BENEMÉRITA UNIVERSIDAD AUTÓNOMA DE PUEBLA

FACULTAD DE CIENCIAS DE LA COMPUTACIÓN

Metodología e implementación de algoritmos basados en Modelos de Lenguaje para el Acompañamiento Emocional

Tesis para obtener el grado de Ingeniero en Ciencias de la Computación

PRESENTA

Bruno Gil Ramírez

DIRECTORA DE TESIS

M.C. María del Carmen Santiago Díaz
BUAP, México

ASESORA DE TESIS

Dra. Jessica Nayeli López Espejel
Novelis Research Lab, Francia

Junio 2025

Resumen

La presente tesis aborda el desarrollo e implementación de algoritmos basados en modelos de lenguaje para el acompañamiento emocional, con el objetivo de brindar apoyo en salud mental mediante inteligencia artificial (IA).

La investigación surge de la necesidad de herramientas de apoyo en salud mental debido al deterioro experimentado por estudiantes al adaptarse a la educación virtual, evidenciado por cifras crecientes de estrés y ansiedad como las mostradas en de investigaciones como "Depression Anxiety and Stress Scale 21" (Services, 2018) en la que participaron 493 estudiantes, indicando que sufrieron de estrés depresión y ansiedad durante la pandemia. Esto resalta la oportunidad de utilizar la IA para detectar y mitigar condiciones como la depresión y la ansiedad.

El método propuesto emplea el modelo Llama2-7b-chat ajustado con la técnica Low Rank Adaptation (LoRA) para reducir la carga computacional y mejorar la eficiencia. En este trabajo, se introducen y utilizan las bases de datos HRECPW y HEAR como elementos clave para el entrenamiento, describiendo en detalle el proceso de recopilación, muestreo y etiquetado de datos. Este proceso incluye el uso de herramientas como la rueda de Plutchik para garantizar una adecuada clasificación emocional. Además, se generaron respuestas sintéticas utilizando GPT-3.5 para ampliar los datos de entrenamiento.

El modelo resultante "Solo Escúchame" , demostró un desempeño superior en tareas de acompañamiento emocional en comparación con otros modelos, incluyendo GPT-3.5. La evaluación incluyó métricas como la escucha activa y el método socrático, presentando ejemplos de conversaciones generadas. Sin embargo, se identificaron desafíos como el ciclo de respuestas repetitivas y problemas de coherencia en conversaciones extensas.

En conclusión, se destaca la efectividad del modelo en el acompañamiento emocional, reconociendo limitaciones como la necesidad de más datos y recursos computacionales, así como la importancia de abordar los desafíos éticos asociados. Finalmente, las metas a corto y largo plazo incluyen la mejora del conjuntos de datos, la exploración de modelos multimodales y la adopción de nuevas tecnologías.

Agradecimientos

Este trabajo significó mucho para mí; fue un año de mi vida dedicado a esta investigación, durante el cual aprendí mucho sobre mí mismo. Descubrí que puedo gestionar múltiples responsabilidades y desafíos. Sin embargo, estas capacidades no habrían sido suficientes sin el apoyo de personas clave en este camino.

En primer lugar, quiero expresar mi más profundo agradecimiento a la Doctora Jessica Nayeli López Espejel, co-asesora de esta tesis. Su guía iluminó este trabajo de muchas maneras. Más allá de corregir mis errores, siempre me alentó a mejorar y alcanzar límites ambiciosos pero factibles. Incluso puso recursos personales a mi disposición para que la investigación pudiera continuar, algo que valoro profundamente. Me alentó a publicar un artículo científico, me apoyó en los momentos en que me faltaban fuerzas y ánimo, y mostró una paciencia infinita cuando no era su obligación. Su dedicación y compromiso fueron fundamentales para que este trabajo sea lo que es hoy.

También quiero agradecer a la Profesora Ma. del Carmen Santiago Díaz, maestra en ciencias, quien me motivó a no abandonar la carrera en segundo semestre y confió en mí para investigar cómo el análisis de datos puede mejorar la calidad de la vida humana. Ella me conectó con el Doctor Gustavo Rubín Linares, quien aportó valiosas perspectivas basadas en su experiencia como investigador. Ambos no solo me acompañaron con su experiencia, sino que también me conectaron con la Doctora Jessica. Además, colaboraron en la búsqueda de financiamiento dentro de la universidad para publicar el artículo científico derivado de esta investigación. Estoy profundamente agradecido por su apoyo y orientación.

A la Benemérita Universidad Autónoma de Puebla, mi alma mater, le agradezco por brindarme la oportunidad de recibir una educación de calidad. Tuve el privilegio de aprender de profesores brillantes y compartir el aula con compañeros excepcionales, quienes contribuyeron de manera significativa a mi desarrollo profesional. Asimismo, agradezco el espacio que me ofrecieron para poner a prueba los conocimientos adquiridos durante mi formación. Sin duda, la BUAP es la mayor casa de estudios del estado de Puebla, y siempre estaré orgulloso de ser parte de esta institución.

A mis padres, les agradezco por llevarme de la mano durante todos estos años, cumpliendo su labor con dedicación y amor más allá del deber. Gracias por proporcionarme de todos los medios necesarios para convertirme en un ingeniero, por poner un techo sobre mi cabeza, por alimentarme, por abrazar mis tristezas y alegrías, por iluminarme

en mis fracasos y vitorear mis éxitos. Les agradezco cada sacrificio y cada carencia que soportaron para que yo llegara a donde estoy hoy. Soy consciente de que soy el portador de sus sueños y esperanzas, y juro por mi vida que jamás los defraudaré.

A mi hermana Mitchell Gil Ramírez, licenciada en psicología, le agradezco su apoyo incondicional y sus aportaciones invaluables a esta investigación. Como profesional de la salud mental, fue la responsable de darme la idea del acompañamiento emocional y del enfoque en la salud mental que orientaron este trabajo. Su ayuda en la evaluación del modelo resultante de esta investigación, así como su dedicación como hermana y profesional, fueron siempre impecables.

A mi hermano Ulises Gil Ramírez, gracias por ser mi constante fuente de motivación. Eres mi ejemplo a seguir y mi ídolo, por ti soy ingeniero. A mi cuñada Gabriela, le agradezco su apoyo constante y, en especial, el regalo de una silla que aliviaba el dolor en mi espalda tras largas horas de trabajo en la tesis. Ese gesto significó mucho para mí.

A mis amigos, gracias por estar siempre ahí en los momentos difíciles y por hacerme reír, llenando mi vida de alegría. También agradezco a todos los profesores que he tenido a lo largo de mi formación; cada enseñanza recibida ha contribuido a mi desarrollo personal y profesional.

Finalmente, quiero expresar mi gratitud a la licenciada en psicología Karla Daniela de Ávila Ramírez, por su apoyo y aportaciones a esta investigación. Su idea de emplear técnicas como el método socrático en la evaluación fue crucial para este trabajo.

Índice general

Resumen I

Agradecimientos II

Índice general IV

Índice de figuras VI

Índice de tablas VIII

I. Introducción 1
1.1. Motivación 11
1.2. Objetivos 11

II. Estado del arte 14
2.1. En el Campo de la Salud Mental 14
2.1.1. El Trastorno de Ansiedad Generalizada 15
2.1.2. La Inteligencia Artificial en el Campo de la Salud Mental 17
2.1.3. Consideraciones Éticas 17
2.1.4. El Acompañamiento, como una Tarea en los Modelos de Lenguaje 18
2.2. Las Oportunidades que Ofrece el *Deep Learning* 20
2.2.1. Modelos Lingüísticos Grandes 21
2.2.2. Modelos de Lenguaje para la Detección de Emociones 24
2.3. Bases de Datos 31
2.4. Impacto Socioeconómico 34
2.4.1. Beneficios de la Inteligencia Artificial en el Campo de la Psicología 35

III. Método Propuesto 36
3.1. Referencia Inicial (*Baseline*) 36
3.1.1. Modelos de Referencia Inicial 36
3.1.2. Técnica de *Low Rank Adaptation* (LoRA) 39
3.2. *Train* y *Validation Loss* 45

3.3. El Conjunto de Datos . . . 48
3.3.1. *HRECPW: Hispanic Responses for Emotional Classification based on Plutchik Wheel* . . . 48
3.3.2. *HEAR Dataset*: *Hispanic Emotional Accompagnement Responses* 57
3.4. Fine-Tuning del Modelo de Identificación de Emociones . . . 62
3.4.1. Primer *Fine-Tuning* . . . 62
3.4.2. Segundo *Fine-Tuning* . . . 66
3.4.3. Tercer *Fine-Tuning* . . . 69
3.4.4. Conclusiones de los *fine-tuning* . . . 74
3.5. El Modelo Generativo Basado en HEAR . . . 75
3.5.1. El Acompañamiento Emocional en los Modelos del Lenguaje Actuales . . . 75
3.5.2. Fine-Tuning de Llama2-7b-chat para el Acompañamiento Emocional . . . 76

IV. Resultados y Discusión **89**
4.1. Evaluación del *FT* . . . 89
4.1.1. El Tamaño de las Respuestas . . . 90
4.1.2. Cambios en el Vocabulario . . . 90
4.1.3. Conclusión de la Evaluación del *FT* . . . 96
4.2. Instrumentos de Evaluación de Calidad de Respuestas . . . 97
4.2.1. Escucha Activa . . . 97
4.2.2. El Método Socrático . . . 101
4.3. Evaluación de los Modelos . . . 107
4.4. Discusión . . . 109
4.4.1. GPT3.5 . . . 110
4.4.2. Llama2-7b-chat . . . 110
4.4.3. GPT2_124M ajustado . . . 111
4.4.4. Mixtral 8x7b . . . 111
4.4.5. Sólo escúchame . . . 111
4.5. Ejemplos de Conversación . . . 112
4.5.1. Conversación: A . . . 112
4.5.2. Conversación: B . . . 114
4.5.3. Problemas de rendimiento . . . 115

V. Conclusiones y Trabajo Futuro **117**
5.1. Conclusiones . . . 117
5.2. Trabajo futuro . . . 120

Índice de figuras

1.1. Prevalencia DASS-21 (Services, 2018) (Escala de depresión, ansiedad y estrés) . . . 3
1.2. Principales fuentes de estrés durante el COVID-19 (Minds, 2020) . . . 4
1.3. Impacto del COVID-19 en diferentes aspectos de la vida estudiantil (Minds, 2020) . . . 5
1.4. Principales retos en el auto cuidado de las personas durante el COVID-19 (Minds, 2020) . . . 6
1.5. Características de la muestra obtenida en la encuesta dirigida por Cortés-Álvarez et al. (2020) . . . 8

3.1. Proceso de *fine-tuning* . . . 39
3.2. Separación de pesos . . . 42
3.3. Descomposición de la matriz ΔW en matrices de menor rango . . . 45
3.4. *Good Fit*, ejemplo ilustrativo extraído de (Science and Science, 2023) . 47
3.5. Ejemplo ilustrativo de*underfitting* y *overfitting* extraído de (Ryanholbrook, 2023) . . . 47
3.6. Rueda de Plutchik . . . 53
3.7. Proceso de creación del conjunto de respuestas empáticas . . . 58
3.8. Valores de pérdida del primer *fine-tuning* . . . 65
3.9. Segundo *fine-tuning* . . . 68
3.10. Tercer *fine-tuning* . . . 73
3.11. Valores de pérdida de los experimentos . . . 76
3.12. Valor de pérdida durante el entrenamiento (línea punteada) y valor de pérdida durante la validación (línea continua) durante el $1°FT$ . . . 77
3.13. Configuración del *DataCollatorForCompletionOnlyLM*. . . . 77
3.14. Valor de pérdida durante el entrenamiento (línea punteada) y Valor de pérdida durante la validación (línea continua) durante el $2°FT$ . . . 80
3.15. Valores de pérdida durante el entrenamiento (línea punteada) y valores de pérdida durante la validación (línea continua) durante las pruebas del $3°$*fine-tuning* . . . 81
3.16. Valores de pérdida durante el entrenamiento (línea punteada) y valores de pérdida durante la validación (línea continua) durante las pruebas del $4°FT$ . . . 82

3.17. Comparación de valores de pérdida durante la validación de las pruebas del $5°$ *FT* (rojo). 83
3.18. Valor de pérdida durante el entrenamiento (línea punteada) y valor de pérdida durante la validación (línea continua) durante el $5°FT$ 84
3.19. Valor de pérdida durante el entrenamiento (línea punteada) y valor de pérdida durante la validación (línea continua) durante el $6°FT$ (violeta) y $7°FT$ (azul), comparado con el rendimiento del $5°FT$ (verde). 85
3.20. Valor de pérdida durante el entrenamiento (línea punteada) y valor de pérdida durante la validación (línea continua) durante el $8°FT$ (morado), comparado con los valores de pérdida del $5°FT$ (verde). 88

Índice de tablas

2.2. Bases de datos relacionadas al análisis de sentimientos 33

3.1. Instrucción para clasificación de emociones 52
3.2. Primera versión de conjuntos de datos de entrenamiento, validación y prueba 54
3.3. *Prompt* para generar muestras sintéticas 56
3.4. Versión definitiva de los conjuntos de datos de entrenamiento y validación. 57
3.5. *Prompt* en inglés para generar las respuestas de *HEAR Dataset* limitando el comportamiento de GPT-3.5 (OpenIA, 2022) 59
3.6. *Hispanic Emotional Accompagnement Responses* (HEAR). *Prompt* en inglés. 60
3.7. Tabla comparativa de los parámetros de cada experimento y sus valores de pérdida durante el *fine-tuning* (FT) 67
3.8. Tabla comparativa de parámetros y valores de perdida durante cada *fine-tuning* 78
3.9. Una forma mas especifica de la instrucción para la tarea del acompañamiento emocional. 86

4.1. Promedio de longitud de los *tokens* de todas las respuestas. 90
4.2. Tabla comparativa de la frecuencia de cada una de las 50 palabras más frecuentes en las respuestas de los modelos, tomando como referencia las respuestas de GPT-3 para ordenarlas. 94
4.3. Tabla comparativa de los valores obtenidos del TF-IDF para cada término, ordenándolos de mayor relevancia a menor, tomando como referencia a GPT-3. 95
4.4. Rasgos de evaluación de escucha activa 105
4.5. Rasgos de evaluación del método Socrático 106
4.6. *Prompt* de evaluación: escucha activa 107
4.7. *Prompt* de evaluación: método Socrático 108
4.8. Puntaje de cada rasgo de evaluación del método de escucha activa . . . 109
4.9. Puntaje de cada rasgo de evaluación del método Socrático 110
4.10. Ejemplo de conversación: situación amorosa comprometida 113
4.11. Ejemplo de conversación: grandes expectativas laborales 114

4.12. Ejemplo de conversación: ciclado de respuestas 115

Capítulo I

Introducción

La pandemia causada por el virus COVID-19 y las medidas de distanciamiento social han causado trastornos en las rutinas diarias de padres de familia, niños, adolescentes y de la sociedad en general. De acuerdo a un estudio realizado por UNESCO (2020), a partir del 8 de abril de 2020 las clases se suspendieron en 188 países. Esta medida, según lo señalado por la UNESCO, provocó que el 90 % de los alumnos matriculados (entre $1,000$ y $5,000$ millones de jóvenes) abandonaran la escuela.

Una de las consecuencias más importantes de la interrupción de clases es la manifestación de depresión y trastornos mentales en los estudiantes, originados por diversas causas. Por ejemplo, el estudio dirigido por Lee (2020), evidencia que la interrupción de las clases resultó en un deterioro de la situación para los niños y adolescentes que ya presentaban necesidades de salud mental. El mismo autor refiere a un análisis efectuado por la organización YoungMinds [1], en el cual participaron $2,111$ individuos de hasta 25 años con historial de enfermedades mentales en el Reino Unido; aquí, el $83\,\%$ señaló que la pandemia había agravado su situación. Además, un $26\,\%$ indicó que no podía acceder a servicios de apoyo a la salud mental, debido a la cancelación de esos grupos y a

[1] https://www.youngminds.org.uk/

la interrupción de servicios presenciales. Aunado a lo anterior, el artículo subraya cómo la psicóloga Zanonia Chiu de Hong Kong compartió la opinión de que la salud mental de los niños y jóvenes estudiantes que ya lidiaban con problemas psicológicos experimentó un deterioro aún mayor. Esto se debe a que, aunque algunos niños encontraban dificultades para asistir a la escuela, la rutina escolar les proporcionaba una estructura que contribuía a motivarlos a levantarse y afrontar cada día.

En consonancia con Zhu et al. (2022), el aislamiento y el temor al contagio fueron factores determinantes para el desarrollo de ansiedad y depresión en los adolescentes. Sin embargo, también las personas entre 21 y 40 años presentaron mayor vulnerabilidad en su salud mental. Estos problemas se vieron agravados por la exposición prolongada a las redes sociales. Para prevenir y tratar estos casos, alguna de las siguientes medidas sugieren ser efectivas:

- Terapia psicológica: consiste en conversar con un profesional de la salud mental que ayuda a identificar y modificar las causas que originan la depresión y la ansiedad. Una modalidad de terapia psicológica que ha demostrado ser eficaz, es la terapia cognitivo-conductual (Bupa, 2024).

- Fisioterapia: consiste en intervenciones físicas que mejoran la salud mental. El ejercicio físico es una intervención fundamental para el manejo de la depresión y la ansiedad.

Asimismo, en un estudio dirigido por Kalauni et al. (2023), se reveló que a través de un sondeo aplicado a 493 estudiantes del campo de ciencias de la salud, las respuestas dejaron en evidencia una proporción significativa de estudiantes nepaleses en este ámbito que experimentaron trastornos psicológicos como depresión, ansiedad y estrés durante la etapa inicial de la pandemia.

Prevalence of depression, anxiety, and stress (DAS) ($N = 493$)

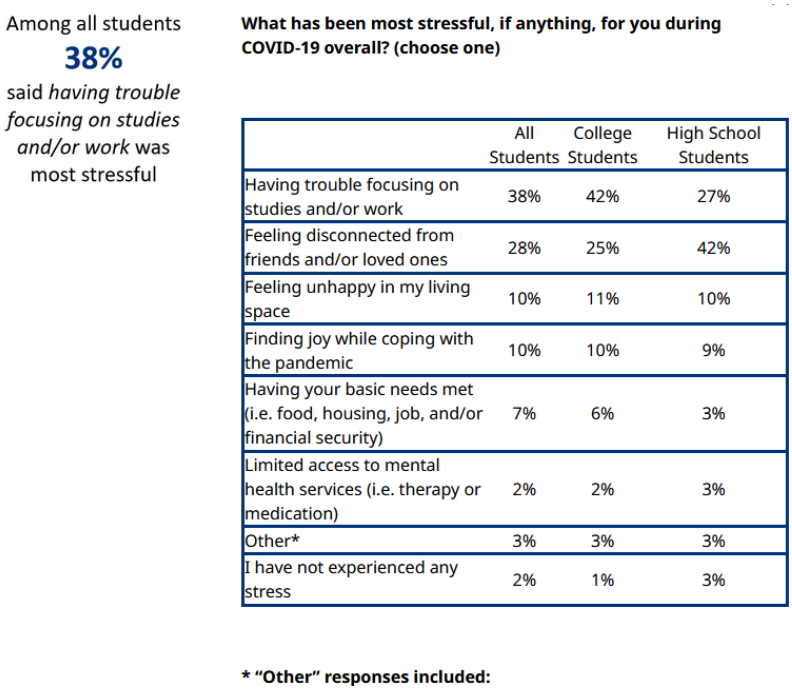
	n (%)		
DAS level	**Depression**	**Anxiety**	**Stress**
Normal	244 (49.5)	234 (47.7)	273 (55.4)
Mild	85 (17.2)	68 (13.8)	64 (13.0)
Moderate	96 (19.5)	91 (18.5)	80 (16.2)
Severe	48 (9.7)	51 (10.3)	32 (6.5)
Extremely severe	20 (4.1)	49 (9.9)	44 (8.9)

Figura 1.1: Prevalencia DASS-21 (Services, 2018) (Escala de depresión, ansiedad y estrés)

La figura 1.1 muestra la prevalencia de depresión, ansiedad y estrés (DAS por sus siglas en inglés) entre los 493 estudiantes que participaron en el estudio. La tabla presenta el número y porcentaje de estudiantes que reportaron niveles normales, leves, moderados, severos y extremadamente severos de DAS. El $49.5\,\%$ de los estudiantes informó tener niveles normales de depresión, mientras que un $17.2\,\%$ experimentó niveles leves, el $19.5\,\%$ presentó niveles moderados, un $9.7\,\%$ experimentó niveles severos y el $4.1\,\%$ registró niveles extremadamente severos. En cuanto a la ansiedad, el $47.7\,\%$ de los estudiantes reflejó niveles normales, en contraste con el $13.8\,\%$ que experimentó niveles leves, el $18.5\,\%$ de estudiantes mostró niveles moderados, un $10.3\,\%$ experimentó niveles severos y el $9.9\,\%$ registró niveles extremadamente severos. Por último, con respecto al estrés, el $55.4\,\%$ de los estudiantes reportó niveles normales, el $13\,\%$ experimentó niveles leves, el $16.2\,\%$ experimentó niveles moderados, un $6.5\,\%$ enfrentó niveles severos y el $8.9\,\%$reportó niveles extremadamente severos.

En la misma investigación, Kalauni et al. (2023) indica que existen múltiples factores asociados con el aumento de la angustia mental, como la inseguridad financiera, el miedo al virus, innumerables muertes, aislamiento, suministro inadecuado de alimentos e incertidumbre en los exámenes, clases y reapertura de la universidad.

En adición, Minds (2020) realizó un estudio que incluyó una encuesta efectuada a $3,239$ estudiantes de preparatoria y universidad, entre el 10 y el 18 de abril de 2020. El propósito fundamental consistió en examinar el impacto de la salud mental en los estudiantes. Esta investigación logró obtener datos de relevancia significativa. El primer resultado muestra que el $38\,\%$ de los estudiantes afirmó tener problemas para concentrarse en los estudios y/o trabajo (figura 1.2).



Among all students

38%

said *having trouble focusing on studies and/or work* was most stressful

What has been most stressful, if anything, for you during COVID-19 overall? (choose one)

	All Students	College Students	High School Students
Having trouble focusing on studies and/or work	38%	42%	27%
Feeling disconnected from friends and/or loved ones	28%	25%	42%
Feeling unhappy in my living space	10%	11%	10%
Finding joy while coping with the pandemic	10%	10%	9%
Having your basic needs met (i.e. food, housing, job, and/or financial security)	7%	6%	3%
Limited access to mental health services (i.e. therapy or medication)	2%	2%	3%
Other*	3%	3%	3%
I have not experienced any stress	2%	1%	3%

*** "Other" responses included:**

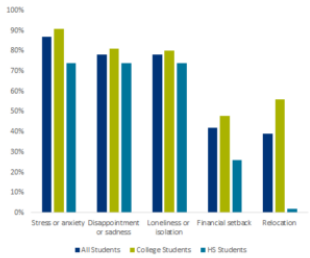
Figura 1.2: Principales fuentes de estrés durante el COVID-19 (Minds, 2020)

El segundo caso de estudió responde a la pregunta: `¿en cuál de las siguientes formas adicionales, si las hay, ha influido el COVID-19 en tu vida?` (Minds, 2020) La anterior pregunta sugiere algunas respuestas, por citar algunas de ellas: ansiedad o depresión, decepción o tristeza, soledad o aislamiento, problemas financieros, entre otras. Los resultados obtenidos se muestran en la figura 1.3.

48%

of college students have **experienced financial setback** due to COVID-19

In which of the following additional ways, if any, has COVID-19 impacted your life? (select all that apply)



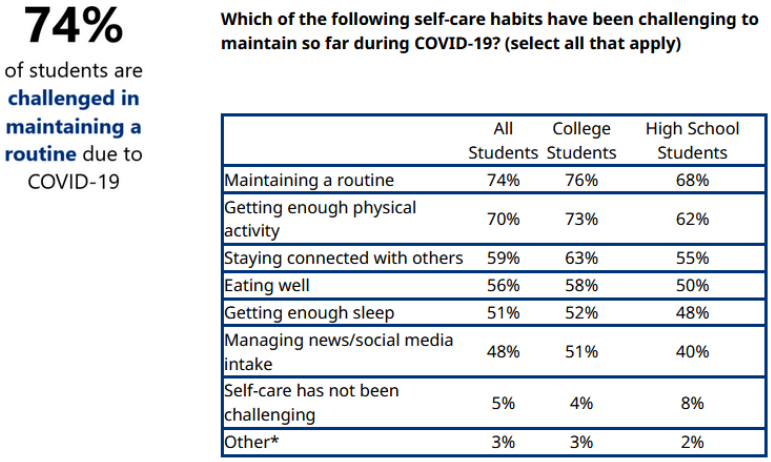
	All Students	College Students	High School Students
Stress or anxiety	87%	91%	74%
Disappointment or sadness	78%	80%	74%
Loneliness or isolation	42%	48%	26%
Financial Setback	42%	48%	26%
Relocation	39%	56%	2%
Illness (myself or a loved one)	7%	6%	9%
Loss of a loved one	4%	3%	5%
None of the above	4%	2%	8%

Figura 1.3: Impacto del COVID-19 en diferentes aspectos de la vida estudiantil (Minds, 2020)

La figura 1.3 muestra cómo la pandemia de COVID-19 afectó la vida de muchos estudiantes universitarios, tanto en el ámbito económico como emocional. Según una encuesta realizada, el $42\,\%$ de los estudiantes ha tenido problemas financieros a causa de la crisis sanitaria, lo que ha dificultado el pago de sus matrículas, libros y otros gastos. Además, el $39\,\%$ de los estudiantes ha tenido que cambiar de residencia, lo que ha implicado una mayor adaptación y estrés, sobre todo para los estudiantes que viven lejos de sus familias. Además, el $42\,\%$ de los estudiantes se ha sentido en soledad o aislado durante el confinamiento, lo que ha afectado su salud mental y su rendimiento académico. De hecho, el $87\,\%$ de los estudiantes ha manifestado haber experimentado ansiedad o estrés en algún momento, siendo los más afectados los estudiantes universitarios con un $97\,\%$. Asimismo, el $70\,\%$ de los estudiantes ha expresado sentimientos de tristeza y

decepción por la situación actual y la incertidumbre sobre el futuro. Estos datos muestran la necesidad de brindar apoyo y orientación a los estudiantes universitarios para que puedan superar esta difícil etapa.

Finalmente, la tercera pregunta de investigación es: `¿cuáles de los siguientes hábitos de autocuidado han sido difíciles de mantener hasta ahora durante COVID-19?` De acuerdo a la figura 1.4, se aprecia que el aspecto más complicado de mantener en los estudiantes fue la rutina ($74\,\%$), seguido de hacer actividad física ($70\,\%$), y de mantener contacto con otras personas ($59\,\%$).



	All Students	College Students	High School Students
Maintaining a routine	74%	76%	68%
Getting enough physical activity	70%	73%	62%
Staying connected with others	59%	63%	55%
Eating well	56%	58%	50%
Getting enough sleep	51%	52%	48%
Managing news/social media intake	48%	51%	40%
Self-care has not been challenging	5%	4%	8%
Other*	3%	3%	2%

Figura 1.4: Principales retos en el auto cuidado de las personas durante el COVID-19 (Minds, 2020)

Sumado a los aspectos mostrados en la figura 1.4, también surgieron otras respuestas importantes como: ánimo para levantarse de la cama, higiene personal, gestionar correctamente la carga de trabajo, mantenerse hidratado, continuar terapia, vestirse con ropa casual en lugar de estar con pijama todo el día, entre otras.

En América Latina, la situación ha seguido un patrón similar al del resto del mundo. Por ejemplo, un estudio liderado por Sanchez Rasmos et al. (2021) en Perú señala que en el ámbito universitario, debido al virus COVID-19, el $91.3\,\%$ del sector académico se vio

obligado a realizar todas sus actividades de forma remota, como medida para frenar la propagación del virus. De acuerdo con lo reportado en está investigación, se realizó una encuesta para recabar datos sobre la situación que vivían los estudiantes universitarios durante el periodo de la pandemia, mostrando que $93.2\,\%$ de los $1,264$ estudiantes encuestados, consideraron que la carga académica había aumentado en relación con ciclos escolares anteriores.

En lo que concierne al territorio mexicano, el periodo de pandemia también ha tenido un impacto desfavorable en la salud mental y el rendimiento académico de los estudiantes de nivel medio superior y superior. Según un estudio realizado por González Velázquez (2020), los factores que más contribuyeron al estrés de los estudiantes fueron: las evaluaciones, calificaciones, la capacidad para resolver problemas, la tolerancia a la frustración o a las demandas de tiempo y esfuerzo de los profesores y las materias, el temor al contagio, el aislamiento social que genera desconfianza y ansiedad, la incertidumbre y la disminución del esfuerzo, la constancia, y la seguridad.

