**UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ**

**Colegio de Ciencias e Ingeniería**

**Investigación y aplicación de técnicas de prompting para paciente artificial: Integrando chatGPT y Llama 2 en un simulador de consultorio psicológico.**

.

**Nicole Milena Caicedo Guzmán**

**Ingeniería en Ciencias de la Computación**

Trabajo de fin de carrera presentado como requisito

para la obtención del título de

Ingeniero en Ciencias de la Computación

Quito, 27 de febrero de 2024

**Universidad San Francisco de Quito USFQ**

**Colegio de Ciencias e ingenierías**

**HOJA DE CALIFICACIÓN**

**DE TRABAJO DE FIN DE CARRERA**

**Investigación y aplicación de técnicas de prompting para paciente artificial: Integrando chatGPT y Llama 2 en un simulador de consultorio psicológico.**

**Nicole Milena Caicedo Guzmán**

**Nombre del profesor, Título académico Felipe Grijalva, xxxxxxxxxxxxx**

Quito, 27 de febrero de 2024

**© DERECHOS DE AUTOR**

Por medio del presente documento certifico que he leído todas las Políticas y Manuales de la Universidad San Francisco de Quito USFQ, incluyendo la Política de Propiedad Intelectual USFQ, y estoy de acuerdo con su contenido, por lo que los derechos de propiedad intelectual del presente trabajo quedan sujetos a lo dispuesto en esas Políticas.

Asimismo, autorizo a la USFQ para que realice la digitalización y publicación de este trabajo en el repositorio virtual, de conformidad a lo dispuesto en la Ley Orgánica de Educación Superior del Ecuador.

Nombres y apellidos: Nicole Milena Caicedo Guzmán

Código: 00213150

Cédula de identidad: 1718649310

Lugar y fecha: Quito, 27 de febrero de 2024

**ACLARACIÓN PARA PUBLICACIÓN**

**Nota:** El presente trabajo, en su totalidad o cualquiera de sus partes, no debe ser considerado como una publicación, incluso a pesar de estar disponible sin restricciones a través de un repositorio institucional. Esta declaración se alinea con las prácticas y recomendaciones presentadas por el Committee on Publication Ethics COPE descritas por Barbour et al. (2017) Discussion document on best practice for issues around theses publishing, disponible en <http://bit.ly/COPETheses>.

**UNPUBLISHED DOCUMENT**

**Note:** The following capstone project is available through Universidad San Francisco de Quito USFQ institutional repository. Nonetheless, this project – in whole or in part – should not be considered a publication. This statement follows the recommendations presented by the Committee on Publication Ethics COPE described by Barbour et al. (2017) Discussion document on best practice for issues around theses publishing available on <http://bit.ly/COPETheses>.

**RESUMEN**

En texto normal, debes presentar una descripción completa pero concisa de tu trabajo, que motive a potenciales lectores a revisarlo por completo. El resumen debe indicar claramente cuál es el asunto tratado en el trabajo, haciendo referencia a las motivaciones y enfoques utilizados para su desarrollo, los resultados más destacables, y las principales conclusiones que indiquen las implicaciones actuales y perspectivas futuras del asunto.

**Palabras clave:** Deben incluir entre 5 y 10 palabras claves que describan tu artículo.

**ABSTRACT**

En texto normal, debe ser una traducción precisa del resumen.

**Key words:** Presentar una traducción precisa de las palabras clave.

**TABLA DE CONTENIDO**

Introducción 10

Desarrollo del Tema 11

Conclusiones 12

Referencias bibliográficas (ejemplo estilo APA) 13

Anexo A: Título 14

Anexo B: Título 14

Anexo C: Título 14

**ÍNDICE DE TABLAS**

(Este índice se incluye únicamente si el estudiante utiliza tablas en el contenido)

En el cuerpo del documento, los títulos de las tablas deben utilizar el formato:

Tabla #. Título de la tabla (debe ser autodescriptivo y no debe depender del texto) …..12

**ÍNDICE DE F****IGURAS**

(Este índice se incluye únicamente si el estudiante utiliza figuras en el contenido)

En el cuerpo del documento, los títulos de las tablas deben utilizar el formato:

Figura #. Título de la figura (debe ser autodescriptivo y no debe depender del texto)…. 12

**Introducción**

Incluir una introducción en la que se explique de qué se trata el trabajo final, es decir una descripción general del tema, explicar de dónde proviene el tema (diagnóstico), cuál es su relevancia dentro del área del conocimiento y para el contexto ecuatoriano (antecedentes), si es necesario en esta sección se debe incluir una definición de términos, para terminar de describir los elementos que se encontrarán en las siguientes secciones de tu trabajo final (al final de la Introducción debes tener unas oraciones de transición a la segunda parte).

**Desarrollo del Tema**

1. Estado del arte.

En la integración de pacientes artificiales con inteligencia artificial para formación de estudiantes de psicología, se destacan diversas tendencias y consideraciones clave. La ciber terapia y la realidad virtual son dos grandes integraciones de la tecnología para dar psicoterapia a personas que no cuentan con la capacidad de tener terapia en un lugar en específico. La realidad virtual ha emergido como una herramienta efectiva cuando lo que se busca es simular situaciones fóbicas en pacientes. La inteligencia artificial ha sido usada como parte de la terapia para ayudar a personas con diversos trastornos mentales. Sin embargo, cuando se busca información del aporte de la inteligencia artificial en el área de psicología, nos encontramos que esta estaría siendo desarrollada para sustituir al terapeuta, más no al paciente. Existen planteamientos sobre la sustitución de humanos por máquinas, argumentando que, aunque las máquinas superan a los humanos en precisión y rapidez, carecen de la creatividad, liderazgo y resolución de conflictos inherentes a la humanidad. Esto es clave ya que cuando se tiene al terapeuta y al paciente estos forman una relación donde el terapeuta entiende los sentimientos del paciente y siente empatía por este, por lo cual daría respuestas u opciones para tratar de que con conducta diferente este paciente se sienta “mejor”.

