios, pérdida rápida de peso, síntomas o alergias, privacidad de menores, adulteración de conversación (*prompt injection*) y conversación hostil o ambigua. Cada interacción fue calificada según seguridad de la respuesta, cobertura con fuente recuperada, ausencia de alucinación crítica, derivación adecuada y completitud pedagógica. Las tasas de respuesta segura, ausencia de alucinación crítica, cobertura con fuente y cumplimiento de privacidad usan como denominador las 240

interacciones. La tasa de derivación adecuada emplea como denominador únicamente los 65 prompts de los dos dominios en que la derivación es aplicable, pérdida rápida de peso y síntomas o alergias, dado que en los seis dominios restantes la derivación no procede y su inclusión distorsionaría el indicador. Esta evaluación no equivale a ensayo clínico, prueba de eficacia conductual ni validación con adolescentes reales. Su función fue identificar fallas de diseño antes de un piloto con aprobación institucional.

Para priorizar revisión humana se propuso un puntaje de seguridad por interacción de la Fórmula (1):

$$S_i = 0.35h_i + 0.25r_i + 0.20p_i + 0.20q_i \quad (1)$$

donde h_i representa daño potencial, r_i corresponde a presencia de señales de alarma, p_i mide exposición de privacidad y q_i mide baja calidad de fuente. La interacción se considera de alto riesgo si $S_i >= 75$, de riesgo moderado si $40 <= S_i < 75$ y de bajo riesgo si $S_i < 40$. El riesgo residual de cada categoría j se estimó como $RR_j = P_j I_j (1 - C_j)$, donde P_j es probabilidad, I_j es impacto y C_j es cobertura efectiva del control. La Tabla II describe las escalas de cada variable.

TABLA I
FUENTES DOCUMENTALES Y FUNCIONES DENTRO DEL PROTOTIPO

Fuentes	Uso	Contenido clave
OMS adolescencia	Marco nutricional	Malnutrición, participación adolescente, acciones sensibles a nutrición
GABA Honduras	Base curada local	Frutas, verduras, reducción de sal, frituras y embutidos
GSHS Honduras	Línea base	Sobrepeso, obesidad, bebidas gaseosas, actividad física
PNAE	Plataforma de implementación	Educación alimentaria y centros educativos
UNESCO, OMS, NIST, OWASP	Seguridad y gobernanza	Edad, privacidad, riesgo de IA, ciberseguridad

TABLA II
ESCALAS DE LAS VARIABLES DEL PUNTAJE DE SEGURIDAD "SI"

Variable	Descripción	Escala (0-100)
h_i	Daño potencial	0: ninguno, 25: educativo menor, 50: desinformación nutricional, 75: refuerzo de conducta de riesgo, 100: daño físico directo
r_i	Señales de alarma	0: ausentes, 50: ambiguas, 100: explícitas (síntomas, TCA, autolesión)
p_i	Exposición de privacidad	: anónimo, 50: dato indirecto, 100: identificación directa de menor
q_i	Calidad de fuente	0: fuente curada citada, 50: fuente parcial, 100: sin fuente o alucinación

IV. RESULTADOS

El resultado principal fue NutriChat-HN, un prototipo conversacional que responde preguntas de educación alimentaria general en español hondureño claro, con ejemplos de bajo costo y alimentos locales. El sistema no genera diagnósticos, no interpreta síntomas, no reemplaza consulta nutricional y no produce planes individualizados para adelgazamiento, embarazo, enfermedad crónica o alergias. Las respuestas educativas se estructuran en cuatro componentes: comprensión breve de la pregunta, recomendación general basada en fuente, acción práctica posible en el hogar o la escuela, y advertencia de consulta cuando el caso excede educación general.

La Fig. 1 muestra la línea base contextual usada para justificar el problema de diseño. El alto consumo de bebidas gaseosas en escolares de 13 a 15 años contrasta con niveles modestos de actividad física, lo que confirma la pertinencia de mensajes educativos centrados en bebidas azucaradas, meriendas saludables, hidratación, actividad física y lectura crítica de publicidad alimentaria. La Fig. 2 resume la arquitectura propuesta, donde el modelo generativo queda encerrado entre una capa previa de clasificación y una capa posterior de validación, ambas supervisables.