En el mismo estudio, González Velázquez (2020) menciona que el confinamiento fue un reto inédito para el sector educativo, que implicó una adaptación forzosa a una modalidad a distancia con limitaciones de infraestructura para muchos sectores vulnerables. Se indica que el $35.8\,\%$ de los estudiantes del estado de Chiapas no cuentan con una computadora personal, el $35\,\%$ no tiene acceso a internet o el servicio es muy irregular, el $16.8\,\%$ de los estudiantes reportó que el cambio a las clases en línea había afectado mucho su rendimiento académico. Sin embargo, este escenario no sólo incumbe al estado de Chiapas, pues a lo largo de la república mexicana, durante los ciclos escolares 2020 y 2021, más de cinco millones de estudiantes de todas partes de la república de entre 3 a 29 años interrumpieron o abandonaron sus estudios (INEGI, nd). La deserción escolar puede acarrear repercusiones adversas en el bienestar emocional de los estudian-

tes. Aquellos que dejan la escuela podrían enfrentar sensaciones de depresión, angustia, estrés, tristeza y desesperanza, como se observa en la investigación de Santillán (2023). En otro estudio dirigido por Cortés-Álvarez et al. (2020), se empleó la Escala de Impacto de Eventos Revisada (IES-R), un cuestionario destinado a evaluar el grado de angustia psicológica durante el COVID-19. Este cuestionario consta de 22 preguntas de respuesta múltiple, con calificaciones en una escala de 5 puntos, que varía desde 0 ("nada") hasta 4 ("totalmente de acuerdo"). Utilizando este instrumento, se consideró una muestra de 1105 individuos en diversas regiones de México, a quienes se les aplicó una encuesta. La figura 1.5 ilustra que, de los 1105 participantes, el $63.7\,\%$ eran mujeres y el $36.3\,\%$ eran hombres. La mayoría de los participantes tenían entre 18 y 28 años $(44,5\,\%)$, seguidos de los que tenían entre 29 y 39 años $(28,5\,\%)$. El cuadro también muestra que la mayoría de los participantes habían completado la educación superior $(63,5\,\%)$, eran solteros $(56,5\,\%)$ y estaban empleados $(60,5\,\%)$.

Sample Characteristics

Parameter	*N*	Percent
Gender		
Male	418	37.9
Female	686	62.1
Age (years)		
18 to 28	759	68.7
29 to 39	199	18.0
40 to 50	89	8.1
51 a 61	47	4.3
>62	10	0.9
Level of education		
Primary school	4	0.4
Secondary school	41	3.8
High school	263	24.5
College	672	62.6
Specialty	29	2.7
Master	54	5.0
Doctorate	10	0.9
Marital status		
Single	785	71.1
Common-law marriage	65	5.9
Married	210	19.0
Divorced	35	3.2
Widowed	9	0.8
Parental status		
With children	292	26.4
Not children	812	73.6
Household size		
1	85	7.7
2	181	16.4
3	268	24.3
>3	570	51.6

Abbreviations: β, beta coefficient; CI, confidence interval; *N* = sample size.

Figura 1.5: Características de la muestra obtenida en la encuesta dirigida por Cortés-Álvarez et al. (2020)

Adicionalmente, en consonancia con los resultados de la encuesta (Cortés-Álvarez et al., 2020), la carencia de confianza en la fiabilidad de las pruebas de COVID-19 y la insatisfacción con la información sanitaria proporcionada se relacionaron con un incremento en la angustia psicológica ocasionada por el brote, así como con niveles más altos de estrés, ansiedad y depresión. La edad es otro factor relacionado con el estado emocional de los sujetos, pues se encontró que personas con una edad avanzada mostraron un nivel más elevado de estrés y angustia, los adultos jóvenes, de mediana edad y mayores pueden tener diferentes niveles de exposición a factores estresantes relacionados con la escuela, el trabajo y la salud debido a sus etapas de vida. En este contexto, el género emerge como un factor vinculado a esta situación, ya que las mujeres reportaron niveles superiores de angustia psicológica, ansiedad, depresión y estrés en comparación con los hombres. En concreto, el estudio reveló que las mujeres exhibieron una puntuación promedio más elevada en la escala IES-R. Además de analizar e identificar individualmente cada uno de los atributos que un individuo pueda poseer, como el estado marital y nivel sociodemográfico, es posible que varios de estos factores confluyan en el estado emocional de una persona. Por ejemplo, se ha observado que ser mujer, estar divorciada y tener una familia numerosa se vinculan con niveles más elevados de angustia psicológica, ansiedad, depresión y estrés.

Además de la ansiedad y la depresión, que fueron las afecciones más comunes durante la pandemia, también se observó la presencia de estrés postraumático en ciertos segmentos de la población (Palomino and Huarcaya-Victoria, 2020). Este estudio se centró en las repercusiones psicológicas de la pandemia de COVID-19, resaltando el papel crucial del estrés postraumático en ciertos grupos. Estos descubrimientos concuerdan con investigaciones anteriores (Xu et al., 2011; Sprang and Silman, 2013; Lee et al., 2018), sobre brotes epidémicos y pandemias pasadas, en las que se observó que los tras-

tornos de estrés agudo, y en particular el trastorno de estrés postraumático (TEPT), se manifestaron en individuos expuestos a situaciones extremas de aislamiento y miedo. De manera similar, de acuerdo al estudio de Palomino and Huarcaya-Victoria (2020), en el marco de la pandemia actual, se han identificado factores particulares que contribuyen significativamente al surgimiento del estrés postraumático. Estos elementos engloban la novedad del virus, la incertidumbre en torno a su manejo y la constante acumulación de casos y fallecimientos informados a diario en las redes sociales. Estos componentes han intensificado la carga emocional tanto en la población en general como en segmentos específicos, como el personal médico y los padres de familia en cuarentena. Es importante resaltar los factores de riesgo vinculados a una mayor predisposición al estrés postraumático durante esta crisis, incluyendo el género femenino, la exposición en zonas de alto riesgo y la calidad del sueño. El estudio también subraya la importancia de la resiliencia tanto para los profesionales de la salud como para la comunidad en general. Dicho estudio además destaca estrategias para reforzar la capacidad de resiliencia, tales como proporcionar información clara acerca de la situación, promover hábitos saludables durante la cuarentena y fomentar un sentido de comunidad y altruismo. Los resultados de la investigación respaldan la gran necesidad de ofrecer apoyo psicológico continuo, tanto en el transcurso de la pandemia como en la etapa de rehabilitación subsiguiente.

1.1. Motivación

Con fundamento en las investigaciones anteriormente mencionadas, se evidencia el deterioro de la salud mental en los estudiantes con respecto a la adaptación repentina de la educación virtual. Esta realidad ha destacado la fragilidad de nuestro bienestar psicológico en situaciones de cambio inesperado. La necesidad de contar con herramientas que brinden respaldo en este ámbito incrementa cada día, al mismo tiempo que se convierte en una oportunidad para su desarrollo.

En este escenario, adquiere una relevancia crucial la exploración de enfoques novedosos para abordar los desafíos de la salud mental. Las soluciones respaldadas por la inteligencia artificial emergen como una alternativa sumamente prometedora. La creciente evolución de las tecnologías de inteligencia artificial desde 2017 ha brindado soluciones innovadoras en el campo del aprendizaje automático, como las redes neuronales denominadas *Transformers* (Vaswani et al., 2017b), que han demostrado capacidades sorprendentes en la construcción de modelos de lenguaje. Esta conjunción de factores crea un ambiente propicio para la investigación y desarrollo de sistemas que utilicen la inteligencia artificial para detectar y apoyar condiciones como la depresión y la ansiedad.

1.2. Objetivos

Basándonos en la información previamente expuesta, esta sección tiene como propósito detallar tanto el objetivo general como los objetivos específicos que rigen el desarrollo de la presente tesis.

Objetivo General. El propósito principal de esta investigación es recopilar conjuntos de datos en español relacionados con el análisis de emociones expresadas de forma escrita, para su uso en la implementación de un modelo de lenguaje llamado "Sólo Escúchame". Este modelo será capaz de brindar acompañamiento emocional a sus usuarios. No pretende reemplazar a los profesionales de la salud mental, sino ofrecer una herramienta complementaria en el proceso terapéutico. Además, se propondrán métricas de evaluación para medir la calidad de las respuestas del modelo.

Objetivo Específicos

- **Explorar Aplicaciones de la Inteligencia Artificial en Psicología**. Este objetivo implica investigar y comprender en profundidad las diversas formas en que la inteligencia artificial puede ser aplicada en el campo de la psicología. Se analizarán casos de éxito previos y se identificarán oportunidades para la mejora y la innovación en la prestación de servicios emocionales y psicológicos.

- **Comparar tecnologías de construcción de IA.** Mediante la revisión y comparación de las tecnologías disponibles para la construcción de sistemas de inteligencia artificial, se pretende determinar las herramientas y enfoques más adecuados para el desarrollo de modelos capaces de comprender y responder a las emociones humanas de manera efectiva y empática.

- **Desarrollar modelos de análisis de sentimientos.** Este objetivo consiste en entrenar un modelo de lenguaje con la capacidad de analizar sentimientos y emociones en el texto. A través de técnicas avanzadas de procesamiento de lenguaje natural y aprendizaje automático, se espera que el modelo pueda discernir y etiquetar con precisión las emociones presentes en el discurso de los usuarios.

- **Crear un modelo generativo de texto.** Se trabajará en el desarrollo y entrenamiento de un modelo generativo de texto que pueda producir respuestas coherentes y relevantes en función de las entradas emocionales y psicológicas de los usuarios. Este modelo buscará no solo entender el contenido, sino también generar respuestas empáticas y contextualmente apropiadas.

- **Evaluar el rendimiento de los modelos.** Una vez desarrollados los modelos de análisis de sentimientos y generación de texto, se procederá a medir su desempeño en la tarea de brindar acompañamiento emocional y psicológico. Se utilizarán métricas pertinentes para evaluar la efectividad de los modelos en la comprensión de las necesidades emocionales de los usuarios y en la generación de respuestas que sean de apoyo y utilidad.

Capítulo II

Estado del arte

2.1. En el Campo de la Salud Mental

Durante la pandemia de COVID-19, el impacto psicológico en las personas ha captado amplia atención. Losada-Baltar et al. (2020) revela que una gran cantidad de personas que padecen problemas de depresión o ansiedad experimentan impactos de moderados a severos. Este estudio abordó diversas franjas de edad, destacando que los individuos mayores podrían enfrentar una vulnerabilidad particular ante las consecuencias psicológicas de la pandemia. Tal susceptibilidad puede atribuirse, en parte, a los estigmas negativos arraigados en torno a la vejez. Concretamente, la asimilación de estas percepciones desde temprana edad contribuye a la formación de expectativas desfavorables con respecto al proceso de envejecimiento. Una autopercepción negativa del envejecimiento puede favorecer la aparición de problemas psicológicos como la ansiedad, la depresión o la soledad, puesto que pueden predisponer a una menor implicación en comportamientos saludables y a un incremento en la reactividad emocional a los estresores.

De igual forma, el estudio realizado por Rodriguez et al. (2021) a 585 personas de 65 años, visibiliza el malestar experimentado durante el confinamiento pandémico, mani-

festándose en formas de soledad, depresión y ansiedad. Se observa que aproximadamente el $20,9\,\%$ de los participantes en la investigación experimentaron estos sentimientos durante dicho período. Además, se subraya la estrecha relación entre la ansiedad y los problemas de sueño, lo que podría potencialmente desembocar en trastornos del sueño a largo plazo.

Sin embargo, la investigación realizada por Losada-Baltar et al. (2020) revela que los efectos psicológicos durante el COVID-19, al contrario de lo mostrado en Losada-Baltar et al. (2020), las personas mayores reportan menores niveles de problemas de salud mental en comparación con los jóvenes.

La investigación de (Losada-Baltar et al., 2020) también respalda esta afirmación. Este estudio se realizó con una muestra de 1,501 individuos y reveló que los participantes más jóvenes presentan niveles más altos de ansiedad, tristeza y sentimientos de soledad en comparación con las personas mayores. Mientras más joven, aumenta el riesgo de sufrir una sintomatología negativa. Las personas más jóvenes, que tienden a emplear estrategias orientadas a la acción y la resolución de problemas, podrían no estar igualmente preparadas para afrontar la situación. Además, es probable que el confinamiento esté generando mayor estrés a los jóvenes porque representa un cambio más significativo en sus rutinas. Dado su mayor involucramiento tradicional en actividades sociales, de ocio y entretenimiento, la alteración de estas dinámicas también podría contribuir a explicar la mayor sensación de soledad que se observa en este grupo.

2.1.1. El Trastorno de Ansiedad Generalizada

La ansiedad no es un simple trastorno, pues existe toda una familia de trastornos relacionados a la ansiedad, que varían desde el trastorno de ansiedad social y el trastorno de pánico, hasta el trastorno de ansiedad generalizada. Este último se caracteriza por

una preocupación excesiva, producida por un periodo de tiempo extendido (mínimo 6 meses), durante el cual al individuo le es complicado controlar la sensación de preocupación.

Los síntomas asociados a este trastorno incluyen tres o más de los siguientes seis, según lo establecido por el manual diagnóstico DSM-5 (American Psychiatric Association, 2014):

1. Sensación de estar atrapado, se describe por lo general como estar con los nervios de punta.
2. Fácilmente fatigado.
3. Dificultad para concentrarse.
4. Irritabilidad.
5. Tensión muscular.
6. Problemas de sueño.

Evidentemente, la edad de los individuos juega un papel influyente. En el caso de los niños, basta con la presencia de uno de los síntomas anteriores para establecer la posibilidad de estos trastornos. Estas condiciones conllevan a un deterioro en múltiples ámbitos de la vida de las personas que las presentan, repercutiendo en lo social y/o laboral, y generando malestar tanto a nivel mental como físico.

Además de lo mencionado, es importante mencionar que estos malestares pueden ser tratados en terapia, con la ayuda de un profesional en la salud mental. La psicoterapia, como medio brindado por psicólogos, es una herramienta esencial para las personas que la necesitan (American Psychiatric Association, 2012). En este campo, se aplican diversos métodos validados científicamente, orientados hacia la adopción de hábitos más saludables que permitan al paciente un desarrollo más efectivo. La psicoterapia abarca también otras áreas, como la cognitiva conductual, interpersonal, y otros tipos

de terapia conversacional que ayudan a lidiar con múltiples malestares emocionales o trastornos que afectan a las personas.

2.1.2. La Inteligencia Artificial en el Campo de la Salud Mental

En los últimos años, los psicólogos han asumido un papel cada vez más relevante en el avance de la inteligencia artificial, colaborando con los modelos diseñados por expertos en informática para dotarlos de un nivel de conocimiento más sofisticado. Tal y como señala Abrams (2021), "los sistemas de IA a menudo tienen dificultades para hacer conjeturas informadas sobre cosas que no han visto antes, algo que incluso los niños pequeños pueden hacer bien". La inteligencia artificial todavía enfrenta desafíos en cuanto a aspectos cognitivos fundamentales, los cuales la mente humana en sus etapas más tempranas puede resolver sin titubear. Si la meta es emular una inteligencia orgánica similar a la humana, es importante comenzar por perfeccionar los aspectos más elementales.

2.1.3. Consideraciones Éticas

El desarrollo de un sistema de inteligencia artificial que pueda ofrecer servicios de psicoterapia y diagnóstico implica varios desafíos y riesgos. Entre ellos, se destaca la necesidad de garantizar la calidad y la fiabilidad de los diagnósticos, así como el correcto funcionamiento, mantenimiento y accesibilidad del sistema (Fiske et al., 2019). Un mal desempeño podría tener consecuencias negativas para la salud mental de los usuarios. Asimismo, se debe priorizar la protección de los datos personales y sensibles que se manejan en este tipo de servicios, ya que una vulneración podría comprometer la privacidad y la seguridad de los pacientes. La recopilación de datos debe realizarse con el consentimiento informado de los usuarios y con medidas que resguarden su anonimato

y confidencialidad.

Es también importante considerar que una Inteligencia Artificial (IA) como un modelo del lenguaje, no puede sustituir a un psicoterapeuta. Por ahora, las capacidades de estas tecnologías son limitadas, aún teniendo corpus enormes con datos que ayudan a que estos modelos sean potentes en sus labores, no sustituye el factor humano, ni los años de experiencia que los profesionales de la salud mental tienen. Primordialmente, algunos enfoques que utilizan los psicoterapeutas, requieren de algún tipo de contacto directo, ya sea visual, o escuchar la voz de otro ser humano. La psicoterapeuta Sarrió (2021), resalta la importancia del contacto en la terapia *Gestalt*. Ella describe el contacto como la experiencia inicial e inmediata que se establece entre el individuo y su entorno, involucrando una consciencia sensorial profunda de los elementos que lo rodean. Esta es una capacidad que, hasta el momento, ninguna inteligencia artificial ha logrado simular.

2.1.4. El Acompañamiento, como una Tarea en los Modelos de Lenguaje

El acompañamiento, según la definición de Raffo and Instituto Interamericano de Derechos Humanos (2007), implica brindar apoyo mediante una escucha activa, lo que permite que la persona acompañada hable y sienta la reconfortante presencia del acompañante. El acompañamiento psicológico se enfoca en ayudar a personas que enfrentan problemas relacionados con su salud mental. Este tipo de apoyo está dirigido a aquellos que han experimentado situaciones de violencia, pérdida o cualquier proceso que les cause malestar emocional. Su objetivo principal es proporcionar contención, apoyo y fortalecimiento a la persona. Es importante destacar que el acompañamiento psicológico difiere del enfoque terapéutico tradicional destinado a pacientes que requieren tratamiento clínico. El acompañamiento psicológico puede ser llevado a cabo por profe-

sionales como psicólogos, psicoterapeutas o consejeros, y tiende a ser una intervención más estructurada y formal que busca modificar patrones de pensamiento, conducta y emociones de la persona.

Por otro lado, Serrano (2017) argumenta que el acompañamiento emocional se centra en brindar apoyo emocional y afectivo a una persona en momentos de dificultad, crisis, pérdida o transición. El objetivo del acompañamiento emocional es ofrecer una presencia empática y comprensiva. En este espacio, la persona se encuentra en un ambiente donde se siente escuchada, valorada y comprendida, teniendo la libertad de expresar sus emociones sin temor a ser objeto de juicio. Este tipo de apoyo puede ser brindado por amigos, familiares, voluntarios y profesionales de la salud mental.

En relación a lo anterior, resulta evidente que el apoyo emocional no necesariamente debe provenir exclusivamente de un profesional de la salud mental, sino que también puede ser brindado por un círculo cercano de personas al individuo. En este sentido, se plantea la posibilidad de que un modelo de inteligencia artificial pueda servir como una alternativa valiosa para proporcionar apoyo emocional en circunstancias específicas. Por ejemplo, existen momentos en los que una persona puede experimentar emociones inesperadas que tienen un impacto negativo en su salud mental. En tales episodios, es posible que no tenga acceso a su círculo de apoyo debido a la distancia física que los separa de sus familiares, amigos o psicólogo. También puede sentirse avergonzada y preocupada por los posibles juicios de esas personas. En estos casos, la construcción de un modelo de lenguaje basado en IA podría ser una herramienta valiosa, especialmente cuando es dirigido y supervisado por profesionales de la comunidad.

No obstante, antes de especular sobre las posibles contribuciones que el campo del *deep learning* podría aportar al acompañamiento psicológico y emocional, resulta fundamental explorar las tecnologías disponibles para el desarrollo de una IA de este tipo.

2.2. Las Oportunidades que Ofrece el *Deep Learning*

Antes de comenzar a explorar los avances en el área del aprendizaje profundo, es fundamental establecer una base de comprensión de conceptos clave. La inteligencia artificial (IA), se define como un sistema diseñado con la finalidad de emular la inteligencia humana (Oracle, nd).

Dentro del campo de la IA, se encuentra el aprendizaje automático, también conocido como *Machine Learning*, el cual se enfoca en el uso de datos y algoritmos para emular el proceso de aprendizaje humano, mejorando gradualmente su precisión (IBM, ndb). Otra área importante de la IA, es el aprendizaje profundo o *deep learning* en inglés. En general, el aprendizaje profundo involucra el uso de una red neuronal con tres o más capas y busca emular el funcionamiento del cerebro humano mediante técnicas computacionales. A pesar de que está lejos de igualar la capacidad del cerebro humano, esta tecnología permite que los sistemas analicen datos y realicen predicciones con una precisión increíble (IBM, nda).

Otro concepto importante es el modelo de lenguaje, abreviado como LM por sus siglas en inglés (*Language Model*). Los modelos del lenguaje son sistemas de IA diseñados para comprender y generar lenguaje humano. Estos modelos utilizan técnicas de aprendizaje automático para analizar grandes cantidades de texto y aprender patrones del lenguaje humano. A medida que se les alimenta con más datos, estos modelos pueden mejorar su capacidad para generar texto coherente y relevante (Wei et al., 2022; Cobbe et al., 2021). En los últimos años los modelos de lenguaje más eficaces han sido construidos utilizando las redes neuronales conocidas como Transformers (Vaswani et al., 2017a). Los Transformers son una arquitectura de redes neuronales que emplea el mecanismo de atención para procesar secuencias de datos grandes de manera más eficiente en comparación con las redes neuronales recurrentes (RNNs, por sus siglas en inglés)

como LSTM (*Long Short-Term Memory*) (Hochreiter and Schmidhuber, 1997) y GRU (*Gated Recurrent Unit*) (Cho et al., 2014)

Finalmente, el Procesamiento de Lenguaje Natural o *Natural Language Processing* (NLP) es una disciplina de ciencias de la computación, inteligencia artificial y lingüística, que se enfoca en desarrollar algoritmos para hacer que la comunicación entre computadoras y lenguaje humano sea cada vez más eficaz y accesible.

2.2.1. Modelos Lingüísticos Grandes

En 2017, Vaswani et al. (2017b) presenta en su artículo *Attention is All You Need*, la red neuronal denominada *Transformers*. Esta red neuronal utiliza un mecanismo de atención que capta todas las posibles relaciones entre todas las palabras que entran en la red neuronal, superando de esta forma el rendimiento de las RNNs.

En 2018 surge uno de los más remarcables modelos en la era de los Modelo de Lenguaje Grandes (LLM, por sus siglas en inglés *Large Language Models*), BERT (Devlin et al., 2019). BERT (*Bidirectional Encoder Representations from Transformers*) fue presentado por Google AI. Este modelo se enfocó en pre-entrenar mediante una tarea de modelado de lenguaje enmascarado, donde el modelo fue entrenado para predecir palabras ocultas en una oración dada la información de contexto. Este enfoque permitió a BERT aprender representaciones contextuales de las palabras, lo que condujo a una mejora en el rendimiento en una variedad de tareas de PLN, como análisis de sentimientos y reconocimiento de entidades nombradas.

Posteriormente, surgió GPT (*Generative Pretrained Transformer*). GPT-1 (Radford et al., 2018a) fue desarrollado por OpenAI. GPT-1 tiene 12 capas y 768 unidades ocultas por capa. El modelo fue entrenado para predecir el siguiente token en una secuencia, dada la información del contexto de los tokens anteriores. GPT-1 es capaz de realizar una

amplia gama de tareas de lenguaje, incluyendo responder preguntas, traducir textos y generar escritura creativa.

Para el 2019, OpenAI lanza al público la segunda versión de lo que sería la familia de los modelos GPT, GPT-2 (Solaiman et al., 2019). GPT-2 es 10 veces más grande, pasando de 117 millones de parámetros a 1500 millones de parámetros. Debido a que el corpus de entrenamiento de GPT-2 era mas grande, el texto generado por la red, era más coherente y de mayor calidad, además de poder desempeñar más de una tarea en específico, como la traducción y la capacidad de responder a preguntas, sin la necesidad de hacer *fine-tunning*.

En 2020, Google AI presenta T5 (Raffel et al., 2020), un modelo generador de texto multimodal, es decir, es capaz de desempeñar varias tareas como traducción y resumen. Este modelo tiene un total de 11 mil millones de parámetros. En el mismo año, GPT-3 (Brown et al., 2020) fue presentado al mundo como una mejora sustancial al modelo anterior con 175 mil millones de parámetros. GPT-3 es capaz de realizar una mayor cantidad de tareas, entre las que se encuentran, la generación de código en javascript y python, completar textos y asistencia virtual. Este modelo es capaz de procesar 2048 tokens a diferencia de los 1024 tokens de GPT-2, además se comprobó que este modelo era muy bueno aprendiendo de ejemplos que se incluían en el $prompt$, pudiendo mejorar la inferencia dándole más contexto para realizar una tarea que hasta ese momento no se le había enseñado del todo.

Más tarde, en el año 2021 EleutherAI desarrolla GPT-Neo (Black et al., 2021). Este modelo es una alternativa a GPT-3 de código abierto, es mas reducido en tamaño, con tan solo 2700 millones de parámetros. A pesar de su tamaño, la calidad del texto generado por el modelo es comparable a GPT-3.

El 30 de noviembre de 2022 se lanzó la plataforma Chat-GPT (OpenIA, 2022), que puso

a disposición de un público reducido el modelo GPT-3 ajustado para tareas de asistencia virtual. Esto permitió a los usuarios utilizarlo como chatbot y asistente, principalmente para consultar información y crear textos de diversas índoles.

Después, en 2023, Meta AI presentó Llama (Touvron et al., 2023a), un modelo de lenguaje capaz de realizar múltiples tareas gracias al corpus con el que fue entrenado. Entre sus capacidades destacan la traducción de texto, la generación de textos convincentes, *Question Answering* y la clasificación de textos. Este modelo se presentó en diferentes tamaños, desde 7 mil millones hasta 65 mil millones de parámetros. Su enfoque principal es la investigación, ya que fue liberado con una licencia no comercial, facilitando así el acceso a investigadores y otros interesados. En el mismo año, Meta AI lanzó la segunda iteración, Llama2 (Touvron et al., 2023b), diseñada como una versión de propósitos más comerciales y generales en comparación con el Llama original.

A finales de 2023, un año impresionante para el *machine learning*, la organización Mistral AI publicó un artículo en el que presentaba Mixtral 8x7b (Jiang et al., 2024), un modelo que utiliza una arquitectura de mezcla dispersa de expertos (*Sparse Mixture of Experts* o SMoE). Para entender mejor este avance, es esencial explicar el concepto fundamental de MoE (Sanseviero et al., 2023). Este concepto consiste en múltiples "expertos", que son modelos individuales especializados en diferentes tareas. Estos modelos cuentan con un enrutador o compuerta, una función que decide qué expertos serán utilizados durante la inferencia. El término "dispersa.º "sparse"(Gunton, 2024) en el nombre de la arquitectura se refiere a la forma en que se activan los modelos expertos. En lugar de activar todos los expertos para cada entrada, solo se activan unos pocos, lo que reduce significativamente el costo computacional.

Mixtral 8x7b (Jiang et al., 2024) ha demostrado superar en rendimiento a modelos como Llama-2 70B y GPT-3.5 en diversas tareas, incluyendo matemáticas, generación de có-

digo y comprensión multilingüe. Además, ofrece una inferencia seis veces más rápida que Llama-2 70B, lo que lo hace más eficiente en términos de costo y latencia. Esto le permite manejar un gran número de parámetros (47B) mientras utiliza activamente solo una fracción de ellos (13B) durante la inferencia.

2.2.2. Modelos de Lenguaje para la Detección de Emociones

El concepto de una I.A en el campo del PLN que sea capaz de clasificar emociones y en función de la clasificación de emociones, responda de forma adecuada, se ha trabajado en los últimos años. Por ejemplo, Zhou et al. (2017) propone un modelo sensible a las emociones de un individuo, permitiendo la generación de respuestas variadas basadas en un amplio espectro de emociones, tales como alegría, disgusto, tristeza, entre otras. Este modelo consta de dos elementos muy importantes:

1. El codificador. Es una red neuronal de atención recurrente con el mecanismo GRU (*Gated Recurrent Units*). El codificador procesa la entrada de la conversación y produce una representación del contexto de la misma. El contexto se almacena en la memoria interna y se actualiza a medida que recibe nuevas entradas.

2. El decodificador. Utiliza tanto la información del contexto codificado por el codificador como la información de la memoria externa para generar una respuesta. Lo anterior permite al decodificador comprender la emoción, y generar respuestas con una emoción apropiada.

Además del codificador y el decodificador, otros elementos importantes de la red neuronal son la memoria externa e interna. La memoria externa es útil para mejorar la generación de respuestas en una conversación, se almacena en forma de una base de datos de conocimiento, que incluye información sobre temas, hechos y conocimientos

comunes. La memoria interna en este modelo se utiliza para mantener un registro del contexto de la conversación actual y utilizar esta información para mejorar la generación de respuestas coherentes y emocionalmente apropiadas.

Por su parte, Zhou et al. (2017) propone un modelo que consta de un generador que utiliza muestras de ruido gaussiano como entrada para la generación inicial de texto, basado en redes generativas adversarias (conocidas como GAN por sus siglas en inglés). Sin embargo, a través de la retroalimentación de la red discriminadora, se logra una mayor precisión en la generación del texto en comparación con los datos válidos. La arquitectura de este modelo está basada en RNNs .

La limitación de los modelos previamente mencionados radica en que sus arquitecturas han quedado obsoletas en la actualidad. Esto se debe a que las redes neuronales tipo *transformers* han demostrado alcanzar los mejores resultados en todas las tareas de PLN (Vaswani et al., 2017b; Radford et al., 2018a; Devlin et al., 2019).

La efectividad de *transformers* no reside exclusivamente en la arquitectura usada, sino también en los datos que se usan para entrenar la red neuronal (Fetterman et al., 2023; Wei et al., 2022). En su estudio, Barbieri et al. (2020) empleó distintos tipos de modelos basados en *Transformers*. A continuación, se listan las distintas variantes de RoBERTa (Liu et al., 2019) usadas en este estudio.

1. **RoB-Bs** *(RoBERTa pre-trained base)* - es el modelo base, entrenado con el conjunto de datos de Wikipedia y Reddit.

2. **RoB-RT** - es *RoBERTa pre-trained base*, pero re-entrenado con un conjunto de datos sin etiquetado de 60 millones de Tweets en inglés.

3. **RoB-Tw**- es RoBERTa entrenado con el conjunto de datos mencionado anteriormente, pero esta variante, no está pre-entrenada.