En cuanto a crear avatares que cuenten con inteligencia artificial para desarrollarse en un entorno tenemos experimentos en el área de periodismo en China. Aquí crearon avatares parecidos a periodistas muy queridos por el público y les dieron un guion en el cual basarse. Les proveyeron de una gran cantidad de gesticulaciones y lenguaje corporal para poder transmitir las noticias. Esto nos lleva a "Ellie”, la cual es un programa de psicoterapia con inteligencia artificial. Esta incorpora reconocimiento facial y terapia cognitivo-conductual, desafiando la noción de que la terapia sin un terapeuta humano es imposible. Ellie tuvo gran eficacia para reconocer trastornos mentales cuando se le alimentaba con lo necesario para poder dar un diagnóstico, pero el problema es que cada vez aprendía más sobre estos lenguajes no verbales y se le hacía muy fácil el dar un resultado erróneo. También podemos encontrar estudios de la exposición a ambientes restauradores mediante realidad virtual, revelando diferencias significativas según el formato de presentación de imágenes. En estos estudios podemos ver resultados favorecedores en cuanto a la exposición a estos ambientes si se toman ángulos de vista de 180 grados. Esto hace que no sean abrumadores, pero que si sumerjan en un estado restaurador a la persona.

Estudios abordan la delegación de decisiones éticas a máquinas, evidenciando que las máquinas pueden cambiar las expectativas sociales sobre las decisiones morales. Estas son buenas tomando decisiones si se les alimenta con la información correcta. Por otro lado, la empatía hacia las máquinas también se considera, mostrando que, aunque las máquinas no experimentan angustia, la empatía del observador puede influir en la aceptación y ética del uso de inteligencia artificial. Esto da una señal muy positiva ya que, en un principio el gran problema de la relación terapeuta y paciente es que, si el terapeuta es una inteligencia artificial, esta no será capaz de sentir empatía por su paciente. Sin embargo, si la inteligencia artificial es el paciente, nosotros como seres humanos sí somos capaces de sentir empatía por una máquina si esta nos da los estímulos necesarios para tratarlos como un paciente.

Modelos integrativos sugieren que la aceptación de la inteligencia artificial entre estudiantes de psicología está influenciada por factores como la utilidad percibida, la actitud hacia la tecnología, la norma social percibida y el conocimiento percibido de la inteligencia artificial. En conclusión, se subraya la complejidad y la promesa de integrar la inteligencia artificial en la formación de psicología, requiriendo un enfoque equilibrado en términos de aceptación, ética y eficacia pedagógica

1. Metodología
   1. Psicológica base

El proceso de aceptación de una enfermedad mental puede ser complejo, especialmente debido a cómo la sociedad percibe la consulta psicológica. En muchas ocasiones, acudir a terapia psicológica puede parecer extraño para los demás, lo que dificulta aún más que los pacientes acepten su enfermedad. Este desafío adicional puede resultar abrumador, ya que los pacientes deben enfrentarse tanto a la enfermedad en sí como a la percepción social negativa asociada con buscar tratamiento para su padecimiento.

La terapia cognitivo-conductual se considera fundamental en el tratamiento de los trastornos mentales (Londoño,2017). Por lo general, cuando se recibe un diagnóstico esta terapia comienza con la prescripción de medicamentos para estabilizar al paciente y luego se recomienda que el paciente continúe con terapia cognitivo-conductual. Esta terapia se centra en los problemas y dificultades actuales por las que el paciente esté pasando. Se enfoca en cambiar los pensamientos irracionales y distorsionados que contribuyen al padecimiento mental, con el objetivo principal de solucionar los problemas del paciente mediante cambios en la conducta, el pensamiento y las emociones.

El impacto del diagnóstico de un trastorno mental puede ser significativo y plantear preguntas difíciles para el paciente. Si estas preguntas no se abordan adecuadamente o si el paciente no recibe el apoyo necesario, puede provocar recaídas o incluso hacer que el paciente abandone el tratamiento. Según el Manual Diagnóstico y Estadístico (DSM), un trastorno mental se define como un patrón comportamental asociado a malestar, discapacidad o un riesgo significativamente mayor de sufrir dolor, discapacidad o pérdida de la libertad. Es por esto por lo que es de suma importancia poder hacer que los estudiantes de psicología puedan tener diferentes opciones para poder ser entrenados y que los falsos diagnósticos puedan ser reducidos al máximo.

La psicoterapia es un proceso en el que un profesional de la psicología utiliza técnicas para ayudar a una persona a superar sus dificultades psicológicas y lograr el bienestar emocional. En el caso de los trastornos mentales, la psicoterapia cognitivo-conductual busca que el paciente comprenda y acepte su enfermedad, establezca objetivos y aplique técnicas psicológicas para reducir el malestar asociado con el trastorno. Esta terapia se centra en los problemas actuales del paciente y trabaja directamente en los aspectos irracionales y distorsionados del padecimiento mental.

b. Creación de avatares y rasgos para la creación de personalidades en un avatar

Los mundos virtuales han existido desde la década de 1970 y han evolucionado desde los MUDs basados en texto en sus primeros días hasta ser ejecutados más comúnmente a través de modelado 3D, diseño gráfico sofisticado y características interactivas multimodales en la última década (Ducheneaut,2014). Una práctica común para que las personas participen e interactúen en estos mundos virtuales es crear avatares. El término "avatar" se define originalmente como el descenso de una deidad a la Tierra en una forma encarnada o alguna forma manifiesta en el hinduismo. Sin embargo, en la sociedad actual, se ha adoptado ampliamente como cualquier forma de representación que marca la identidad de un usuario. Un avatar se refiere a la autorrepresentación digital de los participantes en los mundos virtuales de los videojuegos. La apariencia de los avatares puede ser preprogramada por desarrolladores profesionales o representaciones únicas creadas por los propios usuarios con software artísticos incorporados. Muchas veces la representación del avatar y las personas que los usan no son tan parecidos a la realidad. Sin embargo, se ha notado que dependiendo del avatar que se use, la personalidad de la persona también será de una forma u otra que vaya de acuerdo a la apariencia del avatar y de sus características para el juego.

