CLIAS

CENTRO DE INTELIGENCIA ARTIFICIAL Y SALUD PARA AMÉRICA LATINA Y EL CARIBE

REPORTE FINAL

Hacia el diseño inclusivo de grandes modelos de lenguaje clínico en español: caracterización y medición de daños representacionales para la salud trans.

INSTITUCIÓN

CIECTI - Centro Interdisciplinario de Estudios en Ciencia, Tecnología e Innovación.

PERÍODO DEL INFORME

15/11/24 - 14/11/25

IECS INSTITUTO DE EFECTIVIDAD CLINICA Y SANITARIA

CIPS IMPLEMENTACIÓN E INNOVACIÓN EN POLÍTICAS DE SALUD

IDRC · CRDI Canadá

INTEGRANTES DEL EQUIPO

Enzo Ferrante
Jocelyn Dunstan
Blas Radi
Marina Elichiry
Tamara Quiroga
Micaela Hirsh
David Restrepo
Verónica Xhardez

CONTENIDO

01. INTRODUCCIÓN

Las personas trans enfrentan barreras persistentes de estigmatización y discriminación en el acceso a la salud sexual y reproductiva (SSyR). En este contexto, la creciente incorporación de grandes modelos de lenguaje (LLMs) en aplicaciones de salud, como chatbots, sistemas de apoyo clínico y agentes automatizados, introduce nuevos riesgos, en particular la reproducción o intensificación de sesgos y daños representacionales hacia poblaciones históricamente vulneradas.

El proyecto se propuso estudiar de manera sistemática cómo los LLMs en idioma español representan a las personas trans en el ámbito de la SSyR, identificando potenciales daños y sesgos implícitos que puedan derivar en consecuencias negativas para el acceso, la calidad de atención y la equidad en salud. La investigación se orientó a generar herramientas conceptuales y metodológicas que permitan auditar estos sistemas desde una perspectiva ética, de género y de derechos.

02. OBJETIVOS

El objetivo general fue comprender si los grandes modelos de lenguaje en idioma español pueden incurrir en sesgos y daños de representación específicos al contexto de la salud, particularmente en cuanto a las representaciones de las personas trans, haciendo especial énfasis en los potenciales perjuicios en el ámbito de la Salud Sexual y Reproductiva.
Los objetivos específicos planteados fueron:

1. Conceptualizar los daños representacionales en que pueden incurrir los LLMs especializados en salud, con especial énfasis en la SSyR de las personas trans.
2. Construir una base de datos para llevar adelante dichas auditorías.
3. Construir una metodología para auditar la existencia de sesgos y daños representacionales en SSyR trans para sistemas de LLMs comerciales, generalistas y/o abiertos en idioma español.
4. Evaluar la existencia de sesgos y daños representacionales en SSyR trans para sistemas de LLMs comerciales, generalistas y/o abiertos en idioma español.

03. ACTIVIDADES REALIZADAS

El proyecto se desarrolló mediante un enfoque interdisciplinario que combinó métodos cualitativos y cuantitativos. En una primera etapa, se realizó una revisión bibliográfica y se organizaron mesas de trabajo participativas con personas trans usuaries del sistema de salud y con expertes en SSyR y análisis de sesgos, orientadas a identificar experiencias de discriminación y daños representacionales.

A partir de estos insumos, se construyó un catálogo de daños representacionales, junto con una rúbrica graduada para su evaluación sistemática. Paralelamente, se diseñó y depuró un corpus de preguntas (*prompts*) sobre SSyR, utilizado para evaluar las respuestas de distintos modelos de lenguaje en español.

Se llevaron a cabo evaluaciones comparativas de seis LLMs mediante juicios de expertes humanos y modelos utilizados como jueces automáticos, analizando niveles de acuerdo y divergencia. Complementariamente, se desarrollaron experimentos cuantitativos para medir sesgos implícitos y daños asignacionales, incluyendo tareas de asociación de conceptos, asignación automatizada de turnos médicos, diagnóstico clínico a partir de texto e imágenes y análisis del uso de lenguaje inclusivo.

Finalmente, se realizaron actividades de formación interna, documentación técnica, producción de informes y difusión de resultados en ámbitos académicos y científicos.

04. RESULTADOS

El proyecto produjo un marco conceptual y metodológico pionero para la identificación, evaluación y medición de daños representacionales hacia personas trans en grandes modelos de lenguaje aplicados a la salud sexual y reproductiva.

Entre los principales resultados se destaca la elaboración de una taxonomía de daños representacionales, una rúbrica de evaluación graduada, un corpus de prompts y respuestas que permiten auditar de manera sistemática el comportamiento de LLMs en español, y un estudio sobre sesgos implícitos hacia personas trans en grandes modelos de lenguaje. Estos instrumentos demostraron ser útiles para detectar formas recurrentes de

estigmatización, patologización, invisibilización y malgenerización en las respuestas de los modelos.

Además, en cuanto a sesgos implícitos, los experimentos cuantitativos evidenciaron la presencia de dichos sesgos persistentes en múltiples LLMs, incluyendo asociaciones negativas hacia identidades trans y patrones de sesgo en tareas de asignación de recursos y agentes automatizados. Si bien en los escenarios de diagnóstico automatizado no se observaron diferencias drásticas, sí se identificaron comportamientos diferenciales relevantes desde una perspectiva de equidad.

El proyecto generó evidencia empírica temprana y reutilizable que contribuye al diseño, supervisión y auditoría de sistemas de IA en salud, fortaleciendo las capacidades institucionales para el desarrollo de modelos más justos, responsables e inclusivos y, además, más fiables en términos de calidad y seguridad clínica al reducir errores y sesgos en los contenidos en SSyR para personas trans. Así, el proyecto también sienta bases sólidas para futuras etapas de investigación e innovación en el área.

CLIAS

CENTRO DE INTELIGENCIA
ARTIFICIAL Y SALUD
PARA AMÉRICA LATINA
Y EL CARIBE

IECS
INSTITUTO DE EFECTIVIDAD
CLINICA Y SANITARIA

CiPS
IMPLEMENTACIÓN
E INNOVACIÓN EN
POLÍTICAS DE SALUD

IDRC · CRDI
Canada