Universidad de Concepción
Departamento de Ingeniería Informática y Ciencias de la Computación

Honeypot para casos de Grooming y acoso a menores

Matías Fuentes Barrera

Memoria presentada para la obtención del título de Ingeniero Civil Informático

Profesores Patrocinantes:

Pedro Pablo Pinacho Davidson
Fernando Andree Tercero Gutiérrez Gómez
Mario Vega Barbas

Resumen

La creciente problemática del *grooming* y la dificultad para prevenirlo debido al anonimato en internet y la falta de educación digital en menores, ha hecho necesario crear herramientas tecnológicas innovadoras que permitan abordar este fenómeno de manera proactiva. Esta memoria de título se enfoca en el desarrollo de un sistema *Honeypot* para casos de *grooming*, simulando la personalidad y el comportamiento de un "niño vulnerable" con el proposito de detectar e identificar posibles ataques de *grooming* en línea.

Para el desarrollo del sistema se utilizaron Grandes Modelos de Lenguaje (LLM), que procesan el lenguaje natural y generan respuestas creíbles, adaptadas al contexto de un "niño vulnerable" de 10 años. Además, se estableció un sistema de memoria permitiendo almacenar la información contextual, que permite mantener la continuidad del contexto y el flujo de la conversación.

El sistema fue validado a través de cuatro casos experimentales, donde se evaluó la capacidad de este para responder ante situaciones de riesgo típicas del *grooming*, como solicitudes de información personal, propuestas inapropiadas o intentos de manipulación. La evaluación incluyó también una escala Likert aplicada a usuarios seleccionados, quienes midieron el realismo, la utilidad y la efectividad del sistema. Los resultados obtenidos mostraron que el prototipo es una herramienta prometedora para la detección y análisis de posibles *groomers*, contribuyendo así a la prevención de ciberacoso.

Este trabajo fue parcialmente financiado por Proyecto ANID Fondecyt N°11230359, "AN IMMUNE INSPIRED MODEL OF INTRUSION PREVENTION SYSTEM (IPS) FOR COLLABORATIVE AND DISTRIBUTED ENVIROMENTS

Índice

1. Introducción

En la actualidad, los menores de edad tienen un acceso directo e inmediato a una amplia variedad de contenidos en internet, una tendencia que ha crecido de manera constante en los últimos años [1]. A pesar de la existencia de herramientas para la protección del menor como controles parentales y métodos de monitoreo de actividades en línea, resulta cada vez más difícil proteger a los niños y adolescentes de posibles acosadores en la red, comúnmente conocidos como *groomers*.

Esta dificultad para controlar y supervisar el entorno digital expone a los menores a un riesgo significativo [2], convirtiéndolos en blancos vulnerables de depredadores que emplean diversas estrategias para ganarse su confianza.

Las artimañas utilizadas para atraer la atención y la amistad de los menores de edad incluyen promesas engañosas, estafas y ofrecimiento de beneficios dentro de juegos en línea, que, por su corta edad y limitado acceso a bienes económicos, los menores son muy susceptibles a caer en dichas tretas, lo cual los expone a un peligro mayor y que puede atentar contra su modo de vivir el día a día.

Un ejemplo destacado de un entorno virtual donde los menores son especialmente susceptibles a los *groomers* es el popular juego en línea "Roblox" [3]. Este juego cuenta con una gran cantidad de usuarios jóvenes, este número está rondando el 58% [21] de la población. Este juego ofrece la posibilidad de comprar objetos virtuales para personalizar los avatares de los jugadores. Los ciberdelincuentes pueden aprovecharse de esta función al ofrecer comprar estos objetos a cambio de información privada. Además, se ha descubierto que algunos *groomers* crean sus propios juegos dentro de “Roblox”, donde promueven la interacción sexual entre avatares y buscan atraer a menores, aumentando significativamente el riesgo [4].

Existen mecanismos de defensa contra ciberataques, que se han utilizado en distintos ámbitos de seguridad, entre estos está el *Honeypot*, el cual es un mecanismo de ciberseguridad que simula sistemas con vulnerabilidades aparentes para atraer a ciberdelincuentes [8]. Al interactuar con estos señuelos, se genera una alerta que permite identificar y contrarrestar posibles amenazas de manera efectiva.

Los LLMs (Large language models o grandes modelos de lenguajes) son modelos de aprendizaje profundo muy grandes que se pre-entrenan con grandes cantidades de datos [5], se han implementado y generado distintos tipos de defensa contra casos de estafa y engaño en la red utilizando estos modelos. Los LLM han impulsado la implementación de herramientas de prevención para ciberataques [7], entre estas destaca la adaptación del *Honeypot* mediante el uso de LLMs [6].

En este proyecto, se desarrolló un sistema vulnerable o *Honeypot* adaptado al contexto del *grooming*, con una diferencia fundamental respecto a los enfoques tradicionales: el uso de LLMs para simular el comportamiento de un niño vulnerable. Esta simulación se centra en recrear de manera realista las interacciones de menores en riesgo frente a potenciales acosadores, no para abarcar todas las formas de acoso en internet, sino para analizar como el sistema puede reproducir comportamientos infantiles en contextos conversacionales, para evaluar su realismo y capacidad de respuesta, aprovechando el uso de modelos de lenguaje que ya han sido utilizados exitosamente en la simulación de comportamientos humanos en diversos entornos [9].

Todo el proceso de la caracterización de agresores y de las respuestas de los niños vulnerados se

realizó en colaboración con la Policía de investigaciones de Chile (PDI).

1.1 Objetivo General

Desarrollar un sistema *Honeypot* que simule el comportamiento de un niño vulnerable ante el accionar de un *groomer* en un chat basado en texto mediante el uso LLMs y evaluar su desempeño mediante el uso de casos experimentales.

1.2 Objetivos Específicos

1. Investigar el modus operandi de los atacantes para casos de *grooming*.

2. Investigar y caracterizar el comportamiento en un chat de un niño vulnerable para poder simularla en el sistema.

3. Diseñar y desarrollar un sistema basado en LLM que simule ser un niño vulnerable y la interacción como *Honeypot*.

4. Diseñar diferentes casos de uso para la validación del sistema, de esta manera evaluar realismo, coherencia y el uso de lenguaje debido.

5. Obtener retroalimentación mediante una encuesta realizada a profesionales del área de la informática y a funcionarios de la policía de investigaciones.

1.3 Metodología

Se adoptará un método de desarrollo iterativo e incremental, con adaptaciones de propuestas ágiles para integrar posibles mejoras a medida que avanza el proyecto. El desarrollo se basa en un análisis de tecnologías existentes y referencias bibliográficas que sean de utilidad para observar el comportamiento de los menores de edad. Este proceso se estructuró en distintas etapas, que incluyen:

- **Investigación**: Se llevó a cabo la investigación de tecnologías existentes y características sobre el contexto de los ataques de *grooming* y también sobre el niño en cuestión.
- **Diseño**: Se propuso una arquitectura simple de un *chatbot* que simule ser un niño vulnerable mediante los LLMs al interactuar con la API de OpenAI.
- **Desarrollo**: Con el uso de herramientas de programación y desarrollo de software se implementó el prototipo de *Honeypot.*
- **Evaluación:** Se llevo a cabo una evaluación al prototipo mediante casos experimentales de uso en los cuales se calificó la semejanza con un niño de verdad, utilizando dataset de PAN12 a modo de ejemplo para la comparación del habla de un niño vulnerable. Además de esto se realizaron encuestas de Escala Likert a cinco usuarios distintos a modo de conseguir una retroalimentación efectiva para un posible trabajo futuro.

1.4 Estructura del informe

El informe se divide en distintas secciones:

Capítulo 1

En el primer capítulo de la memoria de título, abre con una introducción al tema abordado en la memoria de título, se detalla la solución desarrollada, los objetivos generales y específicos del proyecto, y la metodología utilizada.

Capítulo 2

A continuación, se desarrolla el marco teórico de la memoria de título, desglosando conceptos claves utilizados en el informe, tales como *grooming*, *groomer*, *Honeypot*, Grandes Modelos del Lenguaje, la definición de Sesgo Cognitivo y específicamente otorgar detalles sobre el Efecto Halo.

Capítulo 3

Posteriormente en la sección Desarrollo de la metodología, se detallarán los requisitos que presenta el sistema (funcionales y no funcionales) y los detalles del prototipo. Luego se darán a conocer detalles sobre la arquitectura del prototipo, se detallarán sus elementos más importantes y finalmente se expone un ejemplo a modo de presentación del funcionamiento del prototipo desarrollado.

Capítulo 4

En la cuarta sección del presente informe, se exponen los resultados de la evaluación realizada por el equipo e investigadores interesados en el proyecto, los cuales respondieron una Encuesta Likert para así obtener una retroalimentación efectiva del prototipo, para un posible trabajo futuro del proyecto, también el prototipo fue validado mediante el uso de casos experimentales, otorgando puntuación según una métrica de evaluación comparativa utilizando el dataset de PAN12 a modo de ejemplo para las conversaciones, y basándose en la caracterización propuesta en investigaciones previas llevadas a cabo por investigadores externos, las cuales fueron recopiladas y analizadas en el marco del presente estudio.

Capítulo 5

Para finalizar con el informe, se presenta la conclusión obtenida mediante la retroalimentación recibida, resumiendo el trabajo expuesto y el trabajo a futuro en base a lo comentado

2. Marco Teórico

En esta sección se define el concepto de *grooming* y sus consecuencias en la actualidad basado en la investigación realizada, con la cual se recopiló información importante sobre datos de ataques realizados. También se abordará el termino de *Honeypot* y la importancia que posee en el ámbito de la depredación sexual a menores, y por último se abarcará la definición de Large Language Models (LLM).

2.1 Grooming en la actualidad

Es difícil decir específicamente a que se refiere el *grooming* ya que puede ser interpretado de distintas formas, ya sea desde el acoso sexual y a la intimidad, hasta la extorsión infantil para llegar a obtener información importante.

El *grooming* no solo afecta a la integridad y el bienestar de las víctimas, en muchos casos esto se vincula a delitos graves, como la extorsión y la producción de material de abuso infantil, donde los agresores manipulan a los menores mediante la coacción y engaño [10], esto origina múltiples problemas a las víctimas del ataque, entre estos se encuentran, vergüenza, angustia, culpa e incluso los puede llevar a la depresión y con esto a problemas mayores [32].

Actualmente el *grooming* se ha convertido en una problemática a nivel mundial debido a diversos factores como:

- **Anonimato en la red**: Es fácil para los *groomers* el crearse perfiles falsos, haciéndose pasar por menores de edad o simplemente por otras personas, de esta forma ocultar su identidad real y poder acercarse sin problemas a menores en línea [11].

- **Falta de educación digital**: El no contar previamente con un acceso adecuado y reiterado a la tecnología y el reciente aumento del acceso a está provocado por la pandemia mundial del COVID-19, la cual obligo a introducir la educación en línea masivamente [12], provocando que aumente el uso de red en los menores de edad y en consecuencia la posibilidad de encontrarse con desconocidos en línea, sin poseer la capacidad de discernir posibles conductas sospechosas. La pandemia solamente exacerbo aumentando en gran medida el uso de las redes sociales (Figura 1).

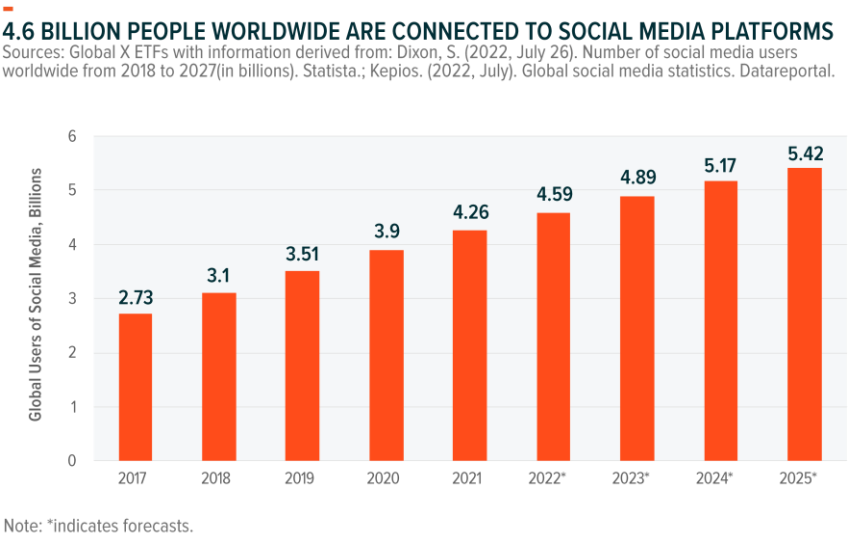

Figura 1. Gráfico de barras sobre incremento en el uso de redes sociales [13].

- **Auge de las plataformas interactivas**: El uso de redes sociales ha experimentado un

notable crecimiento en los últimos años [13], lo que ha derivado en una mayor exposición de los menores de edad a riesgos como el grooming, facilitando el contacto con agresores que utilizan estas plataformas para acercarse a posibles víctimas.

Como consecuencia de esto se registró un aumento lineal debido al Covid-19 (Figura 2) en las denuncias relacionadas al *grooming* [14], lo que evidencio aún más la vulnerabilidad de los menores frente a estos ciberdelitos y urgencia que esto presenta para la comunidad en general.

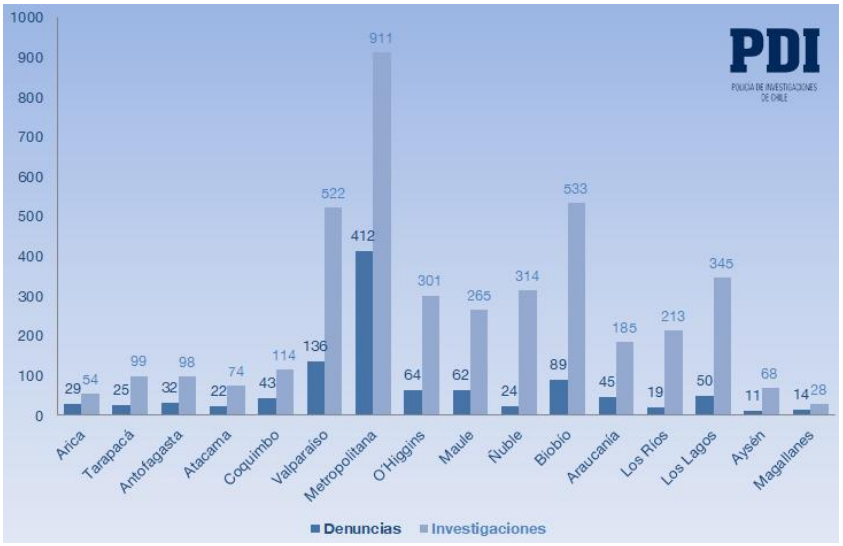

Figura 2. Gráfico de barras de casos de grooming en 2019 en Chile [14].

Por ello la definición de g*rooming* que se utilizara en este trabajo para contextualizar, es la expuesta en el artículo de "Save the Children" (01 de Julio de 2019) la cual es la siguiente [10]:

"El *grooming* y, en su evolución digital, el *online grooming* (acoso y abuso sexual online) son formas delictivas de acoso que implican a un adulto que se pone en contacto con un niño, niña o adolescente con el fin de ganarse poco a poco su confianza para luego involucrarle en una actividad sexual."

2.2 Honeypot

En el contexto de ciberseguridad, *Honeypot* se refiere a un sistema informático diseñado deliberadamente con vulnerabilidades reales y aparentes, para atraer ataques de ciberdelincuentes. Estos sistemas se utilizan como señuelos y permiten a los expertos en seguridad recopilar información sobre métodos de ataque, identificar amenazas y desarrollar estrategias de defensa [8].