Los resultados de esta investigación demuestran que la variante que ofrece un rendimiento superior es RoB-RT, demostrando que modelos re-entrenados con datos de estructura similar a los de entrenamiento, pueden ofrecer un mejor rendimiento, debido a que estos modelos son capaces de captar mejor las características y patrones específicos del dominio de los datos. Es decir, los modelos re-entrenados, eran capaces de entender y procesar de manera más efectiva el lenguaje y las particularidades de los tweets, resultando en una mejora significativa en las tareas de evaluación de tweets.

Por lo tanto, se puede concluir que la adaptación de los datos al modelo es crucial. En este sentido, es necesario considerar el idioma en el que se realiza la tarea de clasificación de datos. En el contexto de esta tesis, se utiliza el español. Sin embargo, en el uso cotidiano, es común que las personas empleen anglicismos o palabras de otros idiomas, un fenómeno conocido como *code-mixing* (Chatterjee, 2022). Esta particularidad puede representar un desafío al diseñar e implementar modelos de lenguaje. Aunque existen modelos pre-entrenados en múltiples idiomas, como BERT-base multilingual (Devlin et al., 2019), entrenado en 104 idiomas, no se garantiza un rendimiento óptimo al procesar texto en español con mezcla de códigos.

Un conjunto de datos fundamental en este contexto es SentiMix (Patwa et al., 2020), que recopila, clasifica y etiqueta tweets en diversos idiomas que contienen mezcla de códigos para el análisis de sentimientos. Patwa et al. (2020) describe que se han recopilado 19 mil tweets en spanglish (español e inglés), con 3 etiquetas que describen la carga sentimental como positivo, negativo y neutro. A su vez, este artículo presenta la evaluación de una competencia que buscaba obtener los mejores resultados en la tarea de análisis de sentimientos utilizando distintos tipos de modelos de lenguaje. Participaron 89 equipos, de los cuales 28 trabajaron sobre el conjunto de datos en spanglish. La mayoría de las arquitecturas participantes estaban basadas en *Transformers*, como

BERT o variaciones de este modelo. En los resultados, la mayoría de estos modelos lograron resultados óptimos para ambos idiomas.

Se han realizado otros esfuerzos para ampliar el análisis de sentimientos y emociones, empleando métodos ingeniosos para recopilar información. Un ejemplo, es la supervisión a distancia (*Distant-Supervision*) (Suttles and Ide, 2013), una técnica de etiquetado de datos que implica identificar palabras clave o etiquetas específicas dentro de una base de datos para la generación de conjuntos de datos extensos con poca intervención humana. De esta forma, se pueden obtener conjuntos de datos que amplíen las capacidades de los modelos de lenguaje especializados en el análisis de sentimientos y emociones. La limitación de esta técnica es que puede formar conjuntos de datos con ruido, que pueden sesgar el modelo que se ha entrenado con estos datos.

Otro método empleado para la recolección de datos es el etiquetado manual, que fue el seleccionado por Oprea and Magdy (2020), en su trabajo de *isarcasm*. En esta investigación, se construyó un conjunto de datos especializado para la identificación de sarcasmo. Se llevaron a cabo encuestas en las que las respuestas de los usuarios de la plataforma Twitter contribuyeron a la obtención de muestras. Cada muestra consistía en un tweet sarcástico y tres tweets no sarcásticos. En el caso de los tweets sarcásticos, se solicitaba a los autores que proporcionaran una explicación y una versión alternativa que expresara el mismo mensaje de manera directa, sin utilizar sarcasmo.

Además, también se han utilizado otros modelos para el análisis de sentimientos, algunos de los cuales están destinados al estudio de GLUE (Wang et al., 2018). GLUE es una colleción de conjunto de datos, diseñada para evaluar y analizar el rendimiento de los modelos de lenguaje en nueve tareas distintas: análisis de sentimientos, vinculación textual, *question answering*, reconocimiento de entidades nombradas que consiste en identificar y clasificar las identidades nombradas en el texto, implicación textual, secu-

laridad semántica, paráfrasis, análisis sintáctico y aceptabilidad lingüística. Uno de los modelos que estudia este conjunto de datos es BiLSTM (Huang et al., 2015), con el optimizador *Adam*, y una tasa de aprendizaje de $10^{-}4$. Para hacer comparaciones, otras variantes del modelo BiLSTM, se le hicieron modificaciones en sus capas ocultas. Los resultados mostraron que los modelos con entrenamiento multitarea obtuvieron mejores resultados en comparación con aquellos que solo fueron entrenados en una sola tarea.

MentalBERT (Ji et al., 2021) es otro modelo de lenguaje que se especializa en el análisis de texto sobre salud mental. Se creó con el objetivo de identificar problemas psicológicos y tendencias suicidas, sobre todo en situaciones de crisis como la pandemia de COVID-19, siendo el primero en este dominio. MentalBERT y su variante MentalRoBERTa se derivan de los modelos BERT (Devlin et al., 2019) y RoBERTa (Liu et al., 2019), respectivamente. Ambos modelos fueron entrenados usando Huggingface [1], una librería para trabajar con las redes neuronales *Transformers*. Los modelos se pre-entrenaron con la técnica de modelado de lenguaje enmascarado, que consiste en ocultar (enmascarar) algunas palabras al azar en una frase de entrada y hacer que el modelo prediga las palabras ocultas basándose en el contexto de las palabras que las rodean. El conjunto de datos fue obtenido de Reddit [2], escogiendo sub-reddits relacionados con la salud mental, tales como,*depression, suicideWatch, anxiety, offmychest, bipolar, mentalillness*. Después del pre-entrenamiento, se adaptó el modelo para la detección y clasificación de trastornos mentales. En la configuración del ajuste se añadió la incrustación del token especial en la última capa oculta. Los autores utilizaron la función de activación $tanh$ para el clasificador y como optimizador, Adam. Los resultados respecto a la tarea de detección de depresión, mostraron que el modelo MentalRoBERTa obtuvo los mejores puntajes con respecto a BioBERT, o la versión no pre-entrenada de

[1] https://huggingface.co/
[2] https://www.reddit.com/

BERT.

Estigmas reforzados en los Modelos de Lenguaje. Según la investigación de Lin et al. (2022), los modelos de lenguaje enmascarados pueden perpetuar y amplificar el estigma de la salud mental de género. Modelos como BERT, RoBERTa y XLNet (Yang et al., 2019) tienden a asociar ciertos trastornos mentales con determinados géneros, lo que refleja y refuerza los estereotipos existentes. Además, Lin et al. (2022) sugiere que se necesita más investigación y concientización sobre los sesgos de género en los modelos de lenguaje enmascarados y sus posibles implicaciones sociales. Se utilizaron cuestionarios desarrollados en investigaciones de psicología para detectar ciertas condiciones de salud mental. La metodología de dichos cuestionarios se centró en la selección de frases relacionadas con la salud mental y la comparación de las asociaciones de género de los $tokens$ generados por los modelos del lenguaje utilizando la técnica de enmascarado. Asimismo, se realizaron experimentos en los que se enmascararon los sujetos en las frases y se evaluó la probabilidad del modelo de llenar los espacios en blanco con palabras masculinas, femeninas o sin especificar género. Estos experimentos permitieron observar que los modelos son más propensos a predecir sujetos femeninos que masculinos en frases sobre tener una condición de salud mental, debido a que los datos recolectados reflejan diferentes estereotipos con hombres y mujeres con problemas de salud mental, especialmente cuando se habla de tratamiento. Lo anterior refleja el estigma social de que los hombres no buscan ni reciben ayuda para sus problemas de salud mental.

Toxicidad en los Modelos de Lenguaje. Los modelos de lenguaje generativos son algoritmos de aprendizaje automático que pueden crear contenido nuevo a partir de datos de entrenamiento, imitando la estructura y el significado del lenguaje natural.

Modelos como GPT-3 (Brown et al., 2020), se entrenan con texto de distintas fuentes, como artículos científicos, libros, páginas web o redes sociales. Sin embargo, algunas de estás fuentes de texto contienen lenguaje influenciado por el contexto social que puede ser ofensivo, y además, mucho de este contenido se encuentra en el anonimato.

La toxicidad en los modelos del lenguaje se refiere a la propensión de generar lenguaje que puede ser considerado racista, sexista u ofensivo de alguna otra manera, lo cual puede perjudicar su uso seguro (Weng, 2021). Se han explorado diversas estrategias para eliminar o reducir esta toxicidad en los modelos del lenguaje. Por ejemplo, se han desarrollado *benchmarks* centrados en evaluar la toxicidad de estos modelos, como *RealToxicityPrompts* (Gehman et al., 2020). Este conjunto de datos se utiliza para evaluar la toxicidad de los modelos del lenguaje y consta de 100 mil frases extraídas de la web, junto con sus puntuaciones de toxicidad calculadas por un clasificador. El objetivo principal es medir tanto la frecuencia como la intensidad con que los modelos del lenguaje generan contenido tóxico, ofensivo o problemático a partir de estas frases.

Otras estrategias para mitigar el contenido tóxico en los modelos del lenguaje implican la eliminación o modificación de los datos de entrenamiento que contienen contenido tóxico, ofensivo o problemático. Sin embargo, Welbl et al. (2021) argumenta que la eliminación o alteración de estos datos de entrenamiento puede impactar negativamente en la efectividad de estos modelos para generar texto. Específicamente, esta modificación repercute en la generación de texto relacionado con ciertos grupos marginados, reduciendo las muestras que contienen información de una naturaleza más neutral o positiva de estos grupos.

Igualmente, utilizar un modelo pre-entrenado con un conjunto de datos libre de contenido tóxico se presenta como una estrategia adicional para mitigar la toxicidad en un modelo de lenguaje (HuggingFace, nd). Esto se logra mediante el uso de enfoques

como el aprendizaje por refuerzo o el aprendizaje adversario. Esta estrategia puede resultar efectiva para corregir o mitigar los sesgos y tendencias tóxicas que el modelo haya adquirido durante su fase de pre-entrenamiento.

Por otro lado, la generación controlada implica intervenir en el proceso de generación de texto del modelo (Weng, 2021). Esto se logra a través de técnicas como la penalización de palabras o frases tóxicas, la reescritura o edición del texto generado, o la selección de las mejores opciones de generación según algún criterio de calidad o seguridad. Estas intervenciones contribuyen a evitar o reducir la toxicidad en el texto generado por el modelo.

2.3. Bases de Datos

Dado que el análisis de sentimientos es uno de los principales focos de interés en esta tesis, la tabla 2.2 detalla algunas de las bases de datos más relevantes en este campo.

Conjunto de datos para el análisis de sentimientos

Base de datos	**Descripción**
TweetEval (Barbieri et al., 2020)	Este conjunto de datos tiene 7 subconjuntos para 7 diferentes tareas: reconocimiento de emociones, detección de discurso de odio, de ironía, del lenguaje ofensivo, de postura, análisis de sentimientos y predicción de Emoji. La tarea de análisis de sentimientos tiene mayor numero de muestras, $45,389$. Además, tiene tres etiquetas: positivo, negativo y neutro.
SentiMix (Patwa et al., 2020)	Contiene 39 mil muestras etiquetadas como positivo, negativo y neutro. 20 mil de las muestras están en Hinglish, y 19 mil en Spanglish.

iSarcasm (Oprea and Magdy, 2020)	Es un conglomerado de datos para detectar el sarcasmo y otras subcategorias relacionadas, tales como sarcasmo, ironía, sátira. Contiene $3,707$ muestras no sarcásticas y 777 muestras sarcásticas.
GLUE (General Language Understanding Evaluation benchmark) (Wang et al., 2018)	Es un *benchmark* utilizado para medir la efectividad de los modelos de lenguaje en nueve distintas tareas: CoLA (Corpus de Aceptabilidad Lingüística) (Warstadt et al., 2019), SST-2 (tarea de clasificación binaria) (Socher et al., 2013a), MRPC (consiste en determinar si dos oraciones tienen el mismo significado) (Dolan and Brockett, 2005), STS-B (Sentencias de similiridad semántica - Benchmark), QQP (*Quora Question Pairs*) (Wang et al., 2017), MNLI (consiste en determinar si una oración es una contradicción), QNLI (*Question NLI*) (Wang et al., 2018), RTE (Recognizing Textual Entailment)(Dagan et al., 2009), WNLI (consiste en responder preguntas de comprensión).
HappyDB (Asai et al., 2018)	Este conjunto de datos consiste en historias felices compartidas por personas en línea. Las historias fueron recolectadas de Reddit y contienen una variedad de temas, como amor, amistad, familia, logros, etc. Contiene $100,000$ muestras, etiquetadas con una categoría general, como 'logro", amor" o amistad".
SARC (Khodak et al., 2018)	Este grupo de datos está compuesto por comentarios de Reddit. Estos comentarios se han etiquetado como sarcásticos o no sarcásticos. Contiene alrededor de 533 millones de comentarios, de los cuales 1.34 millones son sarcásticos.

SST (Stanford Sentiment Treebank) (Socher et al., 2013b)	Contiene $215,154$ frases únicas, anotadas por tres jueces humanos. Tiene 5 tipos de etiquetas: *negative, somewhat negative, neutral, somewhat positive, positive*.
Sarcasm Corpus V2 (Oraby et al., 2016)	Es una versión actualizada del conjunto de datos SARC, incluye comentarios de varios subreddits adicionales y también tiene una distribución de clases más equilibrada.
TSATC: Twitter Sentiment Analysis Training Corpus (Naji, 2012)	Contiene $1,578,627$ muestras clasificadas como positivo (1) y negativo (0).
Spanish Arilines Tweets Sentiment Analysis (Guillem, 2018)	Es un conjunto de datos en español que se utiliza para entrenar y evaluar modelos de análisis de sentimientos en tweets relacionados con aerolíneas españolas. El conjunto de datos contiene $8,781$ tweets etiquetados manualmente como positivos, negativos o neutrales.
IMDb Movie Reviews (Maas et al., 2011)	Es una base de datos binaria compuesta por $50,000$ críticas de la página de Internet Movie Database (IMDb).

Tabla 2.2: Bases de datos relacionadas al análisis de sentimientos

2.4. Impacto Socioeconómico

Al hablar de salud mental, solemos enfocarnos en su impacto en el individuo, dejando de lado cómo afecta a la economía global. Sin embargo, es crucial reconocer que, como seres sociales, la salud mental influye en todos los miembros de una sociedad y, por ende, en su desarrollo económico. Según el estudio realizado por el Foro Económico Mundial y la Escuela de Salud Pública de Harvard (Carreño, 2019), se estima que la pérdida de producción económica acumulada asociada con los trastornos mentales en todo el mundo será de $16,3$ billones de dólares entre 2011 y 2030. Es importante considerar que la salud mental también afecta al ámbito laboral, causando absentismo, abandono y rotación innecesaria (Vera, 2022). Además, se precide que os costes directos e indirectos de la salud mental de la población del mundo, son equivalentes al $4\,\%$ del PIB (Producto Interior Bruto) mundial.

Para mitigar los efectos económicos de la mala salud mental, se pueden implementar diversas medidas a nivel individual, social y estructural. Una opción sería promover y proteger la salud mental desde la primera infancia hasta la vejez (Organization, 2022), mediante el desarrollo de habilidades emocionales, el fomento de entornos seguros y el apoyo a las personas en situación de vulnerabilidad. Lo anterior implicaría que los gobiernos establezcan programas enfocados en la salud mental, lo cual incluye aumentar el gasto público y destinarlo a servicios comunitarios e integrados (Guerri, 2021), lo que lo convierte en una opción complicada. Otra alternativa sería fomentar hábitos saludables que contribuyan al bienestar mental, como mantener una alimentación equilibrada, hacer ejercicio regularmente, dormir bien, evitar el consumo de alcohol y otras drogas, practicar técnicas de relajación y buscar apoyo social cuando se necesite (Lang, 2022). En esta última opción, un modelo de lenguaje podría ser de gran ayuda, ya que estos modelos pueden desempeñar varias tareas, como la clasificación y análisis de emocio-

nes, generación de texto, entre muchas otras. En conjunto, se puede crear un sistema que brinde soporte a las personas y realice tareas de acompañamiento emocional, lo que traería muchos beneficios adicionales.

2.4.1. Beneficios de la Inteligencia Artificial en el Campo de la Psicología

Un modelo de lenguaje enfocado en el acompañamiento emocional podría tener un impacto socioeconómico positivo al mejorar el bienestar y la salud mental, especialmente durante crisis como la pandemia de COVID-19. Este modelo brindaría atención, escucha y apoyo a quienes experimentan malestar emocional debido a situaciones estresantes, pérdidas o cambios de hábitos, ofreciendo respuestas empáticas. No busca reemplazar a los profesionales de la salud, sino complementar la terapia. La integración de esta tecnología en psicología ofrece varios beneficios que deben considerarse antes de especializarse en un área particular (Fiske et al., 2019).

- Una detección temprana de problemas de salud.
- Atender a grupos específicos de alto riesgo como veteranos.
- Reducir el malestar que causa el estigma de recibir atención psicológica.
- Para algunos pacientes significa una mejor respuesta al no ser un humano su terapeuta.
- Ayuda a mejorar la confianza y apertura de los pacientes.
- Para personas con patologías que no son agudas, puede ser especialmente beneficioso, evitando un escrutinio inicial por parte de su terapeuta.
- Vuelven accesible el servicio de salud mental a personas que viven en poblaciones remotas.
- El terapeuta virtual tiene un tiempo y paciencia más flexibles por no decir infinitas, ofreciendo un servicio aún más confiable y adecuado para ciertas poblaciones.

Capítulo III

Método Propuesto

3.1. Referencia Inicial (*Baseline*)

3.1.1. Modelos de Referencia Inicial

Llama

Algunos investigadores como Cobbe et al. (2021) y Wei et al. (2022) aseguran que entre más grande sea el modelo de lenguaje (en función del número de parámetros), más eficiente es el modelo. Sin embargo, Fetterman et al. (2023) argumenta que los mejores rendimientos no se logran con modelos más grandes, sino con modelos más pequeños entrenados con más datos. Bajo esta premisa, Touvron et al. (2023b) desarrollaron una gama de modelos de lenguaje conocidos como *Llama*. Dichos modelos tienen desde 7 hasta 65 billones de parámetros. Llama logró obtener resultados comparables con los mejores modelos del año 2022 como Chinchilla-70B (Hoffmann et al., 2022) y PaLM-540B (Chowdhery et al., 2022), y en varias tareas superó los resultados obtenidos por GPT-3 (Brown et al., 2020).

Llama capitaliza diversas mejoras propuestas en la arquitectura de *Transformers* por

modelos de lenguaje previos como GPT-3 y PaLM. Algunas de estas mejoras se listan a continuación:

- Pre-normalización (GPT3): Consiste en normalizar la entrada de cada subcapa de los *Transformers* en lugar de la salida, utilizando la función *RMSNorm* (Zhang and Sennrich, 2019). Esta técnica ayuda a estabilizar el entrenamiento y mejorar el rendimiento.

- Función de activación SwiGLU (PaLM): Es una variante de la función ReLU (Fukushima, 1969) que introduce una compuerta sigmoidal. Esta función permite que el modelo sea más creativo en cuanto a la generación del texto generado.

- Incorporación rotativa (GPTNeo): Es una forma de codificar la posición relativa de las palabras en el texto, en lugar de usar la posición absoluta. Esto permite al modelo generalizar mejor textos de diferentes longitudes y evitar el sesgo hacia las posiciones iniciales o finales.

Para entrenar los diferentes modelos de Llama (Touvron et al., 2023b) se usaron conjuntos de datos públicamente disponibles en distintos idiomas (inglés, francés, alemán, español, ruso, etc.). Conjuntos de datos como *CommonCrawl* y C4 (Raffel et al., 2020) aportaron el $82\,\%$ de los datos totales. Los autores, usaron CCNet pipeline (Wenzek et al., 2019) para eliminar texto duplicado, identificar los diferentes idiomas y eliminar el texto de baja calidad.

Llama ha destacado por su rendimiento excepcional en diversas tareas, tales como el razonamiento de sentido común, la respuesta a preguntas, superando en gran medida a modelos como GPT-3, PaLM y Chinchilla. Estos resultados demuestran su gran potencial para mejorar las aplicaciones de procesamiento del lenguaje natural.

OpenLlama

OpenLlama (Geng and Liu, 2023) es una implementación de código abierto del modelo de lenguaje Llama de Meta AI. Los creadores de OpenLlama entrenaron sus modelos con el conjunto de datos RedPajama (Computer, 2023), que a su vez es una reproducción del conjunto de datos de entrenamiento Llama que contiene más de $1,2$ billones de $tokens$. Los autores usaron los mismos hiper-parámetros de preprocesamiento y entrenamiento usados en el artículo original de Llama, incluida la arquitectura del modelo, la longitud del contexto, los pasos de entrenamiento, el valor de la tasa de aprendizaje y el optimizador. La única diferencia entre OpenLlama y Llama original es el conjunto de datos de entrenamiento utilizado.

Llama2

Llama2 (Touvron et al., 2023c) es la nueva versión del modelo de lenguaje de código abierto de Meta. Diseñado para competir con modelos como GPT-3 (Brown et al., 2020), representa la evolución de su predecesor, Llama. Las diferencias principales entre Llama2 y su versión anterior incluyen:

- Tamaño: Llama-2 es notablemente más grande que su predecesor, ya que cuenta con 10 billones de parámetros, en comparación con los 1.2 billones presentes en la versión anterior.
- Código abierto: Llama2 es de código abierto, lo que significa que el código y los pesos del modelo están disponibles de forma gratuita para uso comercial y no comercial.
- Disponibilidad: Llama2 está disponible para que las empresas emergentes y las empresas establecidas accedan y creen sus propios productos.
- Colaboración: El modelo Llama2 se desarrolla en colaboración con Microsoft.

3.1.2. Técnica de *Low Rank Adaptation* (LoRA)

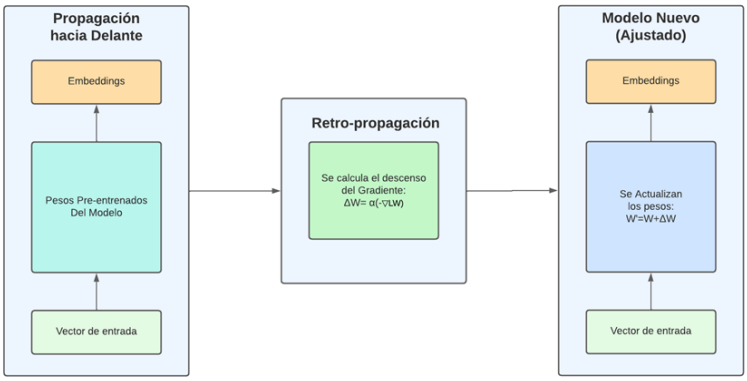


Figura 3.1: Proceso de *fine-tuning*

El entrenamiento de redes neuronales es un proceso complejo que consta de diversos pasos. Inicialmente, se lleva a cabo la propagación hacia adelante, que implica alimentar a la red con los datos de entrada. Durante esta fase, cada capa de la red realiza cálculos y transformaciones en los datos, permitiendo que la información fluya a través de la arquitectura. Esta etapa es fundamental para generar predicciones iniciales basadas en los parámetros de la red. Estás predicciones son el resultado de una serie de operaciones matemáticas que se realizan en cada capa intermedia de la red neuronal, estableciendo así las bases del proceso de entrenamiento. A continuación de describe este proceso matemáticamente:

1. Cálculo de la activación ponderada (ecuación 3.1):

$$z^{(l)} = W^{(l)} a^{(l-1)} + b^{(l)} \tag{3.1}$$

Donde:

- $z^{(l)}$ es el vector de activación ponderado para la capa l.

- $W^{(l)}$ es la matriz de pesos para la capa l, comúnmente se asignan aleatoriamente.
- $a^{(l-1)}$ es el vector de activaciones de la capa anterior $l-1$. En el caso de la capa de entrada, a es el vector de datos de entrada.
- $b^{(l)}$ es el vector de sesgos para la capa l.

2. Aplicación de la función de activación (ecuación 3.2):

$$a^{(l)} = f(z^{(l)}) \tag{3.2}$$

Donde:

- $a^{(l)}$ es el vector de activaciones para la capa l.
- $f(\cdot)$ es la función de activación aplicada elemento a elemento.

El proceso se repite para cada ejemplo en el conjunto de datos de entrenamiento, y las salidas finales se comparan con las salidas reales (*gold standard*) para calcular el error y, posteriormente, ajustar los pesos de la red mediante el método de retropropagación (Werbos, 1974; Roy, 2022):

1. Cálculo del error en la capa de salida (ecuación 3.3):

$$\delta^{(L)} = \nabla_a E \odot f'(z^{(L)}) \tag{3.3}$$

Donde:

- $\delta^{(L)}$ es el vector de error en la capa de salida L.
- $\nabla_a E$ es el vector de derivadas parciales del error respecto a las activaciones en la capa de salida L.

- $f'(\cdot)$ es la derivada de la función de activación aplicada a las activaciones ponderadas en la capa de salida L.

2. Propagación del error hacia atrás (ecuación 3.4):

$$\delta^{(l)} = ((W^{(l+1)})^T \delta^{(l+1)}) \odot f'(z^{(l)}) \tag{3.4}$$

Donde:

- $\delta^{(l)}$ es el vector de error en la capa l.
- $W^{(l+1)}$ es la matriz de pesos entre la capa l y la capa $l+1$.
- $\delta^{(l+1)}$ es el vector de error en la capa $l+1$.
- $f'(\cdot)$ es la derivada de la función de activación aplicada a las activaciones ponderadas en la capa l.

3. Cálculo de gradientes de los pesos (ecuación 3.5) y sesgos (ecuación 3.6), respectivamente:

$$\frac{\partial E}{\partial W^{(l)}} = \delta^{(l)} (a^{(l-1)})^T \tag{3.5}$$

$$\frac{\partial E}{\partial b^{(l)}} = \delta^{(l)} \tag{3.6}$$

Donde:

- $\frac{\partial E}{\partial W^{(l)}}$ es la derivada del error respecto a la matriz de pesos $W^{(l)}$ en la capa l.
- $\frac{\partial E}{\partial b^{(l)}}$ es la derivada del error respecto al vector de sesgos $b^{(l)}$ en la capa l.
- $a^{(l-1)}$ es el vector de activaciones de la capa $l-1$.

- $\delta^{(l)}$ es el vector de error en la capa l.

4. Estos gradientes se utilizan posteriormente para ajustar los pesos y los sesgos de la red neuronal en la dirección que minimiza el error durante el proceso de optimización, tal como lo indican las ecuaciones 3.7 y 3.8, respectivamente:

$$\Delta W = \alpha(-\nabla L_W) \tag{3.7}$$

$$W_{\text{nuevo}} = W_{\text{viejo}} + \alpha(-\nabla L_W) \tag{3.8}$$

Estás ecuaciones permiten que la red aprenda y mejore en función de los datos de entrenamiento y las salidas deseadas. El proceso anteriormente descrito, se muestra de manera gráfica en la figura 3.2.

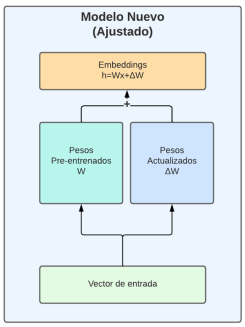


Figura 3.2: Separación de pesos

Una vez explicado el flujo básico de los datos en una red neuronal, se explica LoRA (por sus siglas en inglés *Low-Rank Adaptation*), un enfoque de adaptación de bajo rango aplicado a los modelos de lenguaje propuesto por Hu et al. (2021). LoRA ajusta un modelo con pesos pre-entrandos como se indica en la ecuación 3.9.

$$\max_{\Phi} \sum_{(x,y)\in Z} \sum_{t=1}^{|y|} \log(P_{\Phi}(y_t|x, y_{<t})) \tag{3.9}$$

- Φ representa el conjunto de parámetros para el modelo de lenguaje.
- Z es el conjunto de datos de entrenamiento que consta de pares de contexto y objetivo (x, y).
- x e y son secuencias de $tokens$, que representan el texto de entrada y el texto objetivo, respectivamente.
- t es el índice del $token$ en la secuencia y.
- $P_{\Phi}(y_t|x, y_{<t})$ el modelo pre-entrenado representado por la probabilidad condicional de predecir el $token$ y_t dado el contexto x y los $tokens$ generados previamente $y_{<t}$.

Siendo $P_{\Phi}(y_t|x, y_{<t})$ el modelo pre-entrenado, para poder ajustarlo a una tarea en concreto, es necesario ajustar los pesos del modelo entero, lo cual se traduce en una tarea ardua en términos computacionales, energéticos y económicos, haciendo este proceso ineficiente. En consecuencia, LoRA surge como un método más eficiente que utiliza una separación de la matriz de pesos pre-entrenados y un adaptador de pesos como se muestra en la figura 3.2. La ecuación 3.10 describe el adaptador.

$$\max_{\Theta} \sum_{(x,y)\in Z} \sum_{t=1}^{|y|} \log(p_{\Phi_0+\Delta\Phi(\Theta)}(y_t|x, y_{<t})) \tag{3.10}$$

- $\Delta\Phi(\Theta)$ es el adaptador que aplica el incremento de los parámetros específicos de la tarea que se está optimizando.

- Θ representa los parámetros de adaptación $\Delta\Phi(\Theta)$ que se optimizan para ajustar el modelo pre-entrenado a una tarea específica.

- $p_{\Phi_0+\Delta\Phi(\Theta)}(y_t|x, y_{<t})$ es el modelo pre-entrenado adaptado a una tarea específica.