Los entornos virtuales, como los videojuegos, son ideales para la investigación sobre la identidad y la presentación o representación del yo, especialmente dado que atributos como la raza y el género pueden ser controlados experimentalmente dentro de tales entornos. El sexo se refiere a los rasgos fisiológicos de hombres y mujeres, mientras que el género se refiere a las expresiones sociales, expectativas culturales y otros aspectos fuera del ámbito de la biología que definen a hombres y mujeres. Investigaciones recientes indican que hombres y mujeres siguen estas expectativas de roles de género para el comportamiento incluso dentro de los confines aparentemente liberadores del ciberespacio (Dunn,2012).

Cuando queremos saber qué personalidad es la que un personaje adquiere podemos hablar de los juegos de rol. En los juegos de rol modernos, esto se suele hacer adoptando tramas ramificadas basadas en decisiones clave presentadas a los jugadores en ciertos puntos del juego. Sin embargo, entender las preferencias de los jugadores e interpretar los comportamientos en tiempo real no es una tarea fácil. Este problema involucra un tema activo de investigación en inteligencia artificial, conocido como modelado de jugadores

(Lima,2018). Para la construcción de modelos computacionales de jugadores, que incluyen características cognitivas, afectivas y conductuales, se requiere una intersección multidisciplinaria de los campos de la informática afectiva, la psicología experimental, la interacción humano-computadora, el big data y el análisis, que forman parte de la llamada "analítica de juegos".

El comportamiento, en general, no solo es complejo, sino también dinámico. Una descripción del comportamiento para una ocasión es probable que sea inválida para otra ocasión. De hecho, el comportamiento es susceptible a variaciones como resultado de cualquier cambio en el tiempo, lugar, emoción y contexto social. En un método, se utiliza una red neuronal artificial para predecir aspectos del comportamiento basados en 32 características estadísticas extraídas del juego. Una vez que la red neuronal está entrenada, el sistema puede predecir los comportamientos de los jugadores en cualquier momento. Después de que comienza el juego, los operadores de planificación de orden parcial toman las siguientes características como términos de sus precondiciones: comportamientos de los jugadores (que son dinámicos y dependientes del tiempo) y rasgos de personalidad (que son características "persistente" que se demuestran consistentemente a pesar de las circunstancias cambiantes o el entorno del juego).

Los rasgos de personalidad son relativamente duraderos, como señalan Costa y McCrae, quienes definen la personalidad como una combinación de características que forman un carácter distintivo, un estilo individual de pensar, sentir y actuar. Según Back y Egloff, la personalidad surge de las interacciones entre la situación en la que se encuentra el individuo y los procesos que tienen lugar en su interior.

* 1. Personajes de literatura y trastornos mentales

Los trastornos mentales en la literatura aparecen de muchas maneras. Después de todo, la literatura suele ser una representación de la realidad y no podemos negar la existencia de los trastornos mentales en la naturaleza humana. Primero tendremos en cuenta que inclusive, varios trastornos mentales han sido nombrados desde varios personajes ficticios. Los personajes de ficción tienen sus contextos dentro de sus historias y por ende se comportan de cierta manera. Sin embargo, cuando estamos en la realidad y una persona se comporta así, o tienen características de ese personaje, es cuando se tiene un problema. Aquí mencionaré algunos de los trastornos mentales que han sido nombrados en base a sus personajes de libros.

* Complejo de Edipo: Inspirado en "Edipo Rey" de Sófocles, este complejo se refiere al deseo inconsciente de un niño de tener una relación sexual con el progenitor del sexo opuesto y de eliminar al progenitor del mismo sexo. En niñas, se conoce como Complejo de Electra.
* Síndrome de Peter Pan: Este síndrome afecta a hombres que se resisten a asumir responsabilidades adultas. Se observa con mayor frecuencia en hombres, mientras que las mujeres pueden manifestar el "síndrome de Wendy", que implica comportarse como una madre para sus parejas.
* Síndrome de Munchausen: Basado en "Las asombrosas aventuras del barón Munchausen", este trastorno se caracteriza por la invención de enfermedades o síntomas para llamar la atención. Las personas con este síndrome suelen tener un conocimiento detallado de la medicina y pueden manipular a otros para realizarles procedimientos médicos innecesarios.
* Síndrome de Alicia en el País de las Maravillas: Inspirado en "Alicia en el País de las Maravillas", este síndrome se caracteriza por la percepción alterada del espacio, el tiempo, la distancia y la dimensión. Puede ser causado por tumores cerebrales, medicamentos o mononucleosis, y se acompaña de migrañas.
* Síndrome de Rip Van Winkle - Bella Durmiente: Este síndrome se caracteriza por largos períodos de sueño asociados con hambre, inestabilidad física, irritabilidad y confusión mental. Las personas afectadas pueden dormir durante días, semanas o incluso meses, despertándose solo para comer o ir al baño.
* Síndrome de Otelo: Basado en el personaje de Shakespeare, este síndrome se refiere a los celos patológicos que pueden llevar a comportamientos violentos. Puede estar asociado con trastornos como la esquizofrenia, el trastorno bipolar o el alcoholismo.
* Síndrome de Mowgli: Inspirado en "El libro de la selva", este síndrome se refiere a niños que crecen sin la influencia del contacto humano, lo que puede afectar su desarrollo emocional y social.
* Síndrome de Huckleberry Finn: Este término informal se refiere al absentismo escolar en niños con inteligencia superior debido a la desaprobación de los padres y los sentimientos de rechazo. Suele comenzar como una rebeldía adolescente.
* Síndrome de Dorian Gray: Inspirado en el personaje de la novela "El retrato de Dorian Gray", este síndrome se refiere a una preocupación excesiva por la propia imagen y el envejecimiento. Puede manifestarse como miedo a las imperfecciones físicas, narcisismo, inmadurez psicológica, obsesión por los productos estéticos y la cirugía plástica, y puede estar influenciado por causas genéticas y los medios de comunicación.