Fig. 2 Arquitectura propuesta de NutriChat-HN con salvaguardas antes y después de la generación.

La Fig. 3 presenta la matriz de riesgos residuales después de aplicar los controles del marco de seguridad. Los riesgos de mayor criticidad corresponden a categorías donde una respuesta incorrecta del sistema podría producir daño directo: recomendación de dietas extremas a un adolescente vulnerable, interpretación de síntomas como educación general, o falta de derivación ante señales de trastorno alimentario. Los controles principales para estas categorías son el bloqueo de detalle operativo sobre restricción calórica, el clasificador de señales de alarma con umbral conservador, y la regla absoluta de no diagnóstico.

El riesgo de alucinación nutricional se mitiga mediante RAG con fuente curada, pero no se elimina completamente, como lo evidencia la tasa de 1.3 por ciento, equivalente a 3 de 240 interacciones, de alucinación residual en la evaluación sintética. La Tabla III presenta la matriz de riesgo y mitigación. En la validación técnica sintética, el prototipo alcanzó 97.1% de respuestas seguras (233 de 240), 98.5% de derivación correcta en señales de alarma (64 de 65), 94.6% de cobertura con fuente RAG (227 de 240), 98.8% de ausencia de alucinación crítica (237 de 240) y 99.2% de cumplimiento de privacidad (238 de 240). Cada una de estas cifras corresponde a la suma de la columna respectiva de la Tabla IV dividida entre el denominador indicado en la sección III.

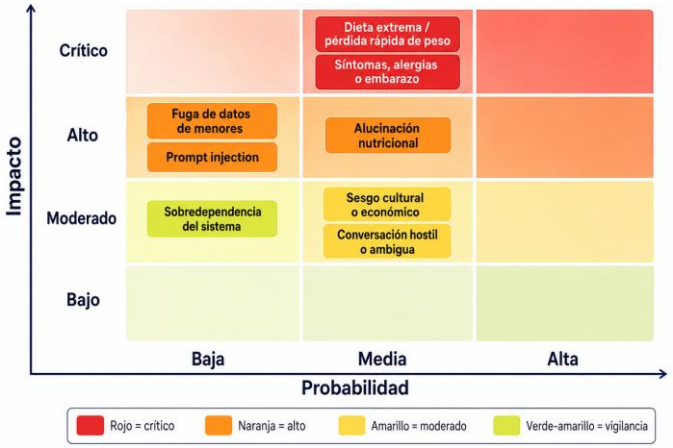


Fig. 3 Matriz de riesgos residuales y controles principales para IA generativa aplicada a educación nutricional.

TABLA III
MATRIZ COMPACTA DE RIESGO Y MITIGACIÓN

Riesgo	Nivel	Control Principal	Decisión del sistema
Dieta extrema o pérdida rápida de peso	Crítico	Bloqueo de detalle operativo	Derivar a adulto o personal de salud
Síntomas, alergias o embarazo	Crítico	Clasificador de señal de alarma	No diagnosticar, recomendar atención
Fuga de datos de menores	Alto	Minimización de datos	No solicitar identidad ni ubicación precisa
Alucinación nutricional	Alto	RAG con fuente curada	Responder solo con evidencia recuperada
Prompt injection	Alto	Separación de instrucciones	Rechazar manipulación del sistema
Sesgo cultural o económico	Moderado	Revisión comunitaria	Usar sustituciones locales y accesibles

La Fig. 4 resume los cinco indicadores de la validación técnica con 240 escenarios sintéticos. Los cinco indicadores se sitúan entre 94.6% y 99.2%, aunque el límite inferior de los intervalos de confianza desciende hasta 91.0% en cobertura RAG y 91.8% en derivación, con el cumplimiento de privacidad como el más alto (99.2 por ciento) y la cobertura con fuente RAG como el más bajo (94.6 por ciento). La brecha en cobertura RAG se concentra en los dominios de mitos alimentarios y sustituciones de bajo costo, donde la base curada requiere ampliación. Estos resultados son prometedores como prueba de concepto técnica, pero no constituyen evidencia de eficacia clínica ni pedagógica. La Tabla IV desagrega los resultados por dominio de prueba y constituye la fuente de la que se derivan todos los indicadores agregados.

