



Detección de depredadores sexuales utilizando grandes modelos de lenguaje

Autor: Yerko Bryan Reyes Avila
Patrocinante: Pedro Pablo Pinacho Davidson
Co-patrocinantes: Fernando Andree Tercero Gutiérrez Gómez
Mario Vegas Barbas

Informe de memoria presentada a la Facultad de Ingeniería de la Universidad de Concepción para optar al título de Ingeniero Civil Informático.

Agosto, 2024
Concepción, Chile



Resumen

La creciente utilización de las redes sociales y plataformas de mensajería ha abierto nuevas vías para la comisión de delitos en línea, entre ellos el grooming, que se refiere al acoso sexual hacia menores de edad mediante el engaño y la manipulación. Esta memoria presenta el desarrollo de un sistema de detección de grooming en conversaciones de chat, utilizando grandes modelos de lenguaje (LLM) para identificar patrones de comportamiento sospechoso y clasificar chats con posibles casos de grooming.

El sistema propuesto emplea una heurística iterativa, apoyada por un módulo de detección de fugas de privacidad (Presidio), que permite analizar y categorizar las conversaciones en diferentes etapas iniciales del grooming: Primer contacto, Generar confianza y Migrar de plataforma. A través de varias iteraciones, el sistema ha sido refinado para mejorar su precisión, reduciendo la dependencia de una única consulta al LLM y ajustando los puntajes de detección para minimizar falsos positivos y negativos.

Los resultados obtenidos muestran que el sistema es capaz de detectar con alta precisión los casos de grooming, especialmente cuando se ajustan los puntajes para dar mayor peso a la clasificación del LLM. Sin embargo, también se destacan los desafíos éticos y técnicos asociados a la implementación de este tipo de sistemas en tiempo real, así como la necesidad de futuras mejoras en la integración de modelos más específicos y el entrenamiento de estos.

Palabras clave: Grooming, Sexual Predator Identification, LLM, Inteligencia Artificial



Índice

Resumen	2
Índice	3
1 Introducción	5
1.1 Objetivos	6
1.1.1 Objetivo general	6
1.1.2 Objetivos específicos	6
2 Marco teórico	7
2.1 Grooming	7
2.1.1 Modus operandi del groomer	7
2.1.2 Detección de depredadores sexuales (SPD)	8
2.1.3 PAN	8
2.2 Grandes modelos del lenguaje (LLM)	8
2.2.1 Token	9
2.2.2 Prompt	9
2.3 Estado del arte de SPD	9
2.3.1 Detección temprana de depredadores sexuales (eSPD)	9
2.3.2 Método de ensamble para SPD	9
2.3.3 Fine-tuning Llama2 para SPD	9
2.4 Herramientas de desarrollo	10
2.4.1 Langchain	10
2.4.2 Heurística	10
2.4.3 Presidio	10
3 Integración con servicios de mensajería instantánea	11
3.1 Dificultades técnicas	11
3.2 Alternativas	11
4 Desarrollo de la propuesta	12
4.1 Patrones identificados por la PDI	12
4.2 Metodología de desarrollo	13
4.3 Desarrollo de la heurística	13
4.3.1 Estados de la conversación	13
4.3.2 Primer enfoque	14
4.3.3 Limitaciones del primer enfoque	16



- 4.3.4 Segundo enfoque 16
- 4.3.5 Limitaciones del segundo enfoque 18
- 4.3.6 Tercer enfoque 19
- 4.3.7 Ejemplo de funcionamiento de la tercera heurística..... 22
- 4.3.8 Modelos..... 22
- 4.3.9 Implementación 23
- 5 Evaluación 24
 - 5.1 Construcción del set de datos 24
 - 5.2 Limitaciones de pruebas de consistencia del clasificador 26
 - 5.3 Diseño de experimentos..... 27
 - 5.3.1 Objetivos de los resultados..... 27
 - 5.3.2 Metodología de evaluación..... 27
 - 5.4 Resultados 29
 - 5.4.1 Detección de grooming con Presidio 29
 - 5.4.2 Análisis de resultados de grooming con Presidio..... 30
 - 5.4.3 Detección de no grooming con Presidio 32
 - 5.4.4 Análisis de no grooming con Presidio 33
 - 5.4.5 Detección de grooming sin Presidio..... 34
 - 5.4.6 Análisis de grooming sin Presidio 35
 - 5.4.7 Detección de no grooming sin Presidio..... 36
 - 5.4.8 Análisis de no grooming sin Presidio 37
- 6 Discusión de resultados..... 38
- 7 Conclusión..... 39
- 8 Anexos 40
 - 8.1 Respuestas entregadas por GPT-4o 40
 - 8.2 Respuestas entregadas por GPT-4o-mini..... 40
 - 8.3 Respuestas entregadas por Llama3-8B 41
- 9 Referencias 43



1 Introducción

Las plataformas de mensajería en línea, como los mensajes de chat y las aplicaciones de mensajería instantánea en las redes sociales, se han convertido en una alternativa a los medios de comunicación estándar. Estas plataformas permiten a las personas intercambiar mensajes de igual a igual sin moderación explícita de contenido, lo que puede ser explotado para diversos fines maliciosos. Los chats dirigidos pueden usarse para difundir odio, manipular a la víctima para propaganda, coordinar actividades criminales o terroristas [1,2], radicalización y, en el peor de los casos, para dirigirse a usuarios en línea menores de edad (menores y niños) para favores sexuales y abuso [3, 4]. A diferencia de los chats públicos o discusiones, los mensajes dirigidos, en la mayoría de los casos, explotan una relación en línea ya existente con otros miembros del grupo en la red [1,2]. En casos donde tal relación previa no existe, el actor malicioso invierte tiempo en construir la relación, a menudo referida como “grooming” en línea, y eventualmente dirige su atención a la víctima [5].

La investigación sobre la visión retrospectiva del grooming en línea experimentado por menores mostró que el 25% de los participantes menores hablaron con adultos desconocidos [6,7]. Más importante aún, el 65% de aquellos que hablaron con un extraño experimentaron una solicitud sexual por parte de un extraño adulto. El 23% de los participantes revelaron que tuvieron conversaciones con extraños adultos que siguieron un patrón de grooming, y alrededor del 38% estableció una relación confidencial con el acosador [6]. En 2019 la Policía de Investigaciones de Chile recibió 1077 denuncias que derivó en 4124 casos investigados bajo la figura penal abuso sexual impropio de mayor 14 años, abuso sexual impropio en menor de 14 años, adquisición o almacenamiento de pornografía infantil, comercialización y/o distribución de este último, en donde más del 10% de ellas las desarrolló la Brigada Investigadora del Cibercrimen (Bricib) Metropolitana. El panorama local en la región del Bío Bío también es desalentador, puesto que, en el mismo artículo, ese mismo año se recibieron 89 denuncias que se abrieron más de 500 casos de investigación [8].

Es fácil detectar tales delitos en comunicaciones públicas como redes sociales y chats públicos, ya que pueden monitorearse con moderadores que evalúen manualmente el contenido o mediante un mecanismo automatizado como el uso de filtros de groserías [9,10]. Los algoritmos automatizados de moderación de chat/discusión pública pueden diseñarse con datos de entrenamiento a gran escala. Sin embargo, varios desafíos pueden preverse en el diseño y uso efectivo de los algoritmos. Por ejemplo, los datos a gran escala pueden no estar disponibles para entrenar algoritmos de moderación, o las regulaciones de privacidad imponen restricciones sobre el uso de dichos datos incluso cuando están disponibles [11]. Tales desafíos tienen un impacto negativo en los avances para prevenir el uso indebido de plataformas de mensajería en línea.

Por otro lado, un sistema de vigilancia eficaz para proteger a los niños en línea requiere entender el comportamiento de los depredadores. Esta comprensión permite mejores métodos de detección y educar a los niños para que sepan cómo reaccionar en situaciones peligrosas [12]. Sin embargo, los casos de informática forense digital necesitan pruebas sólidas para usar en la corte, lo que implica analizar grandes cantidades de datos, lo que aumenta la carga de trabajo en la investigación.

Es por ello, que la rama de machine learning ha desarrollado múltiples investigaciones que aporten a la creación de sistemas de detección de grooming que permiten analizar grandes cantidades de datos de forma automatizada. Muchos de estos acercamientos se han basado en métodos de clasificación de texto que generalmente implican una combinación de dos pasos principales, a saber, extracción de características y diseño de clasificadores. Para que todo el enfoque logre un rendimiento satisfactorio, los investigadores pueden necesitar recorrer una extensa lista de combinaciones que involucran



métodos de extracción de características y métodos de clasificación para encontrar la mejor sincronización o sinergia entre ambos [13]

Recientemente, los emergentes grandes modelos de lenguaje (LLM), que constituyen una gran red neuronal preentrenada en un corpus masivo de datos de texto en diferentes idiomas, tienen el potencial de abordar los problemas mencionados anteriormente de los métodos de clasificación de texto convencionales. Los LLM podrían tener un conocimiento del mundo sustancial, comprensión del lenguaje y capacidades de razonamiento común. Con estas capacidades, los LLM preentrenados pueden ser extremadamente beneficiosos para diversas tareas. Por ejemplo, podrían ser más precisos para casos de uso de diálogo o para resolver problemas de clasificación de texto [14].

1.1 Objetivos

1.1.1 Objetivo general

Desarrollar un sistema de detección de grooming en chats utilizando grandes modelos del lenguaje para identificar y clasificar chats con posibles casos de depredadores sexuales en línea.

1.1.2 Objetivos específicos

- Revisar el estado del arte en relación con la detección de grooming en chats utilizando Machine Learning.
- Identificar set de datos de carácter público que permitan desarrollar y evaluar el sistema de detección y su modelo.
- Desarrollar un sistema que utilice un modelo que analice, clasifique chats y detecte posibles casos de grooming.
- Evaluar el rendimiento del sistema mediante métricas usadas en la literatura.
- Investigar formas de integrar el sistema con plataformas de mensajería existentes para su implementación práctica.



2 Marco teórico

El marco teórico de este trabajo se centra en los conceptos fundamentales que subyacen a la problemática del grooming y en las tecnologías utilizadas para abordarla. Se explora el fenómeno del grooming en línea, describiendo su modus operandi y las características de los depredadores sexuales (groomers). Además, se examinan los grandes modelos de lenguaje (LLM), los cuales constituyen la base tecnológica para la detección automatizada de este tipo de comportamientos. Este marco teórico sienta las bases para la comprensión del enfoque propuesto y la elección de herramientas y metodologías empleadas.

2.1 Grooming

Actualmente no hay unicidad respecto a la definición de grooming, porque, de facto, este solo puede aplicarse cuando la acción ya ha sido realizada, esto es, hablar de grooming en un caso concreto exige retrospectiva, dado que no resulta fácil identificar este delito antes de llevarse a cabo el abuso, ni determinar en qué momento preciso comienza y acaba el acoso [15].

Una definición más consensuada es la que nos da Wood, Wilson y Thorne:

“El grooming puede definirse como acoso sexual a través de internet o como delito de captación de niños y niñas por medio de internet con fines sexuales; o como técnica de acoso y abuso en internet (incluso explotación y trata de menores) en la que los adultos se suelen hacer pasar por menores e intentan establecer un contacto con niños y adolescentes que dé pie a una relación de confianza [16].”

El **groomer** es la persona que ejerce el grooming, y puede ser cualquiera, independientemente de su edad, género o raza. Su acoso puede durar desde semanas hasta años [16]. El modus operandi del groomer incluye estrategias específicas para ganarse la confianza de un menor con el objetivo de explotarlo sexualmente.

2.1.1 Modus operandi del groomer

Se refiere a las estrategias que utiliza una persona para ganarse la confianza de un menor con el objetivo de explotarlo sexualmente.

Según Grosvenor y Lamb [16], el groomer opera de la siguiente manera:

- 1) Crear un perfil falso en una o varias cuentas
- 2) Seleccionar la víctima. Esta selección puede depender de varios factores, como el atractivo de la víctima, la percepción de vulnerabilidad del menor o fácil acceso a internet
- 3) Ganar la confianza del menor, atrayendo emocionalmente a la víctima y desarrollando una confianza engañosa en ella.
- 4) Solicitar fotografías comprometidas, tratar temas sexuales o tener encuentros físicos.
- 5) Chantajear al menor por medio de amenazas o coacción (van Gijn-Grosvenor y Lamb, 2016; Winters y Jeglic, 2017)

O'Connell (2003) diferencia cinco fases en la relación. Su clasificación de estas fases es tomada como punto de partida para muchos estudios de tipificación del grooming y del groomer:

- 1) Fase amistad: el delincuente contacta con el menor para conocerlo y crear una relación de confianza



- 2) Fase sentimental: tanto victimario como víctima empiezan a hacerse confesiones mutuas. El vínculo es más fuerte
- 3) Fase de evaluación de riesgo: a partir de información sobre el entorno cercano a la víctima el delincuente se crea un perfil de ella para considerar si puede ser una víctima real
- 4) Fase de exclusividad: A partir de las fases anteriores, el depredador habla sobre temas sexuales. Aquí intenta conseguir material pornográfico
- 5) Fase sexual: el depredador utiliza material sexual recabado para chantajear a la víctima

2.1.2 Detección de depredadores sexuales (SPD)

La detección de depredadores sexuales, o *Sexual Predator Detection* (SPD), es un campo de investigación en Inteligencia Artificial y Procesamiento de Lenguaje Natural enfocado en la identificación y mitigación de casos de acoso a menores a través de plataformas digitales, como servicios de mensajería instantánea. El objetivo principal es identificar y clasificar patrones de comportamiento y lenguaje en conversaciones en línea que puedan indicar la presencia de un depredador sexual intentando acosar a un menor. Este campo ha sido ampliamente explorado para identificar amenazas potenciales en plataformas en línea. Los enfoques típicos incluyen la recopilación de datos más relevantes, la extracción de las características clave, y el diseño de un clasificador para tomar decisiones precisas. Si estos enfoques se diseñan correctamente, también pueden proporcionar tiempos de procesamiento más rápidos, permitiendo la detección temprana de depredadores y la reducción de amenazas en línea para los jóvenes.