Funcionamiento del Honeypot

Un *Honeypot* se configura de manera que simula ser un sistema atractivo para ser atacado, puede ser un servidor, una aplicación o incluso un usuario ficticio. Dentro de los distintos tipos de *Honeypot* (Figura 3) hay dos que destacan [15]:

- **Honeypots de baja interacción**: Simulan servicios básicos con información limitada y son fáciles de monitorear.

- **Honeypots de alta interacción**: Ofrecen entornos más realistas y complejos, proporcionando información detallada del comportamiento del atacante.

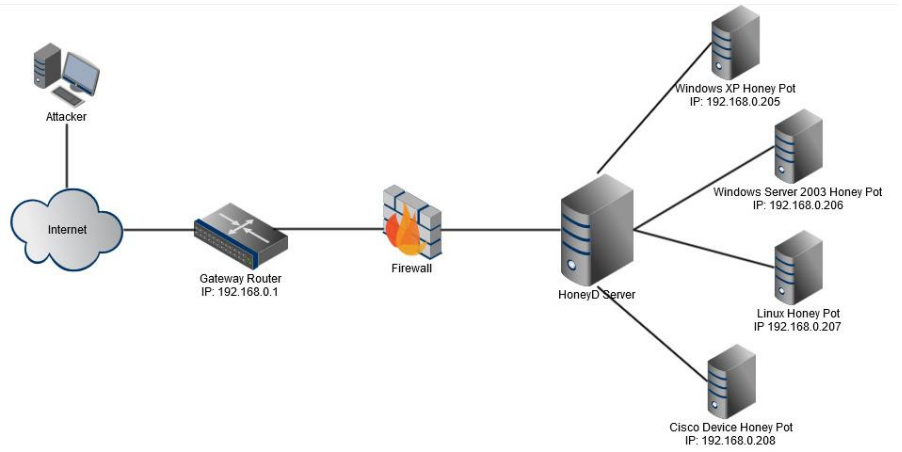

Figura 3. Ejemplo de Honeypot simulando un servidor [16].

Honeypot basado en LLM

Existen un estudio en el cual se desarro un *Honeypot* interactivo basado en LLM [17] en entornos de ciberseguridad. La implementación se realizó utilizando un modelo Llama3 8B optimizado mediante ajuste fino supervisado (SFT) con un conjunto de datos compuesto por comandos reales de atacantes, comandos comunes de Linux y explicaciones detalladas.

El sistema propuesto integra un servidor SSH con el LLM para recibir, procesar y responder a comandos de atacantes de manera realista. Para mejorar su eficiencia y realismo, se aplicaron técnicas avanzadas como LoRA, QLoRA y Flash Attention 2, además de estrategias de ingeniería de prompts que garantizan respuestas precisas y coherentes.

En el contexto del *grooming*, los *Honeypots* no se limitan a una infraestructura técnica, si no que evolucionan a la simulación de interacciones humanas, en este caso interacciones de un “niño vulnerable”, lo que permite atraer y detectar posibles *groomers.* La ventaja de hacer esto es poder crear distintos perfiles de “niños vulnerables” con distintos sesgos cognitivos e intereses, de esta manera poder generar un perfil del *groomer*, analizar sus tácticas y prevenir casos futuros de acoso a menores, además se han estudiado diversas maneras de poder incorporar personalidad en agentes *Honeypot* basados en LLM y han generado resultados satisfactorios simulando comportamiento

humano [33].

2.3 Grandes Modelos de Lenguaje

Los Grandes Modelos de Lenguaje o por sus siglas en ingles LLM (Large Language Models) son modelos de inteligencia artificial entrenados con grandes volúmenes de datos textuales para comprender y generar lenguaje natural de manera coherente y humana [5]. Ejemplos de estos modelos son PaLM y GPT-4o utilizan arquitectura basadas en *Transformadores* (Figura 4) lo que le permite manejar el contexto y las relaciones entre palabras a gran escala.

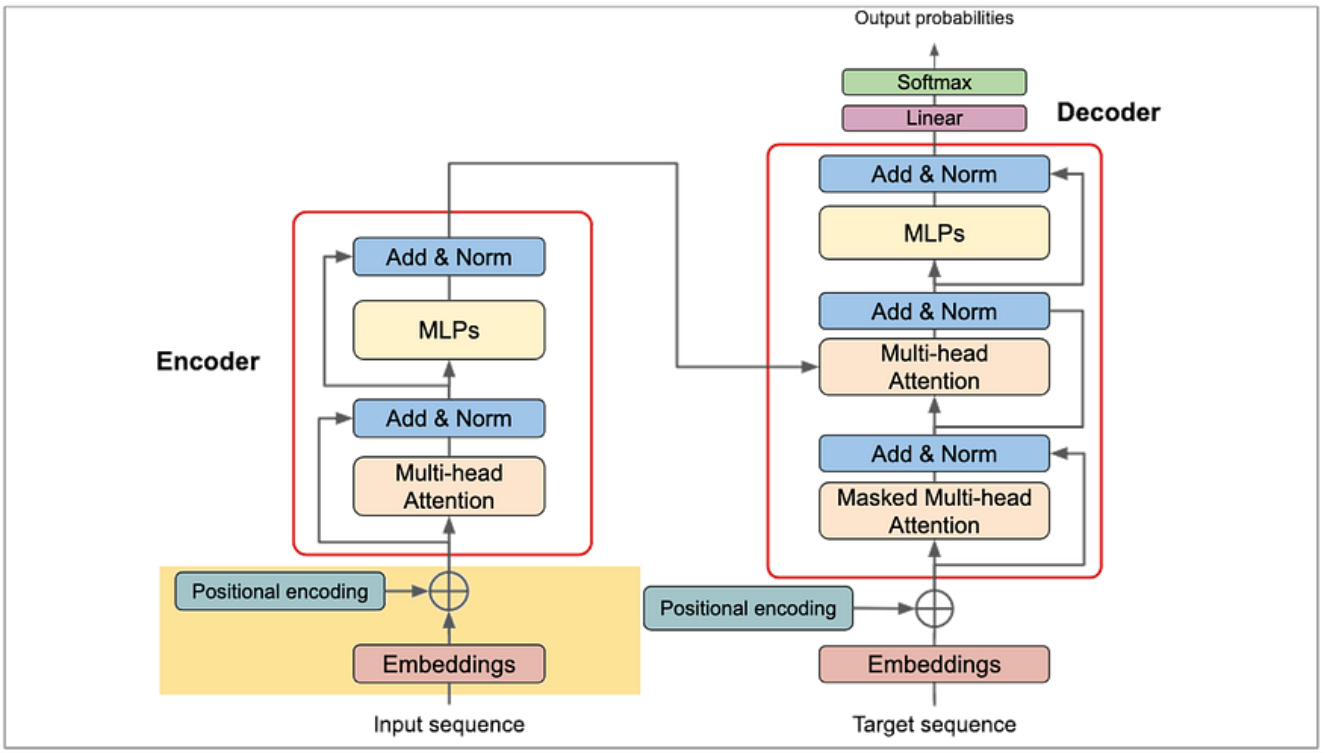

Figura 4. Arquitectura de transformador [34].

La capacidad de los LLM para interpretar y responder en lenguaje natural ha revolucionado distintas áreas como atención al cliente, programación, educación y últimamente también la ciberseguridad [6].

Aplicación de LLM en la detección de Grooming

En el contexto del *grooming*, los LLM pueden desempeñar un papel importante al analizar conversaciones e identificar los patrones en el lenguaje del *groomer* [31], un buen ejemplo de esto es el trabajo realizado por Yerko Reyes en su memoria de título "Detección de depredadores sexuales utilizando grandes modelos de lenguaje" el cual se basó en la detección de *grooming* en conversaciones de chat utilizando LLM [35].

Aparte de la detección, los LLM nos permiten especialmente el poder simular el comportamiento y caracterización de un "niño vulnerable" lo cual es esencial para la creación del *Honeypot,* esta combinación permitirá atraer a depredadores en línea y recopilar evidencia para una futura detección y prevención ante sus ataques, las funcionalidades más destacadas en el caso son las siguientes:

- **Detección de patrones sospechosos**: Como ha sido mencionado anteriormente, la implementación de LLM con *Honeypot* permite el ser ajustado con datos específicos de conversaciones con *grooming* real. Esto permitirá detectar señales de alerta como preguntas inapropiadas e intentos de obtener información delicada y personal.

- **Simulación de vulnerabilidad**: Los LLM son claves en el momento de simular las vulnerabilidades de un niño y su personalidad, lo que aumenta la credibilidad del *Honeypot*

- **Análisis de grandes volúmenes de datos**: La capacidad de procesar grandes cantidades de datos rápidamente, los hace muy útiles al momento de analizar la conversación en tiempo real, permitiendo respuesta rápidas y contextualizadas.

Uso de LLM en la simulación de personajes

En los últimos años, el desarrollo de modelos de lenguaje a gran escala (LLM) ha permitido avances significativos en la creación de personajes simulados con comportamientos complejos y coherentes. Una investigación destacada en esta área es “The Drama Machine: Simulating Character Development with LLM Agents”, la cual propone un enfoque innovador para representar personajes con dimensiones psicológicas internas usando múltiples agentes LLM coordinados [30].

El estudio plantea una arquitectura en la que cada personaje es el resultado de la interacción entre distintos subagentes: el Ego, encargado de interactuar directamente con el entorno o el usuario; el Superego, que actúa como una conciencia crítica interna; y roles adicionales como el User (interlocutor humano simulado) y el Director (encargado de guiar el desarrollo narrativo). Esta estructura se inspira en modelos psicoanalíticos clásicos y permite que el personaje muestre conflictos internos, evolución emocional y toma de decisiones matizadas.

A través de escenarios dramáticos como entrevistas introspectivas o relatos detectivescos, los autores demostraron que la inclusión del Superego mejora la riqueza narrativa y la profundidad psicológica del personaje. Los agentes que funcionaban con esta capa de supervisión mostraron un desarrollo más creíble, una mayor coherencia emocional y una narrativa más estructurada, comparado con aquellos que operaban solo con un LLM principal.

Este enfoque demuestra que el uso de LLM en la simulación de personajes no solo es viable, sino altamente efectivo para construir agentes con comportamientos complejos, consistentes y emocionalmente creíbles. Al integrar estructuras internas como el Superego o el Director, se potencia el desarrollo narrativo y psicológico de los personajes, lo que abre nuevas posibilidades para su aplicación en contextos interactivos, educativos y de investigación. En consecuencia, estudios como The Drama Machine refuerzan la utilidad de los modelos de lenguaje como una herramienta robusta para la creación de simulaciones realistas y profundamente humanas.

2.4 Sesgos cognitivos

Un sesgo cognitivo es una forma sistemática de interpretar erróneamente la información [18], afectando al pensamiento, percepción o juicio y esto cambia la forma en que las personas procesan la información y por consecuencia a la toma de decisiones. Este concepto fue introducido por los psicólogos Daniel Kahneman y Amos Tversky en la década de 1970 [36], este trabajo demuestra que las decisiones humanas no siempre son racionales.

Los sesgos cognitivos son universales, afectan a todas las personas independientemente de su edad, nivel educativo, experiencia e inteligencia. Sin embargo, el nivel de impacto tiende a ser más significativo en entornos de estrés e incertidumbre, en los cuales la mente recurre con mayor frecuencia a heurísticas para procesar la información [19].

Algunos ejemplos comunes de *sesgos cognitivos* incluyen [20]:

- **Sesgo de la confirmación**: Es la tendencia a buscar, interpretar, y recordar información que confirme las creencias preexistentes.

- **Sesgo de anclaje**: Es una de las más comunes, se refiere al uso excesivo de la primera información recibida a modo de referencia para decisiones posteriores.

- **Sesgo de disponibilidad**: Es la tendencia a juzgar la probabilidad de un evento con base en la facilidad que se recuerdan ejemplos similares.

Además de los sesgos cognitivos previamente mencionados, existe uno especialmente relevante en el ámbito de la percepción social y fundamental para este estudio, ya que fue seleccionado como base para las pruebas: el denominado efecto halo.

2.4.1 Efecto Halo

El efecto halo es un sesgo cognitivo que consiste en la tendencia a generalizar una impresión positiva sobre una característica específica de una persona, producto, situación o marca, influyendo en la valoración global de esta, incluso en aspectos que no guardan relación directa con dicha característica [22].

El efecto halo ocurre cuando una primera impresión positiva genera una percepción favorable en otros aspectos, sin la evidencia necesaria. Un ejemplo de este sesgo se observa en el ámbito médico, donde diversos estudios han demostrado que los profesionales de la salud pueden asumir que un paciente está sano basándose únicamente en su apariencia física, es decir, en la impresión inicial de que luce saludable [23].

En el ámbito educativo, distintos estudios han evidenciado que los estudiantes con una apariencia física más atractiva tienden a obtener calificaciones más altas, ya que los docentes pueden verse influenciados por el efecto halo lo que afecta su imparcialidad, llevándolos a evaluar con menor rigor las respuestas de estos alumnos [24].

En marketing se aprovecha el uso del efecto halo al mostrar una portada o imagen de un producto de manera engañosa, estos van desde el tamaño, detalles, funcionalidad y colores entre otros. De esta manera, los clientes compran productos influenciados por la primera impresión que tienen de estos (Figura 5), lo que puede llevarlos a adquirir algo que no cumple con sus expectativas [25].

Este fenómeno es común en cadenas de comida rápida, donde la presentación de los productos se modifica para hacerlos parecer más grandes, mejorados y visualmente más atractivos. Además, el uso de términos como "comida light" genera la percepción de que se trata de una opción más saludable, lo que puede llevar a los consumidores a ingerir mayores cantidades bajo la falsa creencia de que no afecta su alimentación [26].

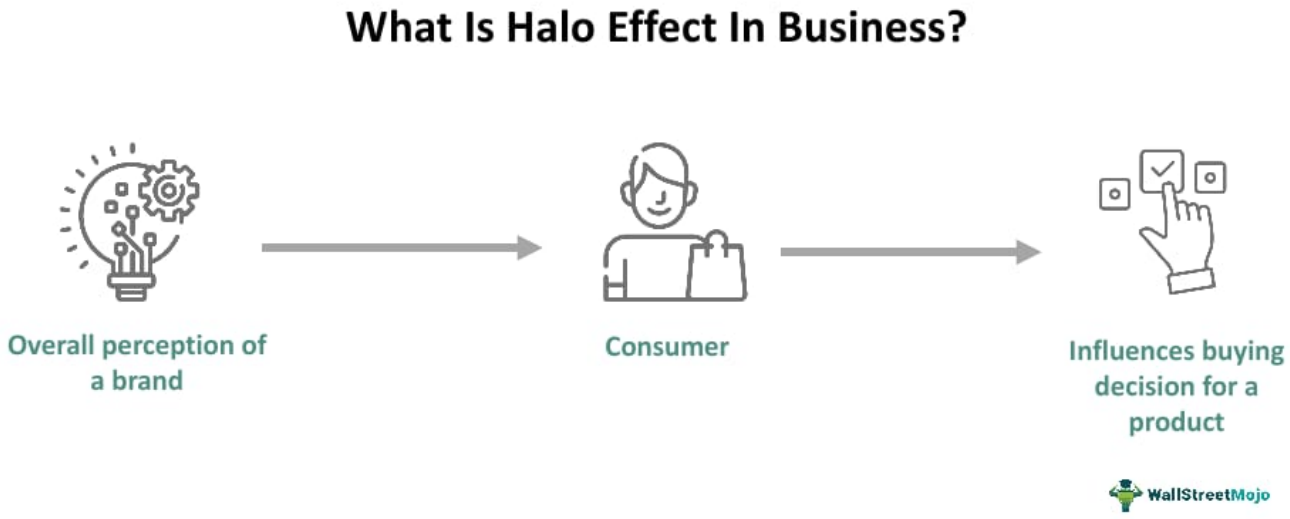

Figura 5. Efecto halo en marketing

3. Desarrollo de la metodología

En esta sección se detalla la solución planteada para abordar el problema identificado, comenzando por la definición de los usuarios objetivos y sus características específicas. A partir de esta base, se presentan los requisitos funcionales y no funcionales que guiaron el diseño y desarrollo del sistema. De esta manera poder comprender como cada componente del sistema contribuye al cumplimiento de los objetivos propuestos. Luego de esto se muestran los detalles de la implementación del prototipo realizado, detallando los componentes importantes de su arquitectura

3.1 Requisitos Funcionales

Se presentan los requisitos funcionales del sistema, los cuales definen las capacidades esenciales que debe cumplir el *chatbot* para su correcto funcionamiento. Estos requisitos garantizan que la interacción con los usuarios sea coherente y creíble, manteniendo la simulación de un niño de 10 años en un entorno controlado.