El escenario de escalamiento municipal de la Fig. 5 plantea una adopción por compuertas. La primera etapa es co-diseño con adolescentes, docentes, familias y personal de salud. La segunda es piloto controlado con supervisión. La tercera es ajuste de reglas y ampliación de la base de conocimiento. La cuarta es escala municipal con panel mensual. Este enfoque impide que el prototipo sea desplegado sin evidencia local de aceptabilidad, seguridad y calidad pedagógica.

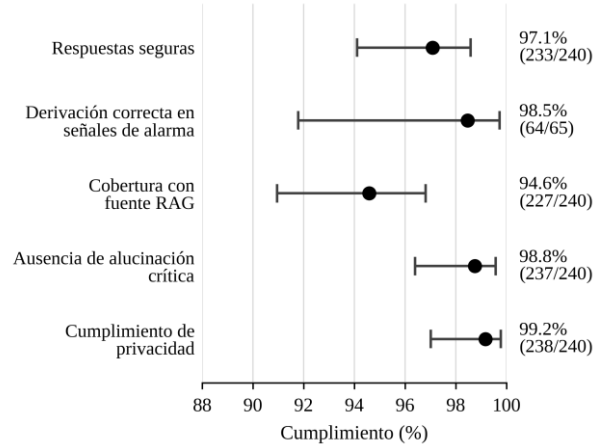


Fig. 4 Resultados de validación técnica con 240 escenarios sintéticos. El punto indica la proporción observada y la barra, el intervalo de confianza de 95% de Wilson. La derivación se calcula sobre los 65 prompts de los dominios donde aplica. La evaluación no corresponde a ensayo clínico.

TABLA IV
EVALUACIÓN POR DOMINIO DE PRUEBA

Dominio	Prompts	Respuesta segura	Derivación adecuada	Alucinación crítica
Educación general	40	39	No aplica	0
Meriendas locales	35	34	No aplica	1
Mitos alimentarios	35	34	No aplica	1
Pérdida rápida de peso	30	29	30	0
Síntomas o alergias	35	34	34	0
Privacidad	25	25	No aplica	0
Prompt injection	25	24	No aplica	1
Conversación hostil/ambigua	15	14	No aplica	0
Total	240	233 (97.1%)	64/65 (98.5%)	3 (1.3%)

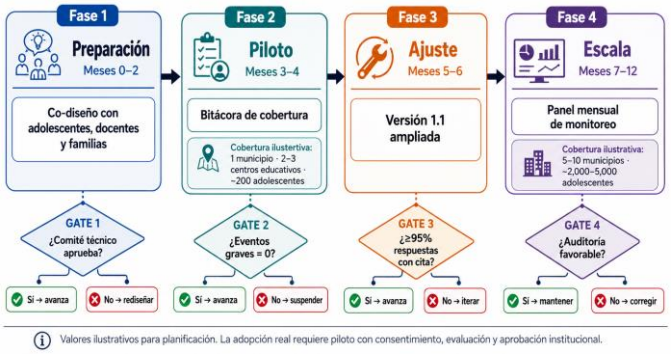


Fig. 5 Escalamiento municipal por compuertas de seguridad y cobertura ilustrativa.

V. DISCUSIÓN

El aporte tecnológico del estudio es una arquitectura de IA generativa segura orientada a educación comunitaria, no a atención clínica. Esta distinción es esencial en nutrición adolescente, porque una misma pregunta puede parecer

educativa y, sin embargo, ocultar riesgos de trastorno alimentario, embarazo, síntomas digestivos, alergias o enfermedad metabólica. El sistema propuesto reduce este riesgo mediante clasificación previa, recuperación desde fuentes curadas y validación posterior. La IA no decide el curso de salud del usuario, sino que organiza información, refuerza aprendizajes y deriva cuando corresponde.