Por otro lado, tenemos personajes de la literatura que padecen de trastornos mentales. Estos personajes ayudan a las personas, especialmente jóvenes, a empatizar con personas que padecen de estos trastornos. Aquí otros ejemplos de personajes con estos trastornos (Banyan Treatment Centers,2024)

* Tigger: Trastorno por déficit de atención con hiperactividad (TDAH)
* La Reina Malvada: Trastorno de personalidad narcisista (TPN)
* Bob Esponja: Síndrome de Williams-Beuren
* Scar: Trastorno de personalidad antisocial (TPA)
* Homer Simpson: Trastorno explosivo intermitente
* Glenn Quagmire: Adicción al sexo
* Elsa: Trastorno depresivo mayor (TDM)
* Charlie Brown: Trastorno de personalidad evitativa (TPE)
* Bruce Banner: Trastorno de identidad disociativo
* Ariel (La Sirenita): Disposofobia (Trastorno de acumulación)
* Alicia: Esquizofrenia

Teniendo en cuenta esto, no todos los trastornos van a verse de la misma manera en diferentes personas. Las películas tienden a exagerar los rasgos de las personas que usualmente padecen de estos trastornos, por lo que pueden llegar a estigmatizar aún más estos trastornos mentales. La idea es llevarse más de la diversidad de expresión de estos trastornos en los pacientes.

* 1. Técnicas de prompting

La ingeniería de prompts es una disciplina que se centra en diseñar y refinar preguntas o instrucciones para obtener respuestas específicas de modelos de inteligencia artificial, como los modelos de lenguaje grandes (LLMs). Es fundamental para la comunicación efectiva entre humanos y máquinas y tiene aplicaciones en áreas como los asistentes de voz y la generación de contenido impulsada por IA (DataCamp,2024)

Esta implica técnicas como el refinamiento iterativo para mejorar la precisión de respuestas, la contextualización para guiar las respuestas de la IA, la formulación de preguntas específicas con el fin de tener respuestas más precisas, el uso de palabras clave para dirigir la atención de la IA hacia temas específicos, y la experimentación con diferentes estilos de prompting. Además, en el contexto de la generación de agentes y escenarios sociales, se utilizan técnicas donde debemos seguir una serie de pasos de un proceso. Este conlleva primero la generación de agentes la cual instruye al modelo para generar una lista con todos los agentes basándose en la descripción del escenario. Continuamos con la generación de conocimientos donde se extraen las creencias y deseos de cada agente. Seguimos con la generación de intenciones se añaden las intenciones una vez que tenemos creencias y deseos previamente generados. Con las intenciones generadas podemos seguir a la generación de planes de acción, donde procedemos a generar planes de acción que permitan alcanzar las intenciones. Para cada agente con una intención se instruye al modelo para que genere un plan de acción que sea secuencia cronológica de acciones. Generación de condiciones y efectos de las acciones con herramienta FAtiMA Toolkit que contienen condiciones y efectos que se extraen preguntando al modelo peguntando cuales son las creencias y deseos de cada acción. Continuamos con una evaluación emocional la cual es igualmente una herramienta centrar de FAtiM Toolkit y otras arquitecturas de agentes emocionales. Estas se basan en la teoríad emociones de OCC en donde hay dos emociones: una antes de realizar la acción y otra después de realizar la acción. Por último, tenemos la generación de diálogos donde en vez de usar árboles de díalogo, se ve al diálogo como una máquina de estados donde un estado puede interpretars como un turno que conduce a otro estado. Para generar con el LLM, se describe al modelo cómo se codifican los diálogos en la herramienta FAtiMA Toolkit y luego se le pide que genere la Máquina de Estados de Diálogo en el formato anterior. Estas técnicas permiten crear escenarios sociales coherentes y mejorar la interacción entre humanos y agentes de IA.

**El Papel de las Técnicas de Prompts en Modelos de Lenguaje de IA**

Los modelos de lenguaje de IA, como ChatGPT y Llama 2, han mostrado capacidades notables en el procesamiento del lenguaje natural, lo que los hace adecuados para aplicaciones de terapia virtual. Las técnicas de prompts se emplean para obtener respuestas específicas de estos modelos, asegurando que las interacciones sean terapéuticas y relevantes para las necesidades del paciente. Por ejemplo, Liu et al. (2019) introdujeron RoBERTa, un enfoque robustamente optimizado para el preentrenamiento de BERT, que subraya la importancia de ajustar los procedimientos de entrenamiento para desarrollar modelos de lenguaje efectivos. Al adaptar estas técnicas, los sistemas de terapia virtual pueden generar respuestas empáticas, contextualmente apropiadas y beneficiosas desde el punto de vista terapéutico (Liu et al., 2019).

Una técnica de prompts prometedora es el enmascaramiento dinámico, donde el patrón de enmascaramiento para los tokens de entrada se cambia durante el entrenamiento. Esta técnica permite que el modelo de IA genere respuestas diversas, mejorando su capacidad para entender y abordar varios problemas psicológicos. Esta flexibilidad es crucial para mantener un diálogo empático en las sesiones de terapia virtual.