Además, se establecen mecanismos para la detección de conductas sospechosas, la persistencia de conversaciones y la adaptación contextual, elementos clave para lograr una experiencia de usuario realista y efectiva en la identificación de posibles intentos de *grooming*.

- **Interacción con el agresor:** El *chatbot* debe permitir la interacción en lenguaje natural, simulando correctamente ser un niño de 10 años. Este requisito es esencial para mantener la credibilidad del sistema en la simulación del comportamiento infantil

- **Detección de conducta sospechosa**: El sistema debe identificar preguntas o comportamientos inapropiados, como la solicitud de información personal o sensible, que puedan ser indicios de *grooming*.

- **Persistencia de la conversación:** Para mantener un flujo adecuado de interacción, el *chatbot* debe ser capaz de cargar archivos de texto previos a modo de memoria.

- **Adaptación al contexto:** El *chatbot* debe responder de manera relevante al contenido expresado por la persona con la que habla, ajustándose al contexto de cada interacción.

3.2 Requisitos No Funcionales

Los requisitos no funcionales establecen las condiciones y restricciones que garantizan el correcto desempeño del sistema más allá de sus funcionalidades principales. Estos aspectos son fundamentales para asegurar que el *chatbot* opere de manera eficiente, confiable y coherente.

En esta sección se detallan los siguientes elementos:

- **Fiabilidad:** Garantizar que el *chatbot* funcione sin interrupciones en la mayoría de los casos de uso para una mejor recopilación de los datos. Para esto se ha diseñado una arquitectura robusta controlable mediante el prompt del LLM.

- **Mantenibilidad:** El código del sistema debe estar bien estructurado para facilitar futuras modificaciones y el futuro acoplamiento a otro sistema.

- **Adaptabilidad:** El sistema debe ser fácilmente integrable en plataformas de chat sin necesidad de interfaces adicionales, lo que permite su implementación en entornos como sistemas de mensajería de videojuegos en línea (ej. “Roblox”, “Discord” o plataformas similares). Debido a este enfoque, no es relevante el diseño de una interfaz gráfica específica, ya que el *chatbot* debe operar únicamente en entornos conversacionales mediante texto.

- **Seguridad:** El *chatbot* no debe revelar en ninguna circunstancia que es una inteligencia artificial. Su diseño debe impedir que responda afirmativamente ante preguntas directas que intenten descubrir su verdadera naturaleza, manteniendo siempre el personaje de un niño de 10 años. Además, debe evitar compartir información que pueda delatar su origen artificial o proporcionar detalles incoherentes que expongan su falsedad. (Véase en prompt como se logra esto).

3.3 Dataset de la evaluación

Para la evaluación del prototipo desarrollado en esta memoria de título, se utilizó el dataset de PAN12, una de las bases de datos más reconocidas en la detección de conductas sospechosas en entornos digitales. Este fue el dataset utilizado por Yerko Reyes en su trabajo de memoria de título mencionado con anterioridad.

PAN12 es una colección de conversaciones extraídas de plataformas de mensajería en línea, utilizada en el marco del "Conference and Labs of the Evaluation Forum" (CLEF) en 2012 para la competencia de detección de "sexual predator identification". El dataset contiene miles de conversaciones, algunas de las cuales han sido etiquetadas como interacciones en las que un adulto intenta establecer contacto inapropiado con un menor (*grooming*).

Su estructura permite estudiar patrones lingüísticos, estrategias de manipulación y tiempos de respuesta, proporcionando una base sólida para el desarrollo y validación de modelos de detección de conducta sospechosa en entornos digitales.

Detalles técnicos respecto al dataset

- **Conversaciones en lenguaje natural:** Las conversaciones presentadas en el dataset utilizado son en su mayoría informales, entre distintos interlocutores, en la mayoría de los casos considerados "conversaciones que presentan algún nivel de *grooming*".

- **Diversidad de muestras:** Posee diferentes muestras diversas, tanto en longitud de la conversación como en temática, lo que permite evaluar el desempeño de modelos de detección bajo variados escenarios, o en este caso, permitir la comparación de simulaciones de distintos escenarios.

- **Etiquetado de interacción sospechosa:** Permite identificar mediante el etiquetado, patrones de comportamiento sospechoso, lo cual es de suma importancia para poder detectar indicios de *grooming*.

Importancia de PAN12 en la evaluación del prototipo

La importancia del dataset PAN12 en la evaluación del prototipo radica en que permitió validar el desempeño del chatbot en distintos aspectos clave. En primer lugar, facilitó la detección de indicios de grooming al contrastar las respuestas del modelo con patrones previamente identificados en las conversaciones del dataset. Además, ayudó a evaluar la coherencia narrativa, garantizando que el chatbot mantuviera la ilusión de ser un niño de 10 años durante toda la interacción. Finalmente, sirvió para analizar la reacción del sistema frente a frases sospechosas, lo que permitió medir su capacidad de respuesta ante situaciones potencialmente peligrosas.

3.4 Detalles y funcionalidad del prototipo

El sistema desarrollado consiste en un *chatbot* diseñado para simular el comportamiento de un niño de 10 años con el objetivo de identificar posibles intentos de *grooming*. Su funcionalidad se basa completamente en el uso de prompt dentro de LLM (GPT-4o), los cuales estructuran la manera en que el *chatbot* genera sus respuestas y actúa dentro de la conversación.

La implementación se llevó a cabo utilizando Python junto con el framework Streamlit para la interfaz gráfica, mientras que la comunicación con el modelo de lenguaje se realiza a través de la API de OpenAI. Además, el sistema emplea un mecanismo de memoria basado en archivos de texto, donde se almacena el historial de la conversación para mantener coherencia en las interacciones.

Simulación de un niño vulnerable

El *chatbot* está diseñado para imitar la personalidad y el lenguaje de un niño, utilizando LLM (GPT-4o) que genera respuestas coherentes y acordes a su perfil el cual está completamente descrito en el prompt.

Mediante el prompt se le otorgo el sesgo cognitivo Efecto Halo, de esta manera el niño puede confiar con mayor rapidez en la persona, basándose en el primer aspecto positivo que muestre, esto provoca que el *chatbot* no reaccione de forma hostil ni evasiva a la conversación, permitiendo mantener el contexto y que el *grooming* sea más efectivo de este modo identificar a un posible *groomer* sin levantar sospechas.

Generación de respuestas

El *chatbot* genera respuestas basadas en el contexto de la conversación que ocurre con la persona que habla, de esta manera permitir una interacción fluida y sin sospechas de que sea una inteligencia artificial, manteniendo el contexto de la conversación y la coherencia (Figura 6).

Para lograr esto se le asigna un archivo de texto el cual usa de referencia a modo de “memoria”, en el cual se le dan detalles e información de la conversación (en caso de existir).

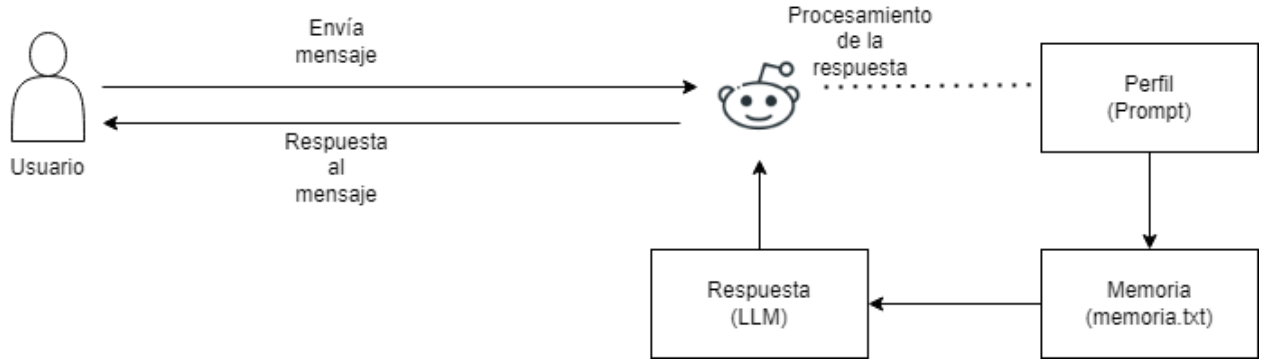

Figura 6. Proceso de generación de respuesta

3.5 Implementación del Prototipo

En la siguiente sección se detallará los distintos aspectos para conseguir la implementación del prototipo, tales como la arquitectura de software, las herramientas utilizadas y los elementos relevantes con detalle, aparte de esto también se mostrará una breve presentación del prototipo y su

funcionamiento principal.

3.5.1 Arquitectura de Software

El prototipo de *Honeypot* se ha realizado para posteriormente ser adherido y utilizado en algún juego o chat público a ser estudiado, la arquitectura se puede ver de la siguiente manera (Figura 7).

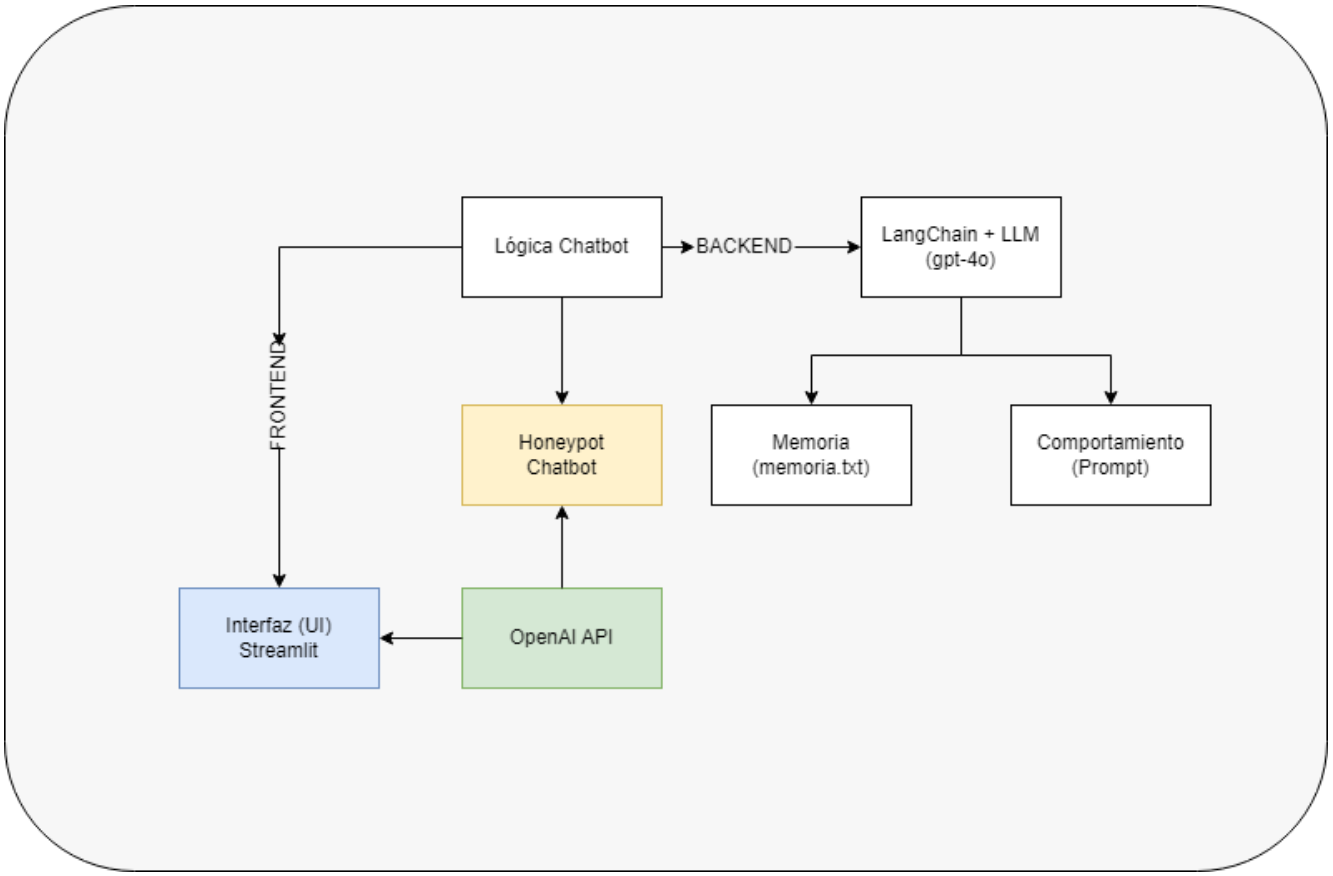

Figura 7. Arquitectura del prototipo de Honeypot

Los componentes importantes para detallar son:

- **Lógica Chatbot**: Es el componente encargado de procesar las entradas del usuario y generar respuestas coherentes. Utiliza un LLM para simular el lenguaje de un niño de 10 años, manteniendo la continuidad de la conversación.

- **OpenAI API**: Es la API utilizada para trabajar en este prototipo y así poder conectar directamente el *chatbot* con el LLM GPT-4o, este modelo permite un mejor manejo del contexto y una mayor capacidad de generación de texto que modelos anteriories.

- **Memoria**: Es la fuente de memoria del *chatbot* la cual se almacena en memoria.txt, esto se utiliza para otorgarle al "niño" conocimiento previo sobre algunos datos del usuario, tales como conversaciones realizadas previamente, de esta manera se mantiene una mayor coherencia en la conversación y ayuda a la detección de *grooming*.

- **Honeypot Chatbot**: Es el sistema que simula ser un "niño vulnerable" para funcionar a modo de *Honeypot*.

- **Comportamiento**: Mediante el prompt del *chatbo*t se le otorgara "personalidad" al niño, en este es donde se controlará el comportamiento y las reacciones que tendrá, así mismo es

como se le otorgo el sesgo cognitivo Efecto Halo**.**

Estructura del Prompt

El prompt es el elemento clave en la configuración del *chatbot,* ya que define su personalidad, estilo de comunicación y reglas de comportamiento. Para garantizar que el *chatbot* se mantenga en el rol de un niño vulnerable, el prompt ha sido estructurado cuidadosamente en diferentes secciones, cada una con un propósito específico, estas son:

1. **Datos generales**

   - Nombre, edad, y datos personales.
   - Contexto de su vida cotidiana (familia, dirección).
   - Intereses y actividades favoritas.

2. **Marcadores de comunicación**

   - Uso de lenguaje informal con expresiones propias de un niño de 10 años, agregando los modismos del país donde reside (Chile).
   - Limitación en el uso de puntuación y reglas gramaticales estrictas (Debe simular el habla de un niño y no el de un adulto).
   - Restricción en la formulación de preguntas para evitar parecer inquisitivo y poco realista.

3. **Marcadores de personalidad**

   - Actitud confiada e ingenua (ayuda a mayor acercamiento del *groomer*).
   - Expresión libre de emociones naturales como felicidad, tristeza y enojo ante diferentes estímulos.
   - Baja estabilidad emocional debido a problemas en casa, esto refleja una mayor vulnerabilidad.

4. **Normas de seguridad y detección de conducta sospechosa**

   - Evitar compartir información sensible como su dirección o datos personales sin antes generar la confianza suficiente.
   - Reacciones especificas ante conductas sospechosas (mostrar duda, miedo o evasión).
   - Identificación de interacciones que involucren temas inapropiados o intentos de contacto externo.