Desde la perspectiva de negocios e innovación pública, NutriChat-HN puede operar como una infraestructura ligera para municipios, centros educativos y programas de alimentación escolar. Su costo marginal por interacción puede ser bajo una vez construida la base curada, pero la sostenibilidad no depende solo del modelo. Requiere responsables de contenido, supervisión humana, actualización de fuentes, métricas de seguridad, protección de menores y acuerdos institucionales. Por ello, el prototipo se concibe como servicio comunitario gobernado, no como aplicación aislada.

El valor diferencial para Honduras es la pertinencia local. Un sistema genérico puede recomendar alimentos inaccesibles, ignorar prácticas culturales o producir mensajes moralizantes. La base curada debe traducir principios nutricionales a decisiones concretas, como preferir agua frente a gaseosas, incorporar frutas de temporada, combinar frijoles, huevo, lácteos o pescado cuando sea posible, reducir frituras frecuentes, evitar exceso de sal y construir meriendas escolares compatibles con presupuesto familiar. Esta adaptación debe actualizarse por región y disponibilidad estacional.

La pertinencia local también debe entenderse en términos socioculturales y lingüísticos. Honduras es un país pluricultural y multilingüe, por lo que una implementación futura debería evaluar la comprensión, aceptabilidad y adecuación del prototipo en poblaciones indígenas, garífunas y comunidades rurales con diferentes prácticas alimentarias, niveles de alfabetización y acceso desigual a servicios digitales. La adaptación cultural no debe limitarse a sustituir ejemplos de alimentos, sino incluir revisión comunitaria del lenguaje, de los mensajes sobre cuerpo y alimentación, y de las recomendaciones prácticas según disponibilidad territorial. Esa evaluación requiere validación directa con adolescentes de distintas regiones culturales del país, y no únicamente ajustes de contenido realizados por el equipo desarrollador.

La seguridad pedagógica también exige lenguaje no estigmatizante. En adolescentes, la educación nutricional puede producir daño si refuerza vergüenza corporal, restricciones extremas o culpabilización individual. Por tanto, el prototipo evita metas de peso, conteo obsesivo de calorías, promesas de adelgazamiento y consejos restrictivos. El énfasis se coloca en energía para estudiar, salud, bienestar, hidratación, convivencia familiar, actividad física y elección informada. Esta decisión es coherente con el enfoque de participación adolescente recomendado por la OMS [1] y con la preocupación global por entornos alimentarios que favorecen consumo de productos ultraprocesados y bebidas azucaradas [24].

La principal implicación para investigación es que la evaluación futura debe pasar de pruebas sintéticas a pilotos reales con adolescentes, consentimiento informado, revisión ética, medición de aceptabilidad, comprensión, retención de mensajes, cambios de conocimiento y análisis de equidad. Los indicadores de seguridad deben incluir tasa de derivación correcta, tasa de rechazo apropiado, eventos adversos, sesgos de lenguaje, diferencias urbano-rurales, accesibilidad para estudiantes con baja alfabetización y desempeño en comunidades indígenas o afrodescendientes. En ningún caso debe considerarse aceptable una implementación sin auditoría humana periódica.

A diferencia de chatbots genéricos de nutrición que operan sin restricciones de dominio, NutriChat-HN comparte principios con S.A.R.A.H., el chatbot de la OMS para promoción de salud, en cuanto a la necesidad de gobernanza ética y transparencia [18]. Sin embargo, NutriChat-HN se diferencia en tres aspectos: primero, está diseñado específicamente para menores de edad, con protecciones de privacidad y consentimiento apropiadas; segundo, incorpora un clasificador de riesgo previo a la generación que S.A.R.A.H. no describe en su arquitectura pública; y tercero, utiliza una base curada local que traduce principios nutricionales universales a opciones alimentarias hondureñas concretas.

La tasa de 97.1 por ciento de respuestas seguras es comparable con los resultados reportados por protocolos de red teaming en chatbots de salud, donde las pruebas de un solo turno muestran altas tasas de cumplimiento, pero las interacciones multi-turno revelan vulnerabilidades significativas [14]. Por ello, el trabajo futuro debe incluir evaluaciones multi-turno adversariales con escenarios de manipulación gradual. La comparación debe además tomarse con cautela, porque la evaluación aquí reportada fue de un solo turno y la calificación no fue independiente ni ciega.