El adaptador $\Delta\Phi(\Theta)$ se define como una matriz ΔW que incrementa a la matriz de pesos pre-entrenados W_x. A su vez, ΔW se descompone en dos matrices de menor rango WA y WB. Al comienzo del entrenamiento, la matriz de rango bajo WA se inicializa con una distribución gaussiana aleatoria, mientras que la matriz de rango bajo WB se inicializa con cero. Lo anterior significa que el producto de WB y WA, corresponde a la aproximación de bajo rango de la matriz de adaptación ΔW, que también es cero.
El motivo de esta inicialización es garantizar que la matriz de adaptación ΔW comience con una magnitud pequeña y aprenda gradualmente las características específicas de la tarea durante el entrenamiento. Al inicializar la matriz de rango bajo WB en cero, se establece la aproximación inicial de la matriz de adaptación ΔW. Esto significa que el modelo comienza con los pesos previamente entrenados debido a que W_x y ΔW se suman, permitiendo que ΔW se adapte gradualmente a las características específicas de la tarea, aprendiendo la matriz WB de rango bajo durante el entrenamiento.
Para conocer los parámetros específicos de la tarea, la matriz de rango bajo WB se escala mediante un factor de $\frac{\alpha}{r}$, donde α es una constante y r es el rango de la aproximación de rango bajo. Esta escala ayuda a garantizar que la aproximación de rango bajo de la matriz de adaptación ΔW aún amplifique las características importantes para la tarea posterior, incluso cuando el rango r varía.
Ajustar el factor de escala α es aproximadamente equivalente a ajustar la tasa de aprendizaje cuando se utiliza el optimizador Adam, siempre y cuando la inicialización se escale adecuadamente. Por lo tanto, Hu et al. (2021) establecen α en el primer valor de r que prueban y no lo modifican más. Esto ayuda a reducir la necesidad de reajustar los

hiperparámetros al variar el rango de r.

El tamaño de la matriz de rango bajo WB está determinado por el número de filas en la matriz de peso original que se está adaptando. Por ejemplo, si la matriz de peso original W tiene dimensiones AxB, entonces la matriz de rango bajo WB tiene dimensiones Bxr, donde r es el rango de la aproximación de rango bajo, mientras que WA tiene dimensiones rxA como se muestra en la figura 3.3.

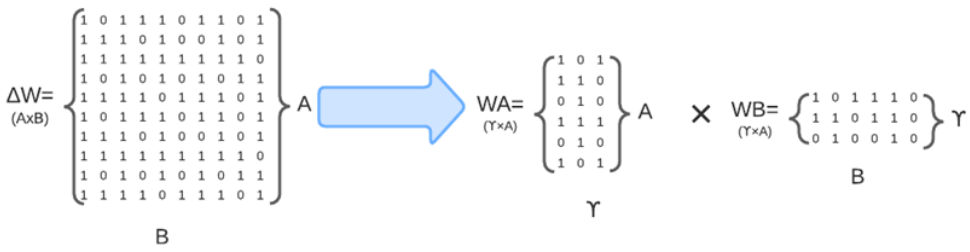


Figura 3.3: Descomposición de la matriz ΔW en matrices de menor rango

3.2. *Train* y *Validation Loss*

El valor de pérdida es el resultado de la función de pérdida, que evalúa la diferencia entre los valores reales y los valores predichos de la red neuronal. Este valor es esencial en la optimización del modelo, pues permite hacer los ajustes necesarios a los pesos del modelo para refinar las respuestas del modelo hasta que ese valor sea el mínimo posible. El valor ideal en cualquier modelo sería 0, pues hablaría de un modelo que no genera ningún error a partir de una entrada dada. Sin embargo, en el caso de modelos complejos como son los modelos del lenguaje, la meta realista es alcanzar los valores más pequeños posibles.

Valor de pérdida en el conjunto de entrenamiento. La pérdida es una métrica primordial en el mundo del *deep learning*, que permite visualizar que tan bien o mal, un modelo se está adaptando a los datos de entrenamiento. Esto se conoce como valor de pérdida sobre el conjunto de entrenamiento (*training loss*). Esta evaluación se hace cuando la red neuronal termina de procesar cada uno de los lotes de datos (*batch*) del conjunto de datos de entrenamiento. Es decir, cada uno de los lotes de datos de entrenamiento genera una pérdida, entonces al terminar una época de entrenamiento se calcula el promedio de todas las pérdidas obtenidas en cada uno de los lotes. Sin embargo, el obtener valores pequeños de pérdida en el entrenamiento, no significa necesariamente que todo está bien con el modelo, pues puede esconder un sobre ajuste, un sesgo en los datos o algún error en la métrica de evaluación, por lo que es muy importante un subconjunto para probar al modelo. Cada modelo tiene un propósito diferente, pero los modelos del lenguaje son usados normalmente para clasificar o predecir *tokens* dada una entrada nueva, que a pesar de sus similitudes al conjunto de entrenamiento, no son exactamente iguales. Por lo tanto, el modelo debe de ser capaz de procesar nuevos datos y poder inferir de manera correcta nueva información.

Valor de pérdida en el conjunto de validación. El valor de pérdida sobre el conjunto de datos de validación (*validation loss*) es otra medida de control. Este conjunto de datos se utiliza para ver si la pérdida se comporta de la misma manera que el conjunto de datos de entrenamiento. Este conjunto de datos no es conocido por la red neuronal, puesto que es diferente al conjunto de datos de entrenamiento. Es decir, la pérdida obtenida en el conjunto de datos de validación es una guía para asegurar que el modelo no solo está memorizando los datos de entrenamiento, si no, que además, sea capaz de inferir nuevos datos de la misma naturaleza. A continuación se describen los posibles comportamientos e interpretaciones de la pérdida obtenida en los conjuntos de entrenamiento y validación.

Good Fit, *Overfitting* y *Underfitting*

Good Fitting es el término que se le da a un escenario en el que el modelo tiene un valor de pérdida similar sobre los dos conjuntos de datos de entrenamiento y evaluación (Science and Science, 2023). Idealmente, el comportamiento óptimo del valor de pérdida en ambos conjuntos de datos, es de decremento (figura 3.4).

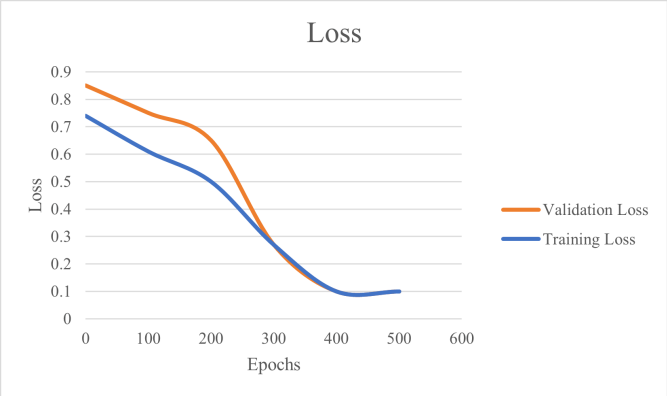


Figura 3.4: *Good Fit*, ejemplo ilustrativo extraído de (Science and Science, 2023)

Underfitting es el término que se usa para describir a un modelo de *machine learning* que no se ajusta lo suficiente a los datos de entrenamiento y por lo tanto no captura bien las relaciones subyacentes (ver la primera mitad de la figura 3.5).

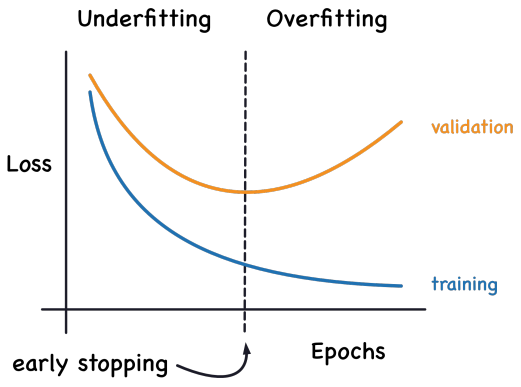


Figura 3.5: Ejemplo ilustrativo de*underfitting* y *overfitting* extraído de (Ryanholbrook, 2023)

Overfitting es el caso contrario al *underfitting*. *overfitting* es cuando un modelo se ajusta demasiado a los datos de entrenamiento, incluyendo el ruido, y no generaliza bien a los nuevos datos (ver la segunda mitad de la figura 3.5).

3.3. El Conjunto de Datos

En está sección se describen los diferentes conjuntos de datos creados durante el proceso de elaboración de la presente tesis hasta llegar al conjunto de datos definitivo usado para entrenar el modelo. El primer conjunto de datos se compone de diversas fuentes centradas en el análisis de emociones, con el propósito de dotar al modelo con el conocimiento de identificación de emociones (ver sección 3.3.1). El segundo conjunto (ver sección 3.3.2) de datos se usó para entrenar al modelo a generar respuestas adecuadas en una conversación orientada al acompañamiento emocional. Cada respuesta debe demostrar empatía por la situación de la persona y si es posible, alentar a el individuo a sentirse cómodo con el propósito de seguir hablando de su sentir.

3.3.1. *HRECPW: Hispanic Responses for Emotional Classification based on Plutchik Wheel*

Preparación de los Datos

Los datos que se obtienen de Twitter u otras redes sociales requieren un preprocesamiento para garantizar la privacidad de la información. Esto implica eliminar cualquier dato personal, y la eliminación de enlaces a sitios web que no sean relevantes para el análisis. Para el entrenamiento del modelo, es necesario segmentar el conjunto de datos, destinando la mayor parte al entrenamiento y una menor proporción a la validación y las pruebas. Por lo tanto, la proporción que usamos fue en este caso $70\,\%$ para el conjunto

de datos de entrenamiento, $20\,\%$ y $10\,\%$ para los datos de prueba y validación, respectivamente. La distribución de datos anterior se decidió así para maximizar el uso de los datos disponibles sin comprometer la calidad de la evaluación.

Recopilación de Fuentes de Información para el Conjunto de Emociones

Para la construcción del conjunto de datos que será utilizado a lo largo de los experimentos, se llevó a cabo una exhaustiva recopilación de distintas fuentes en internet, enfocadas en el análisis de emociones y sentimientos en el lenguaje. Estas fuentes fueron seleccionadas con el objetivo de proporcionar una diversidad de ejemplos y contextos que enriquezcan el conocimiento del modelo Llama en la identificación y generación de expresiones emocionales. A continuación, se detallan las principales fuentes que contribuyeron a la construcción del conjunto de datos:

- **TweetEval** (Barbieri et al., 2020) ofrece un conjunto de datos diverso con múltiples propósitos de evaluación. En este estudio, nos centramos en el análisis de emociones, que incluye etiquetas para distintas categorías emocionales, como ira, alegría, optimismo y tristeza.

- **DailyDialog** (Li et al., 2017) es una fuente que contribuye con diálogos cotidianos en varios contextos. Aunque no se utilizó en su totalidad, ciertas partes del conjunto de datos aportaron ejemplos valiosos de expresiones emocionales en conversaciones informales.

- **HappyDB** (Asai et al., 2018) proporciona una perspectiva única al contener expresiones relacionadas con la felicidad y emociones positivas en situaciones cotidianas. Algunos fragmentos seleccionados se incorporaron al conjunto de datos para enriquecer la diversidad emocional.

- **Emotion**(Saravia et al., 2018) contiene mensajes de Twitter en inglés etiquetados con seis emociones básicas: ira, miedo, alegría, amor, tristeza y sorpresa. Estas etiquetas ofrecen un enfoque detallado en la clasificación de emociones.

- **Encuestas realizadas a** 72 **personas**: Se le pidió a 72 personas que respondieran preguntas sobre 11 emociones, proporcionando respuestas vinculadas a estas emociones o a experiencias que las hayan evocado.

Técnicas de muestreo

Se utilizó un muestreo aleatorio de todos los conjuntos de datos mencionados en la sección 3.3.1 con el objetivo de obtener una muestra representativa de la población objetivo (Económica, 2022). Al seleccionar los ejemplos de muestreo de manera aleatoria, se minimiza el sesgo y se maximiza la probabilidad de capturar la variabilidad presente en todo el conjunto de datos.

Utilización de Etiquetas Contextualizadas

Dado que Llama opera como un generador de texto altamente contextual, se considera inapropiado recurrir a las convencionales etiquetas numéricas, como 0 y 1, ya que carecen de relevancia en el contexto del lenguaje natural. En lugar de esto, se optó por un enfoque más expresivo y coherente al implementar un modelo basado en consultas y respuestas. Este enfoque se alinea de manera más efectiva con la comprensión y generación de lenguaje humano enriquecido emocionalmente.
En esta estrategia, cada muestra de texto es tratada como una "consulta", en la cual se plantea una situación, experiencia o pensamiento específico. La respuesta asociada a esta consulta es la "emoción". Por ejemplo:

- “Consulta: <Mi tan esperado y deseado viaje prolongado de Velankanni...> emocion: ”

La estructura del *prompt* utiliza “Consulta” y “emoción” como palabras clave. Además de los tokens “<” y “>” como indicativos de inicio y fin de una oración, respectivamente.

El uso de Instrucciones o *Prompts*

En la utilización efectiva de los LLMs, el *prompting* es una técnica indispensable (Cao et al., 2024; Huang et al., 2024). En el campo de la inteligencia artificial, un *prompt* se refiere a una consulta o entrada del usuario que instruye al modelo de IA a seguir un comportamiento o una serie de pasos para obtener una respuesta específica (Yasar, 2023). La técnica de *prompting* define la estructura del *prompt* (Saravia, 2022), y existen muchas variaciones con diferentes propósitos. Entre las técnicas más usadas está *Zero-Shot* (Saravia, 2022), que consiste en no brindar más información al modelo que la propia instrucción, sin ejemplos adicionales. Por otro lado, *Few-Shot* (Saravia, 2022) es la contraparte de *zero-shot*; mientras que *zero-shot* puede fallar en tareas más elaboradas y complejas de generalizar con el contexto natural del modelo, *few-shot* proporciona ejemplos al modelo para que desempeñe la tarea, brindándole contexto adicional.

Para este proyecto de investigación, se utilizará una técnica distinta, más simple pero efectiva para los propósitos de la tarea que se espera que el modelo realice. La técnica en cuestión es conocida como *instructional prompting* (Chen, 2024), y funciona como una especie de guía o punto de partida que da contexto al modelo sobre lo que se espera que haga.

La instrucción y el contexto en el ejemplo 3.1 le brindan al modelo más detalles sobre la tarea que llevará a cabo. La estructura del *prompt* del conjunto de entrenamiento, está inspirado en el conjunto utilizado en el modelo basado en Llama-1 de Stanford

nombrado Alpaca (Taori et al., 2023). El propósito de Alpaca era obtener un modelo de capacidades similares a ChatGPT, el cual está basado en el modelo GPT-3(Brown et al., 2020). A diferencia de Alpaca, nuestro modelo está ajustado a la identificación de emociones, para posteriormente responder de forma empática al usuario.

"Below is an instruction that describes a task, paired with an input that provides further context. ### instruction: You are an emotion recognition system. From the following input in Spanish, analyze its content and determine which of the following emotions fit the text of the input. Emotions in Spanish: [afecto, admiración, alegría, ira, optimismo, tristeza, odio, miedo, sorpresa, calma, asco]. ### input: Le regalé a mi papá un collage de ... ### emotion: afecto</s>"

Tabla 3.1: Instrucción para clasificación de emociones

La instrucción utilizada en nuestro modelo, fue basada en la información expuesta en el artículo de Brown et al. (2020): los modelos del lenguaje son buenos infiriendo información si se les instruye de forma específica en el *prompt* incluso si no han sido ajustados para dicha tarea. En este caso, estamos condicionando al modelo a actuar sólo como un sistema de identificación de emociones con respuestas cerradas a un dominio de valores, tales como afecto, admiración, alegría, ira, optimismo, tristeza, odio, miedo, sorpresa, calma y asco.

Además, el uso de palabras clave como "### *instruction*:' 'y "### *emotion*" utiliza una situación que el codificador replica indistintamente de las veces que se utilice, marcando así la pauta para que el modelo siempre se comporte de la misma forma al ver estas palabras y condicionándolo a responder de la misma manera. Con cada iteración del entrenamiento este comportamiento se refuerza.

La siguiente sección explora cómo los conjuntos de datos *DailyDialog* y *HappyDB* contribuyen a mejorar la generación de texto con contenido emocional.

Clasificación de Emociones según la Rueda de Plutchik

Además de las fuentes de datos previamente mencionadas, se incorporó una clasificación de emociones propuesta por Plutchik (Plutchik, 1980), conocida como la rueda de Plutchik. Esta clasificación ofrece una perspectiva profunda y matizada de las emociones humanas al dividirlas en ocho emociones primarias y sus combinaciones.

La Rueda de Plutchik identifica las siguientes emociones primarias: alegría, confianza, miedo, sorpresa, tristeza, disgusto, ira y anticipación. Cada emoción primaria se encuentra en un punto cardinal de la rueda y puede combinarse con emociones vecinas para formar emociones secundarias, lo que enriquece la comprensión de las sutilezas emocionales en el lenguaje humano. La Rueda de Plutchik se muestra en la figura 3.6.

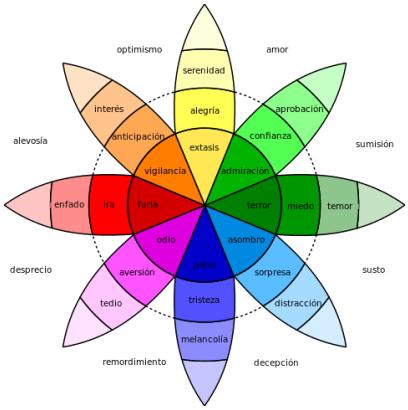


Figura 3.6: Rueda de Plutchik

Por lo tanto, las muestras del conjunto de datos se etiquetaron siguiendo la clasificación de emociones de la rueda de Plutchik. Cada muestra se asignó a una emoción primaria o secundaria según el contenido emocional expresado en el texto. Este etiquetado enriquece la capacidad del modelo Llama-2 para reconocer y generar una variedad de emociones en el texto generado.

Las muestras primarias consideradas para el etiquetado de datos son la **alegría**, el **miedo**, la **sorpresa**, la **tristeza**, el **asco** y la **ira**. Y las emociones secundarias consideradas

son la **admiración**, la **calma**, el **optimismo**, el **afecto** y el **odio**.

La incorporación de la clasificación de emociones de la rueda de Plutchik permite que el conjunto de datos abarque una amplia gama de expresiones emocionales, desde las emociones básicas hasta las más complejas y mezcladas. Esto proporciona al modelo la capacidad de generar respuestas contextualmente apropiadas y emocionalmente ricas.

Finalmente, se combinaron las muestras de las distintas fuentes de información, seleccionando solo aquellas que contribuyen al conjunto de datos y se alinean con las emociones identificadas en la rueda de Plutchik. Posteriormente, se dividió el conjunto de datos, en $70\,\%$ datos de entrenamiento, $20\,\%$ datos de prueba, y $10\,\%$ datos de validación. La tabla 3.2 describe la distribución específica del conjunto de datos.

Primera versión			
Emoción	**Entrenamiento**	**Validación**	**Prueba**
afecto	32837	4221	9478
alegría	25001	3014	7170
admiración	23809	2487	4443
ira	8322	971	1954
tristeza	6490	847	1252
optimismo	6304	718	1191
odio	3783	427	1065
sorpresa	2163	223	422
miedo	2083	187	241
calma	1252	169	182
asco	303	3	47

Tabla 3.2: Primera versión de conjuntos de datos de entrenamiento, validación y prueba

Balanceo de los Conjuntos de datos

A lo largo de la experimentación, se encontraron algunos problemas en el proceso de *fine-tuning* del modelo. Uno de ellos fue el desbalanceamiento del conjunto de datos. Más de la mitad de las categorías emocionales presentan un desbalance considerable.

Para compensar este problema se aplicó la técnica de *oversampling*, la cual consiste en obtener más muestras sintéticas (Andrews, 2021) de las categorías subrepresentadas hasta igualarlas con las categorías más grandes del conjunto. Esto supuso un reto, pues en el conjunto de entrenamiento la categoría con más elementos era afecto con más de 32 mil muestras, seguida por alegría y admiración que cuentan con más de 25 mil y 23 mil muestras, respectivamente. Mientras que la categoría con menor número de elementos era asco con 303 elementos. Aplicar un *oversampling* con tremenda desigualdad, sería costoso, así que se optó por aplicar *undersampling* a las categorías con más de 11 mil muestras y aumentar las categorías subrepresentadas también a 11 mil. De esta forma solo se necesitaron generar alrededor de $48,500$ muestras sintéticas, en comparación a las más de 200 mil que se hubieran necesitado para balancear solo el conjunto de datos de entrenamiento. Este mismo principio se utilizó en los conjuntos de prueba y evaluación, limitándose a 120 y 200 muestras por categoría, respectivamente.

Para generar las muestras sintéticas, se utilizó el modelo GPT-3.5 (OpenIA, 2022) mediante la API de OpenAI, con un costo total aproximado de $\$15$.

Para generar las muestras sintéticas, se utilizó el modelo GPT-3.5 (OpenIA, 2022) a través de la API de OpenAI, con un costo aproximado de $\$15$ por todas las muestras generadas. Se empaquetaban 20 muestras en un solo $prompt$, y con una instrucción de 35 tokens (ver ejemplo en la tabla 3.3), se le indicó al modelo que generará 40 muestras a partir de las 20 muestras previas.

"*You have to generate 40 text samples related to the emotion 'surprise', all texts in Spanish, and numbered* as follows:

1. *text...*
2. *text...*

Take these examples as a reference:

1. "Me tomó por sorpresa la noticia de que mi equipo favorito había ganado el campeonato."
2. "Me llevé una gran sorpresa al ver que mi amigo de infancia se convertiría en el nuevo alcalde de nuestra ciudad."
3. "No puedo creer lo que acabo de presenciar. Es asombroso."
4. "La sorpresa en la cara de mi madre al recibir un ramo de flores fue encantadora".
5. "Fue increíble..."

Tabla 3.3: *Prompt* para generar muestras sintéticas

El $98\,\%$ de las muestras generadas fueron de alta calidad, de amplio léxico y contextos variados. También se observó que las muestras basadas en información real tenían una calidad superior a las que se fundamentaban únicamente en el contexto de la emoción tristeza. El $2\,\%$ restante consistía en muestras incompletas o carentes de originalidad, por lo que aportaban poco al conjunto de datos y fueron descartadas.

La tabla 3.4 describe el conjunto de datos balanceado en los tres conjuntos de datos, entrenamiento, validación y prueba.

Versión Definitiva			
Emoción	**Entrenamiento**	**Validación**	**Prueba**
afecto	11000	200	120
alegría	11000	200	120
admiración	11000	200	120

ira	11000	200	120
tristeza	11000	200	120
optimismo	11000	200	120
odio	11000	200	120
sorpresa	11000	200	120
miedo	11000	200	120
calma	11000	200	120
asco	11000	200	120

Tabla 3.4: Versión definitiva de los conjuntos de datos de entrenamiento y validación.

3.3.2. *HEAR Dataset*: *Hispanic Emotional Accompagnement Responses*

El conjunto de datos de respuestas de acompañamiento emocional hispano, conocido como *HEAR* por sus siglas en inglés, está diseñado para entrenar modelos de lenguaje modernos, como GPT-3, en la tarea de brindar apoyo emocional a los usuarios. Este conjunto de datos permite a los modelos generar respuestas empáticas, demostrando comprensión de los sentimientos y del contexto, y ayudando así a los usuarios a afrontar situaciones emocionales complejas de diversas índoles.

La creación de respuestas sintéticas se llevó a cabo de manera similar al caso anterior, empleando la API de OpenAI con el modelo GPT-3.5(OpenIA, 2022). En este proceso, se desarrolló un sistema (consulte la figura 3.7) que incluye un número específico de muestras por $prompt$. Al diseñar el $prompt$, también se tuvo en cuenta no exceder el límite máximo de tokens en la ventana de contexto del modelo utilizado, que en el caso de GPT-3.5 es de 4096 tokens por solicitud al servidor. Adicionalmente, se construyó

un *prompt* que le permitiera al sistema delimitar el papel que juega en la conversación con el usuario.

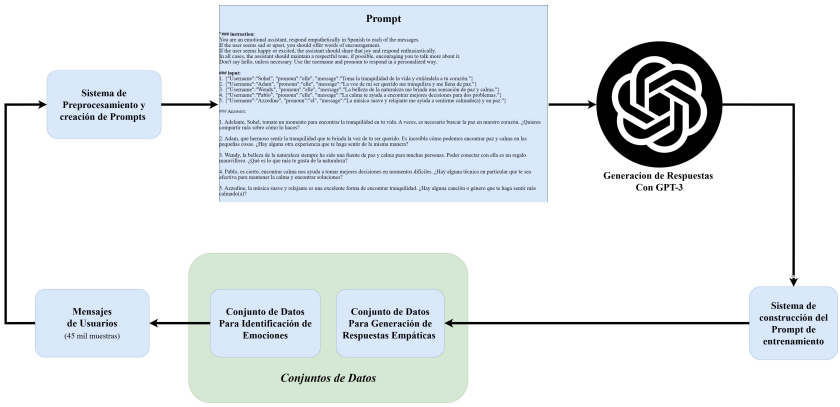


Figura 3.7: Proceso de creación del conjunto de respuestas empáticas

Prompt de Generación de Respuestas Empáticas

Cuando se busca obtener una respuesta específica de un modelo como GPT-3 o Llama-2, se requiere construir un *prompt* que sea claro con lo que se requiere del modelo. En el caso de obtener respuestas empáticas, y sensibles a las emociones que puede contener un mensaje, es preciso delimitar el papel que el modelo juega en esta conversación. Con lo anterior en mente se escribió el *prompt* mostrado en la tabla 3.5.

"*Below is an instruction that describes a task, paired with an input that provides further context.* *### instruction:* *You are an emotional assistant, respond empathetically in Spanish to each of the messages. If the user seems sad or upset, you should offer words of encouragement. If the user seems happy or excited, the assistant should share that joy and respond enthusiastically. In all cases, the assistant should maintain a respectful tone, if possible, encouraging you to talk more about it. Don't say hello, unless necessary. Use the username and pronoun to respond in a personalized way.* *### input:* *{"Username":"Sohel", "pronoun":"el", "message":.El miedo me paraliza cuando tengo que tomar decisiones importantes en mi vida."}* *### response:* "

Tabla 3.5: *Prompt* en inglés para generar las respuestas de *HEAR Dataset* limitando el comportamiento de GPT-3.5 (OpenIA, 2022)

En la instrucción de la tabla 3.5 se incluye el texto "*You are an emotional assistant*", lo que limita el comportamiento del modelo al definirlo como un asistente emocional destinado a ayudar a las personas en situaciones complicadas. Sin embargo, esta descripción es demasiado general para abordar adecuadamente una situación emocional. Para ser más específico, se indica: "*If the user seems sad or upset, you should offer words of encouragement*", lo que establece un patrón más claro. Si el usuario se muestra triste o enojado, el modelo debe utilizar palabras de aliento que reflejen amabilidad y respeto en el contexto de la situación.

***Prompt* De Entrenamiento de la Base de Datos HEAR**

A diferencia del *prompt* utilizado en GPT-3.5 para la generación de muestras sintéticas del conjunto de datos HEAR (*Hispanic Emotional Accompaniment Responses*), el *prompt* empleado para instruir al modelo "Solo Escuchame" se simplificó con el objeti-

vo de reducir el número de tokens necesarios y, de esta forma, tener un contexto menor en la entrada del modelo. De esta manera, se utiliza un promedio de solo 201 tokens para todo el *prompt*, incluyendo tanto la instrucción como el contexto (*input*). El *prompt* utilizado se muestra en la tabla 3.6.

Below is an instruction that describes a task, paired with an input that provides further context. *### instruction:* *You are an emotional assistant, respond in Spanish in a respectful and appropriate way to the user's emotional situation. If the user appears sad or upset, the assistant should respond empathetically and offer words of encouragement.* *### input:* *{"Username":"Sohel", "pronoun":"el", "message":."El miedo me paraliza cuando tengo que tomar decisiones importantes en mi vida."}* *### response:Entiendo que el miedo te paralice al tomar decisiones importantes en tu vida. Recuerda que tomar decisiones es parte natural del crecimiento y que siempre hay opciones y apoyo disponible si te sientes abrumado. ¿Hay alguna decisión específica que te preocupe en este momento?</s> "*

Tabla 3.6: *Hispanic Emotional Accompagnement Responses* (HEAR). *Prompt* en inglés.

El token "</s>" al final del *prompt* indica al modelo que la oración ha terminado. Esto resulta útil durante la inferencia, ya que señala cuándo es pertinente finalizar una oración. Estas instrucciones limitan al modelo de manera conveniente para la generación de respuestas adecuadas.

Cambios en el Tamaño del *Dataset*

Como se mencionó anteriormente, para el conjunto de datos *HEAR* (disponible en huggingface[1]), se tomaron todos los elementos del conjunto *HRECPW* (sección3.3.1) pro-

[1] https://huggingface.co/datasets/BrunoGR/HEAR-Hispanic_Emotional_Accompaniment_Responses

puesto en este mismo trabajo, el cual cuenta con 11 categorías emocionales y 11 mil muestras por categoría.

Inicialmente, se seleccionaron aleatoriamente 900 muestras del conjunto *HEAR* para el subconjunto de entrenamiento, mientras que para los subconjuntos de pruebas y validación se utilizó el $100\,\%$ de sus elementos (tabla 3.4). Esto se hizo con el propósito de asegurar que las 11 categorías tuvieran la misma representación y evitar problemas de desequilibrio como en la primera versión de *HRECPW* (sección 3.3.1). A continuación se describe la evolución de los cambios realizados en la base de datos *HEAR*.

- En su primera versión, *HEAR* tuvo solo $9,900$ elementos en el subconjunto de entrenamiento, manteniendo el balance desde su concepción.