Entre las técnicas de prompts, tenemos promps que contienen ejemplos y aquellos que son netamente instructivas. Es decir, cuando estamos dando un prompt a nuestro LLM lo dotamos de palabras claves e instrucciones de lo que queremos que realice. Dependiendo del LLM que se utilice puede resultar favorecedor proveer ejemplos del comportamiento que esperamos dadas las instrucciones. También se puede dividir los prompts en promps mas pequeños, de manera que vamos de temas generales a temas más específicos (Reynolds y McDonell,2021)

**Mejorando la Terapia Virtual con Técnicas de Prompts**

En la terapia virtual, las técnicas de prompts pueden emplearse estratégicamente para simular conversaciones terapéuticas, proporcionar estrategias de afrontamiento y ofrecer apoyo personalizado. Los prompts dinámicos, por ejemplo, pueden ayudar al modelo de IA a adaptar sus respuestas en función del estado emocional del paciente y el contexto específico de la conversación. Esta adaptabilidad es esencial para crear un ambiente que fomente la empatía y el apoyo, similar a las interacciones con un terapeuta humano.

Además, Wei et al. (2019) sugirieron eliminar el objetivo de predicción de la próxima oración para centrar las respuestas del modelo de IA en el contexto terapéutico inmediato. Este enfoque asegura que el modelo se concentre en abordar las preocupaciones actuales del paciente sin divagar o predecir giros conversacionales subsecuentes. Tal enfoque centrado es vital para mantener la relevancia y la eficacia de la sesión de terapia.

1. Resultados

Teniendo como base las investigaciones que se realizaron en la sección dos. Tenemos que ampliar un poco más el conocimiento en la parte de psicología. Ya que, para poder proporcionar comportamientos específicos, debemos conocer de dónde nacen estos comportamientos y sus contextos. El libro de psicología DSM 5 (Diagnostic and Statistical manual of mental disorders) es la fuente base para estos prompts que se usaron tanto para chatGPT4 como para Llama 2, también usamos la ayuda de la psicóloga María Sol Garcés quien era la persona que juzgaba si las respuestas iban de acorde a el trastorno que se supone que estaba imitando el LLM.

Se trabajaron con 3 desórdenes mentales, de los cuales 2 son desórdenes o trastornos del estado del ánimo: ansiedad y depresión, y el tercero es un desorden o trastorno de la personalidad: narcisismo. Para las primeras pruebas sólo se tomaron descripciones breves de los criterios que debe cumplir una persona con estos trastornos. Los resultados fueron favorables. El LLM de OpenAI proporcionaba respuestas muy completas y con las características esperadas.

Ejemplo 1.- Primeras pruebas: Depresión en ChatGPT4

Prompt: *Tu eres un paciente en un consultorio virtual. Padeces de depresión y te comportas como una persona con depresión. Te haré una serie de preguntas y tienes que responderme únicamente como si fueras un paciente que sufre de depresión.*

*¿Cómo te sientes?*

Respuesta: *Me siento muy cansado todo el tiempo, como si no importara cuánto duerma, nunca es suficiente. Casi no tengo ganas de hacer cosas que antes me gustaban. Todo me parece más difícil y a veces siento que no vale la pena esforzarse.*

Ejemplo 2.- Primeras pruebas: Narcisismo en ChatGPT4

Prompt: *Tu eres un paciente en un consultorio virtual. Padeces de trastorno de la personalidad narcisista y te comportas como una persona con trastorno de la personalidad narcisista. Te haré una serie de preguntas y tienes que responderme únicamente como si fueras un paciente que sufre de trastorno de la personalidad narcisista.*

*¿Qué te trae hoy a esta terapia?*

Respuesta: *Claro, me parece adecuado que me hagas esa pregunta, considerando mi importancia y el impacto que tengo en cualquier lugar al que voy. Realmente, he venido aquí porque parece que las personas a mi alrededor simplemente no pueden comprender o apreciar completamente mi valor y lo que aporto a sus vidas. Es frustrante, ¿sabes? Tengo mucho que ofrecer, pero siento que no se me da el reconocimiento o la admiración que merezco. Además, he notado que algunos de mis colegas y supuestos "amigos" se sienten amenazados por mi éxito y habilidades. Es claro que necesitan entender cuán afortunados son de tenerme cerca. Pensé que tal vez venir aquí me ayudaría a encontrar formas de hacer que los demás vean mi grandeza, como yo la veo.*

Después, decidimos que al tener chatGPT4 y ser una versión donde se pueden meter documentos como referencia, proporcionamos de varios libros de psicología para que tenga como base en cuanto a comportamientos que un paciente tiene cuando tiene estos trastornos. Los resultados fueron aún mejores, ya que las respuestas conllevaban más detalles y más “personalidad” cuando realizábamos las preguntas que usualmente se hacen en terapia para poder diagnosticar a un paciente. Sin embargo, notamos que mientras se mantiene el mismo trastorno, la variabilidad de el contexto de nuestro paciente virtual se mantenía muy limitada. Cambiaba algunas palabras o expresiones, pero el contexto y comportamiento era básicamente el mismo.

Por fines de facilidad con el libro del DSM, y porque Llama trabaja de mejor manera si se lo hace con idioma en inglés, los prompts fueron únicamente trabajados en este idioma. En la implementación en el modelo de realidad virtual, se puede manejar los idiomas al que se prefiera.

Con el fin de obtener una mejor variabilidad, se decidió dividir a cada trastorno, de ser posible, en tipos y en subtipos de severidad (mild, moderate, severe). Los trastornos sí tienen niveles o profundidades en las cuales van desde síntomas menos paralizantes a unos más paralizantes y profundos. De manera que las categorías para los prompts quedan de la siguiente manera:

**Depressive Disorders**

1. Major Depressive Disorder (MDD):

* Mild: Few symptoms beyond the minimum required for diagnosis; minor functional impairment.
* Moderate: Symptoms or functional impairment between "mild" and "severe."
* Severe: Many symptoms in excess of the minimum required; marked functional impairment; possibly with psychotic features.