2.1.3 PAN

PAN (Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection) es una competencia anual que forma parte de la conferencia CLEF (Conference and Labs Evaluation Forum). La competencia se centra en el desarrollo y evaluación de métodos y herramientas para la detección de plagio, la identificación de autoría y la detección de documentos casi duplicados.

2.1.3.1 PAN12

En la edición de 2012 de la competencia PAN, se incluyó una tarea llamada "Sexual Predator Identification", que se centró en la identificación de depredadores en línea. Esta tarea se evaluó utilizando un conjunto de datos con el mismo nombre, compuesto por registros de chats provenientes de canales IRC y del sitio Omegle. Estos chats incluían cibersexo, conversaciones casuales y conversaciones con *grooming*. Los chats fueron divididos en segmentos cada vez que la conversación era interrumpida por más de 25 minutos, y se filtraron todos los segmentos con más de 150 mensajes. Este proceso resultó en un total de 220 mil segmentos, de los cuales el 2,58% correspondía a chats con *grooming*. El objetivo fue imitar una distribución realista de las conversaciones en línea en este conjunto de datos.

2.2 Grandes modelos del lenguaje (LLM)

Un modelo de lenguaje grande (LLM) es un tipo de programa de inteligencia artificial (IA) que puede reconocer y generar texto, entre otras tareas. Estos modelos se entrenan con enormes conjuntos de datos, lo que justifica el adjetivo "grande". Los LLM se basan en el aprendizaje automático, utilizando específicamente un tipo de red neuronal conocida como modelo transformer. En los últimos años, ha



habido una proliferación de LLM, como ChatGPT de OpenAI o Gemini de Google, debido a sus potentes capacidades. Estas capacidades hacen que sea interesante explorar su habilidad para reconocer patrones de *grooming* en chats. Los LLM procesan la entrada en forma de tokens, que son la unidad más pequeña de texto que el modelo puede manejar. La longitud de un token varía según el modelo específico en cuestión.

2.2.1 Token

Los tokens son las unidades mínimas de texto que un modelo de lenguaje puede procesar. Cada palabra, puntuación o incluso parte de una palabra puede ser un token, dependiendo del modelo. En el caso de los modelos grandes de lenguaje (LLM) como GPT-4, la longitud de los tokens es crucial, ya que determina la capacidad del modelo para procesar entradas largas.

2.2.2 Prompt

El "prompt" es la instrucción o entrada que se le da al modelo de lenguaje para que realice una tarea específica. En el contexto de la detección de grooming, el prompt puede incluir descripciones de patrones de comportamiento sospechoso o preguntas dirigidas para guiar al modelo a identificar indicios de grooming en una conversación.

2.3 Estado del arte de SPD

A continuación, se revisan las investigaciones y desarrollos más importantes y/o recientes en la detección de grooming utilizando tecnologías de inteligencia artificial.

2.3.1 Detección temprana de depredadores sexuales (eSPD)

Matthias Vogt, Ulf Leser y Alan Akbik investigaron el problema desde un punto de vista de la prevención. Su objetivo fue analizar el chat desde inicio a fin utilizando un modelo del lenguaje BERT y predecir lo antes posible un posible caso de grooming. Dado que necesitaban de un chat completo, que PAN12 no proporciona, también ellos decidieron crear un nuevo set de datos PANC, que consistía en la unión de los sets de datos ChatCoder2 y PAN12. Con ello, se entrenó a BERT, dando buenos resultados y su posible implementación en móviles [15].

2.3.2 Método de ensamble para SPD

El estudio de Muhammad Ali Fauzi y Patrick Bours, propone un método efectivo para identificar depredadores sexuales en chats en línea mediante una clasificación en dos etapas. La primera etapa distingue conversaciones depredadoras de las normales, y la segunda identifica al depredador y a la víctima dentro de una conversación depredadora. Se utilizaron varios clasificadores de aprendizaje automático, incluyendo Naive Bayes, SVM, y Redes Neuronales, junto con técnicas de ensamble para mejorar el rendimiento [13].

Los resultados que obtuvieron fueron los mejores del estado del arte en su momento [13].

2.3.3 Fine-tuning Llama2 para SPD

El enfoque que proponen Thanh Thi Nguyen, Campbell Wilson y Janis Dalins es más automatizado y sin necesidad de una búsqueda manual de sinergia entre obtener las características de los datos y el diseño del clasificador. Realizaron fine-tuning del modelo de Meta llamado Llama2 de 7 millones de parámetros con el set de datos PAN12. Hasta la fecha, ellos lograron obtener los mejores resultados del estado del arte [14].



2.4 Herramientas de desarrollo

El desarrollo de la propuesta se apoya en varias herramientas clave que permiten implementar y optimizar el sistema de detección de grooming. Se describen las características y funcionalidades de Langchain, un framework que facilita la integración de LLM en aplicaciones, así como de Presidio, una herramienta de código abierto de Microsoft que detecta y anonimiza información sensible en los datos textuales.

2.4.1 Langchain

Langchain es un framework diseñado para facilitar la implementación de aplicaciones que utilizan LLM. Su propósito principal es proporcionar herramientas y estructuras que permitan a los desarrolladores integrar, manejar y optimizar el uso de estos modelos de lenguaje en diversas aplicaciones, desde chatbots hasta análisis de texto y automatización de tareas.

2.4.2 Heurística

Una heurística es un enfoque o método utilizado para resolver problemas de manera más rápida y eficiente cuando los métodos tradicionales no son los óptimos. Las heurísticas no garantizan una solución óptima o perfecta, pero suelen ser suficientemente efectivas para alcanzar una solución aceptable en un tiempo razonable.

En este proyecto, el propósito es generar una heurística con un LLM, que nos permita identificar la presencia de grooming a través de patrones, por ejemplo, la presencia de un adulto y un menor de edad en la conversación o preguntas por horarios en donde los padres no estén, entre otros.

2.4.3 Presidio

Microsoft Presidio es una herramienta de código abierto diseñada para proteger información sensible mediante la identificación y anonimización de datos privados en texto e imágenes. Utiliza técnicas avanzadas de procesamiento de lenguaje natural (NLP) para detectar automáticamente entidades como números de tarjetas de crédito, nombres, direcciones y números de teléfono.

En el contexto de la detección de grooming, Presidio puede ayudar a identificar la información sensible en el flujo de conversaciones.



3 Integración con servicios de mensajería instantánea

Para que el sistema de detección de grooming sea efectivo en la práctica, es crucial integrarlo con las plataformas de mensajería instantánea donde ocurren la mayoría de estas interacciones. Esta sección aborda los desafíos técnicos de dicha integración, considerando las restricciones de privacidad de las aplicaciones y proponiendo alternativas como el uso de aplicaciones de control parental. Esta integración es esencial para aplicar el sistema en escenarios reales, asegurando así la protección efectiva de menores en línea.

3.1 Dificultades técnicas

- **Acceso restringido a mensajería:** Las aplicaciones de mensajería instantánea como WhatsApp, Messenger, y Telegram están diseñadas para proteger la privacidad de los usuarios, lo que significa que no permiten que otras aplicaciones accedan directamente a los mensajes o al contenido que se intercambia dentro de ellas.
- **Uso del Servicio de Accesibilidad:** El servicio de accesibilidad de Android puede usarse para monitorear lo que ocurre en pantalla, lo que permitiría capturar y analizar conversaciones en tiempo real. Sin embargo, esta técnica es invasiva y complicada, ya que se necesita interpretar dinámicamente la interfaz de usuario de aplicaciones que no fueron diseñadas para ser monitoreadas por terceros.

3.2 Alternativas

- **Aplicaciones de Control Parental:** En lugar de intentar acceder directamente a las aplicaciones de mensajería, se podría considerar la integración del sistema con aplicaciones de control parental existentes. Las aplicaciones de control parental son herramientas diseñadas para ayudar a los padres o tutores a monitorear y controlar el uso que los niños hacen de sus dispositivos digitales, como teléfonos móviles, tabletas y computadoras. Entre sus características principales:
 - **Monitoreo de actividad en línea:** pueden registrar las páginas web visitadas y bloquear el acceso a sitios web inapropiados basados en categorías (como pornografía, violencia, drogas, etc.) o en listas negras definidas por los padres.
 - **Monitoreo de Mensajería y Redes Sociales:** Permite a los padres monitorear la actividad en redes sociales como Facebook, Instagram, y Snapchat. Esto incluye la visualización de publicaciones, comentarios, y mensajes privados en algunos casos.
 - **Control de Tiempo de Pantalla:** permiten a los padres establecer límites de tiempo sobre cuánto pueden usar sus hijos el dispositivo. También pueden programar horarios en los que ciertas aplicaciones o funciones del dispositivo estén bloqueadas.
 - **Supervisión de Llamadas y Contactos:** Algunas aplicaciones permiten ver el registro de llamadas, incluyendo la duración y frecuencia de las llamadas entrantes y salientes. En algunos casos, también permiten bloquear números específicos.

Un ejemplo popular es Qustodio, un servicio de control parental.

Entre las características que ofrece está la monitorización de actividad lo cual permite visualizar de manera fácil y en tiempo real el historial de navegación, las publicaciones de Facebook, historial de reproducción de Youtube y demás.

Por consiguiente, un camino viable para la implementación de esta herramienta es a través de la construcción de una aplicación de control parental o integración con una ya existente



4 Desarrollo de la propuesta

En esta sección se describe el proceso de desarrollo del sistema de detección de grooming en conversaciones de chat utilizando grandes modelos de lenguaje (LLM). A partir de patrones identificados por la Policía de Investigaciones (PDI) y un enfoque metodológico iterativo, se detallan las distintas fases de diseño, implementación y optimización de la heurística que guía al sistema. El objetivo central fue crear una herramienta precisa y eficiente que pueda identificar posibles casos de grooming en sus etapas iniciales, evaluando diferentes enfoques y modelos para lograr un equilibrio entre la precisión de la detección y la minimización de falsos positivos.

4.1 Patrones identificados por la PDI

Este análisis se enfoca en comprender el fenómeno del grooming en el contexto chileno, utilizando información proporcionada por la Policía de Investigaciones (PDI). Se explora la dinámica del grooming en línea, donde adultos se aprovechan de plataformas de juegos y redes sociales para manipular a menores de edad con intenciones sexuales. La PDI ofrece una visión detallada de las características tanto de los groomers como de sus víctimas en Chile, identificando patrones clave que han sido fundamentales para el diseño de las heurísticas en el sistema propuesto. Esta sección también considera las necesidades específicas de las autoridades en términos de la recolección rápida y eficaz de evidencia digital, crucial para la persecución penal de estos delitos.

Durante una reunión entre el equipo de LLM Security y personal de la PDI, se identificaron varios patrones relevantes. Se determinó que las víctimas de grooming en Chile generalmente tienen entre 10 y 14 años, y suelen ser abordadas en plataformas de videojuegos en línea. Los groomers atraen a sus víctimas ofreciéndoles ítems de juego o tarjetas de regalo a cambio de favores sexuales. Aunque no hay una diferencia significativa en cuanto al género de los groomers, su rango de edad se sitúa entre los 20 y 40 años. Estos individuos a menudo persuaden a los menores para que borren las conversaciones y se trasladen a plataformas más privadas y sin monitoreo, lo que incrementa el riesgo de abuso.

Además, en la reunión se abordaron distintos escenarios de grooming, que van desde los niños que son atacados inocentemente en videojuegos hasta aquellos menores que, por diferentes razones, están en búsqueda de este tipo de interacciones. En el caso de los niños más vulnerables, muchos groomers utilizan videojuegos como herramienta de contacto, ofreciendo recompensas virtuales o muestras de afecto para captar la atención del menor. Este tipo de grooming es común en aquellos niños que desconocen los riesgos en las plataformas de juego en línea.

Por otro lado, se discutieron también casos donde los menores, ya familiarizados con este tipo de comportamiento, pueden llegar a buscar activamente este tipo de interacciones. Esto ocurre en situaciones donde el menor, quizás por una necesidad emocional o curiosidad, termina siendo más vulnerable ante los groomers, quienes aprovechan esa búsqueda de atención para iniciar conversaciones de confianza que pueden llevar a abusos más graves.

Asimismo, la PDI ha señalado un aumento en la práctica de menores que venden contenido sexual, lo cual incrementa la vulnerabilidad ante los groomers. Esto subraya la necesidad urgente de desarrollar herramientas que permitan la recolección rápida y eficiente de evidencia digital, facilitando así la intervención de las autoridades y la protección de los menores involucrados.



4.2 Metodología de desarrollo

El desarrollo de la propuesta se llevó a cabo mediante un enfoque iterativo, centrado en proponer y evaluar distintas alternativas para mejorar la detección de grooming en conversaciones de. Cada alternativa fue cuidadosamente evaluada en conjunto con los patrocinantes del proyecto, permitiendo así definir la solución más efectiva basada en los datos obtenidos.

El proceso iterativo involucró los siguientes pasos:

- 1) **Definición del enfoque:** Cada enfoque del desarrollo comenzaba con la definición clara del objetivo a alcanzar, que puede incluir la definición una heurística base, mejoramiento de la precisión, la integración de nuevas variables, etc.
- 2) **Desarrollo de la solución:** Se desarrollaban ajustes y refinamientos en la heurística o en la configuración del LLM, integrando nuevos patrones de comportamiento sospechoso o nuevas tecnologías.
- 3) **Revisión con los patrocinantes:** Antes de realizar cualquier prueba experimental, los resultados del desarrollo o solución propuesta eran revisados en sesiones de análisis con los patrocinantes. Se discutían los cambios realizados, su potencial efectividad, y cualquier posible impacto negativo.
- 4) **Consenso para proceder:** Solo tras alcanzar un consenso entre todos los involucrados, se decidía si el enfoque propuesto estaba listo para ser evaluado experimentalmente. Este consenso incluía la aprobación de los patrocinantes sobre la efectividad y la justificación de los cambios realizados.
- 5) **Evaluación experimental:** Una vez aprobado, se procedía a las pruebas utilizando conjuntos de datos previamente seleccionados y etiquetados, evaluando la capacidad de la heurística para detectar grooming y minimizando errores de clasificación.