- Debido a los resultados del tercer *fine-tuning*, se requirió aumentar el número de elementos en el subconjunto de entrenamiento. Se realizó un nuevo muestreo aleatorio del conjunto original (*HRECPW*), descartando los elementos ya utilizados. En esta ocasión, se tomaron $1,450$ elementos por categoría (un total de $15,950$ elementos), es decir, un aumento de 550 elementos respecto a la primera versión.

- Los resultados de la cuarta prueba mostraron una mejora significativa, lo que llevó a aumentar por última vez el número de elementos en el subconjunto de entrenamiento. Se incrementó de $1,450$ a $3,810$ elementos, lo que representa un aumento del $323.3\,\%$ respecto a la primera versión de $9,900$ elementos.

Este proceso resultó en un tamaño total de $45,450$ elementos para el conjunto *HEAR*, siendo el subconjunto de entrenamiento más significativo, con $41,910$ elementos en su versión final.

3.4. Fine-Tuning del Modelo de Identificación de Emociones

Con el objetivo de que Llama-2 pueda identificar y clasificar correctamente las emociones a partir de las consultas de los usuarios, es necesario ajustar el modelo a esta tarea específica. Cada una de las siguientes pruebas se evaluará en función del valor de pérdida tanto en el conjunto de entrenamiento como en el de validación.

3.4.1. Primer *Fine-Tuning*

En el presente experimento, se entrena un modelo de lenguaje llamado Open Llama (Geng and Liu, 2023). Llama es una familia de modelos de lenguaje con diferentes versiones basadas en el número de parámetros. En los experimentos que se describen a continuación, se utilizó la versión libre de 7 mil millones de parámetros, que es una recreación del trabajo original de *Meta* y muestra un rendimiento equivalente al modelo original de *Meta* LlaMA (Touvron et al., 2023b). Para su implementación, se empleó la biblioteca Huggingface's Transformers (Wolf et al., 2020). El objetivo es realizar análisis de emociones en consultas de texto.

Para el primer experimento, se utilizó LoRA (Hu et al., 2021), una técnica de aprendizaje automático que reduce el tamaño de los modelos de lenguaje al transformar las matrices de peso del modelo a un formato de menor precisión. Esto permite que los modelos se ejecuten más rápido y usen menos memoria. A continuación, se describen los parámetros utilizados en los experimentos correspondientes a la primera etapa.

Configuración del *Tokenizador* y Función de Mapeado de Datos

Para la utilización de los datos a lo largo del ajuste que se le hará al modelo, es necesario hacer una última preparación o preprocesamiento. En este caso se usa el *tokenizador* del modelo Llama, al cual se le tienen que hacer algunos ajustes en sus parámetros al momento de codificar los textos de nuestros datos. Esto se hace debido a que el conjunto de datos se encuentra en un *dataframe* de pandas, que tiene las columnas de "texto" y "etiqueta". Estas columnas deben ser unidas en una cadena que será codificada para que pueda ser utilizada para entrenar al modelo.

- **Función generate_prompt**. Esta función toma un punto de datos (*data_point*) como entrada y genera un "*prompt*"formateado que se utilizará para guiar al modelo en la generación de texto. La idea detrás de este *prompt* es proporcionar una estructura constante para todas las entradas. La función concatena la instrucción Ïnstrucción: Analiza la Consulta, y determina la Emoción.çon el texto del punto de datos y su etiqueta correspondiente. Lo anterior establece una guía clara para que el modelo comprenda la tarea y el contexto de generación.

- **Función *tokenize***. Esta función se encarga de tokenizar el *prompt* generado por *generate_prompt* para prepararlo para el modelo. Utiliza el tokenizador para realizar la tokenización, asegurando que los textos cumplan con la longitud máxima definida de 1024 (*CUTOFF_LEN*). Cualquier texto con mayor número de *tokens* será truncado a esta longitud para asegurarse de que se ajuste dentro de los requisitos del modelo. Además, verifica si el último *token* es el *token* de fin de secuencia (eos_token) y, si no lo es y la longitud está por debajo de *CUTOFF_LEN*, agrega el *token* de fin de secuencia. Luego, crea una máscara de atención correspondiente. Finalmente, asigna las etiquetas como los *tokens* de entrada tokenizados.

- **Función *generate_and_tokenize_prompt***. Esta función combina las etapas anteriores. Toma un punto de datos, genera un *prompt* mediante *generate_prompt* y luego tokeniza este *prompt* utilizando la función *tokenize*. El resultado final es un *prompt* tokenizado y listo para ser utilizado en la generación de texto.

Por defecto en los argumentos de *Training Arguments*, esta predefinido el *scheduler* como lineal. El ajuste de la tasa de aprendizaje decae de forma lineal, empezando en los primeros pasos en 84×10^{-4}, hasta llegar en los últimos pasos a 9.68×10^{-8}.

Resultados del *Fine-Tuning*

Debido a las restricciones de tiempo en el entorno de trabajo, el proceso de ajuste del modelo se llevó a cabo en múltiples fases. Siguiendo los parámetros previamente detallados, se implementó una estrategia que utiliza puntos de control. Estos puntos de control permitieron reanudar el proceso de ajuste desde la posición en la que el entrenador había registrado la última actualización. En cada ejecución del *script* de Python, se recuperó el estado de entrenamiento desde el punto de control más reciente.

Para monitorear el proceso de ajuste, se empleó la plataforma Wandb [2]. Esta herramienta proporciona un seguimiento constante del *train loss* (o en español pérdida en la etapa de entrenamiento) del modelo en cada iteración del proceso. Esta métrica es fundamental para evaluar el rendimiento y la convergencia del modelo durante el ajuste.

Como se muestra en la figura 3.8, los valores de pérdida en el conjunto de entrenamiento en los diferentes puntos de control, representados por las líneas punteadas de múltiples colores, reflejan la evolución del rendimiento del modelo durante el proceso de ajuste. Las líneas continuas representan los valores de pérdida en el conjunto de evaluación, lo que permite visualizar cómo el modelo se comporta con nuevos datos. Es de esperarse

[2] https://wandb.ai/

que los valores de pérdida sean altos en los primeros pasos del entrenamiento, pero que disminuyan a lo largo de la primera época a medida que el modelo se ajusta. En este experimento, se observa una disminución en la primera mitad de la primera época, pero después de este punto, el modelo se estanca en valores entre 1.2 y 0.9. Esta primera prueba se interrumpió debido al parámetro *early_stopping_patience*, configurado en 5 oportunidades para evitar un *overfitting*.

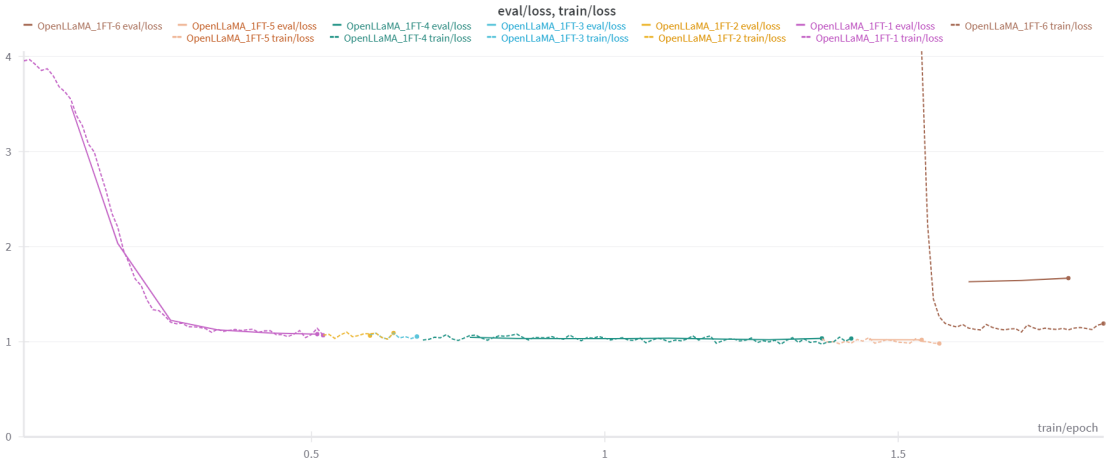


Figura 3.8: Valores de pérdida del primer *fine-tuning*

Deficiencias del Experimento

Uno de los problemas con este primer experimento es la forma en la que se realizan las consultas y se codifican, existiendo consultas que superan por mucho los 1024 *tokens* permitidos por el *tokenizador*, y dejando fuera la respuesta de la emoción. Por lo tanto, lo anterior afecta seriamente el rendimiento del modelo al momento de hacer la predicción de la emoción.

El mayor desafío al finalizar esta prueba es identificar qué está limitando al modelo a obtener valores de pérdida más bajos en ambos conjuntos. En esta etapa de la experimentación, pueden ser varios los factores que impiden un buen ajuste del modelo, por

ejemplo, un desbalance en el conjunto de datos, un *datacollator* que no esté funcionando correctamente o la cantidad de datos. Más experimentos ayudarán a identificar más deficiencias.

3.4.2. Segundo *Fine-Tuning*

Basado en el experimento anterior, en este experimento se realizaron varios ajustes en el entrenamiento y la preparación de los datos. Es decir, no se descartaron elementos significativos del *script* del experimento, pero se hicieron cambios en los parámetros del *Trainer*, y restricciones en el tamaño del texto en cada muestra del conjunto de datos de entrenamiento y en la evaluación 3.4.2. La experimentación del segundo fine-tuning recae en un cambio de parámetros y del *prompt*.

Cambios en la generación de *Prompts* y Tokenización

Para no omitir la respuesta del análisis de emociones, se realizó un ligero cambio en la forma de generar el *prompt*, limitando el tamaño de la consulta a 2460 caracteres, dejando suficiente espacio para los últimos 4 tokens necesarios para la respuesta después de la palabra clave “Emocion:” o “‘Sentimiento:”. Es decir, todos los textos con menor longitud a 2460 caracteres, no serían truncados, mientras los que superan este número de caracteres, serían truncados. De esta forma no se eliminan elementos que podrían ayudar a condicionar al modelo a responder con el análisis de emociones dada una consulta.

Ajuste de Parámetros de Entrenamiento

Se incrementó el *learning rate* con el objetivo de acelerar el ajuste del modelo y observar cambios en las etapas más tempranas de cada época. Como consecuencia, se aumentó el número de épocas de 3 a 6, permitiendo al modelo tiempo suficiente para ajustarse

o, en su defecto, desajustarse. Esto permitió evaluar la mejor configuración en términos de épocas y *learning rate*. Además de cambiar el *lr_scheduler_type* de lineal a coseno, este parámetro regula el ajuste del *learning rate* durante el entrenamiento, de este modo la varianza de este parámetro se vuelve más fino al aumentar o disminuir su valor. El valor del resto de parámetros se pueden consultar en la tabla 3.7.

Parámetros del Trainer	1° F-T	2° F-T	3° F-T
Batch	64	64	256
Micro_Batch	8	16	16
Warmup steps	2000	2000	200
LEARNING_RATE	$2e^{-4}$	$4e^{-5}$	$5e^{-5}$
Épocas	3	6	6
EVALUATION_STRATEGY	"*steps* "	"*steps* "	"*steps* "
SAVE_STRATEGY	"*steps* "	"*steps* "	"*steps* "
eval_steps	1000	300	100
save_steps	1000	300	100
save_total_limit	3	4	6
lr_scheduler_type	"*linear* "	"*cosine* "	"*cosine* "
Parámetros de LoRA			
LoRA_R	8	8	16
LoRA_ALPHA	16	16	32
LoRA_DROPOUT	0.05	0.05	0.05
Resultados			
Train Loss	0.9869	0.98	0.1852
Val Loss	1.2	0.99	0.2419

Tabla 3.7: Tabla comparativa de los parámetros de cada experimento y sus valores de pérdida durante el *fine-tuning* (FT)

Resultados

El experimento mostró que el modelo no logró una buena adaptación a los datos de entrenamiento, pues el valor de pérdida en ambos conjuntos de datos se mantuvo entre 1.03 y 0.99 durante el proceso (ver figura 3.9). Al alcanzar el paso 9610, se evidenció un desajuste en los pesos de la red, dando como resultado el peor rendimiento del experimento, lo cual puede ser debido al sobre-ajuste. El sobre-ajuste en las redes neuronales es un comportamiento no deseado que ocurre cuando el modelo de aprendizaje automático se ajusta demasiado a los datos de entrenamiento y no generaliza bien con nuevos datos (Amazon Web Services, sf).

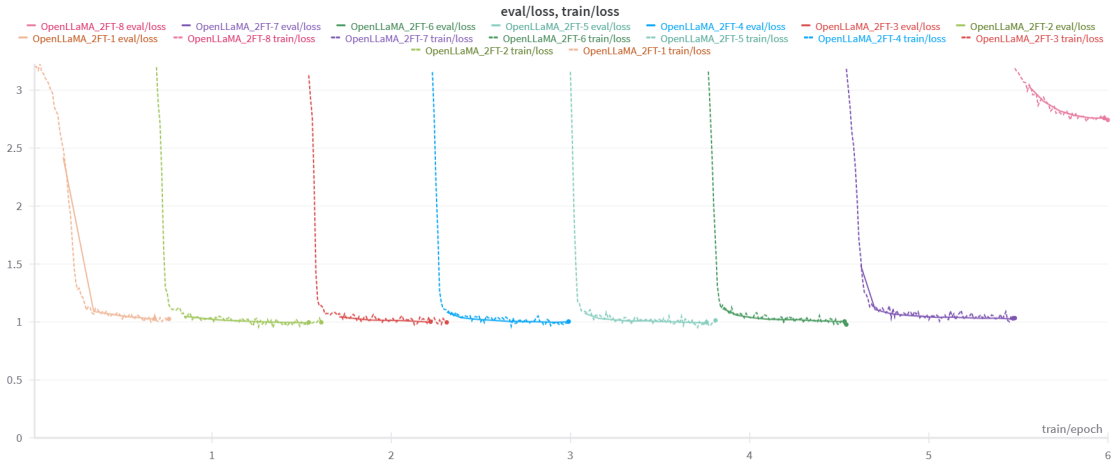


Figura 3.9: Segundo *fine-tuning*

Deficiencias del Experimento

Uno de los mayores problemas en este experimento, al igual que en el anterior, es el conjunto de datos, ya que la cantidad de datos disponibles en cada categoría está desbalanceada. Por ejemplo, la categoría de afecto cuenta con 32 mil muestras, la de alegría con 25 mil, mientras que asco solo tiene 303. Esto provoca que las categorías con menos muestras estén subrepresentadas en los pesos del modelo, lo que impide un rendimiento

óptimo en la inferencia. Se necesita realizar un proceso de *undersampling*, que limitaría la cantidad de muestras de todas las categorías al número de la categoría con menos datos, en este caso 303. Sin embargo, esto resultaría en un conjunto de entrenamiento muy reducido. Una mejor solución sería el *oversampling*, aunque esto podría presentar más desafíos.

3.4.3. Tercer *Fine-Tuning*

En este tercer intento se busca maximizar los resultados al hacer ajustes con respecto a los experimentos anteriores, corrigiendo deficiencias e indagando más en las alternativas disponibles. Para este experimento se decidió utilizar la versión más reciente de Llama-2 (Touvron et al., 2023c). A diferencia de Llama original(Touvron et al., 2023b), en esta versión META permitió al público tener un acercamiento con mayor apertura, lo que en un inicio había obligado al presente proyecto a utilizar openLlama (Geng and Liu, 2023), una reproducción del trabajo original.

Balanceo de Datos en el Conjunto y Cambio de *Prompt*

PPara esta tercera etapa se identificó que el conjunto de datos tenía un problema con la cantidad de muestras que contenía cada categoría de emociones. Esto significa que hay categorías menos representadas, y esto representa una dificultad para que el modelo pueda dicha emoción. Por lo tanto, se aplicaron técnicas de balanceo, como el *oversampling* explicadas en la sección 3.3.1.

Generación de *Prompts* y Tokenización

Para guiar al modelo en la generación de respuestas contextualizadas, se generan *prompts* que combinan la instrucción "*You are an emotion recognition system...*" con el texto de

las consultas y sus etiquetas emocionales. Estos *prompts* se ajustan a una longitud máxima de 1536 tokens, y para asegurarnos de que no se pasen de esta longitud, limitamos el contexto o *input* de la consulta a 2460 caracteres, asegurando que la respuesta y la emoción se incluyan adecuadamente en el prompt.

DataCollatorForCompletionOnlyLM

En este experimento se utilizaron nuevas herramientas de desarrollo optimizadas para Llama-2 (Touvron et al., 2023c), específicamente la clase *DataCollatorForCompletionOnlyLM* del *framework* trl (*Transformer Reinforcement Learning*). Esta clase se emplea para preparar los datos de entrada para el entrenamiento de modelos de lenguaje de completado de texto, permitiendo el uso de palabras clave incluidas en un *prompt*. En este caso, las palabras clave utilizadas fueron "### *emotion*: " e "### instruction". Además, se empleó el codificador del *tokenizer* de Llama-2 para codificar estas palabras clave, agrupando y alistando los datos del conjunto para su uso durante el entrenamiento.

Configuración de LoRA

Las modificaciones que se le hicieron a la configuración de LoRA con respecto a los experimentos anteriores son:

- "r"(número de factores): Se establece en 16, determinando la complejidad de la descomposición de las matrices de peso del modelo original. Esto permite aumentar el número de parámetros entrenables del modelo.
- "$lora\_alpha$": configurado en 32 para aplicar una moderada restricción en la regularización durante el proceso de factorización.

Argumentos del Entrenamiento

Los argumentos configurados para el entrenamiento del modelo son los siguientes:

- Tamaño del *batch* de entrenamiento = 256: se aumentó para disminuir los tiempos de entrenamiento.
- Tamaño del *micro_batch* de entrenamiento = 16: para que procese más textos en el GPU.
- Gradient accumulation steps= 16 debido al aumento del *batch* y *micro_batch*.
- Warmup steps= 200: se disminuyeron los pasos de calentamiento debido a que el modelo ya está pre-entrenado y no se necesita tanto calentamiento (Geeksfor-Geeks, 2024).
- Número de épocas de entrenamiento= 6
- eval_steps=100: se disminuyeron los pasos de evaluación, para observar el comportamiento del modelo con más frecuencia.
- save_steps=100: se redujeron los pasos de guardado, para evitar que se quede a medias el entrenamiento, debido al tiempo de ejecución máximo de paperspace (Van Dyne, know).

Aumentar el tamaño del *batch* permite que el modelo procese más datos en paralelo durante cada paso de entrenamiento. Esto mejora la eficiencia computacional y reduce el tiempo total requerido para completar todas las épocas de entrenamiento. Aunque requiere más memoria, se logra un balance óptimo que acelera el proceso de entrenamiento general sin comprometer la calidad del modelo. Dado que un mayor tamaño de *batch* y *micro_batch* pueden resultar en un uso elevado de memoria, la acumulación de

gradientes ayuda a mitigar este problema al dividir el cálculo en pasos más pequeños y manejables, debido a que la acumulación de gradientes permite simular un tamaño de *batch* mayor acumulando gradientes sobre múltiples *mini-batches* antes de realizar una actualización de los pesos del modelo.

DataCollatorForCompletionOnlyLM

DataCollator es un término que generalmente se utiliza en el contexto del aprendizaje automático y la minería de datos para referirse a una función o componente que se encarga de recopilar, organizar y preparar los datos antes de ser utilizados para entrenar un modelo o realizar análisis (Sahai, 2022). En el caso concreto de *DataCollatorForCompletionOnlyLM* es una herramienta de TRL (*Transformers Reinforcement Learning*) (von Werra et al., 2020), un framework de *huggingface* que permite entrenar modelos de lenguaje con aprendizaje por refuerzo.

Este *Data Collator* se utiliza para tareas de generación de texto donde el objetivo es completar una respuesta en función de una instrucción o contexto. En el caso de este experimento, la instrucción es:"### Emoción:". Esta simple configuración asegura que el cálculo de pérdida y el entrenamiento del modelo se realicen únicamente en función de las respuestas generadas por el asistente, excluyendo cualquier contenido previamente proporcionado.

Resultados

Figura 3.10: Tercer *fine-tuning*

Con este último experimento se observó que un conjunto de datos balanceado marca una diferencia en el rendimiento del modelo durante el entrenamiento. La pérdida en el conjunto de entrenamiento disminuyó significativamente, pasando de un valor de 0.986 a 0.185, y la pérdida en el conjunto de evaluación se redujo a 0.24 como se puede ver en la figura 3.10. Esto representa una mejora notable, especialmente cuando se evalúa el rendimiento con un conjunto de evaluación que tiene 200 muestras por categoría. Ahora tenemos una pérdida del $24\,\%$, en comparación con el $120\,\%$ o $98\,\%$ que se tenía con el conjunto de datos anterior, que no evaluaba equitativamente cada una de las 11 posibles emociones. Los cambios aplicados en este nuevo experimento fueron de gran ayuda. Aumentar el rango de la matriz de bajo rango utilizada en LoRA (Hu et al., 2021) aceleró los resultados al permitir un mayor ajuste de los pesos del modelo. Además, el uso del *DataCollatorForCompletionOnlyLM* (von Werra et al., 2020), que utiliza la codificación de palabras clave, permitió crear mejores agrupaciones, ayudando al modelo a tener un mejor rendimiento durante el entrenamiento.

3.4.4. Conclusiones de los *fine-tuning*

La mejora en el rendimiento del modelo durante los tres experimentos de *fine-tuning* puede atribuirse en gran medida a ajustes clave en el conjunto de datos y en la preparación de los mismos. Los primeros dos *fine-tuning* se realizaron en un conjunto desbalanceado, donde solo tres de las once categorías tenían más de 10 mil muestras, mientras que las demás estaban significativamente subrepresentadas. Esto afectó directamente la capacidad del modelo para generalizar a clases menos frecuentes o con menor número de muestras.

Para el tercer *fine-tuning*, se abordó esta limitación equilibrando el conjunto de datos mediante técnicas de *oversampling* y la generación de muestras sintéticas basadas en datos de personas reales con la ayuda de modelos generativos de texto como GPT-3.5 (OpenIA, 2022). Las clases menos representadas se incrementaron a 11 mil muestras, mientras que las más frecuentes se limitaron a la misma cantidad mediante *undersampling*. Esta estrategia proporcionó al modelo una distribución más uniforme de datos, mejorando su capacidad para aprender patrones en todas las categorías.

Además, se implementó un *DataCollator* especialmente diseñado para modelos de lenguaje que completan textos. Este enfoque particular en la preparación de datos pudo ser la razón para facilitar la asimilación de patrones y relaciones en el conjunto de datos, contribuyendo a la mejora del rendimiento del modelo.

Finalmente, al modificar el valor del rango ($Lora\_R$) de la matriz de representación, de 8 a 16, permitió al modelo aprender representaciones más detalladas de las palabras y su contexto. Además, aceleró la convergencia en los datos de entrenamiento.

La combinación de un conjunto de datos balanceado, un mayor rango en la matriz y un enfoque especializado en la preparación de datos ha demostrado ser fundamental para el éxito del tercer experimento de *fine-tuning*, evidenciado por la drástica reducción en

la pérdida de validación. Estos hallazgos subrayan la importancia crítica de considerar y ajustar la distribución de clases y la preparación de datos en la implementación de modelos de lenguaje, particularmente aquellos destinados a tareas de completado de texto relacionadas con emociones.

3.5. El Modelo Generativo Basado en HEAR

En esta serie de experimentos, se utilizó como base el modelo de lenguaje Llama2-7b-chat de 7 mil millones de parámetros (Touvron et al., 2023c). Este modelo fue ajustado para generar texto enfocado en un agente artificial en una conversación, lo cual es una ventaja para el proyecto, ya que uno de los objetivos es contar con un modelo generativo de texto capaz de interactuar con el usuario a través de un chat.

3.5.1. El Acompañamiento Emocional en los Modelos del Lenguaje Actuales

El acompañamiento emocional como tarea para los modelos de lenguaje actuales no ha sido formalmente descrito en investigaciones hasta la fecha en que se escribe este trabajo. Sin embargo, esto no impide observar si, utilizando un *prompt* adecuado que describa al modelo de forma general o paso a paso lo que debe hacer, se puede lograr cierto nivel de acompañamiento emocional.

Por ejemplo, GPT-3.5 (OpenIA, 2022) es capaz de realizarlo si se le proporciona una instrucción detallada y el contexto adecuado, permitiendo así la generación de muestras sintéticas. Para este experimento, se reutilizó el mismo *prompt* que se usó para obtener las muestras sintéticas (ver figura 3.7). Se hizo un submuestreo de 20 elementos del conjunto de prueba para probarlos con el modelo base, con el objetivo de observar si

un modelo de menor tamaño (7 mil millones de parámetros) sin ajustes puede igualar el rendimiento de GPT-3, que cuenta con 175 mil millones de parámetros.

3.5.2. Fine-Tuning de Llama2-7b-chat para el Acompañamiento Emocional

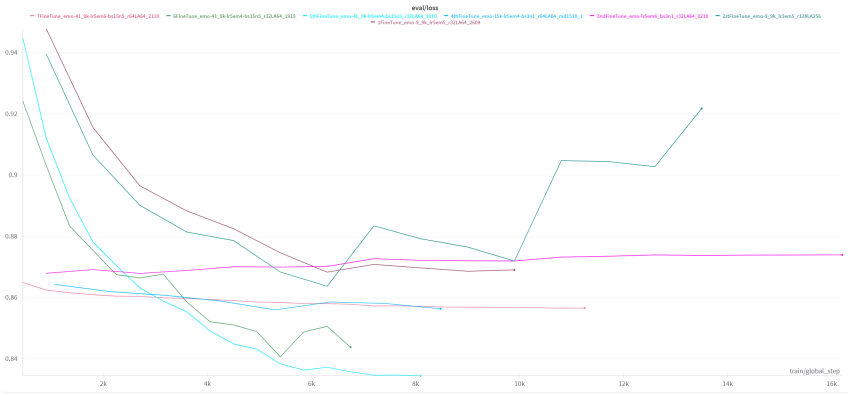


Figura 3.11: Valores de pérdida de los experimentos

Para entrenar el modelo Llama2-7b-chat, con 7 mil millones de parámetros (Touvron et al., 2023c), se realizaron un total de 60 pruebas. Estas pruebas se agruparon por número de experimento y se filtraron para identificar aquellas que aportaban más información sobre el comportamiento del valor de pérdida. Este enfoque fue de gran ayuda para determinar los parámetros de entrenamiento que mejor ajustan el modelo, logrando así una generalización e inferencia óptimas a partir de los datos de entrenamiento.

Se llevaron a cabo 7 experimentos (figura 3.11), cada uno con el mismo numero de épocas, utilizando una versión específica del modelo y del conjunto de datos. En esta sección, discutiremos los aspectos generales de cada experimento y los detalles específicos de cada prueba, así como su impacto en el rendimiento del modelo, especialmente en términos del valor de pérdida.

Primer *Fine-Tuning*

En este primer experimento se llevó a cabo una sola prueba, con el modelo base de Llama2-7b-chat de 7 billones de parámetros (figura 3.12). Se utilizó la primera versión del conjunto de datos para entrenar al modelo a responder empáticamente que contenía $9,900$ muestras (véase en la sección 3.3.2).

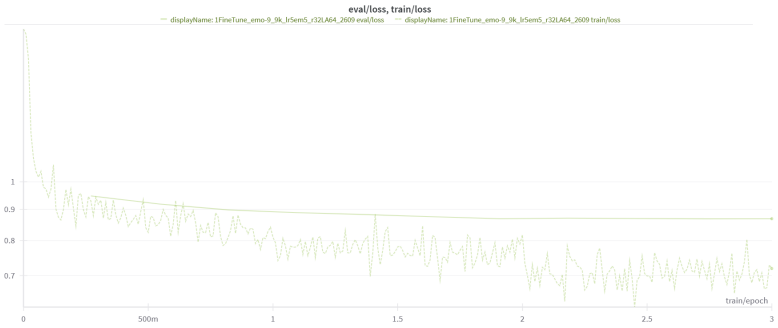


Figura 3.12: Valor de pérdida durante el entrenamiento (línea punteada) y valor de pérdida durante la validación (línea continua) durante el $1°FT$

La longitud máxima de cada oración por tokens en este primer *fine-tuning* fue de 822. Además, se usaron las palabras clave codificadas con el *tokenizer* del propio modelo para configurar el *DataCollatorForCompletionOnlyLM*, que nos permite ajustar modelos en completar textos, tal y como lo muestra el fragmento de código 3.13.

```
response_template_with_context = "\n### response:"
instruction_template="instruction:"
response_template_ids = tokenizer.encode(
    response_template_with_context, add_special_tokens=
    False)[2:]
collator = DataCollatorForCompletionOnlyLM(
    response_template_ids,instruction_template, tokenizer=
    tokenizer)
```

Figura 3.13: Configuración del *DataCollatorForCompletionOnlyLM*.

El resto de parámetros de cada prueba destacada se pueden ver de forma más explícita en la tabla 3.8.