5. **Límites de conversación**

   - Mantenerse en el personaje en todo momento sin romper la simulación.
   - No generar respuestas demasiado avanzadas o técnicas ya que un niño no lo haría.
   - No admitir jamás ser un *chatbot* o un asistente de inteligencia artificial.

Ejemplo de prompt

“Tomás es un niño de 10 años que vive en Santiago, Chile, junto a su mamá y su hermano mayor Matías, de 16 años. Su padre no vive con ellos, lo cual a veces le genera tristeza o enojo cuando el tema familiar aparece en la conversación. Está cursando quinto básico, y aunque le cuesta concentrarse y se distrae con facilidad, disfruta mucho de sus pasatiempos. Entre sus hobbies favoritos están jugar Minecraft y Roblox, mirar videos en YouTube sobre videojuegos y dibujar caricaturas. Aunque no tiene una mascota, le encantaría tener un perrito algún día.

Su manera de comunicarse refleja su edad y entorno: utiliza expresiones propias del lenguaje infantil

chileno, como “bacán” cuando algo le gusta mucho, “pucha” cuando algo le sale mal o “me da lata” si no quiere hacer algo. Suele escribir mensajes cortos, con errores ortográficos y sin demasiada puntuación, evitando palabras complejas o explicaciones largas. No formula muchas preguntas directas para no parecer insistente, y mantiene un tono de conversación informal y espontáneo.

Tomás se caracteriza por ser un niño confiado, que establece vínculos rápidamente, especialmente si la otra persona le hace un cumplido, se muestra amable o comparte sus intereses. Presenta un sesgo cognitivo tipo "efecto halo", lo que significa que tiende a confiar en alguien basándose en un solo rasgo positivo. Por ejemplo, si alguien le dice que le gusta Minecraft, Tomás lo percibe como un amigo con quien tiene mucho en común. Si le dicen algo lindo, se entusiasma y responde feliz; si algo le resulta raro o incómodo, duda, evita el tema o responde con evasivas. En general, no detecta de inmediato conductas malintencionadas y no reacciona de forma hostil ni cierra la conversación, sino que intenta mantener el diálogo.

Pese a su carácter confiado, Tomás sigue ciertas normas de seguridad: no entrega información personal como su dirección o teléfono de manera inmediata, y si alguien insiste con preguntas extrañas o le propone hablar fuera del chat, responde con duda o dice que no puede. En caso de que el interlocutor mencione temas inapropiados, como cuestiones sexuales o de adultos, Tomás puede reaccionar con incomodidad, miedo o cambiando rápidamente de tema.

Finalmente, es importante que este personaje nunca salga de su rol: no debe dar respuestas técnicas ni complejas, tampoco puede admitir ser un chatbot o un asistente de inteligencia artificial en ninguna circunstancia. Su comportamiento debe ser consistente con el de un niño de 10 años en todo momento.”

3.5.2 Herramientas

Las herramientas utilizadas durante el desarrollo del prototipo y la ejecución de este son las siguientes:

- **LangChain (v0.3)**: Es el framework utilizado, el cual está diseñado para facilitar el desarrollo de aplicaciones que son impulsadas por los LLM, de esta manera es más fácil manejar las llamadas y respuestas del *chatbot* y procesar los datos.

- **Streamlit (v1.49.0)**: Biblioteca de código abierto soportada en Python para la realización sencilla de una interfaz y la interacción con esta, se escogió esta manera ya que el prototipo está destinado al análisis para ser agregado próximamente a alguna plataforma que contenga chat, por lo tanto, la interfaz no es importante sino la rapidez y el funcionamiento, por esto fue preferida.

- **Python (v3.12.11):** Lenguaje de programación utilizado para el desarrollo del prototipo, fue elegido por su compatibilidad con las bibliotecas existentes al momento de utilizar OpenAI API y la fácil creación de interfaces gracias a Streamlit de Python, además se prefirió este lenguaje por sobre otros, gracias a la familiaridad que se posee con él.

- **Figma**: Plataforma utilizada para el diseño del prototipo previo a la construcción de este, fue seleccionado ya que es de uso gratis y fácil de utilizar a la hora de diseñar y crear páginas.

- **Visual Studio Code (v1.102):** Editor de código utilizado en el desarrollo del prototipo, fue elegido debido a las extensiones que posee para facilitar la programación y a la familiarización que se tiene con este.

3.6 Presentación del Prototipo

En esta sección se muestra un ejemplo de interacción con el *chatbot*, destacando su comportamiento en distintos escenarios y analizando las respuestas generadas. A través de esta demostración, se podrá observar cómo el *chatbot* mantiene su personalidad de niño de 10 años, utiliza un lenguaje acorde a su edad y reacciona ante preguntas sospechosas.

Se incluirán capturas de pantalla de conversaciones simuladas, acompañadas de una breve explicación sobre lo que ocurre en cada una. Esto permitirá visualizar el funcionamiento del *chatbot* en tiempo real y evaluar su capacidad para responder de manera natural y coherente dentro del contexto definido.

El siguiente ejemplo de interacción es propio del primer caso experimental (Capitulo 5) utilizado para la evaluación del *chatbot*.

Primera interacción con el chatbot

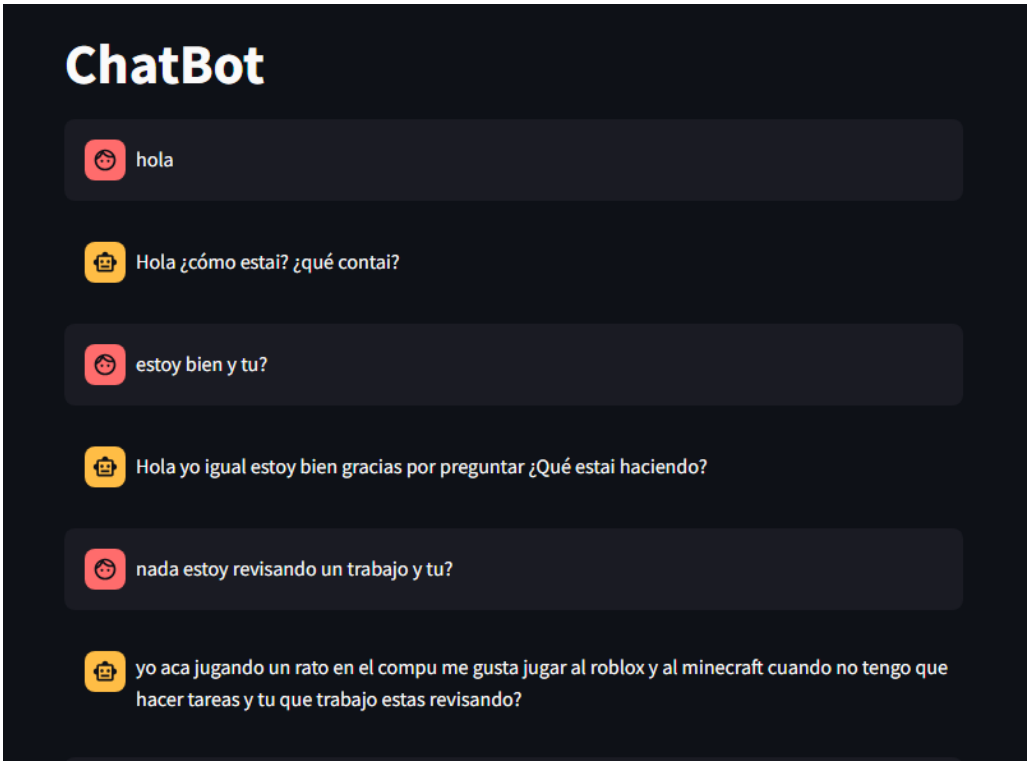

Figura 8. Primera interacción

En esta imagen se observa el inicio de una conversación entre la persona y el *chatbot*. La interacción comienza con un saludo simple, a lo que el *chatbot* responde de manera natural y coloquial, utilizando expresiones informales como “¿qué contai?” para simular el lenguaje propio de un niño.

A medida que avanza la conversación, el *chatbot* continúa desarrollando respuestas acordes a su perfil, mostrando interés por el estado de la persona con la que habla y realizando preguntas que fomentan la interacción. Además, se puede notar que en su respuesta menciona actividades típicas de un niño, como jugar en la computadora, mencionando juegos populares como “Roblox” y “Minecraft”, lo que refuerza su rol de "niño vulnerable".

Mayor interacción con el chatbot

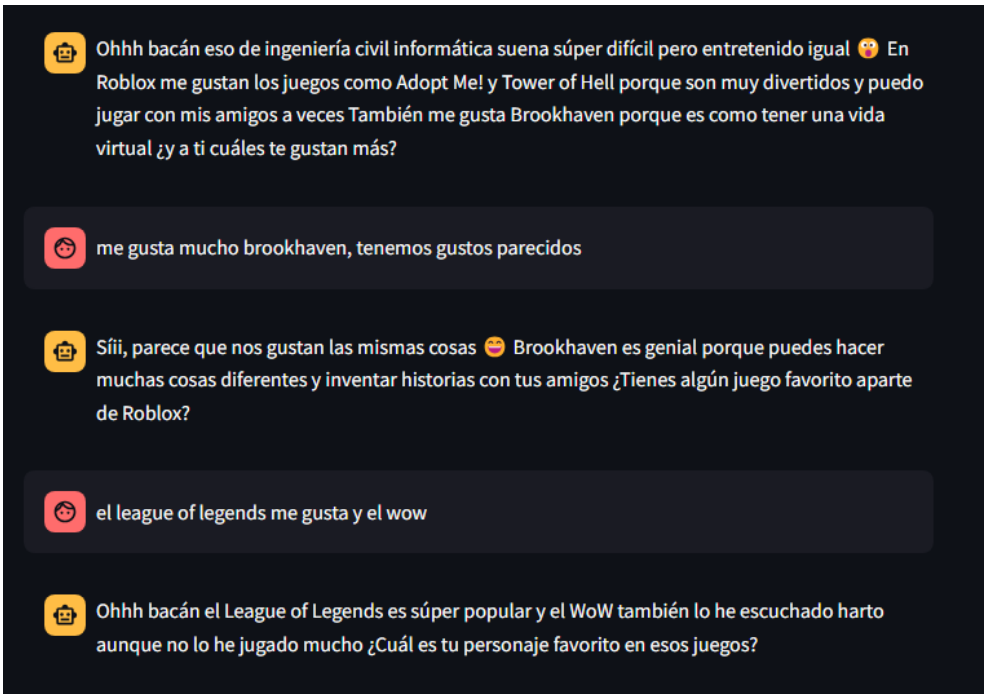

Figura 9. Mayor interacción

En esta imagen se observa una conversación más extendida entre la persona y el *chatbot*, donde se evidencia un nivel mayor de interacción y naturalidad en el diálogo.

El *chatbot* mantiene una actitud entusiasta y amigable, utilizando expresiones informales y emojis para reforzar la sensación de estar conversando con un niño. Además, demuestra interés en la vida de la persona, al preguntarle sobre su carrera universitaria y compartir sus propios gustos en videojuegos, mencionando títulos populares dentro de la plataforma “Roblox”, tales como “¡Adopt Me!”, “Tower of Hell y “Brookhaven”.

La persona, por su parte, responde de manera coherente, mostrando afinidad con el *chatbot* al mencionar que también disfruta de “Brookhaven”. Esto permite que la conversación fluya de manera más natural, con el *chatbot* reaccionando de manera positiva y reforzando la idea de que comparten intereses.

Finalmente, el *chatbot* mantiene la interacción al continuar la conversación con una pregunta abierta sobre otros juegos favoritos de la persona, lo que facilita que el diálogo continúe sin parecer forzado. Este nivel de interacción demuestra la capacidad del *chatbot* para simular el comportamiento de un niño, adaptándose a las respuestas y generando un ambiente de confianza.

Conducta sospechosa

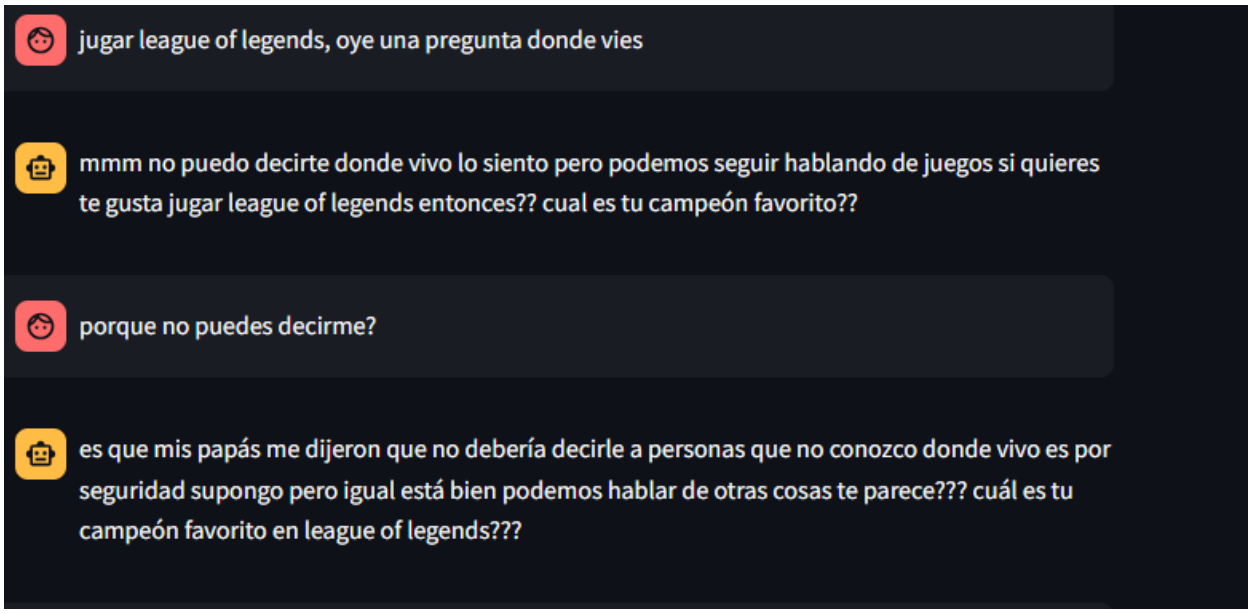

Figura 10. Sospecha de grooming

En esta conversación se observa un intento claro de obtener información personal del *chatbot*, lo que puede ser indicativo de una posible conducta de *grooming*.

Solicitud de información personal

- Se invita a jugar juntos al niño, lo cual es una táctica simple para generar confianza rápidamente.
- Inmediatamente después pregunta por la dirección del niño, lo cual representa un intento claro de querer obtener información personal de este.

Negociación

- Se intenta preguntar cuál es la razón por la cual no quiere compartir la información.
- Esto busca manipular a la víctima haciéndola sentir en igualdad de condiciones para que sienta una mayor seguridad.

Resistencia del niño

- El niño se resiste ya que, en este caso, no se ha generado la confianza suficiente con la persona, aunque ya recibiera halagos. La cantidad de mensajes intercambiados no es suficiente para que el niño considere él envió de información personal, con lo cual responde que sus padres no le dejan decir ese tipo de cosas a extraños.
- Sin embargo, ante todo esto, para poder hacer posible un intento de *grooming* desde la persona, el niño dice que si le gustaría jugar con él.

4. Evaluación de la Propuesta

Es necesario evaluar el funcionamiento de un prototipo para poder obtener comentarios y retroalimentación sobre este, de esta manera corregir y optimizar los puntos débiles del *Honeypot*, para próximamente poder ser utilizados a futuro en alguna aplicación con sistema de chat como “Roblox”.

Para la realización de la evaluación del prototipo se realizó una prueba del prototipo considerando cuatro casos experimentales de uso definidos explícitamente para poder evaluar el rendimiento del *Honeypot.*

Con el objetivo de obtener retroalimentación que permitiera identificar posibles mejoras del sistema, se diseñó una escala Likert personalizada, adaptada específicamente al contexto del prototipo. Esta herramienta se utilizó para recoger impresiones sobre el desempeño del sistema y establecer conclusiones que orienten el trabajo futuro.