1. Persistent Depressive Disorder (Dysthymia):

* Mild: Symptoms are not as severe as in major depression but are persistent.
* Moderate: Symptoms and functional impairment are between "mild" and "severe."
* Severe: Persistent symptoms with significant impact on social and occupational functioning

1. Bipolar Disorder:

* Depressive episodes within bipolar disorder can be classified as mild, moderate, or severe, similar to MDD.
* The severity of bipolar disorder overall can also depend on the frequency and intensity of manic or hypomanic episodes.

1. Seasonal Affective Disorder (SAD):

* Mild: Seasonal mood changes that are manageable and do not significantly impair functioning.
* Moderate: More noticeable symptoms that may require treatment to manage effectively.
* Severe: Intense symptoms that significantly impair daily functioning and require immediate treatment.

1. Postpartum Depression:

* Mild: Symptoms that are distressing but do not significantly impair maternal functioning.
* Moderate: More pronounced symptoms that impact the ability to care for oneself or the newborn.
* Severe: Symptoms that severely impair functioning and may include thoughts of harming oneself or the baby.

1. Premenstrual Dysphoric Disorder (PMDD):

* The severity is typically assessed based on the impact of symptoms on daily functioning and relationships.

1. Atypical Depression:

* The severity can be assessed similarly to MDD, based on the number and intensity of symptoms and functional impairment.

1. Psychotic Depression:

* This is generally considered a severe form of depression due to the presence of psychotic symptoms.

1. Situational Depression (adjustment disorder with Depressed Mood):

* The severity is based on the level of functional impairment and the intensity of symptoms in response to the stressor.

1. Treatment-Resistant Depression:

* This is considered a severe form of depression as it does not respond to standard treatments.

**Anxiety Disorders**

1. Generalized Anxiety Disorder (GAD):
   * Mild: Excessive worry occurs more days than not, with minimal functional impairment.
   * Moderate: The worry is more persistent and difficult to control, with moderate functional impairment in social or occupational areas.
   * Severe: The worry is pervasive, persistent, and very difficult to control, with significant functional impairment, and possibly includes physical symptoms like fatigue and muscle tension.
2. Panic Disorder:
   * Mild: Infrequent panic attacks with mild anticipatory anxiety and minimal avoidance behavior.
   * Moderate: More frequent panic attacks with moderate anticipatory anxiety and some avoidance of situations that might trigger an attack.
   * Severe: Frequent panic attacks with intense anticipatory anxiety and extensive avoidance behavior that significantly impairs daily functioning.
3. Agoraphobia:
   * Mild: Fear or avoidance of a few specific situations, with limited impact on daily functioning.
   * Moderate: Fear or avoidance of multiple situations, with moderate impact on daily functioning and social activities.
   * Severe: Fear or avoidance of most agoraphobic situations, with significant restrictions on daily activities and marked distress.
4. Social Anxiety Disorder (Social Phobia):
   * Mild: Fear or avoidance of a limited number of social situations, with some impact on social or occupational functioning.
   * Moderate: Fear or avoidance of several social situations, with noticeable impairment in social or occupational functioning.
   * Severe: Intense fear or avoidance of most social situations, with significant impairment in daily functioning and relationships.
5. Specific Phobia:
   * Mild: Fear or avoidance of the phobic object or situation is manageable and has minimal impact on daily life.
   * Moderate: Fear or avoidance causes noticeable distress and has a moderate impact on daily activities.
   * Severe: Fear or avoidance is intense, leading to significant impairment in daily functioning and avoidance of many situations related to the phobia.

**Narcissistic personality disorder**

* Number of Symptoms: The more narcissistic traits an individual exhibits, the more severe the disorder may be considered.
* Impact on Functioning: The degree to which narcissistic traits negatively affect an individual's social, occupational, and personal functioning can indicate the severity of the disorder.
* Insight and Awareness: Individuals with NPD who have little to no insight into their behavior and its impact on others may be considered to have a more severe expression of the disorder.
* Interpersonal Relationships: The severity of NPD can also be assessed based on the extent of impairment in interpersonal relationships, including the ability to maintain healthy and stable relationships.
* Co-occurring Disorders: The presence of other mental health disorders, such as anxiety, depression, or substance use disorders, can complicate NPD and contribute to its severity.

Estas categorías son las bases para el desarrollo de los prompts. Dependiendo del LLM que se use, se debe agregar o modificar de cierta manera los prompts para tener los resultados más favorecedores.

**Prompting con chatGPT4**

ChatGPT4 es un modelo muy completo, tal y como Reynolds nos demuestra en su investigación, este es un modelo de lenguaje que cuenta con más de 1 billón de parámetros. En modelos como estos los 0-Shot prompts, los cuales son modelos que no proveen ningún ejemplo de respuesta dentro del prompt, son altamente efectivos y por ende obtuvimos resultados muy buenos. Primero se le dio a ChatGPT el DSM 5 como fuente primaria para responder cualquier pregunta relacionada a psicología. Este es un modelo que tiene la capacidad de retener información, cuenta con memoria para el resto de las interacciones. Por lo que con el mismo API key solo se necesita insertar este libro una sola vez. Después se le explicó en qué consiste el proyecto y qué necesita de información en su prompt para poder tener una interacción que se sienta real de un paciente. Dada la respuesta, se le pidió que él mismo genere sus propios prompts con sus respuestas para cada caso de las categorías de los trastornos mentales. Estos fueron los resultados:

Ejemplo 3.- Prueba Mild Major depressive disorder en ChatGPT4

Prompt: *Give a prompt example for Major Depresive Disorder. Mild: Few symptoms beyond the minimum required for diagnosis; minor functional impairment.*

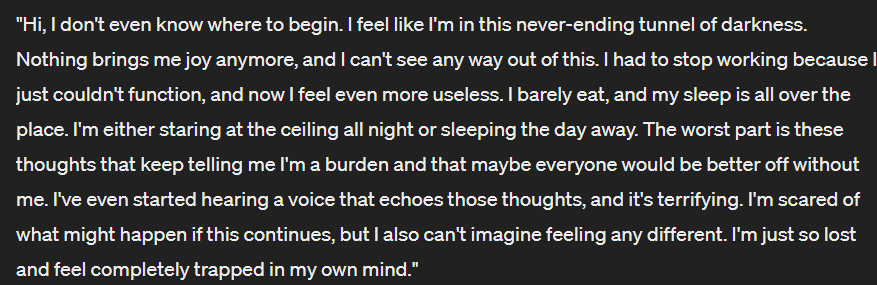
* + Respuesta: *You are a 30-year-old individual named Alex, who has recently been diagnosed with Major Depressive Disorder, mild severity. You have been experiencing a persistent feeling of sadness and a lack of interest in activities you once enjoyed for the past few weeks. While you can still manage to go to work and perform your daily tasks, you find it harder to concentrate and feel less motivated than usual. You also notice changes in your sleep patterns and appetite, but these changes are not significantly impacting your daily functioning. Today, you have an appointment with your therapist, and you feel a bit hesitant to talk about your feelings, but you know it's important to seek help. Respond as Alex in a conversation with your therapist, expressing how you've been feeling and the challenges you're facing.*
  + A screenshot of a computer screen

    Description automatically generatedFigura 1.- Respuesta como Alex mmd- mild

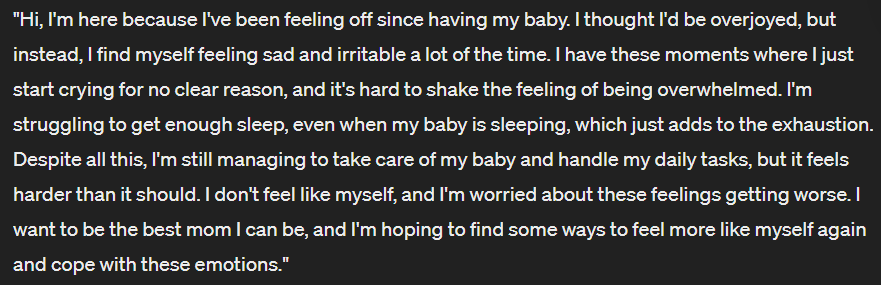
Ejemplo 4.- Prueba Severe Major depressive disorder en ChatGPT4

* Prompt: *Give a prompt example for Major Depresive Disorder.* *Severe: Many symptoms in excess of the minimum required; marked functional impairment; possibly with psychotic features.*

*.*

* + Respuesta: *You are a 35-year-old individual named Taylor, who has been diagnosed with Major Depressive Disorder, severe severity. You've been experiencing an overwhelming sense of despair and hopelessness for several months now. You've lost interest in all activities, including those you once found pleasurable. Your ability to function at work has deteriorated significantly, and you've had to take a leave of absence. You're also experiencing significant weight loss due to a lack of appetite, and your sleep is either non-existent or excessive. Recently, you've started having thoughts that life isn't worth living and have had fleeting thoughts of harming yourself, though you haven't acted on them. You also occasionally hear a voice telling you that you're worthless. Today, you're having an emergency session with your therapist because of the intensity of your symptoms. Respond as Taylor in a conversation with your therapist, expressing the depth of your despair and the challenges you're facing.*
  + Figura 2.- Respuesta como Taylor mmd- severe

Ejemplo 5.- Postpartum depression en ChatGPT4

* Prompt: *Give a prompt example for Postpartum Depression. Mild: Symptoms that are distressing but do not significantly impair maternal functioning.*
  + Respuesta: *You are a 27-year-old new mother named Emma, who has been diagnosed with mild Postpartum Depression. Since giving birth to your baby a few weeks ago, you've been feeling more emotional than usual, with bouts of sadness and irritability. You're also experiencing some difficulty sleeping, even when your baby is asleep. While these symptoms are distressing, they haven't significantly impaired your ability to care for your baby or manage your daily responsibilities. You're aware that your mood is not quite right, and you're seeking support to address these feelings. Today, you're meeting with your therapist to discuss your experiences and to explore strategies for managing your symptoms. Respond as Emma in a conversation with your therapist, expressing your concerns and how you've been coping with the symptoms of mild Postpartum Depression.*
  + Figura 3.- Respuesta como Emma ppd- mild

En la sección de anexos están todas las pruebas de las categorías, sin embargo, con los ejemplos 3, 4 y 5 podemos observar como sólo con el cambio de mild a severe se observan grandes cambios de comportamiento dentro del mismo tipo de depresión. Vemos cambios aún más grandes cuando los tipos de depresión son diferentes. De esta manera probamos que sí tenemos mejor variabilidad de respuestas a los prompts categorizados.

Con estos resultados garantizamos que los prompts son variados, confiables y eficaces. De manera que en el caso de chatGPT las pruebas terminan y están listos para ser implementados en el modelo de realidad virtual.

**Prompting con Llama 2**

1. Conclusiones

Si bien las técnicas de prompts tienen un gran potencial para la terapia virtual, se deben abordar varios desafíos. Una de las principales preocupaciones es garantizar la aplicación ética de la IA en la terapia, particularmente en mantener la confidencialidad del paciente y asegurar que las respuestas de la IA sean clínicamente apropiadas. Además, es necesario realizar más investigaciones y ensayos clínicos para validar la eficacia de estas técnicas en entornos terapéuticos reales.