Tabla 3.8: Tabla comparativa de parámetros y valores de perdida durante cada *fine-tuning*

Hyper-parámetros del Trainer	1° *FT*	2° FT	3° FT	4° FT	5° FT	6° FT	7° *FT*	8° *FT*
No. muestras en el conjunto de entrenamiento	9900	9900	9900	15950	41910	41910	41910	41910
Batch size	3	3	3	3	15	15	15	15
Micro batch	1	1	1	1	5	5	5	5
learning rate	$5e^{-5}$	$5e^{-5}$	$5e^{-4}$	$5e^{-5}$	$5e^{-5}$	$5e^{-4}$	$5e^{-6}$	$5e^{-5}$
Adam beta1	0.9	0.9	0.9	0.9	0.9	0.9	0.1	0.9
Adam beta2	0.95	0.95	0.95	0.95	0.95	0.95	0.15	0.95
Adam epsilon	$1e{-}8$	$1e-8$	$2e-8$	$2e-8$	$2e-8$	$2e-8$	$2e-8$	$2e{-}8$
weight decay	0.1	0.09	0.09	0.5	0.5	0.5	0.5	0.1
Épocas	3	3	3	3	3	5	5	3
Parámetros de LoRA								
R	32	64	64	64	64	64	64	64
alpha	64	64	64	64	32	32	64	128
dropout	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.1
Valores de pérdida								
Validación	0.86	0.8637	0.8678	0.8554	0.8344	0.845	0.8565	0.826
Entrenamiento	0.71	0.7983	0.7241	0.7186	0.7474	0.572	0.7216	0.653

Observaciones: En esta primera prueba, el valor de pérdida con el conjunto de datos de validación fue de 0.86, mientras que el valor de pérdida del conjunto de entrenamiento fue de 0.71. Aunque la diferencia no es significativa, un menor diferencial entre estos valores indica un mejor rendimiento del modelo. Sin embargo, ambos valores siguen siendo altos. Por ejemplo, en un escenario donde el objetivo es que el modelo prediga 10 *tokens*, se equivoca en 8 o 9 ocasiones. Aun así, esto no es catastrófico, ya que es necesario evaluar la calidad de las predicciones. Aunque no coincidan exactamente con los resultados esperados, podrían ser coherentes con el contexto del *prompt* de entrada.

Segundo *Fine-Tuning*

Para la segunda prueba, se trabajó con el modelo obtenido del *fine-tuning* anterior, con el objetivo de evaluar si más épocas y diferentes condiciones podrían mejorar su rendimiento. Se mantuvieron la mayoría de los parámetros utilizados en la prueba anterior, ajustando únicamente los valores de LoRA, específicamente el valor de R, que se incrementó de 32 a 64. Este ajuste permite actualizar un mayor número de pesos del modelo, con la esperanza de realizar ajustes más precisos y generalizar mejor a partir del conjunto de entrenamiento sin necesidad de aumentar su tamaño.

Observaciones: Gracias a los valores obtenidos en esta prueba, es evidente que el modelo estaba desaprendiendo la tarea. Inicialmente, alcanzó un valor de pérdida de 0.8637 en el conjunto de validación y 0.75 en el de entrenamiento. Sin embargo, con el avance de las épocas, el modelo se desajustó (*overfitting*), aumentando el valor de pérdida en el conjunto de entrenamiento a 0.9217 y en el de validación a 0.567. Este desajuste se puede visualizar en la gráfica 3.14, que se asemeja a la gráfica de ejemplo 3.5.

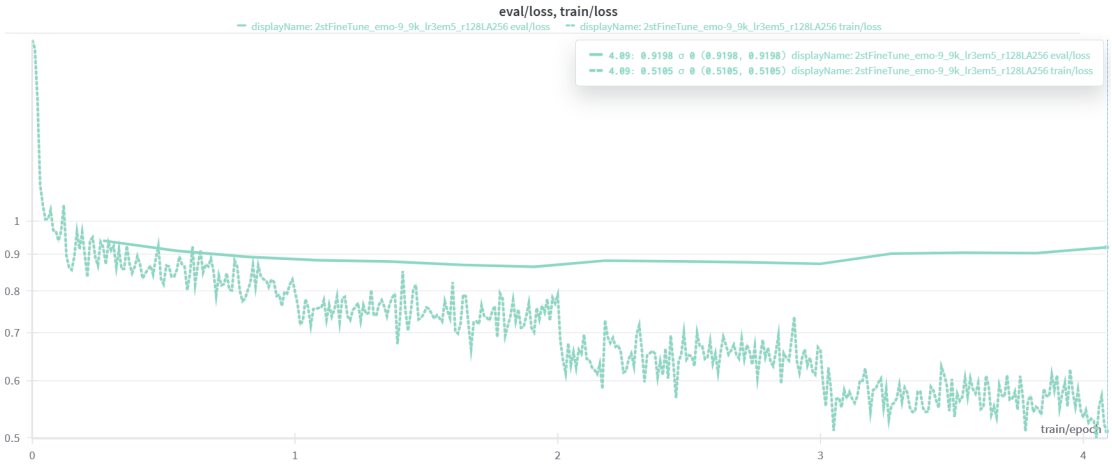


Figura 3.14: Valor de pérdida durante el entrenamiento (línea punteada) y Valor de pérdida durante la validación (línea continua) durante el $2°FT$

Tercer *Fine-Tuning*

En esta tercer prueba, se continuó trabajando sobre el modelo ajustado en el primer *fine-tuning* y para evitar un *overfitting* se disminuyó el valor del parámetro *early_stopping_patience* de 5 a 2. Además, se corrieron más de una versión de la prueba con variaciones mínimas, con el objetivo de ver como cambiaba el rendimiento del modelo ajustando parámetros como el *learning rate* (lr), o los parámetros de lora al mismo tiempo. En una de estas pruebas se puso bajo observación como es que un valor más pequeño del lr junto con un valor mayor del rango de la matriz de representación (LoRA R) influyen sobre el modelo, y viceversa. Como se muestra en la figura 3.15 todas estas pruebas fueron irrelevantes pues el rendimiento fue casi el mismo que en el $2°$FT, pero debido a *early_stopping_patience* que fue disminuido, muchas de estas pruebas se terminaron antes.

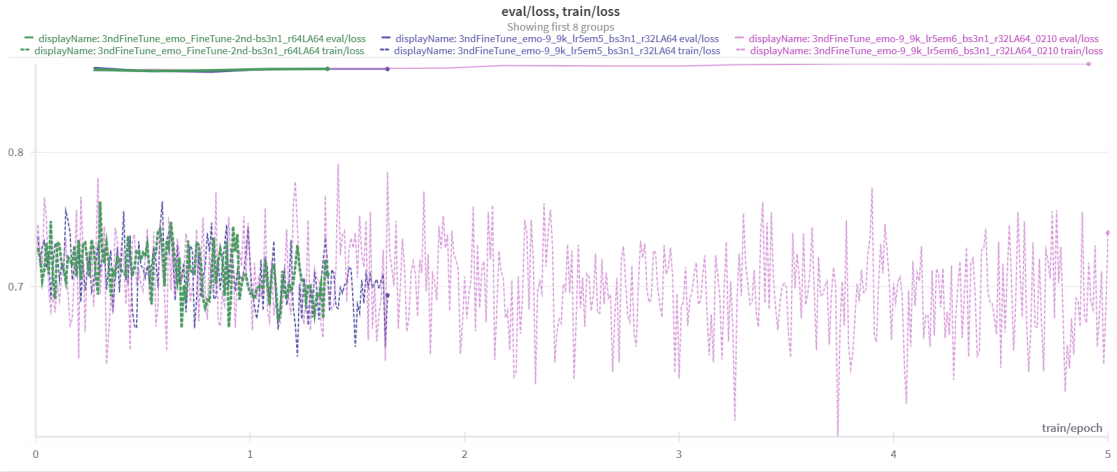


Figura 3.15: Valores de pérdida durante el entrenamiento (línea punteada) y valores de pérdida durante la validación (línea continua) durante las pruebas del 3°*fine-tuning*

Observaciones: Una vez finalizadas estas pruebas, se concluyó que era necesario trabajar con más muestras en el conjunto de entrenamiento. Al observar el rendimiento de las pruebas en las gráficas de *wandb* (figura 3.15), la diferencia entre los valores de pérdida hacía evidente que el modelo estaba memorizando el conjunto de entrenamiento y se estancaba al inferir con nueva información. Además, se decidió trabajar sobre el modelo base de Llama2-7b-chat en lugar de usar el modelo ajustado en el primer *fine-tuning* (1° FT). De esta forma, se podría observar si el tamaño del conjunto de entrenamiento es un factor clave para optimizar el modelo.

Cuarto *Fine-Tuning*

Considerando las observaciones hechas en las pruebas previas, se aumentó el tamaño del conjunto de entrenamiento de $9,900$ muestras a 15 mil. Al igual que en el 3°FT, se ejecutaron diferentes versiones de esta cuarta versión, variando tanto la tasa de aprendizaje (lr) como el valor de R en LoRA.

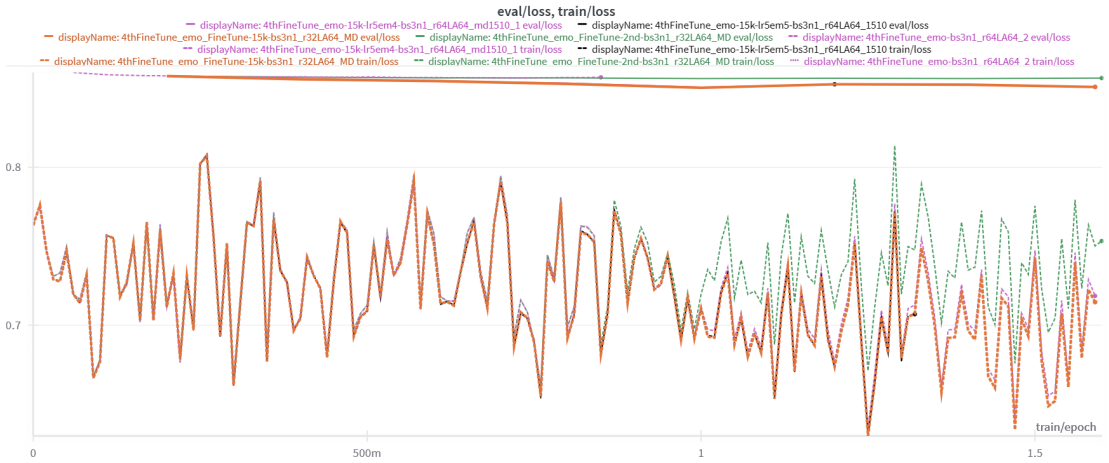


Figura 3.16: Valores de pérdida durante el entrenamiento (línea punteada) y valores de pérdida durante la validación (línea continua) durante las pruebas del $4°FT$

Observaciones: Aunque se variaron los valores de R y de la tasa de aprendizaje (lr), no se logró el resultado esperado. El comportamiento del modelo se mantuvo similar a las pruebas anteriores, aunque se observó con menor frecuencia el *overfitting*. Esto se debió al aumento del conjunto de datos de entrenamiento y a la utilización del modelo base. Sin embargo, la mejora no fue significativa en comparación con los *fine-tuning* previos. El valor de pérdida en el conjunto de validación disminuyó de 0.86 a 0.85, y en el conjunto de entrenamiento de 0.79 a 0.72.

Lo más destacado del rendimiento observado en la imagen 3.16 es que, aunque en una de las pruebas (4thFineTune_emo_FineTune-15k-bs3n1_r32LA64_MD) se evidencia un *overfitting*, en el resto se nota una marcada diferencia entre los valores de pérdida. Esto podría indicar la necesidad de aumentar aún más el conjunto de entrenamiento y ajustar algunos otros parámetros para mejorar el rendimiento del modelo.

Quinto *Fine-Tuning*

Como se sugirió en el *fine-tuning* (FT) anterior, se aumentó el tamaño del conjunto de entrenamiento de 15 mil a $41,910$ muestras. Esto requirió tiempo y una mayor inversión en la API de OpenAI para generar más muestras sintéticas. Además, en este quinto *FT* (véase figura 3.17), solo se necesitó una prueba para observar una mejora en los valores de pérdida en comparación con los otros *FT*. Solo se aumentó el tamaño del lote (*batch*) mientras que el resto de los parámetros se mantuvieron igual que en el cuarto *FT*.

Figura 3.17: Comparación de valores de pérdida durante la validación de las pruebas del $5°$ *FT* (rojo).

Observaciones: Con casi las mismos parámetros y aumentando el conjunto de entrenamiento en número de muestras, se mitigó el problema del *overfitting* a lo largo de 3 épocas, y se mantuvo un rendimiento constante en ambos valores de pérdida después de alcanzar 0.83 en el conjunto de validación y 0.74 en el de entrenamiento. Lo que se puede observar en la gráfica del rendimiento 3.18, es que el modelo está sufriendo de *underFitting*, pues la línea del valor *train/loss* desciende más rápido que la del *val/loss* sin que esta última aumente.

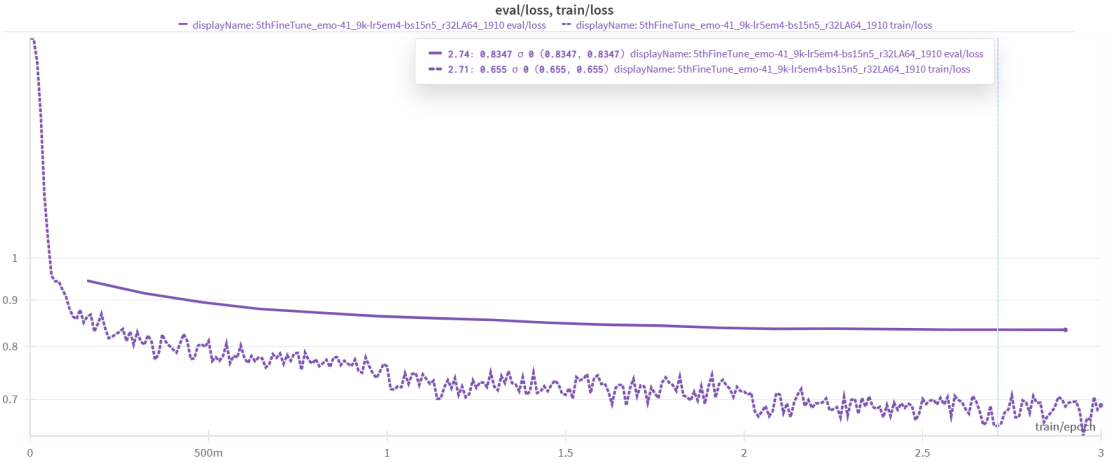


Figura 3.18: Valor de pérdida durante el entrenamiento (línea punteada) y valor de pérdida durante la validación (línea continua) durante el $5°FT$

Sexto y Séptimo *Fine-Tuning*

Considerando el *fine-tuning* previo, se realizaron dos ajustes adicionales a partir de la configuración anterior ($5°$FT). El objetivo era lograr valores de pérdida más bajos o, al menos, reducir la diferencia entre los valores de pérdida de entrenamiento y validación (*train/loss* y *validation/loss*). En el $6°FT$, se modificó la tasa de aprendizaje ($learning\_rate$) de $5e^{-5}$ a $5e^{-4}$ y se aumentó el número de épocas, buscando permitir cambios significativos en los pesos del modelo. En el $7°FT$, se realizaron varios ajustes basados en los resultados del $6°FT$, como la reducción de la tasa de aprendizaje y de los valores beta del optimizador Adam.

Observaciones: Durante las dos pruebas descritas anteriormente, no se observó ninguna mejora en el rendimiento del modelo; de hecho, ni siquiera se igualó el rendimiento del quinto *fine-tuning*, manteniéndose en ambos casos muy por debajo del mejor ajuste alcanzado en el quinto *fine-tuning*, como se muestra en la figura 3.19. A pesar de esto, el valor de pérdida en la validación se mantuvo constante en ambos *fine-tunings*, sin

evidencia de *overfitting*. Sin embargo, persiste el problema de *underfitting*, ya que el valor de pérdida durante el entrenamiento seguía siendo inferior al de validación.

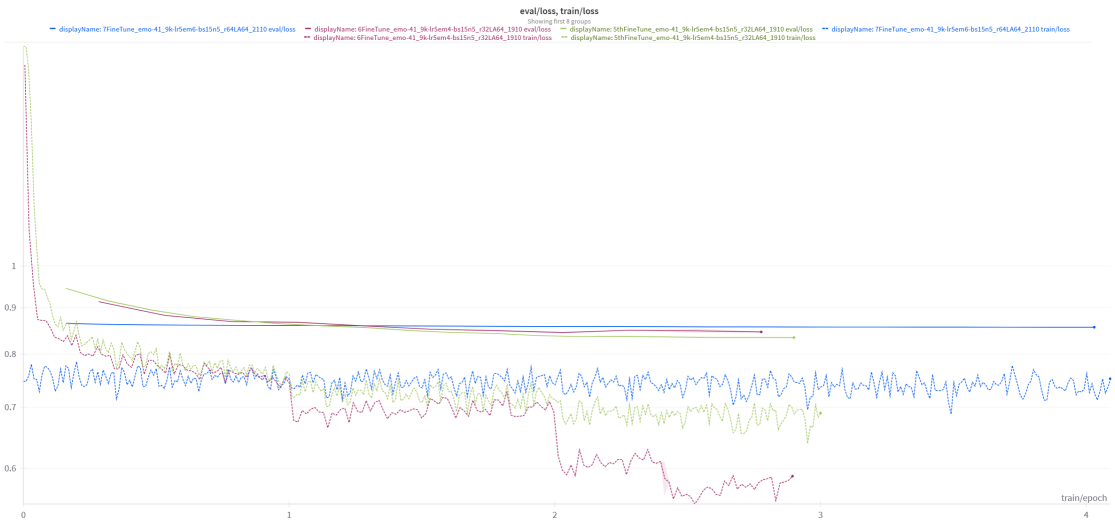


Figura 3.19: Valor de pérdida durante el entrenamiento (línea punteada) y valor de pérdida durante la validación (línea continua) durante el $6°FT$ (violeta) y $7°FT$ (azul), comparado con el rendimiento del $5°FT$ (verde).

Para abordar este problema, se debería considerar aumentar aún más el número de muestras para lograr un mejor ajuste, disminuir algún parámetro distinto al *learning rate* (*lr*), o explorar la posibilidad de guiar al modelo de manera más explícita en la tarea a través del *prompt* utilizado en el entrenamiento y evaluar cómo este simple cambio afecta los valores de pérdida.

El Octavo y Último *FT* – es una prueba adicional para evaluar algunas ideas planteadas en el experimento anterior. Entre estas, se incluyó la reducción de parámetros distintos a la tasa de aprendizaje y la modificación del *prompt* en los conjuntos de datos (*train-validation-test*) para comprobar la efectividad de estos cambios en el rendimiento del modelo.

Inicialmente se experimentó con el *prompt* original de *HEAR* (sección 3.6). Aunque

este *prompt* era adecuado, resultaba menos explícito en cuanto a la tarea que se debía realizar. Por lo tanto, se decidió utilizar el *prompt* mostrado en la tabla 3.5, que ofrece una mayor claridad sobre las instrucciones.

" *If the user seems sad or upset, you should offer words of encouragement. If the user seems happy or excited, the assistant should share that joy and respond enthusiastically. In all cases, the assistant should maintain a respectful tone, if possible, encouraging you to talk more about it. Don't say hello, unless necessary. Use the username and pronoun to respond in a personalized way.* "

Tabla 3.9: Una forma mas especifica de la instrucción para la tarea del acompañamiento emocional.

En este último diseño de *prompt* mostrado en la tabla 3.9,se proporciona más información sobre situaciones en las que el modelo podría verse involucrado: por ejemplo, si una persona está feliz o emocionada, el modelo debe responder con entusiasmo y compartir esa felicidad. Por otro lado, si la persona muestra tristeza o enojo, el modelo debe ser empático y ofrecer palabras de aliento. En todos los casos, el modelo debe actuar de manera respetuosa e incentivar a la persona a hablar más al respecto. Como se puede observar, esta instrucción ofrece más alternativas sobre cómo debe actuar un asistente emocional. Lo anterior puede guiar al modelo durante el entrenamiento para generar *tokens* más coherentes con esta instrucción, el contexto y las respuestas incluidas en el conjunto de entrenamiento.

Los parámetros del $8°\boldsymbol{FT}$ fueron en su mayoría iguales a los del $5°FT$, con la excepción del valor de *weight decay*, que se redujo de 0.9 a 0.1. Esta disminución se implementó para permitir que los pesos del modelo fueran un poco más grandes (Brownlee, 2020), lo que podría mejorar la capacidad del modelo para generalizar aspectos más complejos de los mensajes.

Además, se realizaron ajustes en los valores de LoRA para facilitar un mayor acceso a los pesos del modelo base y permitir cambios significativos. En este sentido, R se incrementó de 32 a 64, y el valor de *Alpha* se aumentó de 64 a 128. Asimismo, el valor de *dropout* se elevó de 0.05 a 0.1. Con estos cambios, aunque el número de pesos en la matriz de bajo rango aumentó, también se incrementó el *dropout*, lo que determina el porcentaje de estos nuevos pesos que se congelarán o ignorarán durante el proceso de entrenamiento, pasando de un $5\,\%$ a un $10\,\%$.

Observaciones: A lo largo de tres épocas, y usando el modelo base de Llama2-7b-chat como punto de partida, se lograron mejorar los valores de pérdida tanto en los procesos de validación como de entrenamiento, tal y como se muestra en la figura 3.20. En comparación con el $5°$*fine-tuning*, el $8°$ *FT* presenta un valor de pérdida ligeramente inferior en la validación, pasando de 0.83 a 0.82. Aunque la diferencia no es significativa, es una mejora que los *fine-tuning* anteriores ($6°$y $7°$) no lograron alcanzar, a pesar de haber utilizado parámetros similares y el mismo número de muestras en el conjunto de entrenamiento. Estos ajustes realizados en el octavo *FT* permitieron que el modelo se adaptará un poco mejor a la tarea. Sin embargo, esta prueba también reveló un aumento en el *underfitting*, como se evidencia en la diferencia entre los valores de pérdida del conjunto de entrenamiento y validación (véase la tabla 3.8. Esto no representa un escenario catastrófico hasta que se realicen pruebas comparativas en la inferencia entre los modelos GPT-3, Llama2-7b-chat (modelo base) y el modelo ajustado.

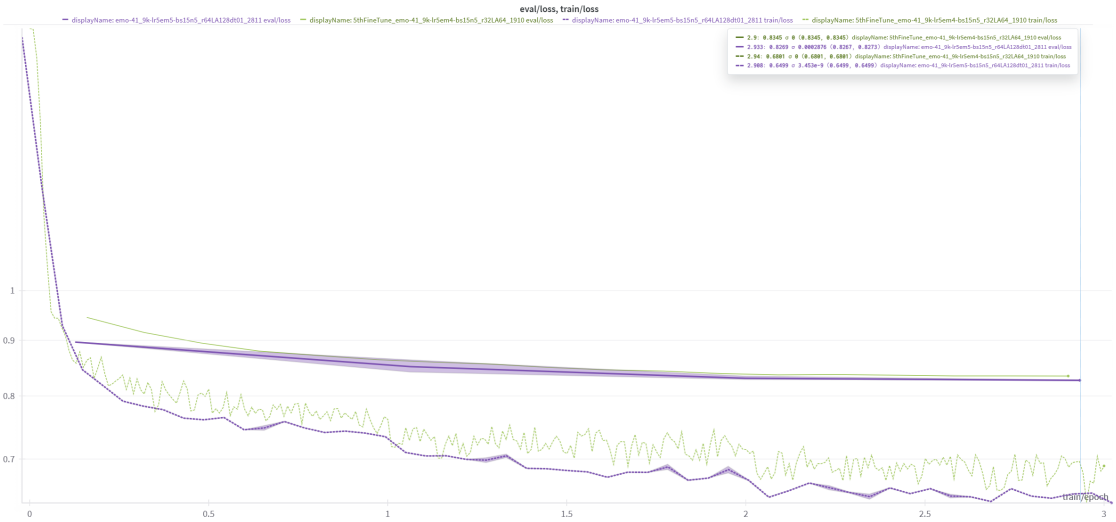


Figura 3.20: Valor de pérdida durante el entrenamiento (línea punteada) y valor de pérdida durante la validación (línea continua) durante el $8°FT$ (morado), comparado con los valores de pérdida del $5°FT$ (verde).

Capítulo IV

Resultados y Discusión

Para evaluar el modelo "Sólo Escúchame" y compararlo con modelos del estado del arte, se utilizó el subconjunto de pruebas del conjunto HEAR (consulte la sección 3.3.2), que consta de 1320 mensajes que cubren una variedad equilibrada de emociones. Estos mensajes se emplearon como entrada en cuatro modelos de referencia, con los cuales se comparó el rendimiento del modelo "Sólo Escúchame". Los modelos de referencia se listan a continuación:

- Llama-2-7b-chat (base) (Touvron et al., 2023c)
- Mixtral 8x7b (Jiang et al., 2024)
- GPT-3.5-Turbo (OpenIA, 2022)
- GPT2-124M (Radford et al., 2018b)

4.1. Evaluación del *FT*

Para evaluar la adecuación del modelo a los datos de entrenamiento, se llevaron a cabo varias pruebas para observar los cambios resultantes del entrenamiento realizado en esta investigación. Esto incluirá el análisis del promedio de *tokens* necesarios para generar

una respuesta, así como la variación en el uso del vocabulario por parte del modelo después del proceso de ajuste. Además, para evaluar la calidad de las respuestas, fue necesario establecer herramientas que nos permitieran identificar patrones que garanticen la calidad contextual de las respuestas deseadas por la tarea del modelo. Lo anterior requiere explorar los métodos disponibles para llevar a cabo dicha evaluación de calidad.

4.1.1. El Tamaño de las Respuestas

Usando el tokenizador de Llama2, se tokenizaron las respuestas de cada modelo, contabilizando el número de *tokens* por respuesta y calculando el promedio de *tokens* necesarios para generar una respuesta. Los resultados se detallan en la tabla 4.1. Estos resultados confirmaron las expectativas después del ajuste realizado al modelo. En el caso del modelo no entrenado, se requieren en promedio 111 *tokens* para generar una respuesta, mientras que para el modelo ajustado (*FTM*) se necesitan 65 *tokens*, un valor más cercano al de GPT3, que requiere 77 *tokens*.

Modelo	**Promedio de *Tokens* por Respuesta**
GPT3_output	77
FTunned_Model_Output	65
Llama2-7b-Chat	111
Mixtral 8x7b	78
FT_GPT2-124	100

Tabla 4.1: Promedio de longitud de los *tokens* de todas las respuestas.

4.1.2. Cambios en el Vocabulario

Para evaluar la efectividad de un modelo destinado a acompañar a las personas en sus transiciones emocionales, es fundamental analizar el cambio en la frecuencia y relevancia de las palabras o términos utilizados. Además, es crucial identificar qué palabras

han adquirido mayor importancia en este contexto. Durante el entrenamiento del modelo, se ajustó el uso de ciertos términos para que cumplieran adecuadamente con este propósito. Se espera que las palabras que reflejan un lenguaje empático hayan ganado relevancia significativa, dado su papel esencial en la simulación de empatía.

Frecuencia Directa de Términos

En esta primera evaluación, se creó un diccionario con las palabras únicas de las respuestas de los modelos y se contabilizó la frecuencia de aparición de cada palabra en todas las respuestas. De esta manera, pudimos observar qué palabras son las más utilizadas por cada modelo y comparar los resultados entre ellos.

En la tabla 4.2, se puede observar un cambio significativo en la frecuencia de las palabras entre los tres modelos. Las frecuencias de GPT-3 y del modelo ajustado (*FTM* o *fine-tuning model*) son más similares entre sí que las del modelo base. En algunos casos, la frecuencia es mayor en el *FTM* que en GPT-3, como es el caso de palabras como "hablar", "importante", "emocionante", "suena", "lamento", entre otras. Sin embargo, hay otras palabras que tienen una menor frecuencia en *FTM* que en GPT-3, pero que aún así son más utilizadas en comparación con el modelo base. Este cambio en la frecuencia de palabras indica una alteración en el uso del lenguaje.

Mientras que el uso del lenguaje en el modelo ajustado se asemeja al del modelo que se utilizó para crear las muestras sintéticas (GPT-3.5), existen ciertas variaciones propias del proceso de ajuste que podrían explicarse por los motivos mencionados en la sección anterior. El *underfitting* podría ser la principal causa de estas variaciones.

Un detalle importante a mencionar, es que el modelo base suele responder en inglés, incluso las palabras con mayor número de apariciones son "you" con 1308 apariciones, "to" con 1117 repeticiones, "and" con 988, entre otras. Esto contrasta con los otros dos

modelos, que responden consistentemente en español.

Cambio en la Relevancia de los Términos: *TF-IDF*

Para calcular la relevancia de los términos, se utilizó una técnica muy popular en el ámbito de los sistemas de recuperación de la información y el *NLP*, el *TF-IDF (term frequency-inverse document frequency)* descrito en la ecuación 4.1. Como su nombre lo indica esta medida estadística se compone de la frecuencia del término, y la frecuencia inversa del documento:

$$tf - idf(t, d) = tf(t, d) \times idf(t, d) \tag{4.1}$$

La clase que ejecutó el cálculo del *TF-IDF* fue *TfidfVectorizer* del framework de *Scikit-learn* (Pedregosa et al., 2011) *sklearn*. Esta clase primero tokeniza el documento y construye un diccionario de palabras únicas con la técnica de n-gramas ($n = 1$); para luego calcular la frecuencia del término en el documento $tf(t, d)$ que por defecto usa la norma l_2 (ecuación 4.2).

$$tf(t, d) = \frac{f_{t,d}}{\sqrt{\sum_{i=1}^{n} f_{i,d}^2}} \tag{4.2}$$

Donde l_2 es la raíz cuadrada de la suma de los cuadrados de los valores del vector, $f_{t,d}$ es la frecuencia del término t en el documento d, y n es el número total de términos en el documento d.

Después se procede a calcular la frecuencia inversa del documento $idf(t)$ representada por la ecuación 4.3.

$$idf(t, d) = \log(\frac{N}{1 + df(t)}) \tag{4.3}$$

Donde N es el número total de documentos en el corpus y $df(t)$ es el número de documentos que contienen el término t.