4.3 Desarrollo de la heurística

El desarrollo de la heurística para la detección de grooming se realizó mediante un enfoque iterativo, centrado en ajustar y refinar un modelo que pudiera identificar con precisión las etapas clave en una conversación que potencialmente implica grooming. Estas etapas son fundamentales porque permiten identificar los momentos críticos en etapas iniciales en los que un depredador sexual intenta ganarse la confianza de un menor y llevar la interacción a un entorno más privado o incluso a un encuentro físico.

4.3.1 Estados de la conversación

La definición de los estados de la conversación dentro de este sistema de detección de grooming responde a la necesidad de estructurar y clasificar las interacciones de manera que permita identificar patrones específicos del comportamiento de los groomers. Dado que el grooming puede ocurrir en una amplia variedad de escenarios, para esta prueba de concepto se decidió centrar el desarrollo en un escenario en particular: **el del niño inocente que juega videojuegos en línea**. Este enfoque fue seleccionado porque es una de las formas más comunes en las que los depredadores sexuales se aproximan a menores, y es un contexto en el que los niños pueden ser vulnerables sin estar conscientes de los riesgos.

En este escenario, los groomers suelen aprovecharse de la naturaleza inocente y desprevenida de los menores, ofreciendo ítems virtuales, recompensas, o incluso muestras de afecto como un medio para ganarse su confianza. Este contexto permitió identificar tres estados principales de la conversación que son clave para detectar grooming en sus primeras fases: **Primer contacto, Generar confianza, y Migrar de plataforma**.



- **Primer contacto:** es la etapa inicial donde el groomer establece el primer contacto con la víctima. Aquí, el depredador suele intentar entrar en confianza con el menor, utilizando tácticas como halagos o promesas de regalos.
- **Generar confianza:** en esta etapa, el groomer y la víctima ya han establecido un cierto nivel de confianza. El depredador comienza a hacer preguntas más personales, busca obtener información sobre la vida privada del menor, como sus horarios, la presencia de los padres, y sus emociones.
- **Migración de plataforma:** en esta fase, el groomer sugiere cambiar la conversación a una plataforma más privada o menos monitoreada, como una aplicación de mensajería instantánea. Este cambio suele coincidir con la preparación para un encuentro físico o para acciones que el groomer prefiere mantener en secreto.

Estos tres estados fueron definidos específicamente para este contexto de prueba, con el fin de capturar las dinámicas del grooming que ocurren cuando un menor, jugando videojuegos en línea, es atacado por un depredador. A medida que el sistema evolucione, podría adaptarse a otros escenarios de grooming, pero para esta prueba de concepto se optó por un desarrollo centrado en este contexto, debido a su prevalencia y peligrosidad.

4.3.2 Primer enfoque

En este primer enfoque, el objetivo fue establecer una base para la detección de grooming utilizando un LLM. El enfoque se centró en crear un prompt estructurado que guiara al modelo en el análisis de las conversaciones.

El prompt se diseñó para incluir tres elementos principales:

- (1) Definición de la tarea
- (2) Descripción de la heurística
- (3) Instrucciones para el formato de salida

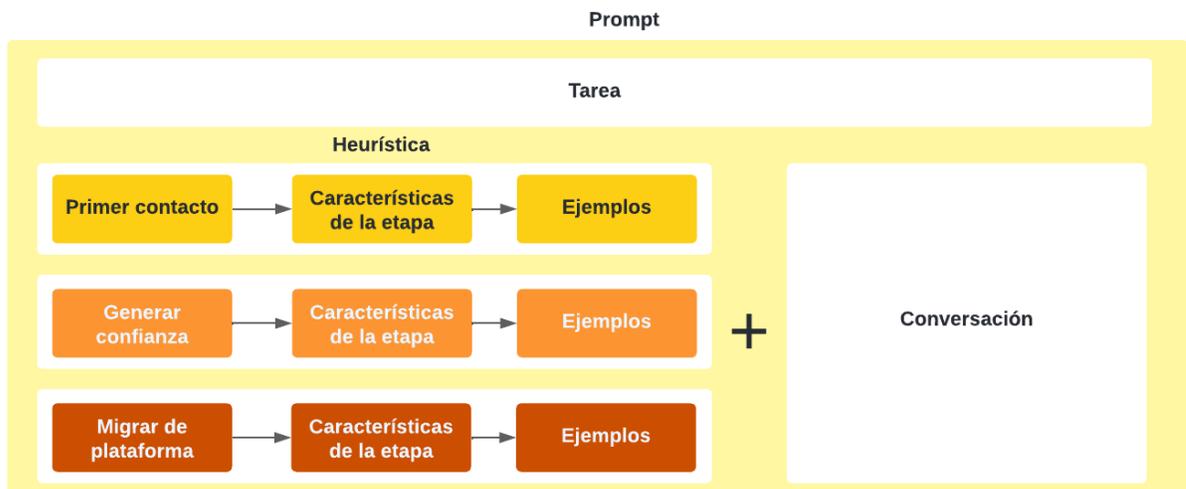


Ilustración 1: Input de primera heurística



En primer lugar, se le define la tarea a realizar:

“Analiza la conversación proporcionada para determinar la probabilidad de que esté ocurriendo grooming. Debes considerar los tres estados definidos en nuestra heurística: 'Primer contacto', 'Generar confianza', y 'Migrar de plataforma'. Proporciona una explicación concisa y detallada de tus hallazgos, destacando cualquier indicio de comportamiento sospechoso y proporcionando una recomendación basada en tu análisis”

Luego, se define la heurística donde por cada etapa se le da la descripción y características principales de cada una de ellas, además se le otorgan ejemplos:

Heurística

1. Primer contacto:

- En esta etapa el groomer ataca al niño, intenta entrar en confianza adulándolo, compartiendo información básica u ofreciendo cosas.
- Palabras clave como: "regalo", "premio", "gratis", "solo para ti".
- Frases típicas como: "¿Quieres una skin?", "Tengo un regalo para ti".

2. Generar confianza:

- En esta etapa el groomer y el niño ya tienen la suficiente confianza para compartir información más sensible, como horarios, presencia de los padres, lugar específico donde vive, sentimientos o problemas emocionales. El groomer también comparte dicha información de él, aquí ya pueden haber establecido algún vínculo afectivo
- Intenta formar una relación de confianza con el niño.
- Palabras clave como: "confía en mí", "no se lo digas a nadie", "somos amigos".
- Frases típicas como: "¿Cuántos años tienes?", "¿Dónde vives?", "¿Tienes novio?".

3. Migrar de plataforma:

- El groomer sugiere cambiar de plataforma (por ejemplo, de una red social a una aplicación de mensajería) o de facto concretar un encuentro físico.
- Insiste en mantener la conversación en secreto.
- Palabras clave como: "hablemos por WhatsApp", "no quiero que tus padres se enteren".
- Frases típicas como: "¿Tienes WhatsApp?", "¿Podemos hablar por privado?".



Ilustración 2: Output de primera heurística

Finalmente, se le pide que como salida nos entregue, cuál es el estado en que se encuentra la conversación, cuál es la probabilidad de grooming en el chat (alta, media o baja), indicadores que justifiquen la probabilidad y la etapa detectada, junto con una recomendación.

4.3.3 Limitaciones del primer enfoque

Aunque este primer enfoque proporcionó una base útil, se identificaron varias limitaciones:

- 1) Alta dependencia de una sola consulta al LLM, lo que podría llevar a resultados inconsistentes.
- 2) Riesgo de alucinación del LLM debido a la complejidad del prompt y el formato de salida requerido

Además, era necesario añadir algún elemento diferenciador al estado del arte, esto llevó a desarrollar un segundo enfoque de la heurística para abordar estos problemas.

4.3.4 Segundo enfoque

Basándonos en las limitaciones identificadas en el primer, desarrollamos una segunda versión de la heurística con el objetivo de mejorar la precisión y reducir la dependencia de una única consulta al LLM.

Tras identificar las limitaciones del primer enfoque, se desarrolló una segunda versión de la heurística con el objetivo de mejorar la precisión del sistema y reducir la dependencia de una única consulta al modelo de lenguaje grande (LLM). Esta iteración introdujo varios cambios significativos en la forma en que se realiza el análisis y se toman las decisiones.

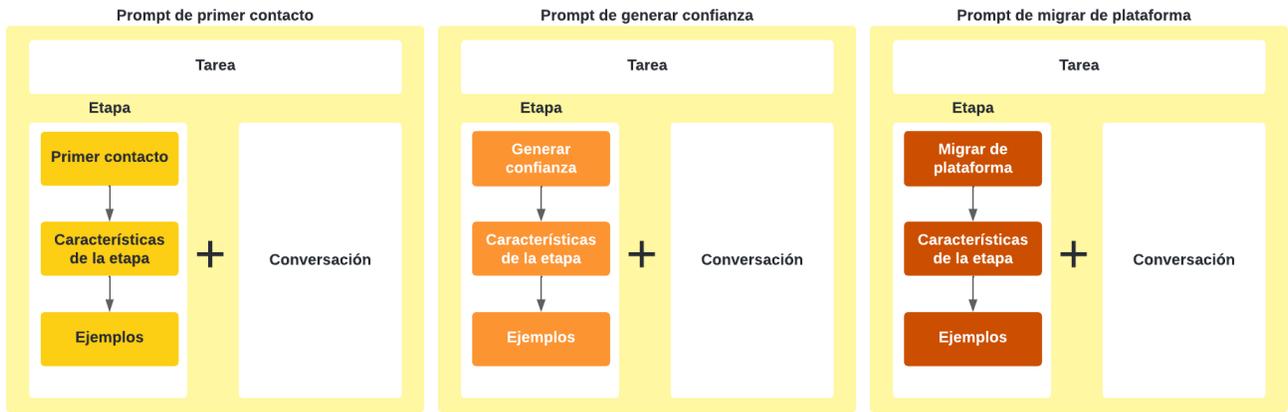


Ilustración 3: Prompt de segunda heurística

Uno de los principales cambios fue la división del proceso de análisis en tres etapas distintas: Primer Contacto, Generar Confianza y Migrar de Plataforma. En lugar de que el LLM decidiera en qué etapa se encontraba la conversación de manera global, ahora se realiza un análisis específico para cada una de estas etapas por separado.

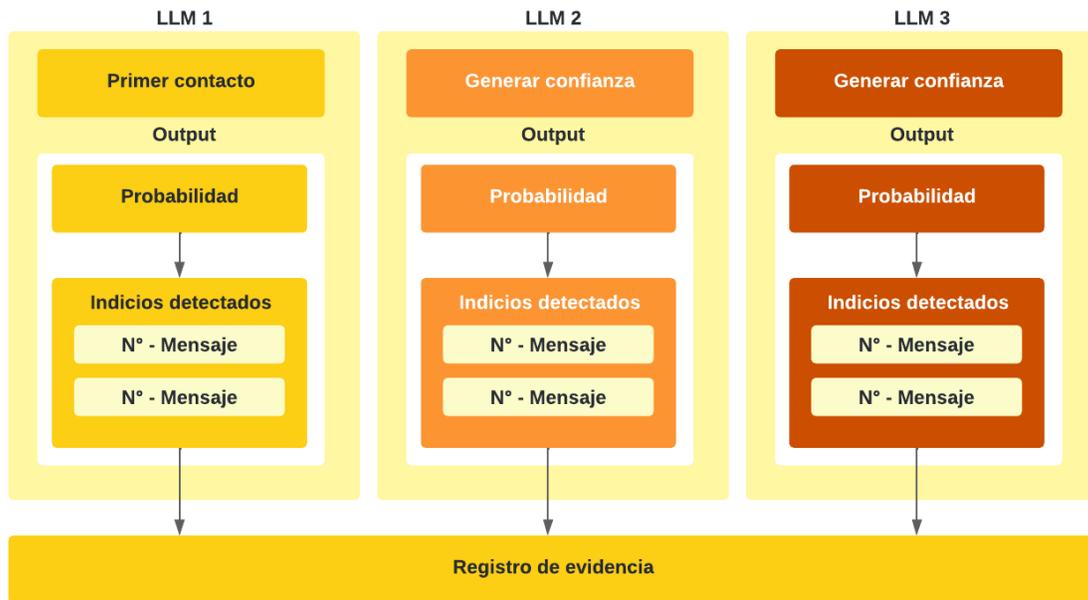


Ilustración 4: Output de segunda heurística

Esto significa que, para cada conversación, el LLM genera tres outputs diferentes, uno por cada etapa. Cada output proporciona una probabilidad de grooming y una serie de indicios que justifican esa probabilidad.

Además de dividir el análisis en etapas, se implementaron mejoras en la forma de manejar las probabilidades y los indicios. La probabilidad ahora se expresa como un valor discreto: 1 indica baja



probabilidad, 2 media, y 3 alta probabilidad. Este cambio permite integrar la probabilidad en una heurística más compleja que combina los resultados de las tres etapas.

Los indicios detectados en cada etapa se almacenan en un diccionario, donde cada entrada del diccionario corresponde a un mensaje específico en la conversación. Este formato estructurado facilita la justificación de las decisiones del modelo, proporcionando un registro de evidencia más detallado que puede ser utilizado para análisis posteriores o en contextos forenses.

Este registro es un recurso valioso en un contexto forense, ya que documenta explícitamente los elementos clave que llevaron a una determinada conclusión, permitiendo que cualquier análisis posterior sea minucioso y verificable.