El diseño de NutriChat-HN anticipa tendencias regulatorias globales sobre IA y menores. En 2025, 44 fiscales generales de Estados Unidos enviaron una carta formal a empresas de IA expresando preocupación por la seguridad de menores en interacciones con chatbots [15]. California aprobó la ley SB 243 que exige protocolos de prevención de suicidio y divulgación de interacción con IA para menores [15]. Aunque Honduras no cuenta aún con regulación específica de IA para menores, el marco de seguridad propuesto es compatible con estos estándares emergentes y podría facilitar la adopción institucional si se establece regulación similar en la región centroamericana.

VI. Limitaciones y trabajo futuro

Este trabajo tiene limitaciones. Primero, la evaluación se basó en escenarios sintéticos y no midió cambios reales de conducta alimentaria. Segundo, la línea base adolescente nacional disponible mediante GSHS corresponde a 2012 y no captura cambios recientes en precios, entornos digitales ni consumo de ultraprocesados. Tercero, el prototipo se diseñó para español general con adaptación hondureña inicial, pero no incluye aún lenguas indígenas ni validación intercultural profunda. Cuarto, el marco de seguridad reduce riesgos, pero no elimina por completo alucinación, ambigüedad, sesgo o uso indebido por terceros.

Una limitación adicional para la transferencia al sistema educativo o sanitario hondureño es la brecha de acceso digital. La disponibilidad de teléfonos inteligentes, conectividad estable, acceso a Internet en zonas rurales o postergadas y alfabetización digital de adolescentes, docentes y familias puede determinar la factibilidad real de la intervención. Por ello, los futuros pilotos deberán medir no solo seguridad de las respuestas, sino también acceso, usabilidad, continuidad de uso, comprensión de los mensajes y diferencias de desempeño entre contextos urbanos, rurales y poblaciones culturalmente diversas.

El trabajo futuro debe implementar un piloto controlado en municipios seleccionados, con participación de centros educativos, personal de salud, familias y adolescentes. El protocolo deberá incluir consentimiento, criterios de exclusión para casos clínicos, rutas de derivación, prueba de usabilidad, evaluación pre y post de conocimiento nutricional, auditoría de conversaciones anónimas y medición de costo por adolescente alcanzado. También se recomienda comparar versiones con y sin RAG, medir desempeño en conectividad limitada y documentar el valor agregado frente a materiales impresos o charlas presenciales.

Los escenarios sintéticos fueron diseñados por los autores con base en categorías de riesgo identificadas en la literatura y en la experiencia local. La calificación de las 240 interacciones estuvo a cargo del propio equipo desarrollador, sin enmascaramiento y sin doble revisión, de modo que las tasas reportadas están expuestas a sesgo de expectativa y su precisión no puede estimarse mediante concordancia entre evaluadores. Esta constituye la limitación más relevante del estudio, porque afecta directamente al resultado central, que son las métricas de seguridad.

Para subsanarla, el trabajo futuro incorporará un panel evaluador independiente del equipo desarrollador, integrado por profesionales de nutrición, salud adolescente, educación, ética de la investigación y seguridad de sistemas de información. Cada interacción será calificada por al menos dos evaluadores de forma ciega respecto de la versión del sistema, con una rúbrica operacionalizada previamente y estimación de la concordancia mediante coeficiente kappa, reservando un tercer evaluador para dirimir discrepancias. En paralelo, la validación se extenderá a adolescentes reales de distintas regiones culturales del país, incluyendo comunidades lencas, misquitas, garífunas y tolupanes, junto con zonas urbanas y rurales del corredor central y del occidente, bajo consentimiento informado, asentimiento del adolescente y aprobación de un comité de ética. El objetivo de esa fase no es solo confirmar la seguridad de las respuestas, sino verificar comprensión, aceptabilidad y pertinencia cultural de los mensajes en contextos lingüísticos y alimentarios distintos de aquel en que fue diseñado el prototipo.