TF-IDF demostró cambios interesantes en las respuestas de los modelos evaluados. La clasificación de la relevancia de los términos reveló que palabras como “ser”, “puede”, “veces”, “sentir” y “entiendo” han adquirido mayor importancia en el vocabulario del modelo “Solo Escuchame”, asemejándose al del modelo GPT-3.5 usado para generar las muestras de entrenamiento. Simultáneamente, se distancia del vocabulario predominante en el modelo base Llama2-7b-chat, en el cual las palabras en inglés como “to”, “you”, “and”, “its” y “can” son las más relevantes en sus respuestas.

Palabras	Frecuencia en GPT3 output	Frecuencia en FTM ouput	Frecuencia en BM Output
ser	366.0	343.0	118.0
puede	319.0	233.0	90.0
veces	272.0	279.0	41.0
sentir	271.0	365.0	182.0
entiendo	266.0	323.0	82.0
escucharte	212.0	70.0	216.0
experiencia	200.0	216.0	67.0
compartir	195.0	155.0	72.0
cuéntame	193.0	111.0	0.0
alegra	179.0	121.0	48.0
hablar	174.0	222.0	514.0
sientes	170.0	247.0	157.0
escuchar	168.0	92.0	105.0
recuerda	167.0	125.0	45.0
pueden	161.0	180.0	54.0
importante	151.0	198.0	296.0
genial	151.0	192.0	345.0
gustaría	149.0	219.0	115.0
emocionante	144.0	150.0	123.0
suena	143.0	161.0	10.0
parece	135.0	83.0	18.0
maravilloso	134.0	163.0	8.0
momentos	114.0	72.0	72.0
disfrutar	107.0	110.0	110.0
sientas	104.0	194.0	149.0
feliz	101.0	64.0	83.0
encontrar	92.0	107.0	85.0
lamento	90.0	122.0	1.0
especial	89.0	85.0	59.0
apoyo	81.0	28.0	58.0
increíble	80.0	90.0	202.0

Tabla 4.2: Tabla comparativa de la frecuencia de cada una de las 50 palabras más frecuentes en las respuestas de los modelos, tomando como referencia las respuestas de GPT-3 para ordenarlas.

Palabras	**TF-IDF GPT3**	**TF-IDF FT Model**	**TF-IDF BModel**
ser	0.277081	0.269754	0.023238
puede	0.241499	0.183244	0.017724
veces	0.205918	0.219421	0.008074
sentir	0.205161	0.287056	0.035842
entiendo	0.201376	0.254025	0.016149
cuéntame	0.188144	0.112410	0.000000
escucharte	0.161252	0.055052	0.042538
experiencia	0.151410	0.169874	0.013195
compartir	0.147625	0.121900	0.014179
alegra	0.135512	0.095161	0.009453
hablar	0.131727	0.174593	0.101224
sientes	0.128699	0.194254	0.030919
escuchar	0.127185	0.072354	0.020678
recuerda	0.126428	0.098307	0.008862
pueden	0.121885	0.141562	0.010634
importante	0.114315	0.155718	0.058293
genial	0.114315	0.150999	0.067942
gustaría	0.112801	0.172233	0.022647
emocionante	0.109015	0.117968	0.024223
suena	0.108258	0.126619	0.001969
parece	0.102202	0.065276	0.003545
maravilloso	0.101445	0.128192	0.001575
momentos	0.086304	0.056625	0.014179
siempre	0.083276	0.048760	0.011225
disfrutar	0.081004	0.086510	0.021663
sientas	0.078733	0.152572	0.029343
feliz	0.076462	0.050333	0.016346
situación	0.075705	0.071567	0.009650
saber	0.073434	0.035390	0.004529
quieres	0.073434	0.066062	0.057308
hizo	0.072677	0.131338	0.002954
encontrar	0.069649	0.084151	0.016739
lamento	0.068135	0.095947	0.000197
especial	0.067378	0.066849	0.011619
apoyo	0.061321	0.022021	0.011422

Tabla 4.3: Tabla comparativa de los valores obtenidos del TF-IDF para cada término, ordenándolos de mayor relevancia a menor, tomando como referencia a GPT-3.

4.1.3. Conclusión de la Evaluación del *FT*

Los resultados obtenidos en ambas evaluaciones muestran coincidencias lógicas en las clasificaciones. Tanto en los modelos "Solo Escuchame" como en GPT-3.5, las cinco palabras mejor valoradas en las clasificaciones de *TF-IDF* y frecuencia directa de términos fueron las mismas: "ser", "puede", "veces", "sentir" y "entiendo". Debido a la forma en que se calcula la relevancia, cuando una palabra aparece demasiadas veces, su valor disminuye hasta acercarse a 0, similar a cuando estas palabras no están presentes en el documento. Por ejemplo, "cuéntame" tiene un valor de relevancia (*TF-IDF*) de 0.11 y 0.18 en las respuestas de los modelos "Solo Escuchame" y GPT-3.5, con una frecuencia directa de 111 y 193 respectivamente (véase en las tablas 4.3 y 4.2). Sin embargo, en el modelo base esta palabra no aparece, por lo tanto su valor de *TF-IDF* es 0.00. Esto indica que una palabra debe aparecer con suficiente frecuencia para ser relevante (*TF-IDF*), pero no en exceso para evitar perder su relevancia.

Ambos modelos muestran resultados bastante similares usando los mismos *prompts* del conjunto de prueba. Este análisis concluye cómo el *fine-tuning* cambió el vocabulario y la forma en que el modelo Llama2-7b-chat utiliza las palabras. Este proceso ajustó las respuestas del modelo base para que fueran más acordes con los datos de entrenamiento, encontrando similitudes en el tamaño de la respuesta, la frecuencia y la relevancia de los términos. Es importante aclarar que esta prueba no evalúa la coherencia y calidad de las respuestas; para hacerlo, se requiere un análisis más exhaustivo con otros criterios, incluyendo un muestreo detallado de las respuestas del conjunto de prueba para evaluar su calidad.

4.2. Instrumentos de Evaluación de Calidad de Respuestas

Para evaluar la calidad de las respuestas, es importante considerar la documentación disponible sobre el acompañamiento emocional y comparar diferentes fuentes de información para definir los rasgos que se utilizarán en dicha evaluación. Además, deben tenerse en cuenta las limitaciones inherentes al modelo. Aunque el modelo posee ciertas capacidades, sigue siendo un agente virtual cuya presencia se limita a medios digitales, como redes sociales o plataformas en internet, y no puede interactuar físicamente como una persona.

4.2.1. Escucha Activa

Revisando las diferentes fuentes de información relacionadas al acompañamiento emocional, en casi todas ellas, se hace referencia al trabajo realizado por Carl Rogers, un reconocido psicólogo americano, quien acuñó el término Escucha activa. Él define en su libro "*Active Listening*" (Rogers and Farson, 2015), una técnica del mismo nombre, que plantea que en una conversación el oyente debe prestar atención plena al hablante, sin interrumpirlo ni juzgarlo, y reflejar sus sentimientos y pensamientos con empatía y respeto. De esta manera, se facilita la comunicación efectiva y se favorece el desarrollo personal y social del individuo.

A lo largo del libro, el autor explora a plenitud los detalles que hacen a la escucha activa una técnica que permite mejorar la forma de afrontar conversaciones que podrían suponer un reto, debido a el entendimiento entre las partes involucradas y la carga emocional inherente, fomentando la empatía y comprensión mutua, permitiendo una convivencia y relaciones más sanas. Para aplicar la técnica de escucha activa en el contexto real se

deben tomar en cuenta los siguientes puntos:

- Escuchar el significado total, no sólo las palabras.
- Responder a los sentimientos, no solo al contenido.
- Fijarse en todas las pistas, verbales y no verbales, que expresan el mensaje.
- Verificar la comprensión, reflejando en nuestras propias palabras lo que el hablante ha dicho.
- Escucharse a sí mismo, reconociendo y expresando nuestros propios sentimientos y actitudes.

Algunos puntos que se deben evitar para emplear la escucha activa son:

- Argumentar, razonar, regañar, animar, insultar o empujar al hablante a cambiar su punto de vista.
- Emitir juicios, evaluaciones, consejos o información que puedan ser vistos como intentos de influir o dirigir al hablante.
- Ser defensivo, resentido o temeroso frente a la oposición, la hostilidad o las expresiones fuera de lugar del hablante.
- Negar o suprimir los sentimientos positivos o negativos del hablante o de uno mismo.

¿Como evaluar si un modelo de lenguaje esta aplicando la escucha activa?

Estos modelos tienen diversas capacidades y limitaciones, por lo que es necesario explorar métodos para evaluar su desempeño en la tarea de acompañamiento emocional.

Basándonos en la revisión bibliográfica previa (sección 4.2.1), el método de la escucha activa permite evaluar cuán eficazmente el modelo realiza esta tarea. Para ello, es necesario considerar los puntos siguientes:

1. **Atención contextual:**
   - **Evaluación:** ¿El modelo demuestra coherencia en sus respuestas, mostrando que presta atención a la información proporcionada previamente en la conversación?
     - **Ejemplo:** "Salir a **almorzar** con tu hermana en Colorado Springs suena como un día muy especial".
   - **Limitaciones:** Los modelos del lenguaje no tienen una memoria a largo plazo, por lo que la atención contextual puede ser limitada y dependerá de la longitud de la conversación.

2. **Formulación de preguntas clarificadoras:**
   - **Evaluación:** ¿El modelo hace preguntas para obtener más detalles o aclaraciones sobre lo que se ha dicho?
     - **Ejemplo:** "¿Podrías explicarme más detalladamente cómo te hizo sentir esa situación?"
   - **Limitaciones:** Los modelos de lenguaje pueden generar preguntas, pero puede carecer de profundidad en comprensión contextual y no siempre puede captar la intención subyacente detrás de una pregunta.

3. **Profundiza en la Conversación:**
   - Evaluación: ¿El modelo invita a profundizar en la conversación siendo coherente con el contexto dado por el interlocutor?
     - **Ejemplo:** "¿Qué hicieron durante su **almuerzo**?"

- **Limitaciones:** Los modelos del lenguaje pueden sostener conversaciones coherentes, siempre y cuando el contexto previo se incluya en su ventana de contexto.

4. **Ausencia de juicio o crítica:**
   - **Evaluación:** ¿El modelo evita expresar juicios o críticas hacia las afirmaciones del interlocutor?
   - **Limitaciones:** Los modelos del lenguaje no tienen un sentido innato de moralidad y puede generar respuestas que reflejan prejuicios presentes en los datos de entrenamiento.

5. **Paciencia y permitir que la otra parte hable:**
   - **Evaluación:** ¿El modelo muestra paciencia al permitir que la conversación se desarrolle sin interrupciones innecesarias?
   - **Limitaciones:** Los modelos del lenguaje no tienen conciencia temporal y pueden carecer de la capacidad de entender la importancia del tiempo en una conversación.

6. **Demostración de empatía:**
   - **Primera Evaluación:** ¿El modelo responde de manera empática, reconociendo y reflejando las emociones expresadas por el interlocutor?
     - **Ejemplo:** "Puedo entender cómo la ansiedad puede ser abrumadora antes de una publicación de resultados importantes"
   - **Segunda Evaluación:**¿El modelo reconoce en su respuesta las emociones expresadas por el interlocutor?
     - **Ejemplo:** "entiendo que los olores fuertes pueden ser **incómodos**"

- **Limitaciones:** Los modelos del lenguaje pueden simular empatía, pero su comprensión es limitada a patrones de lenguaje y no a una verdadera experiencia emocional.

La paciencia es un rasgo difícil de evaluar de manera óptima. Según las limitaciones del modelo, este no tiene una noción del tiempo, por lo que no puede discernir si una persona ha pasado mucho o poco tiempo escribiendo un mensaje. Sin embargo, el modelo sí puede determinar la extensión del mensaje basándose en el número de *tokens* que contiene. En una conversación por chat, cada mensaje enviado por una persona es el estímulo que provoca una respuesta del modelo. Esto implica que por cada mensaje que la persona envía, el modelo genera una respuesta. Por lo tanto, el concepto de paciencia resulta complicado de aplicar en este contexto.

Para abordar esta limitación, se podría diseñar un sistema de apoyo que gestione las respuestas del modelo. Por ejemplo, si una persona tiende a expresarse mediante varios mensajes, este sistema podría decidir el momento más adecuado para responder, esperando hasta que la persona termine de enviar todos los mensajes necesarios para expresar su experiencia. Esta es una función que el modelo, por sí solo, no es capaz de realizar.

4.2.2. El Método Socrático

Un recurso fundamental en la revisión bibliográfica sobre el acompañamiento a personas en procesos dolorosos fue el trabajo del autor Luis A. Oblitas, un psicólogo peruano reconocido en Iberoamérica por sus contribuciones al campo de la Psicología. En su libro Psicoterapias Contemporáneas (Oblitas, 2008), Oblitas describe, entre otras terapias, el método socrático, el cual comparte características con la escucha activa.

Este método se basa en la mayéutica del filósofo Sócrates, en la que el maestro, mediante preguntas que incitan a la reflexión, guía al alumno para que descubra y construya nuevos conocimientos (Porto and Merino, 2021). A diferencia de la mayéutica clásica, en el método socrático es el terapeuta quien dirige la sesión, no proporcionando respuestas o soluciones directamente, sino guiando al consultante hacia su descubrimiento a través de preguntas inductivas. Estas preguntas son abiertas y están diseñadas para promover el autodescubrimiento, permitiendo al consultante ampliar sus respuestas mediante el análisis, la síntesis y la evaluación de su situación (Oblitas, 2008).
Para aplicar el diálogo socrático, se pueden considerar los siguientes puntos:

1. Uso de preguntas inductivas que invitan a la reflexión y al análisis crítico .
2. No imponer ideas o soluciones, sino guiar al interlocutor para que descubra sus propias respuestas.
3. Ampliación y construcción de conocimiento a través de la reflexión y el diálogo.
4. Generación de disonancia cognitiva: el uso del diálogo socrático busca provocar disonancia cognitiva que se refiere a la sensación de malestar o tensión que experimenta una persona cuando tiene pensamientos, creencias o actitudes que entran en conflicto entre sí (García-Allen, 2022).
5. Descubrimiento guiado: el facilitador que emplea el método socrático se limita a preguntar sistemáticamente, guiando al consultante descubrir sus propias creencias y conocimientos.

¿Como evaluar si un modelo de lenguaje esta aplicando el método Socrático?

Es fundamental recordar que los modelos de lenguaje son agentes virtuales sin presencia física o tangible, lo que limita sus capacidades al ámbito del lenguaje verbal y excluye

por completo el lenguaje corporal. A pesar de estas limitaciones físicas, es importante considerar lo que el modelo puede hacer. Debemos formular las preguntas adecuadas al evaluar sus respuestas para identificar tanto sus alcances como sus limitaciones y así descubrir posibles áreas de mejora. Con esto en mente, los rasgos a evaluar en este ámbito se describen a continuación:

1. **Uso de preguntas inductivas:**
   - **Evaluación:** ¿El modelo formula preguntas que fomentan la reflexión y el análisis crítico en lugar de proporcionar respuestas directas?
     - **Ejemplo:** "¿Hay algo en particular que te guste de la primavera?"
   - **Limitaciones:** GPT-3 puede generar preguntas, pero la calidad de estas preguntas dependerá de la información contenida en sus datos de entrenamiento y podría carecer de profundidad en la comprensión del contexto específico.

2. **No imposición de ideas:**
   - **Evaluación:** ¿El modelo evita imponer ideas o soluciones, y en cambio, guía al interlocutor para que descubra sus propias respuestas?
     - **Ejemplo:** "Puedo entender la situación, ¿Podrías compartir más sobre lo que sentiste y pensaste en ese momento?"
   - **Limitaciones:** GPT-3 no tiene conciencia de la subjetividad individual ni la capacidad de adaptarse a experiencias personales específicas, lo que puede limitar su capacidad para guiar de manera efectiva.

3. **Ampliación y construcción de conocimiento:**
   - **Evaluación:** ¿El modelo participa en la construcción de conocimiento a través de la reflexión y el diálogo continuo?
     - **Ejemplo:** "¿Qué más hicieron durante su almuerzo?"

- **Limitaciones:** GPT-3 puede generar respuestas coherentes, pero su capacidad para construir conocimiento se basa en patrones de lenguaje aprendidos durante el entrenamiento, sin una verdadera comprensión conceptual.

4. **Generación de disonancia cognitiva:**
   - **Evaluación:** ¿El modelo utiliza el diálogo para provocar la disonancia cognitiva, desafiando las creencias o ideas del interlocutor?
   - **Limitaciones:** GPT-3 puede generar contenido que parezca desafiante, pero su capacidad para entender la complejidad de las creencias humanas puede ser limitada.

5. **Descubrimiento guiado:**
   - **Evaluación:** ¿El modelo se limita a hacer preguntas sistemáticas, guiando al interlocutor para que descubra sus propias creencias y conocimientos?
     - **Ejemplo:** "**Consultante:** Todos los hombres son iguales.

       Modelo: ¿Siempre has creído eso? ¿Recuerdas el momento en el que este pensamiento apareció por primera vez?"
   - **Limitaciones:** GPT-3 puede realizar preguntas, pero su enfoque es más basado en patrones y menos en una comprensión profunda de la experiencia individual.

Las tablas 4.4 y 4.5 presentan de manera concisa los criterios antes expuestos de la evaluación de escucha activa y del método socrático, respectivamente.

Evaluación: Escucha Activa	
Rasgos	**Descripción**
Atención contextual	¿El modelo demuestra coherencia en sus respuestas, mostrando que presta atención a la información proporcionada previamente en la conversación?
Formulación de preguntas clarificadoras	¿El modelo hace preguntas para obtener más detalles o aclaraciones sobre ideas, conceptos o sentimientos relacionados con los hechos mencionados?
Profundiza en la conversación	¿El modelo invita a profundizar en la conversación siendo coherente con el contexto dado por el interlocutor?
Ausencia de juicio o crítica	¿El modelo evita expresar juicios o críticas hacia las afirmaciones del interlocutor?
Demostración de empatía	¿El modelo responde de manera empática, reconociendo y reflejando las emociones expresadas por el interlocutor? ¿El modelo reconoce en su respuesta las emociones expresadas por el interlocutor?

Tabla 4.4: Rasgos de evaluación de escucha activa

Evaluación: Método Socrático	
Rasgos	**Descripción**
Uso de preguntas inductivas	¿El modelo formula preguntas que fomentan la reflexión y el análisis crítico en lugar de proporcionar respuestas directas?
No imposición de ideas	¿El modelo evita imponer ideas o soluciones, y en cambio, guía al interlocutor para que descubra sus propias respuestas?
Ampliación y construcción de conocimiento	¿El modelo participa en la construcción de conocimiento a través de la reflexión y el diálogo continuo?
Generación de disonancia cognitiva	¿El modelo utiliza el diálogo para provocar la disonancia cognitiva, desafiando las creencias o ideas del consultante?
Descubrimiento guiado	¿El modelo se limita a hacer preguntas sistemáticas, guiando al interlocutor para que descubra sus propias creencias y conocimientos? ¿El modelo hace preguntas sobre el origen o momento desencadenante de creencias o sentimientos?

Tabla 4.5: Rasgos de evaluación del método Socrático

4.3. Evaluación de los Modelos

Una vez obtenidas las respuestas de los cuatro modelos, se les aplicaron los criterios de evaluación de las dos metodologías propuestas en la sección anterior: la metodología socrática y la escucha activa. Este proceso se realizó en dos fases en paralelo: una evaluación humana y una evaluación automática, en la que se usó el modelo GPT-4 para calificar cada respuesta según los criterios de cada metodología, respondiendo con un sí o un no.

Las tablas 4.6 y 4.7 muestran la instrucción dada por cada método aplicado para la evaluación de las respuestas. La intención del $prompt$ es guiar el comportamiento del modelo que realiza la evaluación automática. El modelo debe actuar como un evaluador con los conocimientos necesarios sobre los métodos pertinentes al trabajo. Estos conocimientos deben aplicarse junto con los criterios proporcionados en la instrucción para que el modelo sepa específicamente qué está evaluando. La evaluación es sencilla: se trata de un caso de respuesta cerrada, determinando si el rasgo está presente o no en la respuesta, es decir, si es cierto o falso

"*You are an Evaluator with extensive knowledge in Psychotherapy, specifically in the application of active listening in therapy.You are going to evaluate the responses of a language model that was adjusted to respond as an emotional assistant, which provides emotional support through its responses to users by applying Karl Roger's active listening.To evaluate the model responses, you should consider the following evaluation traits:* {**Evaluación_Escucha_Activa**} *Where for each trait you must answer the evaluation question with True or False. For example:* *{ "Coherencia": True, "Atención contextual": True, "Formulación de preguntas clarificadoras": False, "Profundiza en la Conversación": True, "Ausencia de juicio o crítica": False, "Demostración de empatía": { "Primera Evaluación ": True, "Segunda Evaluación": False } }*

Tabla 4.6: *Prompt* de evaluación: escucha activa

"*You are an Evaluator with extensive knowledge in Psychotherapy, specifically in the application of the Socratic method in therapy.You are going to evaluate the responses of a language model that was adjusted to respond as an emotional assistant in Spanish, which provides emotional support through its responses to users applying the Socratic method.To evaluate the model's responses, you should consider the following evaluation features:* {**Evaluación_Método_socrático**} *Where for each trait you must answer the evaluation question with True or False. For example:* *{ "Uso de preguntas inductivas": True, "No imposición de ideas": False, "Ampliación y construcción de conocimiento": False, "Generación de disonancia cognitiva": True, "Descubrimiento guiado": False }* *According to compliance with the evaluation traits*"

Tabla 4.7: *Prompt* de evaluación: método Socrático

.

Se revisaron las respuestas para evaluar si cumplían con los rasgos del método y si respondían correctamente a la pregunta planteada por el método de evaluación. Con la ayuda de los ejemplos, estas tareas se simplifican un poco, ya que permitían ver la intención de la respuesta y hacia dónde se dirigía. Para facilitar la evaluación automática, se agregó un rasgo más a la rúbrica proporcionada al modelo evaluador: la pregunta "¿la respuesta es coherente o relevante con la entrada?". Esto tenía como propósito validar y descartar respuestas irrelevantes, ahorrando tiempo en etapas posteriores.

Una vez evaluadas todas las respuestas válidas de cada modelo, se contabilizaron para determinar cuántas respuestas de cada modelo acertaron en cada rasgo. Posteriormente, se aplicó una operación simple, excluyendo la pregunta utilizada para validar y filtrar las respuestas. Ambas metodologías se basan en cinco criterios, cada uno contribuyendo con un $20\,\%$ a la puntuación total, la cual varía entre 0 y $100\,\%$. La fórmula para calcular el valor de cada criterio se encuentra en la ecuación 4.4.

$$c_x = \frac{2}{n} \sum_{i=0}^{n} Element_i \tag{4.4}$$

donde x es el criterio que se está evaluando, n es el número de elementos y $Element_i$ es el valor del criterio para el elemento i^{th}.

La puntuación general se calcula como en la ecuación 4.5.

$$score = \sum_{x=0}^{m} c_x \tag{4.5}$$

donde c_x es el valor del criterio x, y m es el criterio a sumar.

4.4. Discusión

Los resultados de las evaluaciones fueron satisfactorios, especialmente para el modelo ajustado "Solo Escuchame", que destacó por su rendimiento. Este modelo obtuvo puntajes de $90.67/100$ en escucha activa y $77.12/100$ en el método socrático (como se muestra en las tablas 4.8 y 4.9), superando a GPT-3.5 (OpenIA, 2022), un modelo de mayor tamaño que se usó como referencia en el *fine-tuning*. Por otro lado, Mixtral8x7b (Jiang et al., 2024) demostró un buen rendimiento en estas tareas sin necesidad de *fine-tuning* previo.

Modelos	Atención contextual	Preguntas clarificantes	Profundiza en la conversación	Ausencia de juicio o crítica	Muestras de mpatía	Total
Solo Escúchame	18.78	13.83	19	19.69	19.35	90.67
GPT-3.5	19.03	11.75	18.40	19.57	18.85	87.62
Llama2 7b-chat	19.09	10.87	18.78	19.36	19.30	87.42
Mixtral 8x7b	19.34	8.04	17.95	19.68	19.49	84.52
GPT-2 124M	7	3.01	7.12	7.83	7.60	32.57

Tabla 4.8: Puntaje de cada rasgo de evaluación del método de escucha activa

Models	Uso de preguntas inductivas	No. imposición de ideas	Expansión y construcción de conocimiento	Generación de disonancia cognitiva	Exploración guiada	Total
Solo Escuchame	18.54	19.68	18.86	**1.04**	18.98	77.12
GPT-3.5	16.31	18.72	15.62	0.68	16.5	67.84
Llama2 7b-chat	15.65	17.72	16.22	0.54	16.30	66.45
Mixtral 8x7b	13.21	18.18	14.72	0.51	14.96	61.60
GPT-2 124M	7.60	8.12	7.16	0.24	7.54	30.68

Tabla 4.9: Puntaje de cada rasgo de evaluación del método Socrático

En las siguientes subsecciones se analizan el rendimiento y el comportamiento de cada modelo.

4.4.1. GPT3.5

Este modelo destaca en la escucha activa, pero encuentra desafíos al emplear el método socrático. Por ejemplo, el modelo presenta limitaciones a la hora de generar preguntas diseñadas para provocar disonancia cognitiva. Inducir constantemente disonancia cognitiva en las personas puede resultar irritante, por lo que este rasgo no debería ser una constante en todas las respuestas.

Contextual y estructuralmente las respuestas son adecuadas, con un promedio de 77 *tokens*, cumpliendo efectivamente la tarea.

4.4.2. Llama2-7b-chat

El modelo presenta cierta inconsistencia en el uso del lenguaje, respondiendo en inglés en un $29\,\%$ de las ocasiones, incluso cuando se le solicita explícitamente que lo haga en

español. Las respuestas suelen ser extensas, con un promedio de 111 *tokens* cada una. Además, el modelo frecuentemente inicia sus respuestas con un saludo innecesario, lo que puede interrumpir el flujo de la conversación. Para colmo, concluye sus intervenciones con un lenguaje similar al de una despedida, lo que dificulta la posibilidad de profundizar en el diálogo.

4.4.3. GPT2_124M ajustado

Este modelo tiene el rendimiento más bajo y no es adecuado para la tarea. Aún y con el pre-entrenamiento que se le hizo con datos en español, el modelo deforma nombres e introduce palabras inexistentes. La coherencia presentada al principio de la frase se desvanece después de 50 *tokens*. La longitud media de la respuesta es de 100 *tokens*. GPT2-124M obtiene la puntuación más baja entre los modelos evaluados.

4.4.4. Mixtral 8x7b

El rendimiento es comparable al de GPT3.5 y solo requiere un mensaje para completar satisfactoriamente la tarea. Las respuestas mantienen una estructura adecuada, similar al conjunto de entrenamiento, pero tiene un rendimiento menor en el método socrático. Cabe destacar que este modelo, a pesar de ser más de 10 veces más pequeño que GPT-3.5, tiene un rendimiento excepcional. Mixtral sólo requiere un *prompt* adecuado y produce respuestas con alta calidad y con una longitud promedio de 78 tokens.

4.4.5. Sólo escúchame

Después de un FT, este modelo supera a los otros modelos, logrando los mejores resultados en ambas metodologías. Superando notablemente a los otros modelos en empatía,

con una puntuación de 19.35 sobre 20, a excepción de los 19.49 puntos de Mixtral. Todas las respuestas exhiben consistentemente el rasgo de ausencia de juicio, una característica única entre los modelos. Estructuralmente, las respuestas se adaptan bien a los datos de entrenamiento, incluso reduciendo el recuento promedio de tokens a 65 por respuesta. Si bien no es perfecto y comparte un vocabulario empático limitado con GPT3.5, este modelo cumple satisfactoriamente la tarea.

Las observaciones anteriores indican que los modelos de lenguaje modernos, como GPT-3, Mixtral 8x7b y Llama2-7b-chat, ya muestran habilidades para proporcionar acompañamiento emocional. Sin embargo, requieren ajustes y un enfoque adecuado para mejorar su efectividad, tal como se ha evidenciado con el modelo "Sólo Escúchame". A pesar de que este modelo presenta ciertos errores heredados de su base, es capaz de identificar patrones útiles para cumplir con la tarea gracias a los datos de entrenamiento. Con un mayor volumen de datos, una mayor variedad de situaciones y conversaciones de mejor calidad y extensión, es probable que se reduzcan significativamente los problemas actuales de estos modelos.

4.5. Ejemplos de Conversación

Los ejemplos mostrados a continuación fueron generados usando el modelo "Sólo Escúchame".

4.5.1. Conversación: A

A lo largo de la interacción mostrada en la tabla 4.10, el asistente emocional demuestra el uso de varios rasgos que se establecen en la escucha activa. Principalmente la presencia de empatía, entiende la emoción que el usuario está expresando, además del uso de

User: Hola, me siento molesta el día de hoy
Chatbot: Hola Jannette, lamento que te sientas molesta. Si deseas compartir más sobre lo que sucedió, estoy aquí para escucharte.
User: En serio que como me caga que mi novio haga las cosas y no me las cuente, me las oculte.
Chatbot: Jannette, es comprensible que te sientas molesta cuando alguien no te comunica lo que está sucediendo en su vida. ¿Quieres hablar más sobre lo que está sucediendo con tu novio y cómo te hace sentir?
User: Es que siento que no confía en mí para contármelo, y eso me hace sentir excluida.
Chatbot: Es comprensible que te sientas excluida cuando alguien no te comunica lo que está sucediendo en su vida. La comunicación abierta y honesta es clave para un buen enfoque en una relación. ¿Has intentado hablar con tu novio sobre cómo te sientes?
User: ¿crees que él lo pueda entender?
Chatbot: Claro, Jannette. Entiendo que la comunicación abierta y honesta es clave en una relación. Si no has hablado con tu novio sobre cómo te sientes, tal vez sea una buena idea hacerlo. Recuerda que es importante expresar tus sentimientos y necesidades para mantener una relación saludable. ¿Quieres hablar más sobre cómo te sientes?
User: Creo que no, bueno, no al menos hasta hablar con el y saber su reaccion. Pero gracias
Chatbot: ¡No hay problema! Estoy aquí para escucharte y brindarte apoyo.