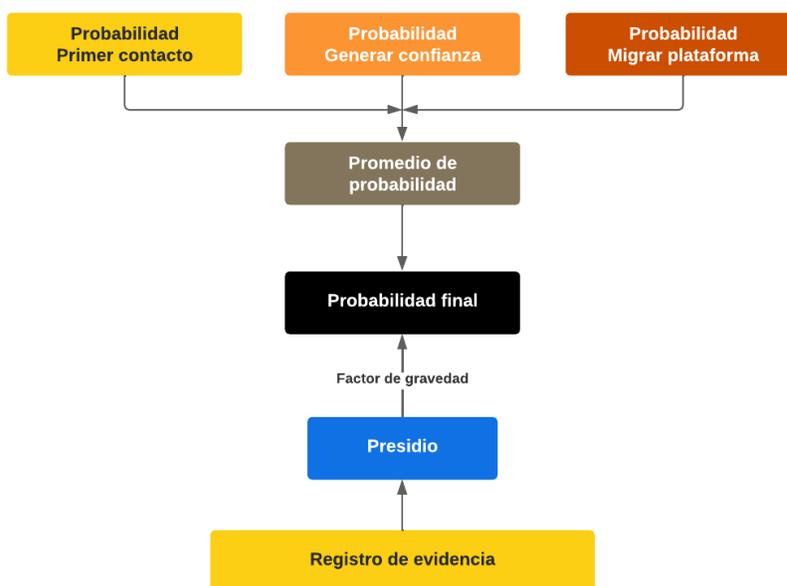


Ilustración 5: Cálculo de probabilidad final

Una innovación clave en este enfoque fue la integración de Presidio para analizar el **Registro de Evidencia**. Presidio identifica entidades de privacidad, como nombres, direcciones, y otros datos sensibles, y el resultado de este análisis se utiliza para calcular un factor de gravedad. Este factor pondera la probabilidad final de grooming en la conversación en función de la cantidad y el tipo de entidades de privacidad detectadas.

Finalmente, la probabilidad final de grooming se calcula combinando el promedio de las probabilidades de las tres etapas con el factor de gravedad derivado del análisis de Presidio. Este enfoque integrado no solo busca lograr un balance entre la precisión en la detección y la minimización de falsos positivos, sino que también asegura que cada decisión esté respaldada por un rastro de evidencias verificable, lo que maximiza la utilidad del sistema en contextos forenses.

4.3.5 Limitaciones del segundo enfoque

La heurística ahora es más compleja y se diferencia del estado del arte, implementando un módulo de fugas de privacidad que influye en la probabilidad final. Sin embargo, se presentan nuevas limitaciones.



Por un lado, Presidio es una herramienta que fue desarrollada antes de la llegada de los grandes modelos de lenguaje (LLM), lo que limita su capacidad para comprender el contexto completo en el que se utiliza una palabra detectada como una entidad sensible para la privacidad. Por ejemplo, si se analiza una conversación en la que los usuarios discuten sobre sus artistas favoritos, la gran cantidad de nombres mencionados en el chat podría hacer que el módulo de privacidad calcule un factor de gravedad elevado, generando un falso positivo. Esto se debe a que Presidio no puede discernir si estos nombres realmente constituyen un riesgo para la privacidad o no en función del contexto.

Por otro lado, si en alguna de las etapas del análisis una alta probabilidad de **grooming** es detectada mientras que las demás arrojan probabilidades bajas, el promedio entre estas puede reducir la alerta potencial, atenuando la señal de advertencia. Además, tampoco es posible otorgar un mayor peso a una etapa específica del análisis, ya que las conversaciones no siempre siguen un desarrollo lineal. Es decir, un chat no necesariamente avanzará de la etapa de primer contacto a la de generar confianza y finalmente a la de migración de plataforma de forma secuencial; en su lugar, las interacciones pueden fluctuar entre estas etapas de manera no predecible.

Este enfoque ilustra las limitaciones tanto del uso de herramientas previas a los LLM como de la dificultad de asignar ponderaciones adecuadas en un proceso no lineal como lo son las conversaciones de grooming.

Estas observaciones llevaron a desarrollar un tercer enfoque de la heurística.

4.3.6 Tercer enfoque

El tercer enfoque de la heurística se desarrolló para abordar las limitaciones identificadas en la segunda versión, con un énfasis particular en mejorar la integración de Presidio y refinar el proceso de determinación de la etapa del chat.

El tercer enfoque de la heurística se desarrolló con el objetivo de abordar las limitaciones identificadas en el enfoque anterior, especialmente en lo que respecta a la integración de Presidio y la precisión en la determinación de las etapas del grooming en las conversaciones.

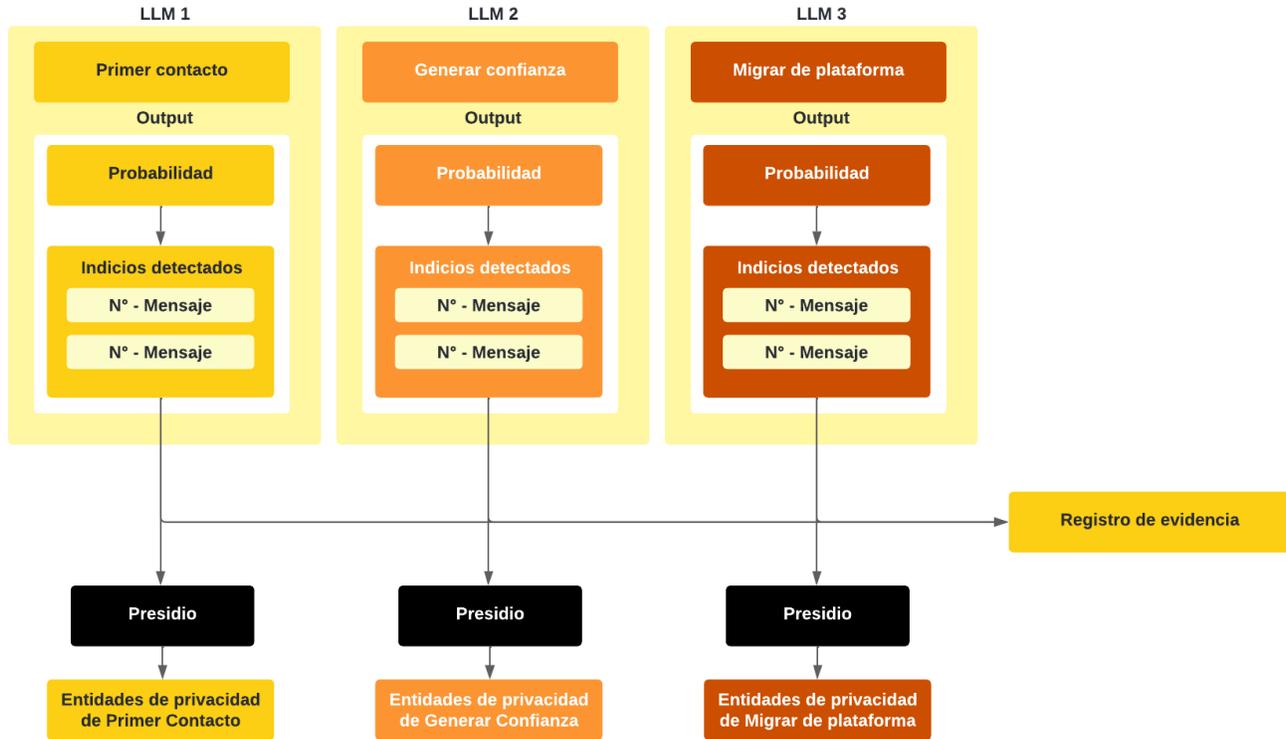


Ilustración 6: Ouput y análisis de Presidio en tercera heurística

En esta oportunidad, se realizó una modificación fundamental en el rol de Presidio dentro de la heurística. En lugar de influir directamente en la probabilidad final de detección de grooming, Presidio se utiliza ahora para analizar las fugas de privacidad de manera independiente en cada etapa del grooming. Esto significa que Presidio examina el contenido de cada etapa (Primer Contacto, Generar Confianza, y Migrar de Plataforma) de forma separada, identificando entidades de privacidad como nombres, direcciones y otros datos sensibles.

Las entidades detectadas por Presidio se agrupan y se hace un conteo en cada etapa, lo que permite calcular un puntaje de fugas de privacidad.

Tabla 1: Ponderaciones en función de la etapa

Entidad	Primer contacto	Generar confianza	Migrar de plataforma
Persona (nombres)	1	0.8	0.4
NRP (Nacionalidad, Religión, Grupo político)	1	1	0.2
Localidad	1	1	1
Estampa de tiempo	0.5	0.5	1



Correo electrónico	0.2	0.5	1
Número de celular	0.2	0.5	1
URL	0.2	0.2	1
Dirección IP	0.2	0.2	1

Este puntaje se pondera de acuerdo con la relevancia de la información en cada etapa, basándose en una tabla de ponderaciones establecida previamente.

Justificación de las ponderaciones:

- La información personal (como nombres) es más relevante en etapas iniciales a diferencia de etapas de migración.
- Información de contacto y plataformas alternativas son más importantes en la etapa de migración.
- Información sobre ubicación es relevante en todas las etapas, en primer contacto es información básica para congeniar, en generar confianza podría tratarse de una ubicación más exacta para demostrar confianza y en migración podría indicar algún lugar de encuentro.

En este momento cada etapa tiene un puntaje o ponderación de fugas privacidad.

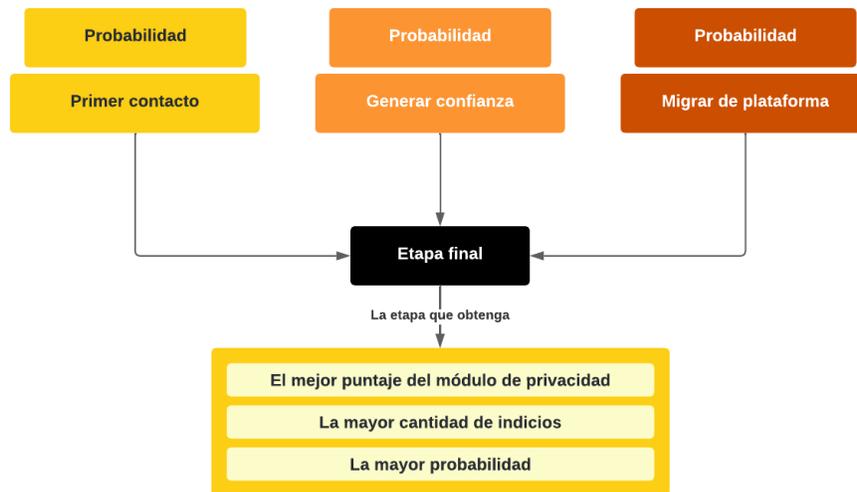


Ilustración 7: Sistema de puntaje y selección de etapa de tercera heurística

También se desarrolló un sistema de puntos para determinar la etapa en la que se encuentra la conversación. Ahora, este sistema considera tres factores clave: la ponderación del módulo de privacidad, la cantidad de indicios detectados por el LLM, y la probabilidad asignada por el LLM. Cada uno de estos factores se evalúa por separado en cada etapa, y la etapa que acumula la mayoría de los puntos se determina como la etapa actual de la conversación.



Proceso de funcionamiento:

- 1) Se asigna un punto al factor con el valor más alto en cada etapa.
- 2) La etapa que acumula la mayoría de los puntos se determina como la etapa actual de la conversación.
- 3) Esta etapa final determina tanto el estado como la probabilidad de grooming en el chat.

4.3.7 Ejemplo de funcionamiento de la tercera heurística

Supongamos una conversación con el siguiente análisis:

Tabla 2: Ejemplo de funcionamiento de tercera heurística

Etapa	Ponderación Presidio	Cantidad de indicios	Probabilidad LLM
Primer contacto	0.8	3	2
Generar confianza	1.2	5	3
Migración de plataforma	0.5	2	1

Se le asignará a “Generar confianza” un punto por obtener la mayor ponderación de Presidio, otro punto por obtener la mayor cantidad de indicios y otro más por obtener la mayor probabilidad dada por el LLM.

Como resultado, la etapa de la conversación será “Generar confianza” y la probabilidad de grooming es “3” o “Alta”.

4.3.8 Modelos

Los modelos que fueron considerados para implementar los clasificadores fueron los siguientes:

- **GPT-4o**: Modelo basado en GPT de OpenAI, es el modelo más capaz de la empresa hasta la fecha y de los más populares. Para utilizar este modelo en aplicaciones se utiliza una API la cual tiene un costo, el cual es por token.
- **GPT-4o-mini**: Modelo basado en GPT de OpenAI, es el modelo que viene a sustituir a GPT-3.5-Turbo, siendo también uno de los más utilizados y baratos en la actualidad. Para utilizar este modelo en aplicaciones se utiliza una API la cual tiene un costo, el cual es por token.
- **Llama3-8B**: Modelo de Meta, bastante popular en la comunidad de Hugging Face y de modelos de código abierto. La principal forma de utilizarlo es de manera local, por lo que se requiere utilizar hardware adecuado.

Para la elección del LLM se le dio a cada modelo 2 conversaciones del dataset PAN12 que contenían grooming de manera explícita, es decir, al menos uno de los participantes era un menor de edad y había patrones de grooming de por medio.

Luego se evaluó la calidad del output que entregaban utilizando un prompt básico que consistía en preguntar la presencia de grooming y entregar los mensajes textuales que justifican la presencia de este.



Tabla 3: Comparación de calidad de respuestas entre distintos modelos

Modelo	Calidad de respuesta
GPT-4o	Detectaba correctamente la presencia de grooming Era capaz de entregar el número del mensaje y su contenido textual respetando el formato esperado para Indicios.
GPT-4o-mini	No detectaba correctamente la presencia de grooming. Pese a detectar posibles patrones de grooming, no quiso asumir que el otro participante era alguien mayor. No respetaba el formato de indicios, puesto que no daba los mensajes textuales, sino más bien una interpretación o resumen de estos.
Llama3 8B	Detectaba correctamente la presencia de grooming. No respetaba el formato de indicios, entrega una interpretación o abreviación de las respuestas

Finalmente, se decidió para esta prueba de concepto utilizar el modelo más capaz de OpenAI, para no tener en cuenta consideraciones de hardware y obtener la mayor calidad de respuesta posible. Las respuestas completas entregadas por los modelos se pueden encontrar en el [Anexo 9.1 – 9.3](#).