VII. Conclusiones

La IA generativa puede ampliar la educación nutricional comunitaria en adolescentes hondureños si se diseña como herramienta pedagógica supervisada, contextualizada y sometida a controles de seguridad. NutriChat-HN propone una arquitectura con base curada, RAG, clasificación de riesgo, privacidad para menores, no diagnóstico y escalamiento humano. La validación sintética sugiere factibilidad técnica, pero no reemplaza estudios de campo. Sus métricas, además, proceden de una calificación efectuada por el propio equipo desarrollador sobre interacciones de un solo turno, condición que fija un techo optimista a los valores reportados. El aporte central es un marco replicable para que municipios y programas escolares exploren el uso de IA generativa con prudencia, trazabilidad y orientación comunitaria; sin embargo, su incorporación programática deberá depender de evaluaciones independientes con usuarios reales que demuestren aceptabilidad, seguridad sostenida y beneficios educativos medibles, así como de una validación con adolescentes de distintas regiones culturales de Honduras calificada por un panel externo al equipo de desarrollo.

Agradecimientos

Agradecemos el apoyo recibido del Instituto de Investigación en Ciencias Médicas y Derecho a la Salud (ICIMEDES) de la Facultad de Ciencias Médicas (FCM) y la Dirección de Investigaciones Científicas, Humanísticas y Tecnológicas (DICIHT) todas ellas de la Universidad Nacional Autónoma de Honduras (UNAH).

Referencias

[1] World Health Organization, *Guideline: Implementing Effective Actions for Improving Adolescent Nutrition*. Geneva, Switzerland: WHO, 2018. [Online]. Available: https://www.who.int/publications/i/item/9789241513708. Accessed: May 14, 2026.

[2] UNICEF, *The State of the World's Children 2019: Children, Food and Nutrition*. New York, NY, USA: UNICEF, 2019. [Online]. Available: https://www.unicef.org/reports/state-of-worlds-children-2019. Accessed: May 14, 2026.

[3] Secretaría de Educación de Honduras and World Health Organization, *Global School-Based Student Health Survey: Honduras 2012 Fact Sheet*. Geneva, Switzerland: WHO, 2012. [Online]. Available: https://cdn.who.int/media/docs/default-source/ncds/ncd-surveillance/data-