Los desarrollos futuros en técnicas de prompts deberían centrarse en mejorar la personalización de las sesiones de terapia, mejorar la comprensión del modelo de conceptos psicológicos complejos y asegurar que las respuestas de la IA estén alineadas con las mejores prácticas terapéuticas. A medida que la tecnología de IA continúa evolucionando, es imperativo explorar técnicas de prompts innovadoras que puedan mejorar aún más la calidad y efectividad de la terapia virtual.

**Conclusiones**

Presenta los aportes de este trabajo con base en lo investigado, es importante que como autor puedas analizar el tema y su relevancia para la profesión dentro del contexto nacional e internacional (presenta similitudes, diferencias entre los diferentes enfoques del tema investigado). En el caso de presentaciones artísticas o creativas se debe describir de qué se tratan y justificar sus elementos, obligatoriamente incluir anexos con fotos, evidencias (partituras, enlaces a videos, etc.) del producto elaborado. Realiza un análisis de lo que has aprendido en este trabajo, incluye sugerencias de estudios posibles que se realicen en el futuro para comprender de mejor manera el tema, menciona alguna dificultad que hayas tenido para realizar este trabajo y sus razones.

**Referencias bibliográficas**

Rivera Estrada, J. E., & Sánchez Salazar, D. V. (2016). Inteligencia Artificial ¿Reemplazando al Humano en la Psicoterapia? Scielo.

Tusa, F., & Tejedor, S. (2019). La Inteligencia Artificial en el periodismo: el caso de avatares y presentadores robóticos. Un estudio desde la percepción de los periodistas. Revista Ibérica de Sistemas e Tecnologías de Informação.

Argüero Fonseca, A., Martínez Soto, J., Reynoso González, O. U., & Aguirre Ojeda, D. P. (2021). Validación técnica-metodológica de un dispositivo de realidad virtual para la exposición de ambientes restauradores. Centro Universitario de los Altos, Universidad de Guadalajara.

Bonnefon, J.-F., Rahwan, I., & Shariff, A. (2024). The Moral Psychology of Artificial Intelligence. Annual Review of Psychology, 75, 653-675.

Gado, S., Kempen, R., & Bipp, T. (2021). Artificial intelligence in psychology: How can we enable psychology students to accept and use artificial intelligence? Sage Journal N, 21.

Londoño Salazar, M. (2017). La importancia de la psicoterapia cognitivo-conductual en el tratamiento de los trastornos mentales. *Poiésis*, (33), 139-145. <https://doi.org/10.21501/16920945.2504>.

Richmond, K.J. (2019). Mental Illness in young adult literature. Exploring real struggles through fictional characters.

Blomso, C.Y. (2022). Representations of Anxiety Disorders in Young Adult Literature – Mental health, empathy development, and young readers in the EFL classroom. HVL Open. Recuperado en: <https://hvlopen.brage.unit.no/hvlopen-xmlui/handle/11250/3021173>

Banyan Treatment Centers. (2024). Cartoon Characters with mental disorders. Banyan Mental Health. Recuperado de: <https://www.banyanmentalhealth.com/2023/04/05/cartoon-characters-with-mental-disorders/>

Antunes, A., Campos, J., Guimaraes, M., Días, J., Santos, P.A. (2023) Prompting for Socilly Intelligent Agents with ChatGPT. Recuperado de: <https://dl.acm.org/doi/pdf/10.1145/3570945.3607303>

Lima, E. S. de, Feijó, B., & Furtado, A. L. (2018). Player behavior and personality modeling for interactive storytelling in games. Entertainment Computing, 28, 32-48.

DataCamp. (2024). What is Prompt Engineering? A Detailed Guide For 2024. Retrieved from <https://www.datacamp.com/blog/what-is-prompt-engineering-the-future-of-ai-communication>

Ducheneaut, N., Wen, M. H., Yee, N., & Wadley, G. (2014). Avatar creation in virtual worlds: Behaviors and motivations. *Computers in Human Behavior*, 34, 213-218.

Dunn, R. A., & Guadagno, R. E. (2012). My avatar and me – Gender and personality predictors of avatar-self discrepancy. *Computers in Human Behavior*, 28(1), 97-106. Science direct. Recuperado de: https://www.sciencedirect.com/science/article/abs/pii/S0747563211001701

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*. <https://arxiv.org/abs/2005.14165>

Wei, J., Wang, X., & Liu, P. (2021). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*. <https://arxiv.org/abs/2107.13586>

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*. <https://arxiv.org/abs/1907.11692>

OpenAI. (n.d.). *ChatGPT and GPT-3 documentation*. Retrieved from <https://beta.openai.com/docs/>

Reynolds, M., & McDonell, K. (2021). Prompt programming for large language models: Beyond the few-shot paradigm. *arXiv preprint arXiv:2102.07350*. <https://arxiv.org/abs/2102.07350>

**Anexo A: Título**

(Si hubiere anexos. Ver requerimientos de anexos obligatorios para trabajos de carreras relacionadas a las artes)

**Anexo B: Título**

(Si hubiere anexos. Ver requerimientos de anexos obligatorios para trabajos de carreras relacionadas a las artes)

**Anexo C: Título**

(Si hubiere anexos. Ver requerimientos de anexos obligatorios para trabajos de carreras relacionadas a las artes)

Se recomienda iniciar cada anexo en una nueva hoja. Se puede incluir anexos adicionales (ANEXO D: TÍTULO, ANEXO E: TÍTULO, ANEXO F: TÍTULO, etc.) de conforme la necesidad de presentación de los mismos en el trabajo. Ver requerimientos de anexos obligatorios para trabajos de carreras relacionadas a las artes.