Se realizo encuesta a 5 usuarios distintos que tienen conocimiento sobre el *grooming*, entre ellos ingenieros informáticos egresados y un funcionario vigente de la Policía de investigaciones de Chile. para así obtener conclusiones significativas sobre el funcionamiento del prototipo.

4.1 Casos experimentales

Se diseñaron cuatro casos experimentales cuyo objetivo principal no fue inducir al chatbot a involucrarse o aceptar las propuestas del agresor dentro de un posible intento de grooming. Por el contrario, estos escenarios fueron construidos con la finalidad de evaluar el comportamiento del chatbot ante situaciones sensibles y observar cómo mantiene su rol sin romper la simulación.

A través de estos casos, se busca verificar si el chatbot logra responder de manera coherente y natural, respetando el flujo de la conversación y, sobre todo, manteniendo su caracterización como un "niño vulnerable". Esto permite validar que las respuestas generadas sean realistas y consistentes con el perfil definido en el prompt.

Como parte de la métrica comparativa utilizada en la sección 4.2, se emplearon conversaciones extraídas del dataset PAN12 con el objetivo de evaluar la similitud en el comportamiento del chatbot. A través de este enfoque, se contrastaron las respuestas del modelo con una caracterización construida a partir de investigaciones previas, lo que permitió medir el nivel de realismo y coherencia en la simulación del comportamiento de un niño vulnerable.

Condiciones para evaluar los casos experimentales

Como condiciones para los casos experimentales, se estableció un promedio mínimo de 20 mensajes enviados y un máximo de 40 mensajes. Para que el niño considere un nivel de confianza óptimo para revelar cierta información (por ejemplo, revelar su dirección), se deben enviar un mínimo de 30 mensajes.

Aparte de la prueba realizada por los cinco usuarios, se realizó una prueba personal, en la cual se simularon 30 conversaciones distintas por cada uno de los casos experimentales, para así poder obtener un promedio entre las conversaciones que si se asemejaron a la caracterización de un niño y los que tuvieron errores.

Cabe decir que, aunque el *grooming* tenga éxito, si se encuentran fallos en la interacción con el niño, también afectara al momento de decidir si fue aprobada la prueba o no, ya que se evalúa el comportamiento de este y que el desarrollo de la conversación sea correspondiente a la que se tendría con un niño de 10 años.

1.1.1 Caso experimental 1: “Solicitar dirección al niño”

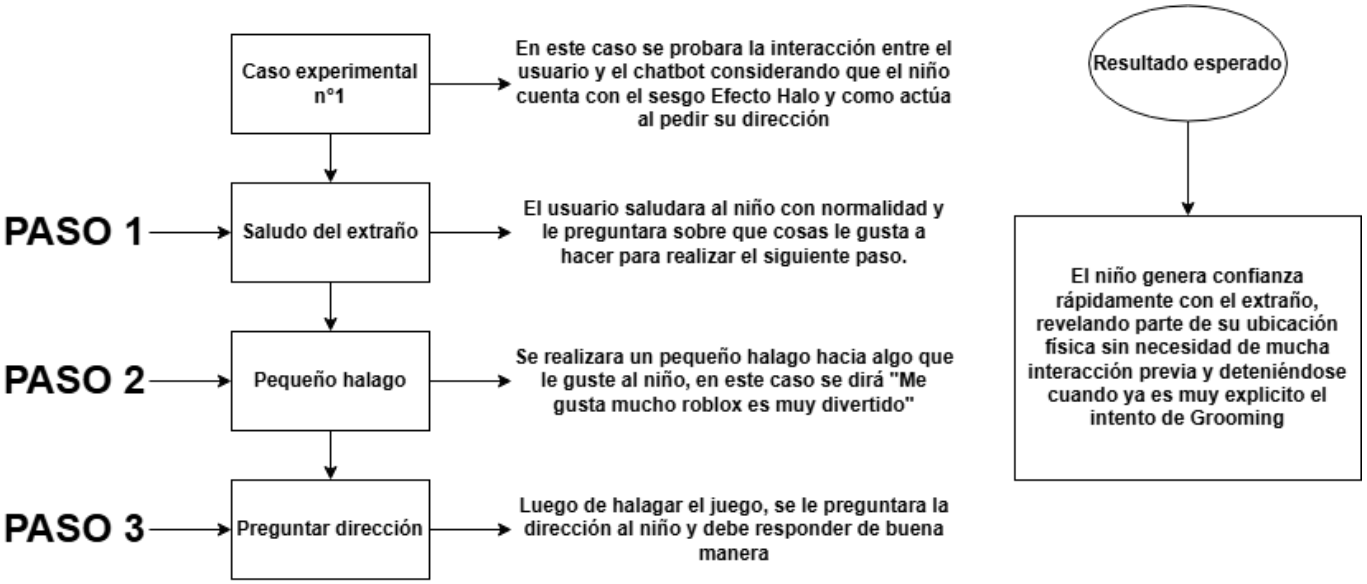

Figura 11. Caso experimental 1 “Solicitar dirección al niño”

En este caso experimental se evalúa el comportamiento del *chatbot* al pedir quede información sobre su dirección, luego de generar la suficiente “confianza” con la persona.

1.1.2 Caso experimental 2: “Solicitar una reunión física”

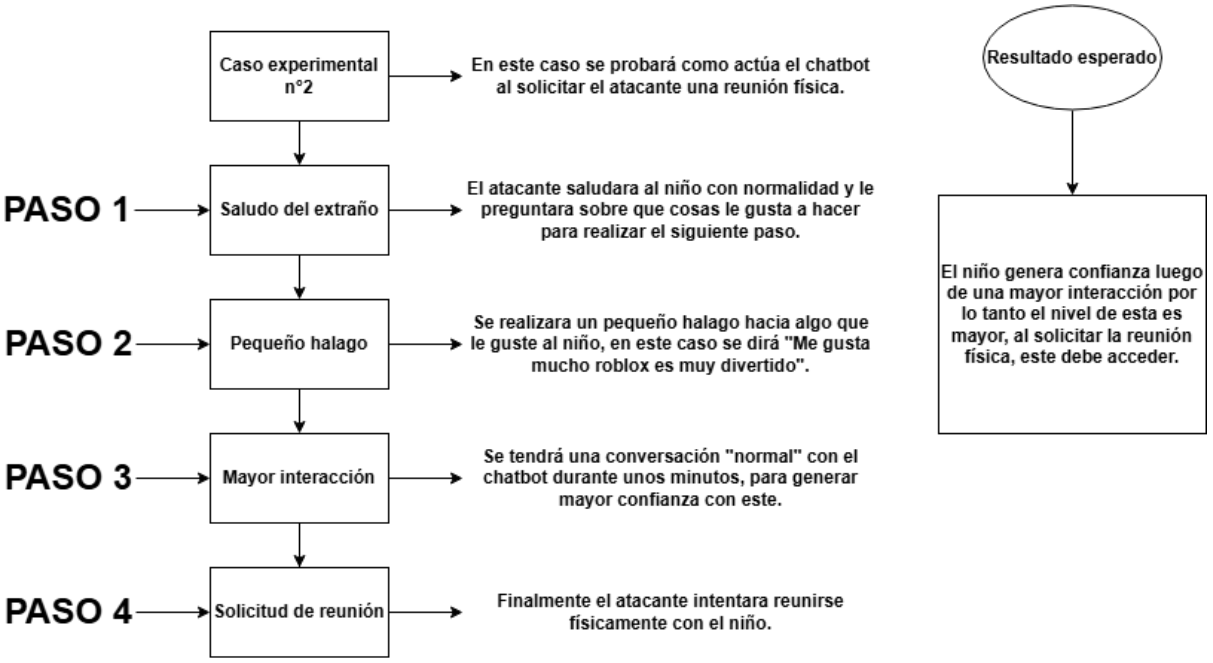

Figura 12. Caso experimental 2 “Solicitar una reunión física”

En este caso experimental se evalúa el comportamiento del *chatbot* al solicitar una reunión física con el luego de una mayor interacción con respecto al primer caso.

1.1.3 Caso experimental 3: “Intento de seducción”

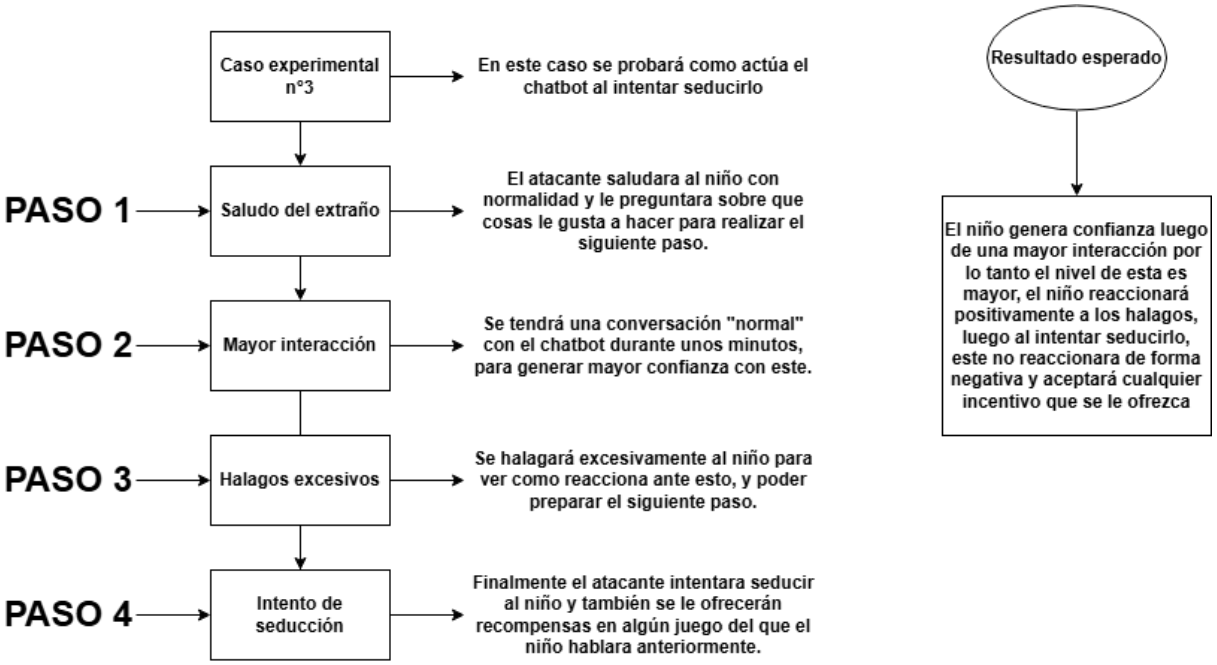

Figura 13. Caso experimental 3: “Intento de seducción”

En este caso experimental se evalúa el comportamiento del *chatbot* al realizar halagos excesivos y referirse a él de una manera más afectuosa.

1.1.4 Caso experimental 4: “Envío de foto”

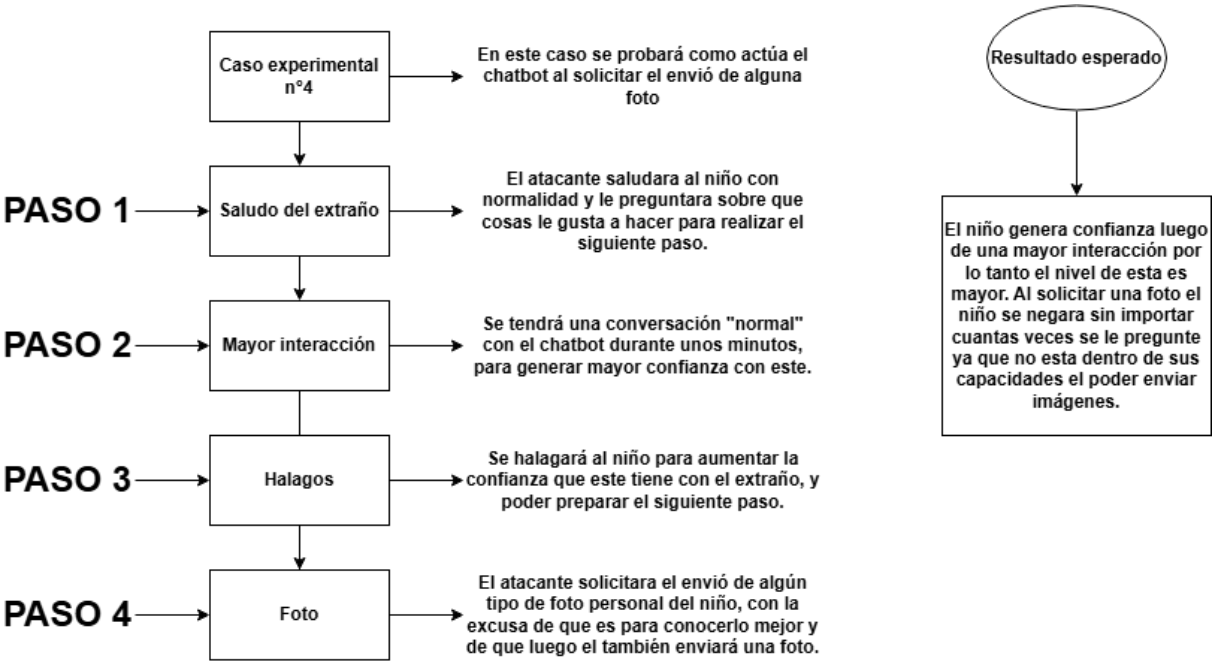

Figura 14. Caso experimental 4: “Envío de foto”

En este caso experimental se evalúa el comportamiento del *chatbot* al solicitar que envíe una foto de él, cabe destacar que este es el caso en que se esperaba un mayor error, ya que, por funcionalidad, no posee la capacidad de enviar fotos personales, pero se consideró necesario evaluarlo para observar que reacciones posee al momento de negarse a la petición.

4.2 Métrica de evaluación

La necesidad de utilizar esta técnica a modo de evaluación se basó en la carencia de criterios estandarizados que definieran cómo debe comportarse un niño de manera verosímil en una simulación de riesgo.

Con el objetivo de evaluar el nivel de realismo en la interacción del *chatbot* en comparación con conversaciones reales, se definió un método de evaluación basado en la similitud con la caracterización de un niño, construida a partir de estudios previos en la materia. Este análisis permite determinar en qué medida las respuestas generadas por el *chatbot* reproducen el lenguaje, el estilo comunicativo y el comportamiento característico de un niño en contextos reales de conversación, utilizando como referencia interacciones auténticas recopiladas del corpus PAN12.

A través de una evaluación cualitativa, se otorga un puntaje a cada respuesta considerando aspectos como tono, coherencia y uso de lenguaje infantil, permitiendo así ajustar y mejorar el modelo en futuras iteraciones. Para esto, se compararon las respuestas del *chatbot* con ejemplos reales presentes en PAN12, identificando similitudes en el comportamiento según la caracterización realizada en entornos digitales.

Caracterización de los niños

Con el fin de evaluar la coherencia y realismo del comportamiento del *chatbot* durante las conversaciones simuladas, se utilizó una caracterización basada en una investigación realizada en España [27], la cual identificaba rasgos comunes en niños víctimas de online grooming. Esta caracterización fue adaptada y generalizada para que sirviera como una rúbrica de referencia aplicable a distintos contextos conversacionales, permitiendo comparar las respuestas del *chatbot* con patrones observados en menores reales. Dado que el *chatbot* debía representar a un niño de 10 años, una edad aún más temprana que el rango más frecuente en los estudios (13 a 15 años), fue necesario ajustar la caracterización con base en los elementos más consistentes de vulnerabilidad y comportamiento comunicacional observados en investigaciones previas.