4.3.9 Implementación

Para la implementación de la heurística con el uso de LLM, se utilizó el lenguaje de programación Python con el framework Langchain, debido a la facilidad que nos permite conectar la API de OpenAI con nuestro módulo de Presidio y con la lógica de la heurística.



5 Evaluación

En esta sección, se describe el proceso de evaluación, incluyendo la construcción del conjunto de datos, las limitaciones de las pruebas y los métodos utilizados para analizar los resultados. Para la evaluación de dicho objetivo se utiliza una versión reducida y etiquetada de PAN12, que consistirá en conversaciones seleccionadas específicamente para identificar patrones de grooming. Este conjunto de datos incluirá tanto casos de grooming como no grooming, con el fin de probar la eficacia del sistema y su capacidad para discriminar correctamente entre ambos tipos de conversaciones.

5.1 Construcción del set de datos

```
<conversations>
  <conversation id="e621da5de598c9321a1d505ea95e6a2d">
    <message line="1">
      <author>97964e7a9e8eb9cf78f2e4d7b2ff34c7</author>
      <time>03:20</time>
      <text>Hola.</text>
    </message>
    <message line="2">
      <author>0158d0d6781fc4d493f243d4caa49747</author>
      <time>03:20</time>
      <text>hi.</text>
    </message>
  </conversation>
</conversations>
```

Ilustración 8: Ejemplo de una conversación de PAN12

PAN12 consiste en una colección de mensajes de canales de IRC y chats de Omegle, cada conversación y participante contenía un ID único. Un participante o autor podía estar en más de una conversación.

El objetivo de PAN12 era identificar a los depredadores sexuales mediante los ID de los participantes, evaluados a través de un archivo proporcionado por los creadores del dataset que contenía una lista de todos los depredadores identificados. Sin embargo, aunque se incluían los ID de los depredadores, el dataset no proporcionaba una manera clara de corroborar cuáles conversaciones específicas correspondían a un ataque de grooming por estos groomers. Esto se debía a que los depredadores podían participar en múltiples conversaciones y no todas eran necesariamente casos de grooming, lo que dificultaba asociar de forma precisa las conversaciones predatorias con los ID de los participantes.



El set de datos se compone de la siguiente manera:



Ilustración 9: Composición del dataset PAN12

Hay un total de 357.622 chats, donde 11.350 son de tipo P (3% aproximadamente) y 346.272 son de tipo A y N.

Para este nuevo set de datos, el objetivo es obtener conversaciones predatorias y no predatorias de manera de poder evaluar si el modelo y heurística es capaz de detectar con precisión el grooming en caso de haberlo y no lanzar una alerta innecesaria en caso contrario.

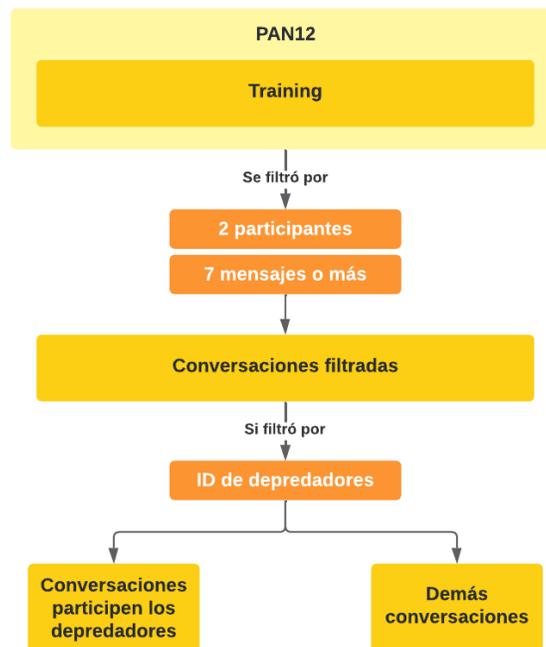


Ilustración 10: Proceso de construcción del subdataset de PAN12

Para cumplir con ello, se solo se utilizó la porción de training y se filtró con aquellas conversaciones en donde haya exactamente 2 participantes ya que el grooming no es común verlo en conversaciones grupales y también que dichas conversaciones tuvieran más de 7 mensajes, luego se filtra nuevamente



con el archivo que contiene los depredadores obteniendo conversaciones en donde hayan participado y aquellas en las que no.

Dada la naturaleza de PAN12 descrita anteriormente, un depredador puede estar en múltiples chats dando los siguientes casos:

- El depredador se queda hablando solo.
- El depredador habla de temas sexuales, pero con otro adulto.
- El depredador está cometiendo grooming.

A raíz de lo anterior, se procede hacer un etiquetado manual de las conversaciones predatorias y no predatorias obteniendo 100 conversaciones con grooming y 100 conversaciones sin contenido sexual ni grooming.

De las conversaciones con grooming obtenidas podemos hacer una caracterización de ellas:

- En su mayoría las víctimas y depredadores ya se conocían entre ellas y por lo tanto ya había un vínculo.
- Era predominante el lenguaje amoroso y cercano.
- Se hablaban de posibles encuentros físicos, mas no en su mayoría.
- No hay presencia de ofrecimientos de ítems o compra de artículos a cambio de favores sexuales.
- Es común preguntas acerca del horario escolar y horarios donde se ausentan los padres.

Con respecto a la caracterización de las conversaciones no grooming, podemos destacar lo siguiente:

- En su mayoría son conversaciones breves o donde se despiden o termina abruptamente.
- Hay presencia de spam.
- Hay presencia de lenguaje de odio.
- Hay presencia de peleas y discusiones.
- En su mayoría son conversaciones casuales con preguntas sobre gustos y redes sociales.

5.2 Limitaciones de pruebas de consistencia del clasificador

Una limitación importante que considerar en este estudio es la naturaleza no determinista de los Modelos de Lenguaje Grande (LLM). Debido a su arquitectura y proceso de generación de texto, los LLM pueden producir resultados ligeramente diferentes para la misma entrada en diferentes ejecuciones. Esta variabilidad inherente hace necesario realizar múltiples pruebas para obtener una comprensión más precisa del rango de variación en las respuestas del modelo.

Idealmente, se realizarían numerosas pruebas para cada conversación en el conjunto de datos, lo que permitiría un análisis estadístico robusto de la consistencia y precisión del modelo. Sin embargo, esta aproximación se enfrenta a una limitación práctica significativa: el costo asociado con el uso de la API del modelo.

El uso de modelos como GPT-4 a través de APIs conlleva un costo por token procesado. Realizar un gran número de pruebas para cada conversación resultaría en gastos considerables, especialmente dado el tamaño del conjunto de datos y la longitud potencial de las conversaciones analizadas.

Teniendo en cuenta estas consideraciones, se decidió establecer un equilibrio entre la necesidad de evaluar la variabilidad del modelo y la gestión eficiente de los recursos. Por lo tanto, se ajustó el número de pruebas a 4 por conversación. Este número proporciona una muestra suficiente para observar



posibles variaciones en las respuestas del modelo, al tiempo que mantiene los costos dentro de límites manejables.

Es importante reconocer que este compromiso entre exhaustividad y practicidad puede limitar en cierta medida la generalización de los resultados. Sin embargo, incluso con estas limitaciones, las 4 pruebas por conversación deberían proporcionar información valiosa sobre el rendimiento y la consistencia del modelo en la tarea de detección de grooming.

5.3 Diseño de experimentos

El objetivo principal de los experimentos es evaluar la eficacia de la heurística desarrollada para la detección de grooming en conversaciones de chat, tanto con la integración de Presidio como sin ella. Estos experimentos tienen como finalidad determinar si la integración de Presidio, como un módulo de detección de fugas de privacidad, mejora la precisión de la detección de grooming, o si, por el contrario, introduce sesgos que pueden afectar la tasa de falsos positivos o negativos.

5.3.1 Objetivos de los experimentos

Los experimentos diseñados tienen varios objetivos específicos que buscan evaluar diferentes aspectos del sistema de detección de grooming. Estos objetivos son:

- **Evaluar la precisión del sistema:** El primer objetivo es determinar si la heurística desarrollada es capaz de identificar correctamente las etapas del grooming en las conversaciones analizadas.
- **Comparar el impacto de Presidio:** Además de medir la precisión general, se pretende analizar cómo la integración de Presidio, una herramienta de detección de entidades sensibles influye en los resultados de la detección. Esto implica evaluar si Presidio contribuye a mejorar la identificación de patrones de grooming o si, por el contrario, introduce falsos positivos debido a la detección de información sensible que no necesariamente está relacionada con grooming.
- **Medir la robustez del sistema:** Finalmente, se busca comparar la consistencia de los resultados obtenidos con y sin la integración de Presidio.

5.3.2 Metodología de evaluación

Para alcanzar los objetivos mencionados, se diseñó una metodología de evaluación específica que se basa en el uso de un conjunto de datos etiquetado y en la configuración de dos variantes de la heurística desarrollada.

En primer lugar, se elaboró una versión reducida y etiquetada del dataset PAN12, el cual contiene conversaciones previamente clasificadas como casos de grooming y no grooming. La selección de este dataset se realizó con el propósito de garantizar que la muestra utilizada en los experimentos esté balanceada, lo que permite una evaluación más precisa de la heurística. Este set de datos ofrece un marco para probar la eficacia del sistema en escenarios realistas.

Para comparar de manera efectiva la heurística con y sin la integración de Presidio, se diseñaron dos configuraciones experimentales distintas. La primera configuración consiste en utilizar la heurística sin Presidio, donde el análisis de las conversaciones se realiza exclusivamente a través del modelo de lenguaje grande (LLM). En esta configuración, el modelo se encarga de identificar las diferentes etapas del grooming (Primer contacto, Generar confianza, y Migrar de plataforma) y de calcular la probabilidad de que esté ocurriendo grooming en la conversación, sin considerar la detección de entidades sensibles.



La segunda configuración introduce la integración de Presidio como un módulo adicional. En esta variante, además del análisis realizado por el LLM, Presidio se utiliza para detectar entidades de privacidad, como nombres, direcciones, o números de teléfono. La influencia de estas detecciones es observada para determinar cómo afectan la probabilidad final de detección de grooming. Esta configuración busca evaluar si Presidio añade valor en la identificación de patrones de grooming o si, por el contrario, introduce falsos positivos que puedan comprometer la eficacia del sistema.

Finalmente, la evaluación de los resultados se realizará midiendo tanto la tasa de aciertos como la de errores en ambas configuraciones. La tasa de aciertos incluirá las detecciones correctas de grooming, mientras que la tasa de errores considerará tanto los falsos positivos como los falsos negativos. Además, se llevará a cabo un análisis detallado de los casos en los que la integración de Presidio haya tenido un impacto significativo en los resultados, lo que permitirá evaluar si su inclusión es beneficiosa o contraproducente para el sistema.

5.4 Resultados

En esta sección se presentan los resultados obtenidos de la evaluación del sistema de detección de grooming basado en grandes modelos de lenguaje (LLM). El objetivo principal de esta evaluación es determinar la eficacia del sistema para identificar conversaciones que contienen patrones de grooming, así como su capacidad para distinguir correctamente las conversaciones sin grooming. A continuación, se detallan los resultados obtenidos en ambos escenarios: con y sin la integración de Presidio, destacando los aspectos más relevantes de cada uno.

5.4.1 Detección de grooming con Presidio

A continuación, se muestran la distribución de las probabilidades utilizando la heurística con la integración de Presidio.

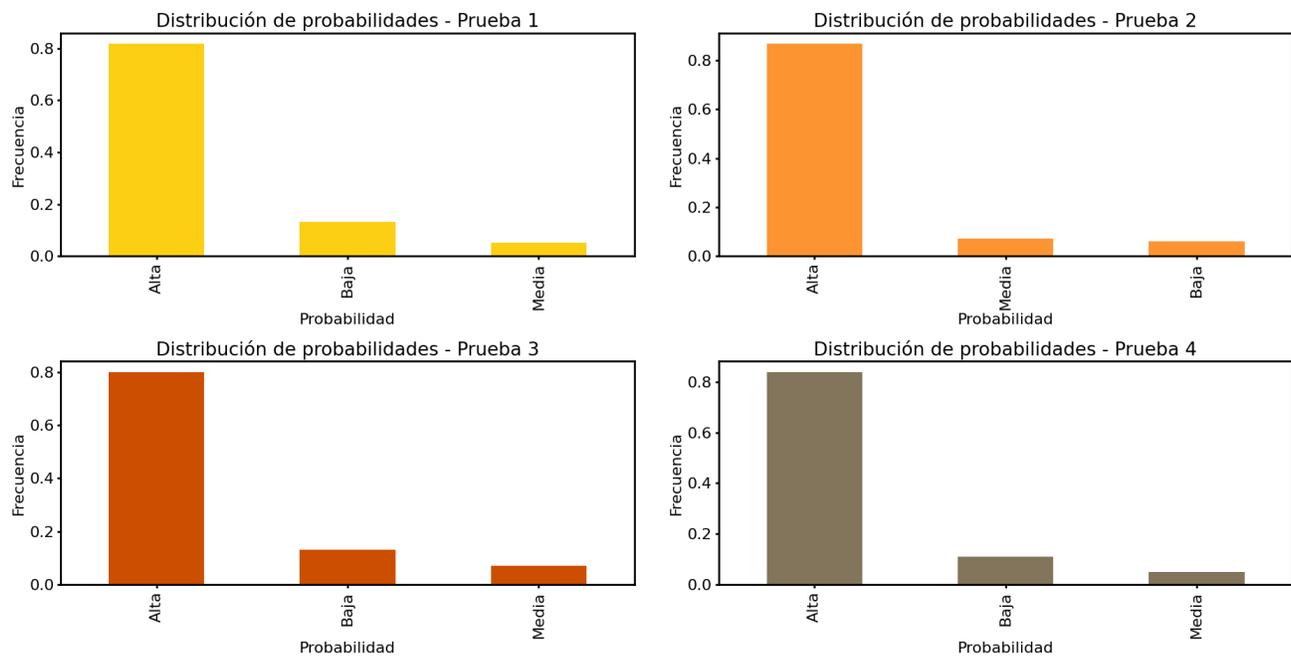


Ilustración 11: Gráficos de distribución de probabilidades para conversaciones grooming con Presidio

En este escenario, se observó que más del 83% de las conversaciones con grooming fueron clasificadas con una alta probabilidad de grooming, mientras que menos del 20% fueron detectadas con probabilidades bajas o medias. Esto sugiere que, aunque el sistema es capaz de identificar correctamente la mayoría de las conversaciones sospechosas, aún existe un margen de error que resulta en falsos negativos.