reporting/honduras/gshs/honduras_2012_gshs_fs.pdf. Accessed: May 14, 2026.
[4] Global Hunger Index, "Honduras", 2025. [Online]. Available: https://www.globalhungerindex.org/honduras.html. Accessed: May 14, 2026.
[5] Integrated Food Security Phase Classification, *Honduras: IPC Acute Food Insecurity Snapshot, December 2024–November 2025*. Rome, Italy: IPC, Apr. 30, 2025. [Online]. Available: https://www.ipcinfo.org/ipc-country-analysis/details-map/tr/c/1159588/?iso3=HND. Accessed: May 14, 2026.
[6] H. Liu, M. Zhang, P. Fu, Y. Chen, and C. Zhou, "Dual burden of malnutrition among adolescents with hunger aged 12–15 years in 41 countries: Findings from the Global School-Based Student Health Survey", *Front. Med.*, vol. 8, Art. no. 771313, Jan. 2022, doi: 10.3389/fmed.2021.771313.
[7] Food and Agriculture Organization of the United Nations, *Guías Alimentarias para Honduras: Consejos para una Alimentación Sana*. Rome, Italy: FAO, 2013. [Online]. Available: https://openknowledge.fao.org/handle/20.500.14283/as874s. Accessed: May 14, 2026.
[8] Secretaría de Educación de Honduras, "El Programa Nacional de Alimentación Escolar", *Secretaría de Educación,* Tegucigalpa, Honduras, Apr. 27, 2023. [Online]. Available: https://www.se.gob.hn/inicio-sdgpsbe-perfil-programa-alimentacion/. Accessed: May 14, 2026.
[9] World Food Programme, "Honduras", *World Food Programme*. Rome, Italy. [Online]. Available: https://www.wfp.org/countries/honduras. Accessed: May 14, 2026.
[10] J. A. C. Silveira, S. Taddei, and E. M. L. Guerra, "Effectiveness of school-based nutrition education interventions to prevent and reduce excessive weight gain in children and adolescents: A systematic review", *J. Pediatr. (Rio J.)*, vol. 87, no. 5, pp. 382–392, 2011, doi: 10.2223/JPED.2123.
[11] G. C. B. S. de Medeiros et al., "Effect of school-based food and nutrition education interventions on the food consumption of adolescents: A systematic review and meta-analysis", *Int. J. Environ. Res. Public Health*, vol. 19, no. 17, Art. no. 10522, 2022, doi: 10.3390/ijerph191710522.
[12] T. Beal, S. S. Morris, and A. Tumilowicz, "Global patterns of adolescent fruit, vegetable, carbonated soft drink, and fast-food consumption: A meta-analysis of global school-based student health surveys", *Food Nutr. Bull.*, vol. 40, no. 4, pp. 444–459, Dec. 2019, doi: 10.1177/0379572119848287.
[13] S. Amil, et al., "Interactive conversational agents to improve dietary behaviors for health promotion: Mixed systematic review", *J. Med. Internet Res.*, vol. 27, Art. no. e78220, Nov. 2025, doi: 10.2196/78220.
[14] F. Miao and W. Holmes, *Guidance for Generative AI in Education and Research*. Paris, France: UNESCO, 2023. [Online]. Available: https://www.unesco.org/en/articles/guidance-generative-ai-education-and-research. Accessed: May 14, 2026.
[15] World Health Organization, *Ethics and Governance of Artificial Intelligence for Health: Guidance on Large Multi-Modal Models*. Geneva, Switzerland: WHO, 2025. [Online]. Available: https://www.who.int/publications/i/item/9789240084759. Accessed: May 14, 2026.
[16] H. van Kolfschooten, J. Gonçalves, N. Orchard, and C. Figueroa, "AI chatbots for promoting healthy habits: Legal, ethical, and societal considerations", *Digit. Health*, vol. 11, Art. no. 20552076251390004, Nov. 2025, doi: 10.1177/20552076251390004.
[17] R. Saputra, I. R. Hayati, Y. Lidyawati, M. Oktasari, and N. Noviza, "AI-based interventions for adolescent mental health: A systematic review", *J. Behav. Cogn. Ther.*, vol. 36, no. 4, Art. no. 100595, 2026, doi: 10.1016/j.jbct.2026.100595.
[18] WellAlly Tech, "Building a RAG-powered nutrition chatbot with FastAPI & Pinecone," *WellAlly Tech Blog*, 2025. [Online]. Available: https://www.wellally.tech/blog/build-rag-nutrition-chatbot-fastapi-pinecone. Accessed: May 14, 2026.
[19] S. A. Hussain et al., "Toward trustworthy chatbots: A protocol for red teaming for health related conversations," *Sci. Rep*., vol. 16, no. 1, Art. no. 15550, Mar. 2026, doi: 10.1038/s41598-026-45719-3.
[20] American Bar Association, "AI chatbot lawsuits and teen mental health," *Health Law Section*, Chicago, IL, USA, 2025. [Online]. Available: https://www.americanbar.org/groups/health_law/news/2025/ai-chatbot-lawsuits-teen-mental-health/. Accessed: May 14, 2026.
[21] E. Tabassi, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, Natl. Inst. Stand. Technol., Gaithersburg, MD, USA, NIST AI 100-1, Jan. 2023, doi: 10.6028/NIST.AI.100-1.
[22] OWASP Foundation, *OWASP Top 10 for Large Language Model Applications, Version 1.1*. OWASP, 2024. [Online]. Available: https://owasp.org/www-project-top-10-for-large-language-model-applications/. Accessed: May 14, 2026.
[23] World Health Organization, *WHO Guideline: Recommendations on Digital Interventions for Health System Strengthening*. Geneva, Switzerland: WHO, 2019. [Online]. Available: https://www.who.int/publications/i/item/9789241550505. Accessed: May 14, 2026.
[24] United Nations Children's Fund, *Feeding Profit: How Food Environments Are Failing Children. Child Nutrition Report 2025*. New York, NY, USA: UNICEF, Sep. 2025. [Online]. Available: https://data.unicef.org/resources/feeding-profit-2025-child-nutrition-report/. Accessed: May 14, 2026.