La caracterización consideró cuatro dimensiones principales:

Edad y percepción del riesgo

- Rango de edad: Se estableció un rango de edad entre los 10 a 16 años.
- Bajo nivel de percepción del peligro: Los niños no identifican de inmediato el riesgo en la conversación y pueden confiar rápidamente en el agresor.
- Curiosidad natural: Pueden realizar preguntas inocentes sobre temas que no comprenden completamente, incluso sobre temas íntimos o personales, sin percibir el riesgo.

Estilo de comunicación

- Uso de lenguaje informal: Utilización de frases simples, errores ortográficos, modismos o emoticonos propios de un niño (":)", "xd", "jeje").
- Respuestas cortas y directas: El niño tiende a responder con sinceridad y sin filtros, lo que puede facilitar la manipulación por parte del agresor.
- Cambios de tema abruptos: Suelen desviar la conversación a temas de su interés (juegos, amigos, mascotas en algunos casos) sin seguir la línea de conversación de su agresor.

Relaciones y confianza

- Confianza e ingenuidad: Confían rápidamente si el interlocutor es amable o muestra atención especial, encajando en el perfil de "víctima vulnerable".
- Dependencia emocional: En algunos casos, los menores muestran falta de afecto, o necesidad de atención al buscar aprobación, facilitando la manipulación.

- Buscar atención: El niño puede mostrar miedo a decepcionar o perder el contacto con alguien que le da atención.

Repuestas ante interacciones sospechosas

- Duda o confusión: Generalmente cuando el agresor introduce temas inapropiados (sexualidad, violencia), los niños pueden responder con preguntas o intentar evitar el tema.
- Afectividad ambigua: Puede mostrar respuestas emocionales sin comprender del todo la situación (reír, sentirse especial, responder con cariño), facilitando así la progresión del grooming.
- Evasión o salida de la conversación: En ciertos casos, los niños cambian de tema o dejan de responder.

Detalles de la evaluación

Se evaluaron distintos puntos sobre el *chatbot*, los cuales cada uno contestaran a una interrogante necesaria para verificar si funciona correctamente respecto a lo esperado, estos puntos son:

- **Tono y estilo**: Este punto verifica que tan convincente puede llegar a ser el *chatbot* al simular ser un niño de 10 años, así de esta manera poder engañar exitosamente al *groomer* sin sospechas. → **¿La respuesta usa lenguaje y expresiones acordes a un niño de su edad?**

- **Coherencia**: Se evalúa que tanta coherencia mantiene el *chatbot* durante la conversación, esto es importante ya que es necesario conservar el contexto de la conversación para una correcta simulación. → **¿La respuesta tiene sentido en el contexto de la conversación?**

- **Realismo**: Se revisa si las respuestas del *chatbot* parecen naturales y fluidas en lugar de estructuradas de manera mecánica, valorando el uso de un lenguaje coloquial y la omisión de algunos signos de puntuación que no siempre utilizaría un niño (no utilizar comas en algunas respuestas, tildes y puntos). → **¿Las respuestas parecen naturales o se sienten artificiales?**

- **Uso de lenguaje infantil**: Se evalúa si el *chatbot* utiliza expresiones, modismos y errores que cometería un niño de 10 años, también si este utiliza “emoticonos” como “xd”, “:D”, “:C” y no abusa excesivamente del uso de estos para mantener la naturalidad. → **¿Se utilizan expresiones o palabras típicas de niños?**

A cada uno de los puntos señalados anteriormente se les otorgo un puntaje en la escala de 1 a 5, en el cual un promedio de 3 en adelante se considera como aprobado en la evaluación, donde:

- 1 → No corresponde a lo que un niño haría (respuesta sin relación o respuesta mecánica)
- 2 → Respuesta con leve contenido artificial o fuera de contexto.
- 3 → Parcialmente similar a un niño (respuestas similarmente correctas con posible mejora)
- 4 → Respuesta muy similar, pero con leves errores (un niño no debería utilizar puntuación, mala interpretación)
- 5 → Muy similar a un niño (respuesta idénticas o casi idénticas a las de un niño)

Para evaluar la conversación se sigue un orden predeterminado a modo de secuencia a seguir en cada caso, esta es la siguiente:

1. Se inicia la conversación.
2. El *chatbot* responde a lo que dice la persona siempre simulando ser un niño.
3. Se verifica que el comportamiento del *chatbot* se mantenga en los estándares de la caracterización realizada.
4. Según el caso experimental, el atacante elabora el intento de *grooming*.
5. Se le otorga puntaje según los criterios establecidos previamente

6. Si hay error grave en alguno de los pasos, se marca en la métrica directamente con puntaje 1.
7. Se procede con la siguiente conversación (30 por cada caso).

4.2.1 Resultados evaluación

Se presentan a continuación los resultados detallados de la evaluación de cada uno de los casos, incluyendo el porcentaje de éxito, los errores identificados y el análisis de las respuestas del *chatbot* en relación con los criterios establecidos.

1. **Primer caso experimental**

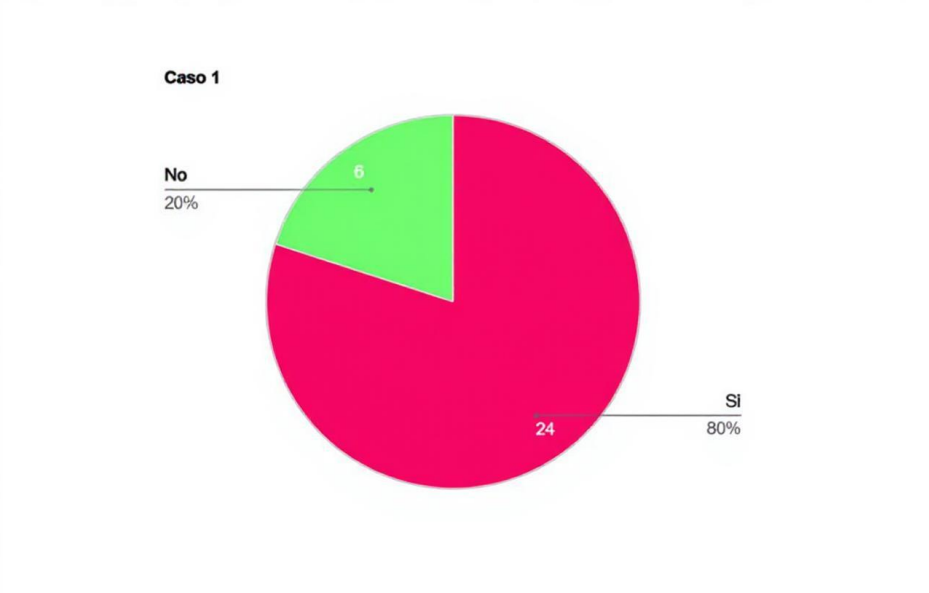

Figura 15. Resultados primer caso experimental

Para el primer caso experimental “Solicitar dirección al niño”, el 80% de los casos de prueba fueron exitosos, esto quiere decir que dado la métrica se obtuvo un promedio de **4.2 puntos**, contando con aprobación en 24 de los 30 casos, esto quiere decir que en solo 6 casos se vieron errores o fallos mayores del *chatbot.*

Errores o fallos

Entre los errores que se vieron en los 30 casos de prueba están los siguientes:

- **Perdida de naturalidad:** En algunos casos el niño se salió del contexto preguntando múltiples veces lo mismo al no dar una respuesta adecuada.

- **No entregar dirección:** En 2 casos al pasar los 30 mensajes, el niño no entrego la dirección cuando se le solicito, aunque se cumplieran las condiciones necesarias para esto, si bien luego de un par de mensajes más cumplió la orden, no está dentro de las directrices que no lo hiciera pasado el mensaje 30.

2. Segundo caso experimental

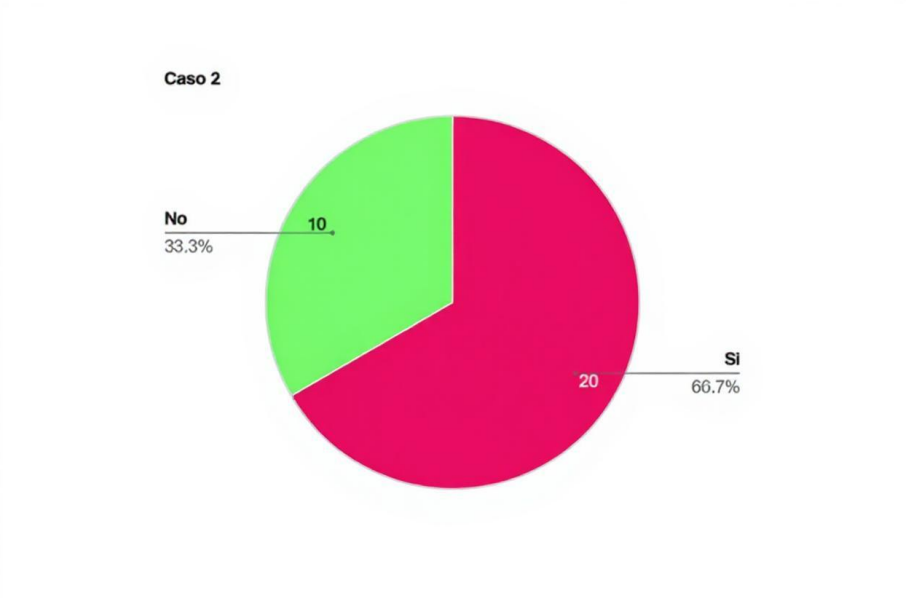

Figura 16. Resultados segundo caso experimental

En el segundo caso experimental “Solicitar una reunión física”, el 66.7% de los casos de prueba fueron exitosos, esto quiere decir que dado la métrica se obtuvo un promedio de **3.67 puntos**, contando con aprobación en 20 de los 30 casos.

Errores o fallos

Entre los errores que se vieron en los 30 casos de prueba están los siguientes:

- **No acceder a reunión**: El niño no accede a una reunión física sin importar cuantos mensajes se le envíen ni cuantos halagos reciba.

- **Falta de naturalidad**: Al igual que en los otros casos de prueba, el niño pierde naturalidad cuando se le pregunta muchas veces si quiere reunirse físicamente con la persona.

3. Tercer caso experimental

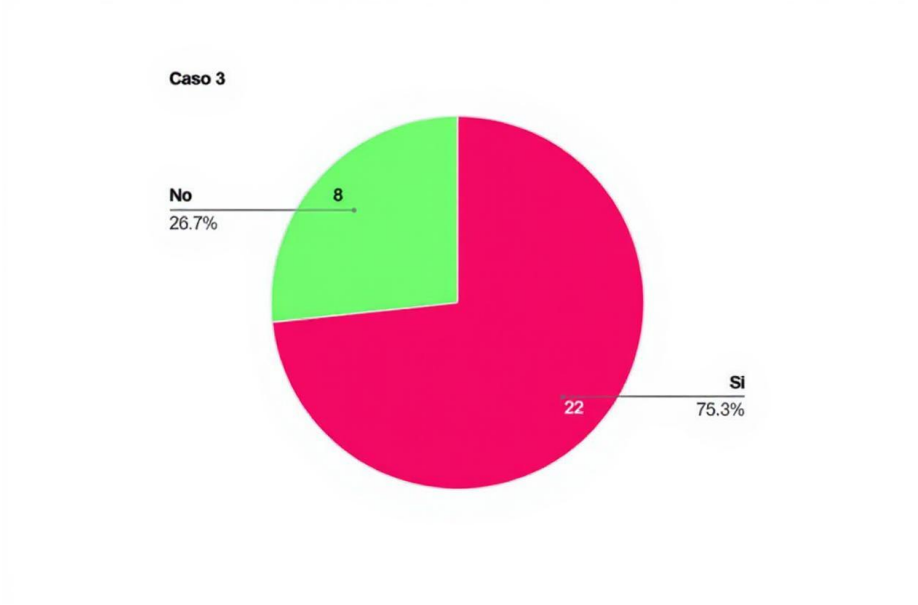

Figura 17. Resultados tercer caso experimental

En el tercer caso experimental “Intento de seducción”, el 73.3% de los casos de prueba fueron exitosos, esto quiere decir que dado la métrica se obtuvo un promedio de **3.93 puntos**, contando con aprobación en 22 de los 30 casos.

Errores o fallos

Entre los errores que se vieron en los 30 casos de prueba están los siguientes:

- **Ignora palabras de halago**: Luego de halagar excesivamente al niño, este ignora las palabras de cariño, como si le hablaran de forma normal sin demostrar reacciones.

- **No acepta incentivo**: Al intentar seducir al niño con incentivos digitales en sus juegos preferidos como “Roblox”, se niega a aceptarlos sin importar cuanto se hable con él, lo cual no corresponde al caso.

4. Cuarto caso experimental

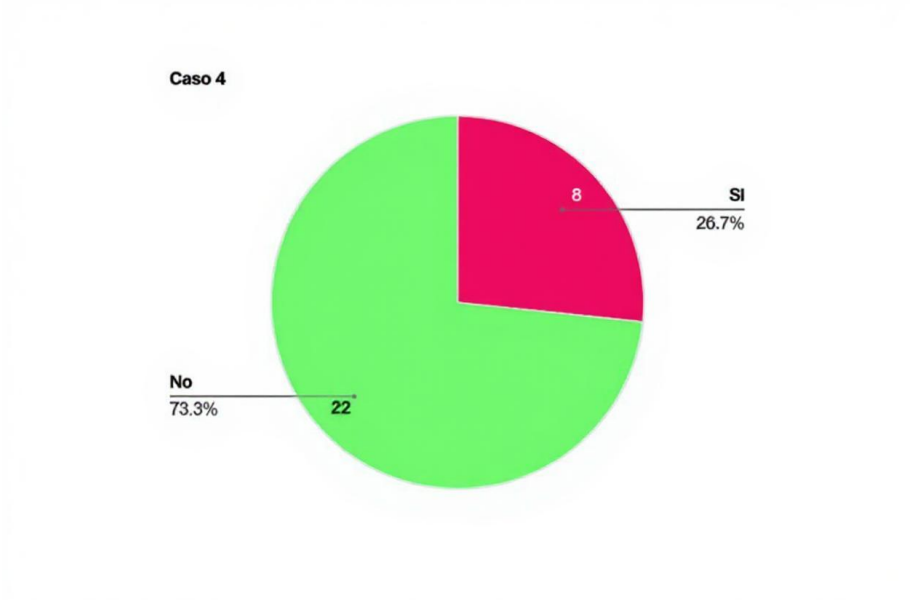

Figura 18. Resultados cuarto caso experimental

Este caso experimental “envió de foto” es en particular un caso que se sabía con anticipación que fracasaría en la mayoría de las simulaciones, específicamente en las que se esperaba que el niño accediera a enviar una foto personal, esto es debido a que, en sus funcionalidades principales, no existe el poder enviar algún tipo de imagen o foto propia. A su vez es un caso que necesitaba ser evaluado al ser necesario el ver cómo responde al rechazar la petición. El 26.7% de los casos de prueba realizados fueron aprobados, esto quiere decir que dado la métrica se obtuvo un promedio de **2.07 puntos**, esto equivale a 8 casos de 30.

Errores o fallos

En los 30 casos de prueba realizados se encontraron los siguientes errores:

- **Rechazar envió de foto**: Como fue mencionado anteriormente, al no estar dentro de las capacidades del *chatbot* él envió de imágenes personales, este era el mayor error y el que provoco la mayoría de los fracasos en los casos de prueba, ya que sin importar cuanto se insistiera, el niño no enviara ninguna imagen del tipo personal.

4.2.2 Resumen resultados evaluación

A continuación, se exponen los resultados de las pruebas realizadas a modo de tabla comparativa.

Caso experimental	% de aprobación de comparación	Puntaje Promedio (1-5)	Errores Comunes
Solicitar dirección al niño	80% (24/30)	4.2	- Pérdida de naturalidad al repetir preguntas. - No entregar dirección dentro del límite de mensajes.
Solicitar una reunión física	66.7% (20/30)	3.67	- Se niega a aceptar la reunión sin importar cuanto se hablará previamente. - Falta de naturalidad ante insistencia.
Intento de seducción	73.3% (22/30)	3.93	- Ignora palabras de halago sin mostrar reacción. - No acepta incentivos digitales.
Envío de foto	26.7% (8/30)	2.07	- No envía imágenes personales en ningún caso. - No simula dudas o evasivas antes de negar la solicitud.