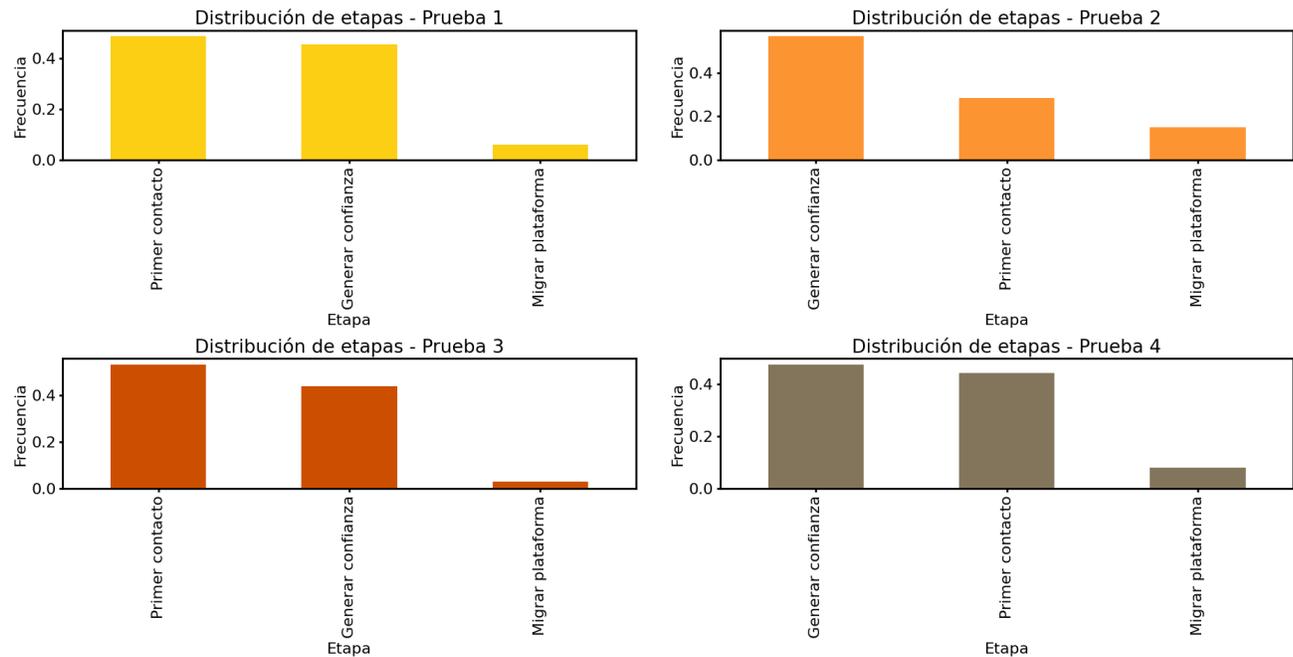


Ilustración 12: Gráficos de distribución de etapas para conversaciones grooming con Presidio

El análisis reveló que más del 90% de las conversaciones se clasificaron en las etapas de "Primer Contacto" o "Generar Confianza". Un hallazgo interesante es que algunas conversaciones avanzadas fueron clasificadas erróneamente en la etapa de "Primer Contacto". Esto podría deberse a la detección de saludos cariñosos que el modelo interpretó incorrectamente como indicativos de un primer acercamiento, sugiriendo una necesidad de refinamiento en el análisis de contexto.

5.4.2 Análisis de resultados de grooming con Presidio

Con respecto a la distribución de probabilidades, es clara la presencia de falsos negativos, es por ello por lo que se realizó un análisis de estos casos y se determinó lo siguiente:

- El LLM pese a detectar una alta probabilidad de grooming, no necesariamente será quien entregue la mayor cantidad de indicios.
- La cantidad de indicios tiene una estricta relación con la ponderación de Presidio, a mayor cantidad de indicios mayor probabilidad de obtener un mayor puntaje.



Lo anterior se puede observar en el siguiente ejemplo:

Tabla 4: Ejemplo de falso negativo dado por el sistema de puntaje

Etapa	Ponderación Presidio	Cantidad de indicios	Probabilidad LLM
Primer contacto	1	2	2
Generar confianza	1.2	3	3
Migración de plataforma	2	5	1

Pese a que el LLM de Generar confianza detectó una Alta probabilidad, la heurística en base al sistema de puntajes escogió la etapa de Migración de plataforma y por consiguiente una Baja probabilidad.

En relación con los falsos negativos de probabilidad Media no fueron afectados por la heurística, sino más bien la máxima probabilidad dada por los modelos de cada etapa fue Media.

Por otro lado, respecto a la distribución de las etapas es de esperar que la gran mayoría de ellas fueran detectadas en la etapa Generar confianza, dada la caracterización realizada anteriormente.

También una gran mayoría es detectada como Primer contacto, pese a que no es correcto, una posible razón es que la mayoría de las conversaciones comienzan con saludos muy cariñosos y adulaciones, comportamientos característicos de esta etapa, pero el modelo no es capaz de inferir que hubo una conversación anterior o que ya había un vínculo formado.

5.4.3 Detección de no grooming con Presidio

A continuación, se muestran la distribución de las probabilidades utilizando la heurística con Presidio en conversaciones sin grooming:

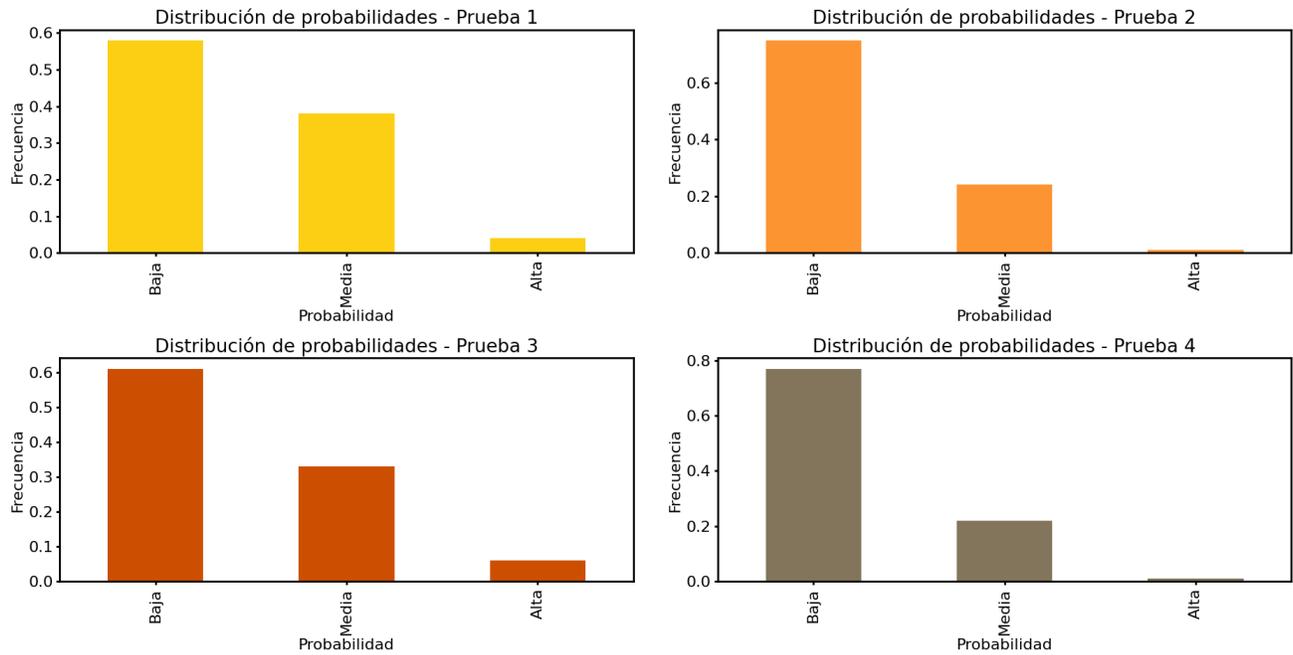


Ilustración 13: Gráficos de distribución de probabilidades para conversaciones sin grooming con Presidio

En el escenario de conversaciones sin grooming, la heurística original también mostró resultados interesantes. Aproximadamente el 99% de estas conversaciones fueron clasificadas con probabilidades bajas o medias, y menos del 1% se clasificaron con una alta probabilidad, lo que indica una baja incidencia de falsos positivos, lo cual es deseable para evitar alarmas innecesarias en entornos de monitoreo continuo.

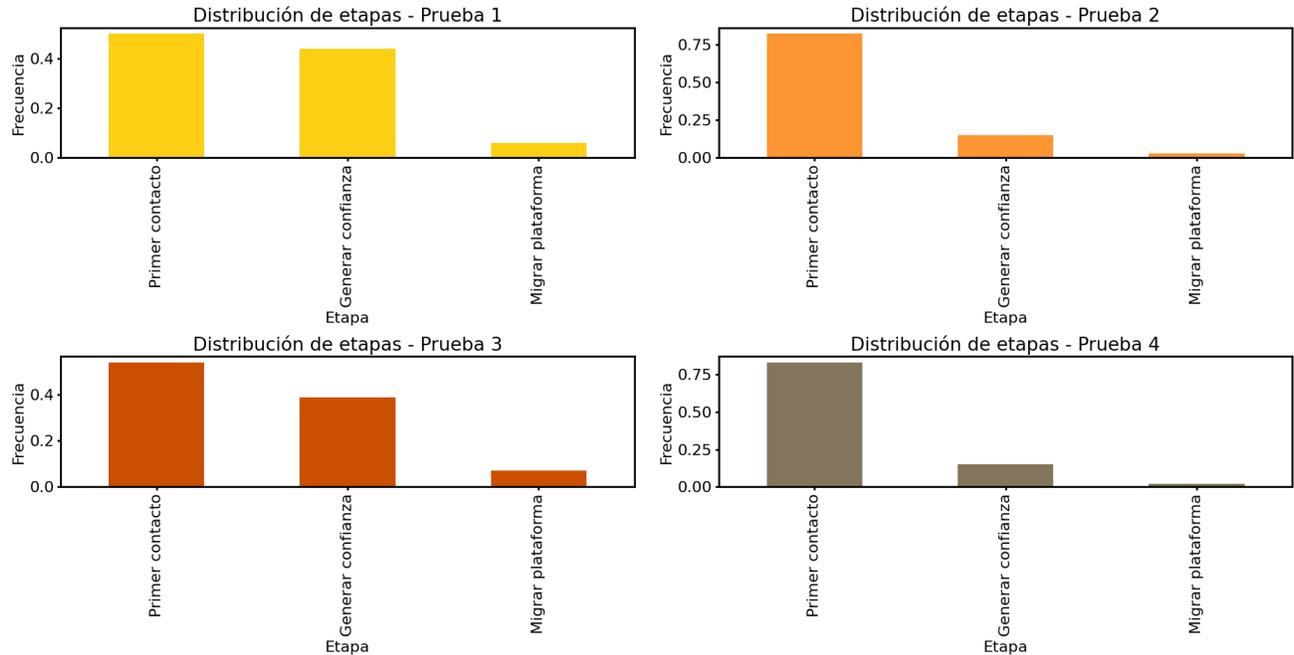


Ilustración 14: Gráficos de distribución de etapas para conversaciones sin grooming con Presidio

La mayoría de estas conversaciones se identificaron correctamente en la etapa de "Primer Contacto", lo cual es consistente con la naturaleza breve y formal de las conversaciones no predatorias. Sin embargo, es relevante destacar que un pequeño porcentaje fue erróneamente clasificado en etapas más avanzadas, lo que podría reflejar una sobreestimación de la intimidad o confianza en ciertos intercambios de información.

5.4.4 Análisis de no grooming con Presidio

En relación con la distribución de probabilidades, se observó la presencia de falsos positivos, aunque en una medida mucho menor que en el escenario con grooming. El análisis de estos casos particulares determinó que, en su mayoría, al menos dos LLM detectaron una alta probabilidad, lo que aumentó la probabilidad de que la heurística seleccionara alguna de estas etapas y, por lo tanto, asignara una alta probabilidad general. En casos minoritarios, solo un LLM detectó una alta probabilidad, pero fue el que proporcionó más indicios, obteniendo así un mayor puntaje con Presidio. En la mayoría de los casos, la alta probabilidad fue asignada por el modelo en la etapa de "Generar confianza", posiblemente debido a que detectaba la construcción de un vínculo y el intercambio de información más personal y/o sensible.