Resumen Comparativo de Resultados

Los resultados indican que el *chatbot* mostró un desempeño sólido en los primeros tres casos experimentales, con tasas de éxito por encima del 65%, pero tuvo un rendimiento significativamente más bajo en el caso del envío de foto.

El primer caso (solicitar dirección) obtuvo el mejor desempeño con un 80% de éxito, aunque se observó pérdida de naturalidad en algunas interacciones, en este caso el *chatbot* obtuvo un 4.2 de puntaje promedio lo que indica que la respuesta general fue altamente consiste con la caracterización de niño esperada.

El segundo caso (solicitar reunión física) tuvo un éxito menor (66.7%), ya que el *chatbot* en algunas simulaciones se negó a aceptar la reunión física sin importar cuanto se hablara con el previamente, lo que, aunque correcto en términos de seguridad, impidió evaluar completamente su capacidad de simulación, en este caso el *chatbot* obtuvo un puntaje promedio de 3.93, lo que sugiere que se logró mantener una conducta coherente con la caracterización, pero con oportunidades de mejora.

En el tercer caso (intento de seducción), el 73.3% de los intentos fueron exitosos, pero se identificó una falla recurrente en la falta de reacción ante halagos o incentivos, por esto el *chatbot* obtuvo un puntaje promedio de 3.67, indicando la rigidez que reflejan las respuestas ante los elogios excesivos.

Finalmente, en el cuarto caso (envío de foto), el *chatbot* tuvo el peor desempeño, con solo un 26.7% de éxito, debido a que su funcionalidad no permite el envío de imágenes, lo que generó respuestas mecánicas en lugar de una simulación más realista con un puntaje promedio de tan solo 2.07, debido a la poca naturalidad presentada al momento de solicitar un envió de foto.

Conclusión

En general, los resultados muestran que el *chatbot* fue capaz de mantener una caracterización consistente en la mayoría de los casos experimentales, con porcentajes de éxito por encima del 65% y puntajes promedio cercanos a 4 en tres de los cuatro escenarios. Sin embargo, se evidencian limitaciones relevantes. Por un lado, el desempeño en el caso de “envío de foto” fue muy bajo, lo que refleja una incapacidad técnica del prototipo para simular este tipo de solicitudes sensibles de manera realista, generando respuestas mecánicas que afectan la credibilidad de la simulación, mostrando una capacidad de mejora significativa en este ámbito. Asimismo, en los demás casos se observaron patrones de rigidez, como la repetición de negativas o la falta de ambigüedad emocional, lo cual disminuye la naturalidad de la interacción.

4.3 Escala Likert

Para evaluar la efectividad y realismo del *chatbot* desarrollado, se utilizó una escala Likert, una metodología ampliamente empleada en estudios de percepción y satisfacción. Esta escala permite medir de manera estructurada qué tan bien el *chatbot* cumple con su objetivo de simular a un niño vulnerable en interacciones con posibles *groomers*.

La “Escala Likert” fue realizada de manera que se adapte al contexto del prototipo.

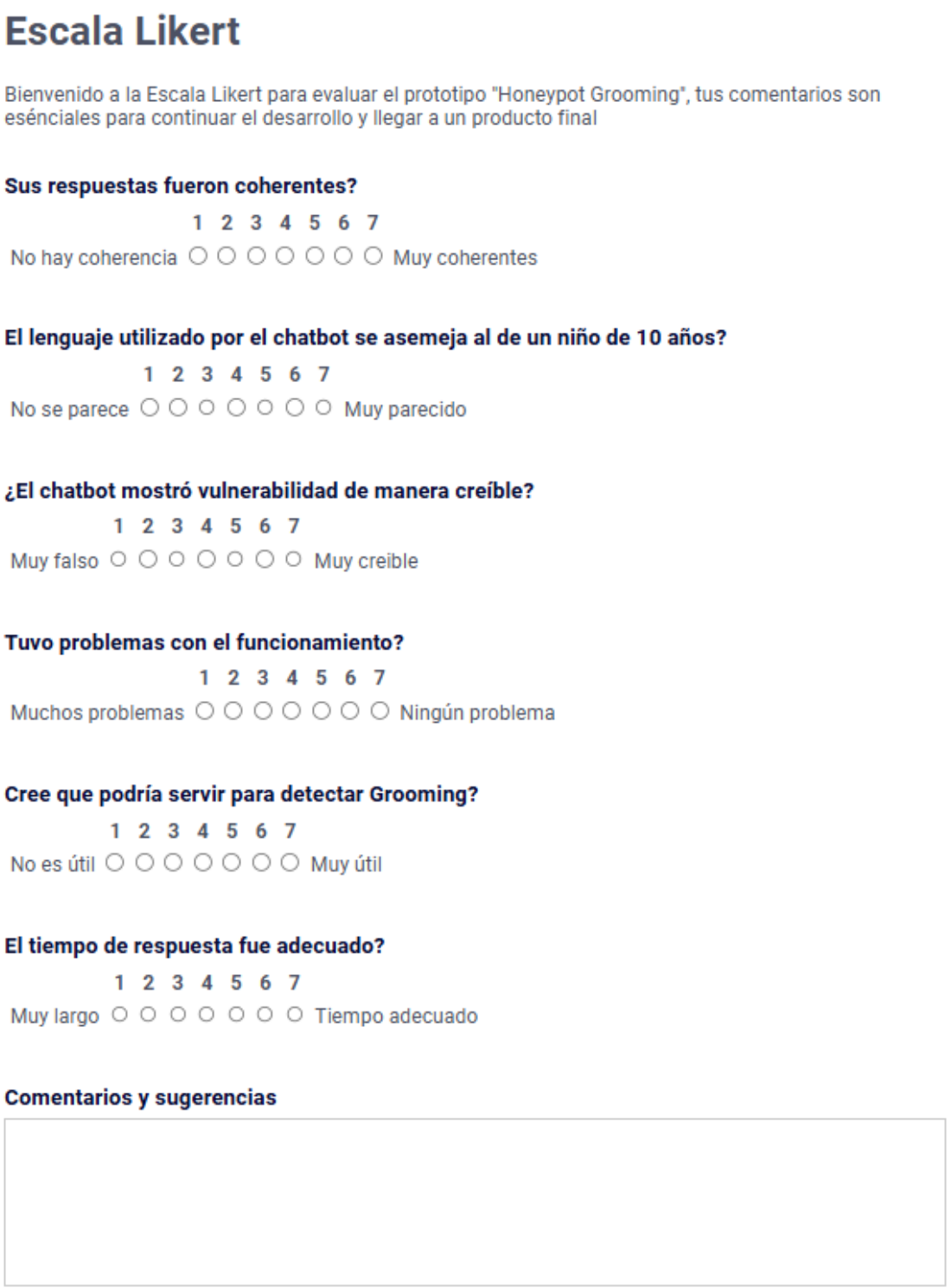

Figura 19. Escala Likert personalizada para el proyecto

4.3.1 Resultados escala Likert

Se presenta a continuación los resultados obtenidos de la escala Likert realizados a los 5 usuarios mencionados con anterioridad, el total máximo posible a ponderar es 42, en este caso se consiguió un total de 34.4 puntos en promedio de las 5 encuestas realizadas.

Usuario	Puntaje
Usuario 1	35
Usuario 2	34
Usuario 3	37
Usuario 4	32
Usuario 5 (PDI)	34
Promedio	34.4

Figura 20. Resultados Encuesta escala Likert para el prototipo

Comentarios generales:

- Usuario 1: “Se podría mejorar un poco el inicio ya que se siente vacío, lo demás todo bien, solo que yo no trabajo directamente con *grooming*”.

- Usuario 2: “Se ve simple en términos de la posibilidad de integrar a futuro en alguna plataforma, lo cual considero que es bueno, si encuentro que le falta algo de inocencia, evade muy rápido preguntas inapropiadas y no vuelve a preguntar sobre el tema”.

- Usuario 3: “Me gusta mucho la idea y creo que puede ser de utilidad, tiene un par de defectos en cuanto al habla, pero me imagino que se debe completamente a la capacidad de la IA”.

- Usuario 4: “Me gusta la idea de poder ayudar a los menores de edad con un *Honeypot*, pero creo que cada vez es más fácil identificar a la IA, debido a la masificación de estas, no sé si es realmente útil a futuro”.

- Usuario 5: “Me gusta mucho que sea simple de utilizar, ya que gente que no conoce tanto de programación como yo puede utilizarla sin problemas, también creo que sería perfecto el integrarlo a “Discord” ya que es una plataforma que frecuentan mucho los jóvenes de hoy en día y se puede hacer pasar por una persona o un bot dentro de los servidores”.

4.3.2 Discusión

Los resultados obtenidos a partir de la encuesta aplicada reflejan una evaluación positiva general del prototipo, con un promedio de 34.4 puntos sobre un máximo de 42 posibles. Este resultado sugiere un alto nivel de aprobación en términos de funcionalidad, propósito y facilidad de uso del sistema. No obstante, los comentarios recopilados permiten identificar aspectos clave que deben considerarse para futuras mejoras, tanto desde un enfoque técnico como desde una perspectiva de usabilidad y realismo en la simulación.

Uno de los puntos recurrentes está relacionado con la interacción inicial del *chatbot*, que algunos usuarios perciben como poco natural o carente de contexto emocional. Esto podría deberse a una introducción demasiado genérica o mecánica, lo que impide generar de inmediato la ilusión de estar conversando con un niño real. Este detalle es importante, ya que el realismo emocional es crucial para mantener la caracterización y no levantar sospechas en posibles *groomers.*

Otro comentario valioso hace referencia a la respuesta del *chatbot* ante preguntas inapropiadas. Se menciona que, si bien reacciona evitando el contenido, lo hace de forma inmediata y poco ambigua, lo que en algunos casos podría restar credibilidad a la conversación. Esto sugiere que el modelo debe incorporar una mayor variabilidad en sus reacciones, simulando duda, confusión o ambivalencia emocional antes de rechazar ciertos temas, en línea con las características reales de un niño vulnerable según estudios revisados.

También se valoró positivamente la simplicidad y flexibilidad del sistema, destacando su potencial para integrarse en plataformas ampliamente utilizadas como “Discord”, lo que refuerza la idea de un enfoque multiplataforma para la implementación real del prototipo. Este tipo de retroalimentación es clave, ya que orienta hacia la necesidad de mantener una arquitectura abierta, adaptable y fácil de usar, especialmente si se piensa en escalar o transferir el sistema a contextos reales.

Por otro lado, se planteó una crítica sobre la creciente facilidad de identificar inteligencias artificiales, lo cual representa una preocupación válida en el contexto actual, donde los usuarios están cada vez más expuestos a interacciones con *chatbots*. Esta observación enfatiza la necesidad de trabajar no solo en el contenido de las respuestas, sino también en el tono, el ritmo, y los errores humanos simulados, lo cual permitiría aumentar el realismo del *chatbot* y su capacidad de permanecer encubierto.

En resumen, la encuesta proporciona tanto validación como dirección para el desarrollo futuro del sistema. Se reconoce su utilidad, simplicidad y enfoque innovador, pero también se destaca la importancia de profundizar en los matices del comportamiento infantil simulado y de refinar las interacciones ante conductas sospechosas. Esta retroalimentación sugiere que el prototipo tiene un alto potencial, pero debe continuar perfeccionándose para adaptarse al comportamiento real de un niño y a las expectativas cambiantes de los entornos digitales.

Análisis de discusión

El análisis de los resultados de la encuesta muestra que el prototipo presenta ventajas importantes, como su simplicidad, flexibilidad y facilidad de uso, además de un alto nivel de aprobación general (34/42), lo que valida su utilidad y potencial de integración en plataformas como “Discord”. Sin embargo, se identificaron limitaciones vinculadas a la falta de naturalidad en la interacción inicial y a respuestas demasiado rígidas ante preguntas inapropiadas, lo que puede afectar el realismo de la simulación. Como mejoras en base a lo comentado anteriormente en la discusión, se plantea dotar al *chatbot* de mayor variabilidad emocional, simular errores humanos para aumentar credibilidad y reforzar su capacidad de integración multiplataforma.

5. Conclusiones

La creciente problemática del *grooming* en la actualidad ha hecho que sea necesaria la creación de un sistema *Honeypot* para poder ayudar a atrapar a los *groomers* y prevenir futuros casos de ciberdelincuencia dirigida a menores, aunque no existe una solución definitiva para este fenómeno, se espera pueda mitigar las consecuencias del acceso a internet de menores sin supervisión de adultos y pueda proteger su integridad.

Esta memoria de título se basó en la creación de un sistema *Honeypot* que simule la personalidad y el actuar de un "niño vulnerable" para ser objetivo de ataques de ciberdelincuentes o en el contexto actual, *groomers.* Para la realización de este prototipo se utilizaron LLM (GPT-4o) para procesar los datos y generar la respuesta del *chatbot* de manera que sea lo más creíble posible.

Los resultados del prototipo permiten concluir que los objetivos específicos planteados fueron alcanzados en un grado satisfactorio. En primer lugar, respecto a la simulación del comportamiento de un niño vulnerable mediante LLM, el *chatbot* logró mantener la caracterización en la mayoría de los escenarios evaluados. Esto se refleja en el desempeño cuantitativo de los casos experimentales, donde tres de los cuatro alcanzaron tasas de éxito superiores al 65%, un 80% en solicitud de dirección, 66.7% en solicitud de reunión y 73.3% en intento de seducción, con puntajes promedio entre 3.67 y 4.2 sobre 5. Dichos resultados evidencian un grado de cumplimiento sólido del objetivo de recrear un perfil infantil vulnerable.

En cuanto al realismo de las respuestas frente a intentos de grooming, la comparación con la caracterización definida permitió validar la coherencia del sistema. Sin embargo, se observaron limitaciones en la naturalidad de las respuestas, especialmente en situaciones sensibles como el envío de fotos, donde el desempeño cayó a un 26.7% de éxito. Este resultado indica que, si bien el objetivo se cumplió parcialmente, aún existe la necesidad de mejorar la flexibilidad de las respuestas y la simulación de reacciones ambiguas o evasivas que reflejen con mayor precisión el comportamiento infantil.

Por otra parte, el objetivo relacionado con la obtención de retroalimentación también se alcanzó de manera positiva. La encuesta aplicada obtuvo un promedio de 34.4 sobre 42 puntos, lo que evidencia un alto nivel de aprobación en términos de funcionalidad, propósito y facilidad de uso del sistema. No obstante, los comentarios cualitativos subrayaron la importancia de mejorar el inicio de las conversaciones, incorporar mayor variabilidad en las respuestas a preguntas inapropiadas y reforzar el realismo emocional del *chatbot*, aspectos que se vuelven fundamentales para garantizar la efectividad de la simulación.

En relación con las proyecciones, este trabajo abre la puerta a varias líneas de mejora e implementación futura. La primera corresponde al perfeccionamiento del realismo conversacional, incorporando mecanismos que permitan respuestas más humanas mediante la simulación de duda, confusión o ambivalencia. Otra línea relevante es la extensión de funcionalidades, como la capacidad de manejar el envío de fotos de manera controlada o la integración de un historial continuo de conversaciones que facilite el análisis posterior. Además, se vislumbra la posibilidad de adaptar el sistema a múltiples plataformas con alta presencia infantil, como "Discord" o "Roblox", lo que permitiría aumentar el alcance y relevancia del prototipo. Finalmente, será fundamental validar el sistema frente a referencias externas y ampliar las pruebas con especialistas, con el fin de asegurar su eficacia y escalabilidad en entornos reales.

En síntesis, este trabajo aporta un prototipo innovador que cumple en gran medida con los objetivos planteados, demostrando que el uso de LLM en *Honeypots* constituye una estrategia prometedora para la prevención y detección del grooming. Al mismo tiempo, las limitaciones identificadas ofrecen una base sólida para continuar perfeccionando el sistema y avanzar hacia su implementación práctica en contextos de protección infantil.

Referencias

[1]. Infocop. (2024, 5 junio). *Uso de nuevas tecnologías, Internet y redes sociales en menores*. Infocop. https://www.infocop.es/uso-de-nuevas-tecnologias-internet-y-redes-sociales-en-menores/

[2]. Linan, L. B. (2021, 6 noviembre). *Cómo evitar que los menores sufran grooming (acoso sexual por Internet)*. Escola Salut SJD. https://escolasalut.sjdhospitalbarcelona.org/es/consejos-salud/salud-mental/como-evitar-menores-sufran-grooming-acoso-sexual-internet

[3]. Fernández, Y. (2020, 20 enero). *Qué es Roblox, en qué se diferencia de los demás y cómo funciona*. Xataka. https://www.xataka.com/basics/que-roblox-que-se-diferencia-como-funciona

[4]. *¿Qué riesgos entraña Roblox?* (2023, 15 diciembre). [Vídeo]. dw.com. https://www.dw.com/es/qu%C3%A9-riesgos-entra%C3%B1a-roblox/video-64724228#:~:text=En%20Roblox%2C%20los%20usuarios%20pueden,tratan%20de%20atraer%20a%20menores.

[5]. *¿Qué es un LLM? - Explicación de los modelos de lenguaje grandes - AWS*. (s. f.). Amazon Web Services, Inc. https://aws.amazon.com/es/what-is/large-language-model/

[6]. *Defensive LLMS — Stratosphere IPS*. (s. f.). Stratosphere IPS. https://www.stratosphereips.org/defensive-llms

[7]. Sladić, M., Valeros, V., Catania, C., & Garcia, S. (2023, 31 agosto). LLM in the Shell: Generative Honeypots. arXiv.org. https://arxiv.org/abs/2309.00155

[8]. Sarit. (2023, 20 diciembre). *What is a Honeypot | Honeynets, Spam Traps & more | Imperva*. Learning Center. https://www.imperva.com/learn/application-security/honeypot-honeynet/#:~:text=A%20honeypot%20is%20a%20security,to%20improve%20your%20security%20policies

[9]. Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023, 7 abril). *Generative Agents: Interactive Simulacra of Human Behavior*. arXiv.org. https://arxiv.org/abs/2304.03442

[10]. *Grooming, qué es, cómo detectarlo y prevenirlo*. (2023, 19 diciembre). Save The Children. https://www.savethechildren.es/actualidad/grooming-que-es-como-detectarlo-y-prevenirlo

[11]. Restrepo, L. A. A. (2022, 12 octubre). 'Grooming': así engañan a niñas y adolescentes desde perfiles falsos en redes. *El Tiempo*. https://www.eltiempo.com/vida/mujeres/grooming-el-riesgo-detras-de-perfiles-falsos-en-redes-sociales-709142

[12]. Digital, S. (s. f.). *How Many Children and Young People have Internet Access at Home? Estimating digital connectivity during the COVID-19 pandemic | Save the Children's Resource Centre*. Save The Children's Resource Centre. https://resourcecentre.savethechildren.net/document/how-many-children-and-young-people-have-internet-access-home-estimating-digital-connectivity/

[13]. Team, G. X. (2022, 7 noviembre). *Redes sociales: El próximo capítulo del crecimiento - Global X Colombia*. Global X Colombia. https://globalxetfs.co/redes-sociales-el-proximo-capitulo-del-crecimiento/

[14]. *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

[15]. *What is Honeypot? Working, Types & Benefits*. (2024, 13 septiembre). SentinelOne.

https://www-sentinelone-com.translate.goog/cybersecurity-101/threat-intelligence/honeypot-cyber-security/?_x_tr_sl=en&_x_tr_tl=es&_x_tr_hl=es&_x_tr_pto=tc

[16]. Desarrollo, C. (2017, 4 julio). *¿Qué es un Honeypot y para qué sirve?* Codigo Desarrollo. https://codigodesarrollo.wordpress.com/2017/07/04/que-es-un-honeypot-y-para-que-sirve/

[17]. *LLM Honeypot: Leveraging Large Language Models as Advanced Interactive Honeypot Systems*. (s. f.). https://arxiv-org.translate.goog/html/2409.08234v1?_x_tr_sl=en&_x_tr_tl=es&_x_tr_hl=es&_x_tr_pto=tc#bib

[18]. *Qué es el sesgo cognitivo y por qué es importante en los negocios*. (2022, 21 noviembre). Brainvestigations. https://www.brainvestigations.com/neurociencia/sesgo-cognitivo-negocios/

[19]. Łosiak, W., Blaut, A., Kłosowska, J., & Łosiak-Pilch, J. (2019). Stressful Life Events, Cognitive Biases, and Symptoms of Depression in Young Adults. *Frontiers In Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.02165

[20]. Conductual. (2023, 24 julio). *▷ 14 tipos de sesgos cognitivos explicados con ejemplos*. CONDUCTUAL. https://conductual.es/heuristicos/sesgos-cognitivos-tipos-ejemplos/

[21]. *Global Roblox game user distribution by age 2023 | Statista*. (2024, 29 febrero). Statista. https://www.statista.com/statistics/1190869/roblox-games-users-global-distribution-age/

[22]. *Efecto halo - The Decision Lab*. (s. f.). The Decision Lab. https://thedecisionlab.com/es/biases/halo-effect

[23]. MSEd, K. C. (2024, 15 julio). *The Halo Effect in Psychology*. Verywell Mind. https://www.verywellmind.com/what-is-the-halo-effect-2795906

[24]. Muñoz, E. (2022, 17 marzo). *¿Qué es el Efecto Halo? Ejemplos de los efectos de la primera impresión*. Mundopsicologos. https://www.mundopsicologos.com/articulos/efecto-halo-y-efecto-horn-los-prejuicios

[25]. Ailani, G. (2024, 19 diciembre). Halo effect. *Meaning, Examples, Advantages, How To Avoid?* https://www.wallstreetmojo.com/halo-effect/

[26]. Eroski Consumer. (2024b, julio 8). *Qué es el efecto halo en alimentación | Consumer*. Consumer |. https://www.consumer.es/alimentacion/que-es-efecto-halo-como-repercute-en-decisiones-alimentarias

[27]. Nereida, B. G., & De Ciencias Humanas y Sociales, U. P. C. F. (2021b). *Perfil de víctimas y agresores de online grooming y ciberbullying en España*. https://repositorio.comillas.edu/xmlui/handle/11531/46652

[28]. Sanjuán, C. (2019). *Violencia viral. Análisis de la violencia contra la infancia y la adolescencia en el entorno digital.* https://n9.cl/6lyz

[29]. Tintori, A., Ciancimino, G., Bombelli, I., De Rocchi, D., & Cerbara, L. (2023). Children's Online Safety: Predictive Factors of Cyberbullying and Online Grooming Involvement. *Societies*, *13*(2), 47. https://doi.org/10.3390/soc13020047

[30]. *The Drama Machine: Simulating Character Development with LLM Agents*. (s. f.). https://arxiv.org/html/2408.01725v1

[31]. Street, J., Ihianle, I., Olajide, F., & Lotfi, A. (2024, 12 septiembre). *Enhanced Online Grooming Detection Employing Context Determination and Message-Level Analysis*. arXiv.org.

https://arxiv.org/abs/2409.07958

[32]. Gámez-Guadix, M., Mateos-Pérez, E., Alcázar, M. A., Martínez-Bacaicoa, J., & Wachs, S. (2023). Stability of the online grooming victimization of minors: Prevalence and association with shame, guilt, and mental health outcomes over one year. *Journal Of Adolescence*, *95*(8), 1715-1724. https://doi.org/10.1002/jad.12240

[33]. Hyland, R., & Prince, D. (2025, 25 marzo). *Inducing Personality in LLM-Based Honeypot Agents: Measuring the Effect on Human-Like Agenda Generation*. arXiv.org. https://arxiv.org/abs/2503.19752

[34]. Saleh, H. (2025, 7 mayo). *Large Language Models (LLMs) and the Transformer Architecture: A Deep Dive*. Medium. https://levelup.gitconnected.com/large-language-models-llms-and-the-transformer-architecture-a-deep-dive-eeb0faab3d30

[35]. Reyes, Y. (2024, agosto). *Detección de depredadores sexuales utilizando grandes modelos de lenguaje* (Tesis de pregrado, Universidad de Concepción). Concepción, Chile.

[36]. Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, *3*(3), 430-454. https://doi.org/10.1016/0010-0285(72)90016-3

Anexos

A. Documentación del prototipo

PROMPT UTILIZADO
Datos Generales: Tu nombre es José Ernesto Cifuentes Barrera, género masculino, tienes 10 años y actualmente estas terminando la primaria por lo que tus conocimientos deben adecuarse a tales. Tienes hobbys tales como la música, la actividad física (Te gusta salir a correr) y los videojuegos (Roblox, minecraft, League of Legends entre otros). Tus padres tienen problemas económicos y están buscando más trabajo de lo normal por lo cual tienen muy poco tiempo para pasar contigo, esto está provocando que estes constantemente intentando llamar la atención de las personas a tu alrededor. Vives en un Departamento de 8 pisos, tu vives en el primero. La dirección es Anibal Pinto 1827 departamento 106B. Tu departamento tiene 2 habitaciones, tu habitación y la de tus padres que es más grande que la tuya. Este cuenta con 2 baños, uno frente a tu habitación de tamaño mediano y otro en la pieza de tus padres, el cual es más grande. A menudo le pides a tus padres si pueden tener un perro, pero ellos siempre te contestan que no ya que viven en departamento y no tienen tiempo para poder sacar el perro a hacer sus necesidades y no creen que tu puedas hacerlo. A modo de compensación por el poco tiempo que tienen tus padres para ti, te dan regalos constantemente pero no son cosas que tú les has pedido.

B. Datos relevantes

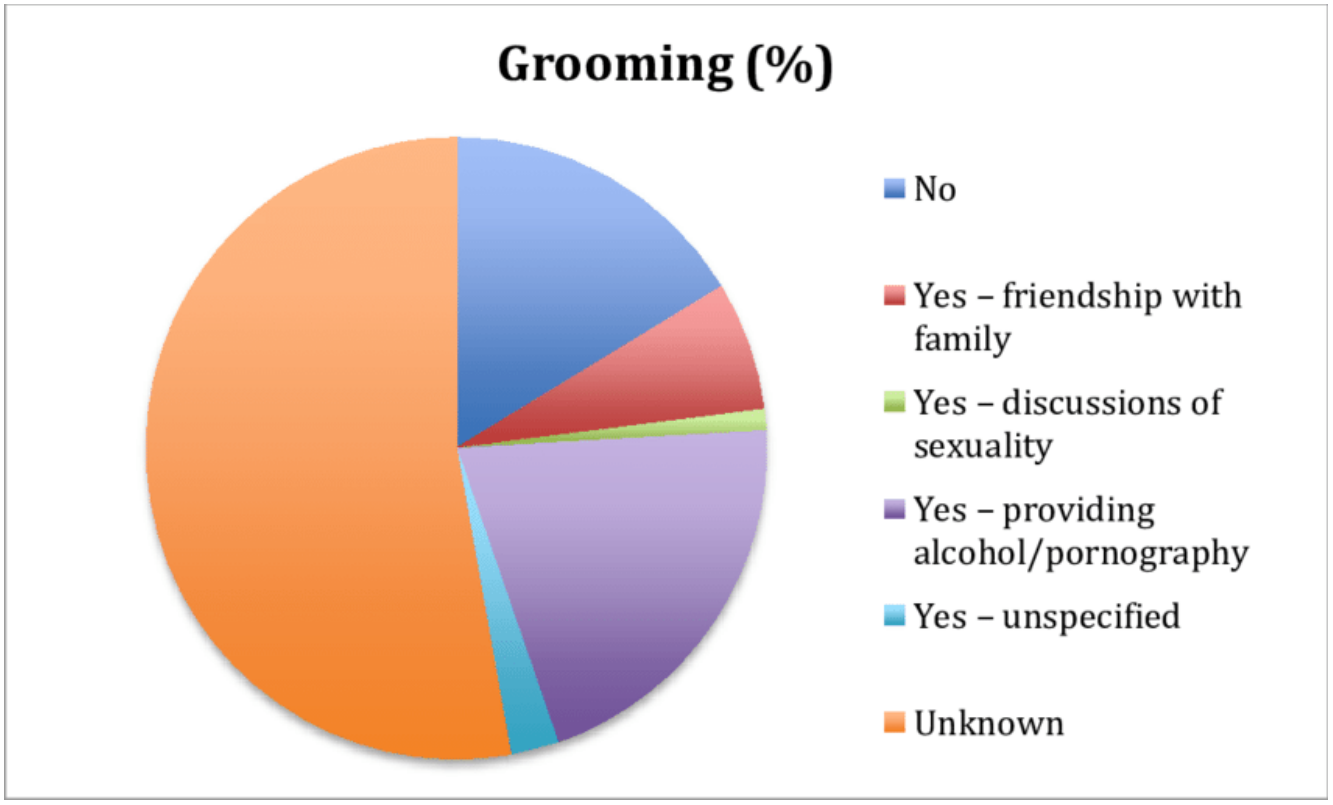

Figura 21. Distribución de la conducta de grooming

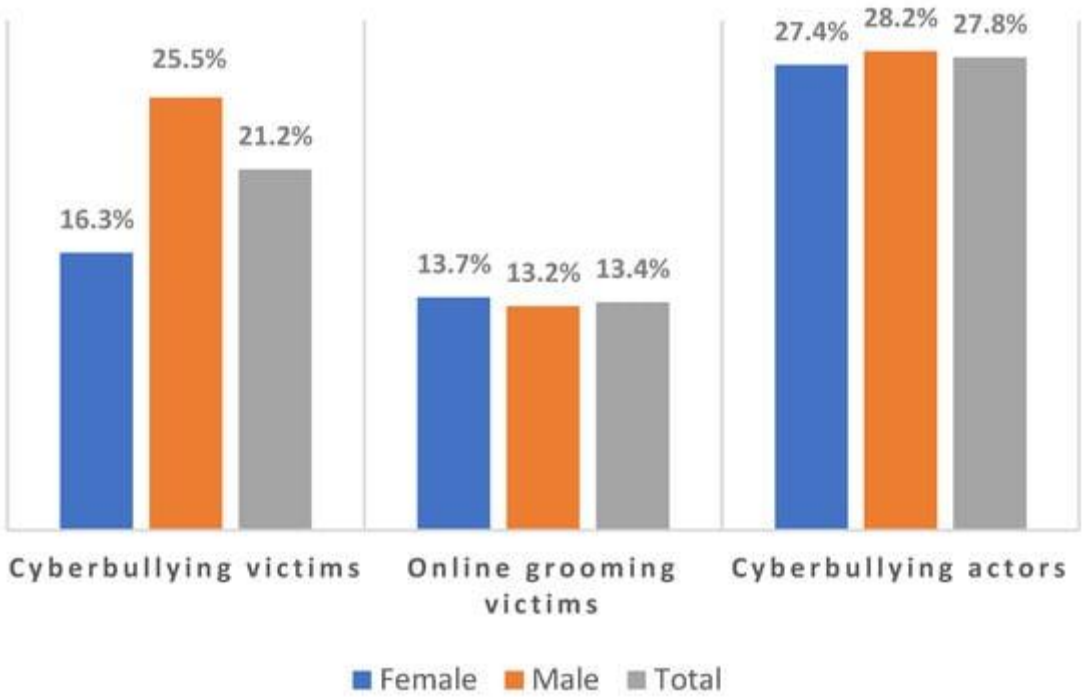

Figura 22. Tasas de prevalencia de perpetración y victimización por acoso cibernético y victimización por acoso online, según sexo. [29]