5.4.5 Detección de grooming sin Presidio

A continuación, se muestran la distribución de las probabilidades utilizando la heurística con los pesos o puntajes ajustados en conversaciones con grooming:

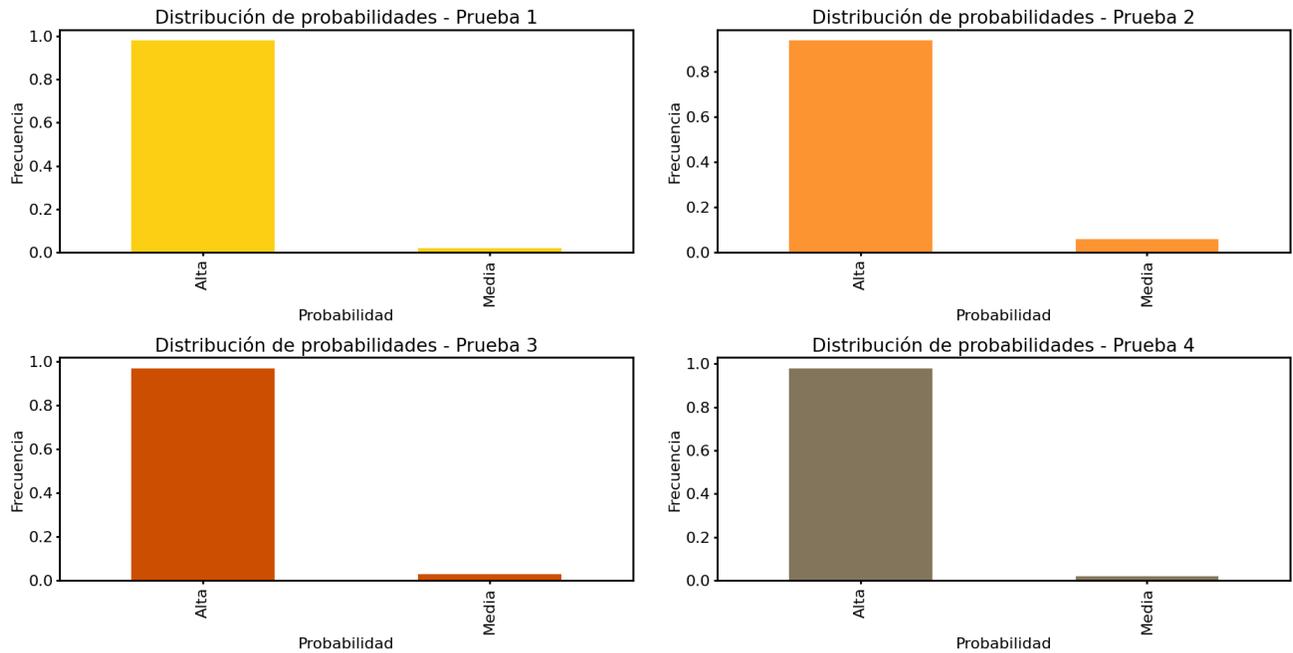


Ilustración 15: Gráficos de distribución de probabilidades para conversaciones grooming sin Presidio

Al remover Presidio del sistema, se observó una mejora en la precisión global, con más del 96% de las conversaciones con grooming clasificadas con alta probabilidad. Esto demuestra que el LLM, por sí solo, es altamente eficaz en la detección de grooming.

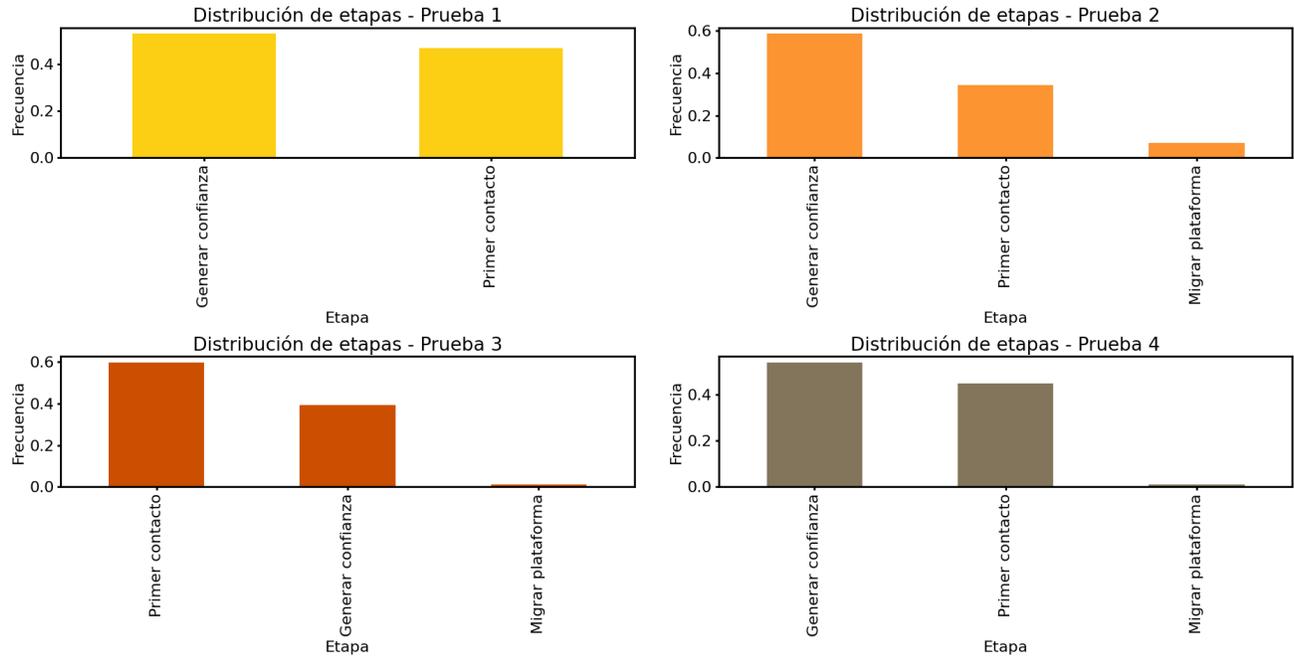


Ilustración 16: Gráficos de distribución de etapas para conversaciones grooming sin Presidio

Similar al análisis con Presidio, las etapas predominantes fueron "Generar Confianza" y "Primer Contacto". Este patrón reafirma que estas son las fases críticas donde el grooming es más detectado, y donde el modelo falla al no poder diferenciar los saludos afectuosos que demuestran un vínculo afectivo con las adulaciones características del Primer Contacto para captar la atención del niño. La ausencia de Presidio, sin embargo, parece haber incrementado la claridad en la clasificación de las etapas.

5.4.6 Análisis de grooming sin Presidio

A partir del enfoque sin la incidencia de Presidio en la elección de la etapa y probabilidad, se ha reducido drásticamente la cantidad de falsos negativos.

En cuanto a las conversaciones clasificadas con una probabilidad media, se determinó que estas eran generalmente breves, con 30 mensajes o menos, y presentaban patrones de grooming sutiles, donde en su mayoría no se utilizaba lenguaje sugerente. Además, se observó que el modelo no siempre clasificaba consistentemente una misma conversación con probabilidad media, lo que sugiere variabilidad en la detección de señales menos evidentes.

A partir de estos hallazgos, podemos aseverar que en ciertas circunstancias dada la naturaleza no determinista del LLM este no es capaz de detectar patrones sutiles o de plano no detectar grooming, aunque son menos del 1% de los casos.

5.4.7 Detección de no grooming sin Presidio

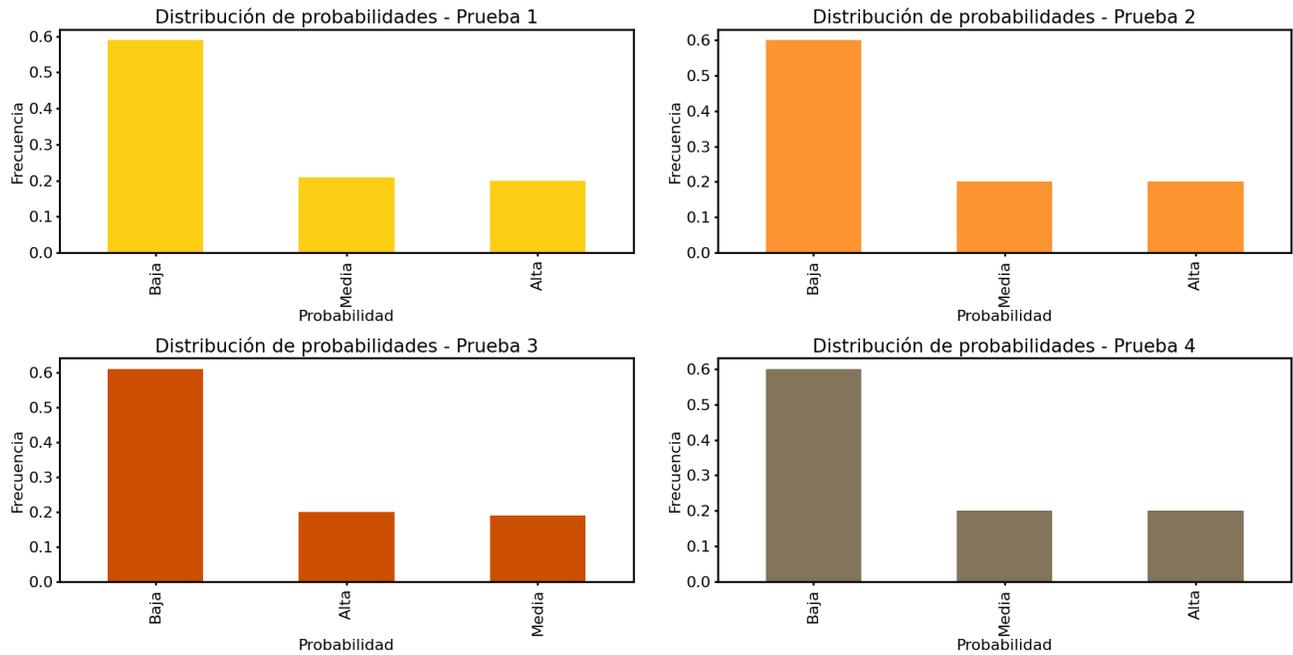


Ilustración 17: Gráficos de distribución de probabilidades para conversaciones sin grooming ni Presidio

Con respecto a la distribución de probabilidades, podemos observar que cerca del 60% corresponde a una probabilidad Baja, entre el 20-30% a la probabilidad Media y el 20% o menos restante a una probabilidad Alta.

Al remover la incidencia de Presidio se observó un aumento en los falsos positivos detectados.

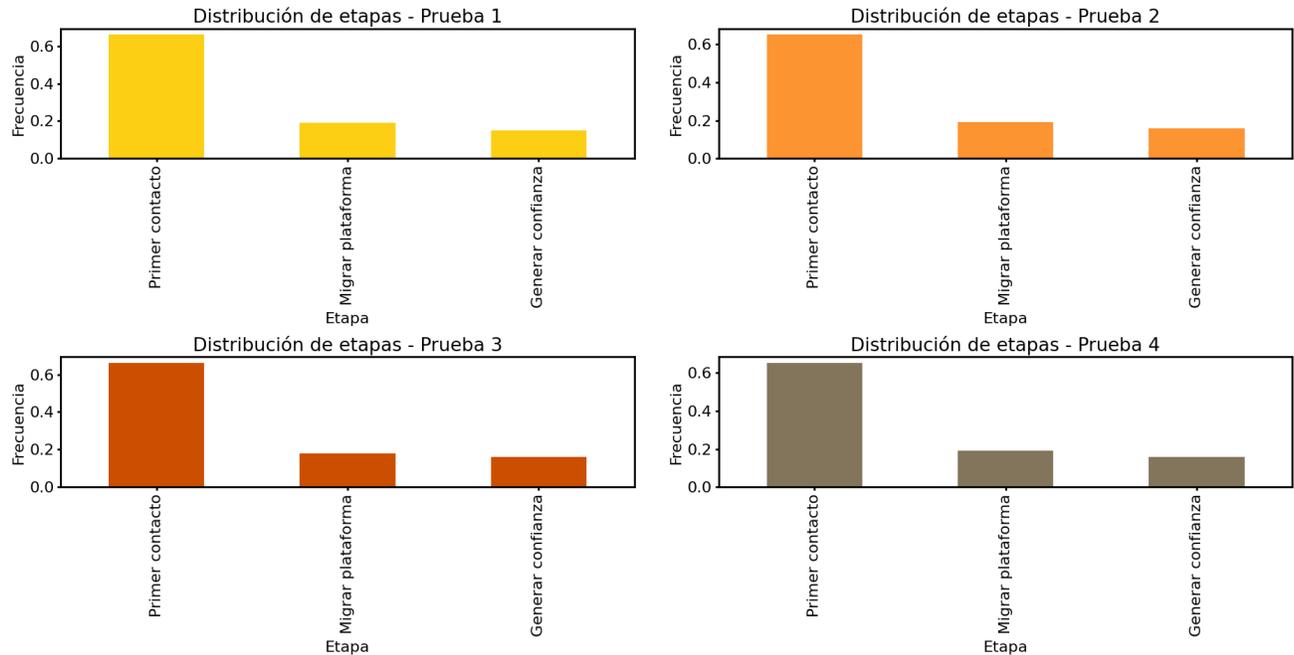


Ilustración 18: Gráficos de distribución de probabilidades para conversaciones sin grooming ni Presidio

Aquí podemos apreciar que se mantiene la tendencia a Primer contacto, seguido de Generar confianza y finalmente por Migrar de plataforma.

A diferencia de la versión con Presidio, hay un leve aumento en las conversaciones clasificadas como Migrar de plataforma.

5.4.8 Análisis de no grooming sin Presidio

En consecuencia, de los ajustes de puntaje, ha habido un aumento en la proporción de conversaciones clasificadas como probabilidad media y alta en comparación con la versión con Presidio. Esto sugiere que el ajuste, si bien ha mejorado la detección de grooming, también ha incrementado los falsos positivos en conversaciones sin grooming.

El aumento en dichas probabilidades se debe a que el LLM asume que al menos uno de los participantes es un menor de edad, por lo que cualquier intercambio de información o algún intento de generar un vínculo amoroso es percibido como un ataque grooming.

Con respecto a la distribución hay una clara mayoría de Primer contacto y se alinea con la caracterización de las conversaciones sin grooming, seguido de Migrar de plataforma y por último Generar confianza.



6 Discusión de resultados

Este estudio exploró el uso de grandes modelos del lenguaje (LLM) para la detección de grooming en conversaciones en línea, con un enfoque particular en el escenario de niños jugando videojuegos en línea. A través de un proceso iterativo, se desarrolló una heurística que combina el análisis de un LLM con un módulo de detección de fugas de privacidad (Presidio).

Los resultados obtenidos muestran que el sistema es capaz de detectar con alta precisión los casos de grooming, especialmente después de ajustar el sistema de puntaje para dar mayor peso a la clasificación del LLM y quitar la incidencia de Presidio.

En el caso de tener en cuenta el módulo de fugas de privacidad, obtenemos los siguientes resultados:

1. Para conversaciones con grooming:
 - a. Más del 83% fueron clasificadas con alta probabilidad de grooming.
 - b. Menos del 10% fueron clasificadas con probabilidad media.
 - c. Cerca del 10% fueron clasificadas con baja probabilidad, dándonos falso negativos.
 - d. La mayoría fue clasificada en la etapa Generar confianza alineándose con la caracterización de estas conversaciones.
2. Para conversaciones sin grooming:
 - a. Más del 99% fueron clasificadas con Baja o Media probabilidad.
 - b. Cerca del 1% fueron clasificadas como Alta, dándonos falsos positivos.
 - c. La mayoría fue clasificada en la etapa Primer Contacto, alineándose con la caracterización de estas conversaciones.

Para el caso de no tener el módulo de fugas de privacidad y simplemente usar la probabilidad entregada por el LLM, tenemos lo siguiente:

1. Para conversaciones con grooming:
 - a. Más del 96% fueron clasificadas con alta probabilidad de grooming.
 - b. Menos del 1% fueron clasificadas con probabilidad media.
 - c. La mayoría fueron identificadas en las etapas de "Generar confianza" o "Primer contacto".
2. Para conversaciones sin grooming:
 - a. Aproximadamente el 60% fueron clasificadas con baja probabilidad.
 - b. Entre el 20-30% fueron clasificadas con probabilidad media.
 - c. El 20% o menos fueron clasificadas con alta probabilidad, dándonos falsos positivos.

Estos resultados sugieren que el sistema es altamente sensible en la detección de grooming, lo cual es crucial para la protección de menores en línea. Sin embargo, también muestra una tendencia a producir falsos positivos en conversaciones normales, especialmente cuando asume que uno de los participantes es menor de edad.

En general, ambas configuraciones mostraron una alta capacidad para detectar grooming. La versión sin Presidio es más directa y presenta menos falsos negativos, mientras que la versión con Presidio es más completa en la detección de datos sensibles, aunque a costa de una mayor complejidad en la interpretación de resultados.



7 Conclusión

Este trabajo se centró en la creación de un sistema de detección de grooming en conversaciones de chat, utilizando grandes modelos de lenguaje (LLM) para identificar y clasificar posibles casos de depredadores sexuales en línea. La creciente utilización de plataformas de mensajería ha hecho que este tipo de sistemas sea esencial para la protección de menores, permitiendo identificar comportamientos sospechosos en etapas tempranas del grooming y obtener información verificable que es de valiosa utilidad para la Policía de Investigaciones.

Durante el desarrollo del sistema, surgieron varias dificultades significativas. La integración del módulo de detección de fugas de privacidad (Presidio) resultó compleja y trajo consigo la posibilidad de generar falsos positivos, especialmente cuando el sistema de puntajes seleccionaba una etapa y por consiguiente probabilidad de manera errónea. Esta tendencia a los falsos positivos planteó un desafío importante en la precisión del sistema. Con el enfoque sin la integración de Presidio, el sistema demostró una mayor eficacia en la detección de grooming, reduciendo significativamente los falsos negativos, pero con un leve aumento en los falsos positivos.

Este trabajo contribuye significativamente a la investigación en el campo de la detección de grooming, proporcionando un marco metodológico robusto que puede ser utilizado y mejorado en futuras investigaciones.

7.1 Trabajo a futuro

El futuro del sistema de detección de grooming desarrollado puede enfocarse en varias áreas clave:

- **Explorar nuevos modelos de lenguaje:** Investigar el uso de otros modelos como Claude 3.5 Sonnet de Anthropic o Gemini 1.5 Pro de Google, además de modelos locales. Esto permitiría optimizar el balance entre calidad de respuesta y costo computacional, obteniendo mejores resultados con menos recursos.
- **Utilización de un dataset que refleje la realidad chilena:** Incorporar un dataset que represente de forma más precisa la situación del grooming en Chile, con datos locales y características que reflejen los patrones y comportamientos específicos de los groomers y las víctimas en este país. Esto permitiría ajustar mejor las heurísticas y mejorar la precisión del sistema en contextos locales.
- **Fine-tuning de modelos:** Realizar un fine-tuning de los modelos con datasets específicos de grooming. Esto mejoraría la precisión de las respuestas al entrenar los modelos con ejemplos más alineados al problema del grooming, reduciendo falsos positivos y negativos.
- **Implementación en tiempo real:** Investigar cómo adaptar el sistema para su uso en tiempo real. Esto permitiría intervenciones más rápidas y preventivas, facilitando la protección de los menores antes de que el grooming avance a fases más peligrosas.
- **Consideraciones éticas:** Ampliar el análisis de las implicaciones éticas del sistema, garantizando que su uso respete la privacidad de los usuarios y minimice los falsos positivos, ya que estos podrían tener repercusiones negativas tanto en las víctimas como en los usuarios investigados.



8 Anexos

En esta sección se incluyen documentos complementarios y detalles técnicos que respaldan el desarrollo y la evaluación del sistema de detección de grooming presentado en este trabajo.

8.1 Respuestas entregadas por GPT-4o

ID de conversación	Respuesta entregada
399bccd8d0df9f08c75325df9c90f8ec	<p>Sí, hay indicios de grooming en esta conversación. Los mensajes específicos que lo respaldan son:</p> <ul style="list-style-type: none"> • Mensaje 19: "so are we on for tonite sweety?" - Usa lenguaje cariñoso inapropiado. • Mensaje 26: "ur parents still there?" - Pregunta si los padres estarán ausentes. • Mensaje 29: "they gone til tom night" - Confirma que los padres estarán ausentes. • Mensaje 59: "hey whats ur address so i can use mapquest" - Insiste en obtener la dirección del adolescente.
853dfe3ffd5f1728f469a37f95a607a2	<p>Sí, hay indicios de grooming en esta conversación. Los mensajes específicos que lo respaldan son:</p> <ul style="list-style-type: none"> • Mensaje 8: "a tall bald guy who is well hung?" - Hace un comentario sexual inapropiado. • Mensaje 12: "dont you want to?" - Refuerza la insinuación sexual anterior, tratando de obtener confirmación. • Mensaje 17: "well ok tell me where you live" - Pide la dirección, lo que es típico en grooming para facilitar un encuentro físico.

8.2 Respuestas entregadas por GPT-4o-mini

ID de conversación	Respuesta entregada
399bccd8d0df9f08c75325df9c90f8ec	<p>No hay evidencia de grooming en esta conversación. Los mensajes son principalmente sobre planes para encontrarse y coordinar horarios. No hay señales de manipulación, coacción o comportamientos típicos de grooming.</p>
853dfe3ffd5f1728f469a37f95a607a2	<p>En esta conversación, hay algunos mensajes que podrían levantar sospechas de grooming, aunque el contexto completo es necesario para una evaluación precisa. Aquí están los mensajes específicos que podrían ser indicativos de comportamiento problemático:</p> <ol style="list-style-type: none"> 1. Comentarios sugestivos y coquetos: <ul style="list-style-type: none"> ○ Líneas 8-9: El mensaje "a tall bald guy who is well hung?" seguido de "lol" es una broma sugestiva que puede ser inapropiada dependiendo del contexto y de la relación entre las personas involucradas. 2. Solicitar información personal:



	<ul style="list-style-type: none"> ○ Línea 17: "well ok tell me where you live" es un mensaje que solicita información personal. Pedir detalles sobre la ubicación puede ser un signo de grooming, especialmente si la relación no es cercana o si el contexto es inapropiado. <p>3. Respuestas a la solicitud de información:</p> <ul style="list-style-type: none"> ○ Líneas 19-22: Las respuestas del usuario "oh sorry my mom was talking to me", "lol", "i have to go help with dinner really quick", y "bbiaf" podrían ser una forma de evitar proporcionar la información solicitada, lo cual es una buena práctica para proteger la privacidad.
--	--

8.3 Respuestas entregas por Llama3-8B

ID de conversación	Respuesta entregada
399bccd8d0df9f08c75325df9c90f8ec	<p>Sí, hay indicios de grooming en esta conversación. A continuación, se presentan algunos mensajes específicos que respaldan esto:</p> <ul style="list-style-type: none"> ▪ El mensaje 19 de 90dfa20487af303d92fe68ec15eb1e3e: "so are we on for tonite sweety?" puede ser visto como un intento de crear una relación íntima y afectiva con c2881109eb31473638f353dc2ce0a66c. ▪ El mensaje 41 de c2881109eb31473638f353dc2ce0a66c: "ubless u wanna come late" puede ser interpretado como un intento de invitar a 90dfa20487af303d92fe68ec15eb1e3e a quedarse en su casa sin supervisión. ▪ El mensaje 59 de 90dfa20487af303d92fe68ec15eb1e3e: "hey whats ur address so i can use mapquest" puede ser visto como un intento de obtener información personal y confidencial de c2881109eb31473638f353dc2ce0a66c. ▪ El mensaje 65 de 90dfa20487af303d92fe68ec15eb1e3e: "so what ur address?" puede ser interpretado como un intento de obtener información personal y confidencial de c2881109eb31473638f353dc2ce0a66c.
853dfe3ffd5f1728f469a37f95a607a2	<p>Sí, hay indicios de grooming en esta conversación. A continuación, se presentan algunos mensajes específicos que respaldan esto:</p> <ul style="list-style-type: none"> ▪ El mensaje 8 de 0b6b05c740a1bf50ca7f9a461598a3b9: "a tall bald guy who is well hung?" puede ser visto como un intento de hacer un comentario inapropiado y sugestivo sobre 149c26e105b5db5b968f634d15c28246. ▪ El mensaje 17 de 0b6b05c740a1bf50ca7f9a461598a3b9: "well ok tell me where you live" puede ser interpretado como un intento de obtener información personal y confidencial de 149c26e105b5db5b968f634d15c28246.



Es importante destacar que estos mensajes pueden ser interpretados de diferentes maneras y que no necesariamente significan que 0b6b05c740a1bf50ca7f9a461598a3b9 esté cometiendo grooming. Sin embargo, pueden ser considerados como indicativos de una relación inapropiada o de un intento de manipulación.



9 Referencias

- [1] M. Gaikwad, S. Ahirrao, S. Phansalkar, K. Kotecha, Online extremism detection: A systematic literature review with emphasis on datasets, classification techniques, validation methods, and tools, in: IEEE Access, Vol. 9, IEEE, 2021, pp. 48364–48404.
- [2] J. Bartlett, L. Reynolds, The State of the Art 2015: A Literature Review of Social Media Intelligence Capabilities for Counter-Terrorism, Demos London, 2015.
- [3] C.M. Steel, E. Newman, S. O'Rourke, E. Quayle, An integrative review of historical technology and countermeasure usage trends in online child sexual exploitation material offenders, in: Forensic Science International: Digital Investigation, Vol. 33, Elsevier, 2020, 300971.
- [4] C. Peersman, C. Schulze, A. Rashid, M. Brennan, C. Fischer, iCOP: Live forensics to reveal previously unknown criminal media on P2P networks, in: Digital Investigation, Vol. 18, Elsevier, 2016, pp. 50–64.
- [5] H.C. Whittle, C. Hamilton-Giachritsis, A.R. Beech, Victims' voices: The impact of online grooming and sexual abuse, in: Universal Journal of Psychology, Vol. 1, Citeseer, 2013, pp. 59–71.
- [6] E.A. Greene-Colozzi, G.M. Winters, B. Blasko, E.L. Jeglic, Experiences and perceptions of online sexual solicitation and grooming of minors: a retrospective report, in: Journal of Child Sexual Abuse, Vol. 29, Taylor & Francis, 2020, pp. 836–854.
- [7] E. Quayle, Prevention, disruption and deterrence of online child sexual exploitation and abuse, in: Era Forum, Vol. 21, Springer, 2020 pp. 429–447.
- [8] Grooming: investigamos más de 4 mil casos en 2019. (2020, 26 febrero). <https://www.pdichile.cl/centro-de-prensa/detalle-prensa/2020/02/26/grooming-investigamos-m%C3%A1s-de-4-mil-casos-en-2019>
- [9] A. Malm, R. Nash, R. Moghadam, Social network analysis and terrorism, in: The Handbook of the Criminology of Terrorism, Wiley Online Library, 2017, pp. 221–231.
- [10] P. Chitrakar, C. Zhang, G. Warner, X. Liao, Social media image retrieval using distilled convolutional neural network for suspicious e-crime and terrorist account detection, in: 2016 IEEE International Symposium on Multimedia, ISM, IEEE, 2016, pp. 493–498.
- [11] P. Wisniewski, The privacy paradox of adolescent online safety: A matter of risk prevention or risk resilience? in: IEEE Security & Privacy, Vol. 16, IEEE, 2018, pp. 86–90.
- [12] T. Ringenberg, K. Misra, K.C. Seigfried-Spellar, J.T. Rayz, Exploring automatic identification of fantasy-driven and contact-driven sexual solicitors, in: 2019 Third IEEE International Conference on Robotic Computing, IRC, IEEE, 2019, pp. 532–537.
- [13] Fauzi, Muhammad & Bours, Patrick. (2020). Ensemble Method for Sexual Predators Identification in Online Chats. 1-6. 10.1109/IWBF49977.2020.9107945.
- [14] Nguyen, T. T., Wilson, C., & Dalins, J. (2023). Fine-Tuning Llama 2 Large Language Models for Detecting Online Sexual Predatory Chats and Abusive Texts
- [15] Matthias Vogt, Ulf Leser, and Alan Akbik. 2021. [Early Detection of Sexual Predators in Chats](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and*



the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 4985–4999, Online. Association for Computational Linguistics.

- [16] Garayzábal, Elena & Guía, Irene. (2021). DETECTION OF SEXUAL PREDATORS IN CHAT ROOMS AND GROOMING ONLINE. THE ROLE OF FORENSIC LINGUISTICS. *Tonos Digital*. 46.