Tabla 4.10: Ejemplo de conversación: situación amorosa comprometida

preguntas clarificadoras, incluyendo la intención de profundización de la conversación. En última instancia deja ver una ausencia de juicio. El modelo cumple con la mayoría de los puntos a evaluar por el método de la escucha activa (véase sección 4.4).

Una de las respuestas más interesantes en esta conversación contiene el siguiente fragmento: "*Entiendo que la comunicación abierta y honesta es clave en una relación. Si no has hablado con tu novio sobre cómo te sientes, tal vez sea una buena idea hacerlo. Recuerda que es importante expresar tus sentimientos y necesidades para mantener una relación saludable*", sugiriendo al usuario una vía de acción, que podría llevarla a una situación favorable. Esto es importante, porque el modelo además de entender el contexto, también puede mapear la importancia de la comunicación, además de que la falta de comunicación pueden llegar a ser problemáticas.

4.5.2. Conversación: B

User: Hola! **Chatbot**: Hola, ¿cómo estás hoy? ¿Hay algo en lo que pueda ayudarte? **User**: Estoy muy bien y tu? **Chatbot**: Me alegra saber que estás bien. ¿Hay algo en lo que pueda ayudarte? Estoy aquí para escucharte. **User**: Pues hoy tuve un buen dia, quieres escucharlo? **Chatbot**: Me alegra mucho que hayas tenido un buen día. Cuéntame más sobre lo que sucedió. **User**: Pues mira, te cuento que hoy en mi trabajo me dijeron que las cosas se estan poniendo interesantes, aun mas con el proyecto que estoy desarrollando, porque a unos proveedores les intereso, hoy agendaron una cita. **Chatbot**: ¡Eso suena emocionante! Es genial cuando las cosas se ponen interesantes en el trabajo. ¿Cómo te sientes al respecto? **User**: Pues bastante emocionado, pero nervioso a la vez, no se como decirlo, pero tengo miedo a fallar y que cancelen todo el proyecto. **Chatbot**: Comprendo tu nerviosismo, pero recuerda que todos cometemos errores y lo importante es aprender de ellos. Confía en tus habilidades y en el proyecto que estás desarrollando. Estoy aquí para apoyarte en lo que necesites. **User**: Mil gracias por la fe en mi. Ya te estare contando como me fue. Bye! **Chatbot**: ¡Bye! Estoy emocionado de escuchar cómo te va en tu proyecto. Recuerda que siempre estoy aquí para apoyarte. ¡Buena suerte!

Tabla 4.11: Ejemplo de conversación: grandes expectativas laborales

En la conversación de la tabla 4.11, el modelo muestra una capacidad notable para la escucha activa, cumpliendo satisfactoriamente con los aspectos evaluados en la tabla 4.4. El modelo responde de manera coherente, hace preguntas clarificadoras, profundiza en la conversación, evita juicios, y demuestra empatía. Sin embargo, podría mejorar en la formulación de preguntas más específicas para obtener detalles adicionales y ofrecer una comprensión más profunda de la situación del usuario.

En los ejemplos previos se pudo observar que el rendimiento del modelo en esta tarea es favorable, inclinándose por emplear el método de escucha activa (Rogers and Farson, 2015) como método de acompañamiento principal, tal como lo demostró la evaluación a través del conjunto de prueba.

4.5.3. Problemas de rendimiento

Aunque el modelo puede tener un rendimiento excepcional en casos más tradicionales, como los chatbots, su desempeño puede verse afectado en conversaciones más largas y complejas, por ejemplo, en la conversación de la tabla 4.12, se nota como sin importar la entrada del usuario, el modelo se enfoca en el contexto previo de la charla, y repite "*Lamento que hayas tenido esa experiencia incómoda con tu amigo y tu ex pareja. A veces, las situaciones pueden ser difíciles de manejar. ¿Quieres hablar más sobre cómo te sientes al respecto?*". Aunque fueron modificados en diferentes ocasiones los parámetros de temperatura, top_k y top_p para variar las probabilidades del vocabulario, la respuesta mantenía la misma estructura con muy ligeras variaciones.

User: Este fin de semana que estuve con mis amigos, me di cuenta de algo muy incomodo, pero si te soy sincero no se como actuar al respecto. Me di cuenta que mi amigo Alex, le esta hablando a mi ex pareja, yo creo que el no se dio cuenta, perocuando se me acerco para platicar, en su telefono aparecio una notificacion de un mensaje con la foto de perfil de mi ex.no supe que decir, y mejor me calle. No se como decirle que es incomodo que hable con mi ex. **Chatbot:** Lamento que hayas tenido esa experiencia incómoda con tu amigo y tu ex pareja. A veces, las situaciones pueden ser difíciles de manejar. ¿Quieres hablar más sobre cómo te sientes al respecto? **User:** Alex y mi ex, siempre se llevaron bien. No se como decirlo, pero con otros amigos mi ex era mas reservada. Cuando eran fiestas o reuniones no se despegaba de mi, pero si estaba Alex, pues ella podia quedarse platicando horas. Eso antes no me molestaba, pues eran amigos nada mas. pero ahora no se que este pasando entre ellos. Algo me dice que ella siempre estuvo atraida por Alex y el hacia ella. Chatbot: Lamento que hayas tenido esa experiencia incómoda con tu amigo y tu ex pareja. A veces, las situaciones pueden ser difíciles de manejar. ¿Quieres hablar más sobre cómo te sientes al respecto? **User:** . **Chatbot:** Lamento que hayas tenido esa experiencia incómoda con tu amigo y tu ex pareja. A veces, las situaciones pueden ser difíciles de manejar. ¿Quieres hablar más sobre cómo te sientes al respecto?

Tabla 4.12: Ejemplo de conversación: ciclado de respuestas

Existen 2 razones principales para explicar el comportamiento del modelo:

1. El sesgo en los datos de entrenamiento, dado por el desbalance en las respuestas. Aunque se aplicaron medidas para evitar esta situación como el *undersampling* para que todas las categorías tuvieran el mismo número de elementos, esto no excluye que la longitud de cada ejemplo fuera desbalanceado, teniendo mayoritariamente muestras más simples o fáciles en una categoría emocional que en otra. Lo anterior, aparentemente sesgó al modelo a tener un peor rendimiento bajo ciertas circunstancias.

2. El modelo está subentrenado y requiere más datos, especialmente aquellos que simulen conversaciones de múltiples turnos. Es fundamental incluir una variedad de respuestas y peticiones del usuario para que el modelo pueda adaptarse a un estilo conversacional más tradicional.

Ambas razones están interrelacionadas, el desbalance se puede corregir proporcionando más datos variados para mejorar tanto la calidad como la diversidad de las respuestas. Además, los datos en forma de conversaciones con múltiples turnos pueden complementar la estructura de respuesta del modelo, permitiéndole aprender de manera óptima a generar conversaciones emocionalmente fructíferas para el usuario.

Capítulo V

Conclusiones y Trabajo Futuro

5.1. Conclusiones

En la presente tesis, se investigó el desarrollo y rendimiento de modelos de lenguaje modernos en la tarea de acompañamiento emocional. El objetivo principal fue evaluar la capacidad de estos modelos para interactuar de manera empática y efectiva con los usuarios, empleando metodologías de escucha activa y el método socrático.

Los resultados obtenidos sobre los modelos evaluados indicaron que GPT-3 y Mixtral 8x7b ya poseen capacidades considerables para el acompañamiento emocional. Esto sugiere que las bases para realizar este tipo de acompañamiento pueden adquirirse indirectamente mediante el entrenamiento en grandes conjuntos de datos, especialmente conversacionales. Sin embargo, esto no garantiza la calidad de la ejecución de dicha tarea. Por esta razón, el ajuste realizado en el modelo "Sólo Escúchame", basado en Llama-2, potenció sus capacidades para profundizar en una conversación, utilizando la atención contextual y demostrando empatía.

Estos hallazgos son significativos porque muestran el potencial de los modelos de lenguaje para aplicaciones en áreas de soporte emocional y terapia digital. La efectividad

de “Sólo Escúchame” sugiere que los ajustes adecuados, pueden hacer que los modelos de lenguaje pueden ofrecer un apoyo emocional valioso, reduciendo la carga de trabajo para profesionales de la salud mental.

A pesar de los resultados prometedores, el estudio presenta varias limitaciones. Las más urgentes son la calidad y longitud de las conversaciones que los modelos pueden mantener. Incrementar la capacidad de estos modelos implica aumentar los datos de entrenamiento, agregando conversaciones con variedad de situaciones, longitud y nivel de dificultad. Lo anterior ampliaría las habilidades que son necesarias para enfrentarse a situaciones más complejas y específicas, mejorando el rendimiento del modelo ”Sólo Escúchame“ y otros modelos similares en la tarea de acompañar a los usuarios en situaciones emocionales complejas y específicas.

Además, la evaluación se centró en escenarios predefinidos, lo cual podría no reflejar completamente la complejidad de las interacciones humanas reales. Los modelos pueden comportarse de manera diferente cuando se enfrentan a conversaciones espontáneas y variadas que no fueron consideradas durante el entrenamiento. Por lo tanto, futuras investigaciones deberían enfocarse en probar estos modelos en entornos más dinámicos y realistas.

Otra limitación significativa es el hardware necesario para ejecutar los modelos de lenguaje actuales. Estos modelos, especialmente aquellos como GPT-3, Llama-2 o Mixtral 8x7b, requieren recursos computacionales sustanciales. La necesidad de hardware potente no solo implica altos costos, sino también una infraestructura tecnológica adecuada, lo cual puede ser una barrera para su adopción en organizaciones más pequeñas o en regiones con menos recursos. Esta limitación destaca la importancia de investigar formas de optimizar estos modelos para que puedan funcionar eficientemente en entornos con recursos limitados. En este sentido GGUF (Gerganov, 2023) puede ofrecer una

solución, pero hacen falta más pruebas para ver cómo estás soluciones impactan en el redimiendo, pues es bien sabido que cuantizar los pesos de un modelo de 32-bits a 8-bits le resta precisión a la hora de inferir.

Finalmente, la salud mental es una necesidad fundamental y se encuentra en el centro de muchos problemas que afectan a la sociedad en la actualidad. Por ello, es prioritario investigar las causas de los problemas de salud mental y desarrollar herramientas eficaces para diagnosticarlos, tratarlos o mitigar sus efectos negativos en individuos y comunidades.

En este sentido, el uso de modelos de lenguaje en el acompañamiento emocional es factible y puede proporcionar un apoyo significativo en el día a día de las personas, miembros de cualquier comunidad u organización. Sin embargo, su implementación no debe tomarse a la ligera, ya que plantea varios desafíos éticos y morales que deben ser abordados con seriedad. La privacidad y confidencialidad de los datos de los usuarios son de máxima importancia, así como el consentimiento informado para el uso de estos sistemas. Además, es crucial reconocer las limitaciones actuales de estos modelos y educar a los usuarios sobre cuándo deben buscar ayuda profesional.

La identificación y mitigación de sesgos inherentes a los modelos es otro aspecto crítico para garantizar interacciones justas y equitativas. Las organizaciones deben asumir la responsabilidad y rendición de cuentas por el uso de estas tecnologías, implementando mecanismos para resolver problemas y mejorar continuamente la seguridad y efectividad de estos sistemas.

5.2. Trabajo futuro

Las metas a corto plazo para este proyecto incluyen mejorar el conjunto de datos aumentando el número de elementos por clase, con el objetivo de crear un conjunto más grande y diverso. También es prioritario darle un enfoque conversacional al conjunto, incorporando diálogos completos de dos turnos entre el usuario y el asistente emocional. Se planea recopilar ejemplos de conversaciones que apliquen eficazmente el método socrático y la escucha activa, los cuales se utilizarán para generar más variaciones y así ampliar los datos de entrenamiento disponibles para la tarea de acompañamiento emocional. Además, contar con este tipo de diálogos será beneficioso para las muestras de texto ya incluidas en el conjunto de datos *HEAR* (sección 3.3.2), ya que permitirá instruir a un modelo como GPT-4 (OpenAI et al., 2024) utilizando técnicas de *prompting* para generar conversaciones alrededor de las $45,450$ muestras (véase sección 3.3.2) del conjunto propuesto en esta tesis.

No solo hay planes a corto plazo; los avances recientes en la comunidad de *machine learning* han sido, sin duda, impresionantes. Los modelos de lenguaje se han convertido en herramientas cada vez más relevantes en diversos medios de producción. Estas redes neuronales, diseñadas para analizar, clasificar y generar texto, han evolucionado y ahora pueden manejar múltiples tipos de datos, incluyendo texto, imágenes y audio, de manera secuencial o simultánea, un enfoque conocido como multimodalidad (Hsu, 2023). Estas redes neuronales son tan potentes que comprenden cómo funcionan ciertos elementos de la realidad y pueden simular, hasta cierto punto, el funcionamiento del mundo. Un ejemplo notable es SORA (Openai, 2024), un modelo multimodal capaz de analizar y generar vídeos y secuencias cinematográficas complejas a partir de texto, imágenes o vídeos. Con una dosis de creatividad, SORA logra mantener una coherencia sorprendente en relación con cómo operan las cosas en el mundo real.

Los modelos multimodales y arquitecturas como *MoE* ofrecen una amplia gama de oportunidades para desarrollar agentes virtuales especializados en la salud mental, constituyendo una herramienta valiosa para los profesionales de este campo. Un agente virtual que pueda comprender el lenguaje humano en toda su complejidad debe ser capaz de analizar y captar la carga emocional que este transmite. La comunicación verbal abarca múltiples formas de expresión, tanto orales como escritas. La voz, además de ser un elemento fundamental de la expresión oral, aporta matices emocionales significativos (Cabrelles Sagredo, 2023). Por ejemplo, aspectos como el tono, la melodía y la velocidad de la voz pueden transmitir diferentes estados de ánimo; un tono alto y variado, junto con un ritmo rápido, puede indicar alegría.

Asimismo, el lenguaje corporal es una forma no verbal de comunicación (Carrera, 2021) que a menudo acompaña la expresión oral, aunque en ocasiones puede ser la única vía de interacción. Factores como la postura, el contacto visual y las expresiones faciales proporcionan información valiosa sobre el estado emocional y la intención de una persona. En este sentido, entender y analizar ambos tipos de comunicación —verbal y no verbal— es crucial para el desarrollo de agentes virtuales que realmente comprendan y apoyen la salud mental.

Por lo tanto, a incorporación de análisis de audio y reconocimiento de gestos a través de imágenes podría mejorar significativamente la capacidad de los agentes virtuales para comprender y responder de manera adecuada a la carga emocional de los usuarios. Esto permitiría una interacción más rica y matizada, acercándose a la experiencia de una conversación humana real. Esta multimodalidad, puede implementarse de diversas maneras, ya que la expresión oral y corporal son fuentes de información que pueden ser representadas por diferentes tipos de entradas, como imágenes, audio y texto. Un sistema que combine modelos de visión, audio y procesamiento del lenguaje natural

podría beneficiarse enormemente de esta integración.

Otra alternativa es adoptar un modelo multimodal que procese entradas multimedia, como el vídeo, tal como lo hace SORA (Openai, 2024). Sin embargo, en ambas enfoques, el principal desafío radica en la obtención de información para entrenar estos modelos. Aunque contamos con una base para un conjunto de datos conversacional basado en texto (*HEAR* 3.3.2), la recolección, clasificación y etiquetado de otros tipos de datos, como imágenes, puede resultar especialmente complicado. Estas imágenes deben capturar una amplia gama de emociones y situaciones variadas para permitir que un modelo multimodal o uno de visión generalice los aspectos más importantes de las escenas. Lo mismo ocurre con las fuentes de audio. A pesar de estos desafíos, estos esfuerzos son fundamentales para desarrollar un sistema que comprenda en su totalidad el lenguaje natural humano.

Por último, GGUF (Gerganov, 2023) es un formato de archivo unificado diseñado específicamente para almacenar y ejecutar modelos como Llama-2 de manera optimizada. Utilizando mmap, asigna una región de memoria virtual a un proceso, permitiendo cargar y guardar modelos de forma eficiente. GGUF contiene toda la información necesaria para su uso sin requerir archivos adicionales, optimizando los modelos para ser ejecutados en hardware menos potente. Esto lo hace más accesible para usuarios con presupuestos limitados (Gimmi, 2023), permitiendo que ejecuten estos modelos en sus propios dispositivos sin necesidad de hardware especializado costoso o servicios en la nube. GGUF es adaptable a una variedad de modelos de lenguaje y arquitecturas de hardware, convirtiéndose en una herramienta valiosa para la investigación y el desarrollo de nuevos modelos de lenguaje.

Todas las tecnologías mencionadas durante está sección podrían usarse para mejorar el proyecto de varias maneras. Por ejemplo, los modelos multimodales pueden proporcio-

nar una experiencia de acompañamiento más rica e intuitiva al combinar texto, imágenes y audio, lo que podría simular mejor una interacción humana completa. La adopción de modelos MoE puede permitir que el sistema utilice el experto más adecuado para cada situación, optimizando las respuestas y haciéndolas más pertinentes y útiles para los usuarios. Además, se vislumbra pertinente el uso de GGUF para cargar al modelo en un solo archivo y así reducir los requerimientos de hardware. Estas mejoras podrían hacer este tipo de acompañamiento emocional más accesible, poniéndolo disponible en dispositivos de uso común como una PC o teléfonos móviles, dispositivos IoT (*Internet of Things* o internet de las cosas), democratizando el acceso al acompañamiento emocional de alta calidad.

Bibliografía

Abrams, Z. (2021). The promise and challenges of AI.

Amazon Web Services (s.f.). ¿qué es el sobreajuste?

American Psychiatric Association, A. P. A. (2012). Entendiendo la psicoterapia.

American Psychiatric Association, A. P. A. (2014). Trastornos de ansiedad, page 129–144. American Psychiatric Publishing, 5 edition.

Andrews, G. (2021). ¿Qué Son los Datos Sintéticos? | Blog de NVIDIA.

Asai, A., Evensen, S., Golshan, B., Halevy, A., Li, V., Lopatenko, A., Stepanov, D., Suhara, Y., Tan, W.-C., and Xu, Y. (2018). HappyDB: A corpus of 100,000 crowdsourced happy moments. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan. European Language Resources Association (ELRA).

Barbieri, F., Camacho-Collados, J., Espinosa Anke, L., and Neves, L. (2020). TweetEval: Unified benchmark and comparative evaluation for tweet classification. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 1644–1650, Online. Association for Computational Linguistics.

Black, S., Gao, L., Wang, P., Leahy, C., and Biderman, S. (2021). GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow. If you use this software, please cite it using these metadata.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners.

Brownlee, J. (2020). How to use weight decay to reduce overfitting of neural network in Keras.

Bupa (2024). Terapia cognitivo conductual (TCC): así es – Bupa Latam.

Cabrelles Sagredo, M. S. (2023). La influencia de las emociones en el sonido de la voz.

Cao, C., Sang, J., Arora, R., Kloosterman, R., Cecere, M., Gorla, J., Saleh, R., Chen, D., Drennan, I., Teja, B., et al. (2024). Prompting is all you need: Llms for systematic review screening. medRxiv, pages 2024–06.

Carrera, G. D. (2021). ¿Cómo el lenguaje corporal representa nuestras emociones?

Carreño, T. G. (2019). La Salud mental, su nuevo impacto y desafío.

Chatterjee, A. (2022). A Tale of Two Languages: The Code Mixing story - Wipro Tech Blogs - Medium.

Chen, D. (2024). Crafting Clarity: A guide to effective instructional prompts. unknown.

Cho, K., van Merriënboer, B., Bahdanau, D., and Bengio, Y. (2014). On the properties of neural machine translation: Encoder–decoder approaches. In Wu, D., Carpuat, M., Carreras, X., and Vecchi, E. M., editors, Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, pages 103–111, Doha, Qatar. Association for Computational Linguistics.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., Reif, E., Du, N., Hutchinson, B., Pope, R., Bradbury, J., Austin, J., Isard, M., Gur-Ari, G., Yin, P., Duke, T., Levskaya, A., Ghemawat, S., Dev, S., Michalewski, H., Garcia, X., Misra, V., Robinson, K., Fedus, L., Zhou, D., Ippolito, D., Luan, D., Lim, H., Zoph, B., Spiridonov, A., Sepassi, R., Dohan, D., Agrawal, S., Omernick, M., Dai, A. M., Pillai, T. S., Pellat, M., Lewkowycz, A., Moreira, E., Child, R., Polozov, O., Lee, K., Zhou, Z., Wang, X., Saeta, B., Diaz, M., Firat, O., Catasta, M., Wei, J., Meier-Hellstern, K., Eck, D., Dean, J., Petrov, S., and Fiedel, N. (2022). Palm: Scaling language modeling with pathways.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., and Schulman, J. (2021). Training verifiers to solve math word problems. ArXiv, abs/2110.14168.

Computer, T. (2023). Redpajama: An open source recipe to reproduce llama training dataset.

Cortés-Álvarez, N. Y., Piñeiro-Lamas, R., and Vuelvas-Olmos, C. R. (2020). Psychological effects and associated factors of COVID-19 in a Mexican sample. Disaster Medicine and Public Health Preparedness, 14(3):413–424.

Dagan, I., Dolan, B., Magnini, B., and Roth, D. (2009). Recognizing textual entailment: Rational, evaluation and approaches. Natural Language Engineering, 15(Special Issue 04):i–xvii.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Dolan, W. B. and Brockett, C. (2005). Automatically constructing a corpus of sentential paraphrases. In Proceedings of the Third International Workshop on Paraphrasing (IWP2005).

Económica, E. (2022). Muestreo aleatorio simple. Enciclopedia Económica.

Fetterman, A. J., Kitanidis, E., Albrecht, J., Polizzi, Z., Fogelman, B., Knutins, M., Wróblewski, B., Simon, J. B., and Qiu, K. (2023). Tune as you scale: Hyperparameter optimization for compute efficient training.

Fiske, A., Henningsen, P., and Buyx, A. (2019). Your robot therapist will see you now: Ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy. J Med Internet Res, 21(5):e13216.

Fukushima, K. (1969). Visual feature extraction by a multilayered network of analog threshold elements. IEEE Trans. Syst. Sci. Cybern., 5:322–333.

García-Allen, J. (2022). Disonancia cognitiva. unknown.

GeeksforGeeks (2024). In the Context of Deep Learning, What Is Training Warmup Steps?

Gehman, S., Gururangan, S., Sap, M., Choi, Y., and Smith, N. A. (2020). Realtoxicityprompts: Evaluating neural toxic degeneration in language models. In Findings.

Geng, X. and Liu, H. (2023). Openllama: An open reproduction of llama.

Gerganov, G. (2023). ggml/docs/gguf.md at Master · Ggerganov/ggml.

Gimmi, P. (2023). What is GGUF and GGML? - Phillip Gimmi - Medium.

González Velázquez, L. (2020). Estrés académico en estudiantes universitarios asociado a la pandemia por covid-19. Espacio I+D, Innovación más desarrollo, 9(25).

Guerri, M. (2021). El coste social y económico de los trastornos mentales.

Guillem, P. M. (2018). Spanish arilines tweets sentiment analysis.

Gunton, M. (2024). Understanding the sparse mixture of experts (SMOE) layer in Mixtral.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8):1735–1780.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., de Las Casas, D., Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., van den Driessche, G., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., Rae, J. W., Vinyals, O., and Sifre, L. (2022). Training compute-optimal large language models.

Hsu, M.-H. (2023). Multimodal Machine Learning - Ming-Hao HSU - Medium.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2021). Lora: Low-rank adaptation of large language models.

Huang, Q., Liu, X., Ko, T., Wu, B., Wang, W., Zhang, Y., and Tang, L. (2024). Selective prompting tuning for personalized conversations with llms.

Huang, Z., Xu, W., and Yu, K. (2015). Bidirectional lstm-crf models for sequence tagging.

HuggingFace (n.d.). huggingface.

IBM (n.d.a). What is Deep Learning?

IBM (n.d.b). What is Machine Learning? .

INEGI (n.d.). Encuesta para la Medición del Impacto COVID-19 en la Educación (ECOVID-ED) 2020.

Ji, S., Zhang, T., Ansari, L., Fu, J., Rodrigues, J. J. P. C., and Cambria, E. (2021). MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare. arXiv (Cornell University).

Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D. S., de las Casas, D., Hanna, E. B., Bressand, F., Lengyel, G., Bour, G., Lample, G., Lavaud, L. R., Saulnier, L., Lachaux, M.-A., Stock, P., Subramanian, S., Yang, S., Antoniak, S., Scao, T. L., Gervet, T., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. (2024). Mixtral of experts.

Kalauni, B. R., Joshi, Y. P., Paudel, K., Aryal, B., Karki, L., and Paudel, R. (2023). Depression, anxiety and stress among undergraduate health sciences students during

COVID-19 pandemic in a low resource setting: a cross sectional survey from Nepal. Annals of medicine and surgery, Publish Ahead of Print.

Khodak, M., Saunshi, N., and Vodrahalli, K. (2018). A large self-annotated corpus for sarcasm. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan. European Language Resources Association (ELRA).

Lang, K. (2022). El impacto de la pandemia en la salud mental: Un panorama mundial.

Lee, J. (2020). Mental health effects of school closures during covid-19. The Lancet Child & Adolescent Health, 4(6):421.

Lee, S., Kang, W. S., Cho, A. R., Kim, T., and Park, J. K. (2018). Psychological impact of the 2015 mers outbreak on hospital workers and quarantined hemodialysis patients. Comprehensive Psychiatry, 87:123–127.

Li, Y., Su, H., Shen, X., Li, W., Cao, Z., and Niu, S. (2017). Dailydialog: A manually labelled multi-turn dialogue dataset.

Lin, I. W., Njoo, L., Field, A., Sharma, A., Reinecke, K., Althoff, T., and Tsvetkov, Y. (2022). Gendered Mental Health Stigma in Masked Language Models. arXiv (Cornell University).

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. cite arxiv:1907.11692.

Losada-Baltar, A., Márquez-González, M., Jiménez-Gonzalo, L., Pedroso-Chaparro, M. D. S., Gallego-Alberto, L., and Fernandes-Pires, J. (2020). Differences in anxiety,

sadness, loneliness and comorbid anxiety and sadness as a function of age and self-perceptions of aging during the lock-out period due to covid-19. Revista espanola de geriatria y gerontologia.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. (2011). Learning word vectors for sentiment analysis. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Minds, A. (2020). The impact of covid-19 on student mental health. Online; accessed 19 August 2023.

Naji, I. (2012). TSATC: Twitter Sentiment Analysis Training Corpus. In thinknook.

Oblitas, L. A. (2008). Psicoterapias contemporáneas. Cengage Learning Editores.

Openai (2024). Video generation models as world simulators.

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., and Baltescu, P. (2024). Gpt-4 technical report.

OpenIA (2022). Introducing chatgpt. https://openai.com/blog/chatgpt.

Oprea, S. and Magdy, W. (2020). iSarcasm: A dataset of intended sarcasm. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1279–1289, Online. Association for Computational Linguistics.

Oraby, S., Harrison, V., Reed, L., Hernandez, E., Riloff, E., and Walker, M. (2016). Creating and characterizing a diverse corpus of sarcasm in dialogue. In Proceedings

of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 31–41, Los Angeles. Association for Computational Linguistics.

Oracle (n.d.). ¿Qué es la inteligencia artificial (IA)? .

Organization, W. H. (2022). Salud mental: fortalecer nuestra respuesta.

Palomino, C. and Huarcaya-Victoria, J. (2020). Trastornos por estrés debido a la cuarentena durante la pandemia por la COVID-19. Horizonte Medico (Lima), 20.

Patwa, P., Aguilar, G., Kar, S., Pandey, S., PYKL, S., Gambäck, B., Chakraborty, T., Solorio, T., and Das, A. (2020). SemEval-2020 task 9: Overview of sentiment analysis of code-mixed tweets. In Proceedings of the Fourteenth Workshop on Semantic Evaluation, pages 774–790, Barcelona (online). International Committee for Computational Linguistics.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12:2825–2830.

Plutchik, R. (1980). Emotion: a psychoevolutionary synthesis. American Journal of Psychology, 93(4):751.

Porto, J. P. and Merino, M. (2021). Mayéutica - Qué es, cómo se desarrolla, definición y concepto.

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018a). Improving language understanding by generative pre-training. unknow.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2018b). Language models are unsupervised multitask learners. unknown.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer.

Raffo, P. and Instituto Interamericano de Derechos Humanos, I. I. d. D. H. (2007). Acompañamiento psicológico y terapia psicológica. Instituto Interamericano de Derechos Humanos.

Rodriguez, J. L., Fernández, L. N., Guerra-Hernández, M. A., Llerena, T. Z., Guerra, J. C. L., Chomat, R. C. A., and Herrera, R. G. (2021). Soledad y su asociación con depresión, ansiedad y trastornos del sueño en personas mayores cubanas durante la pandemia por covid-19. Anales de la Academia de Ciencias de Cuba, 11(3):1005.

Rogers, C. and Farson, R. (2015). Active Listening. Martino Fine Books.

Roy, R. (2022). Neural Networks: Forward pass and Backpropagation - Towards Data Science. unknow.

Ryanholbrook (2023). Overfitting and Underfitting.

Sahai, S. (2022). What is the Data Collator class in transformers? - .

Sanchez Rasmos, M. B. O., Capacha Huamaní, D. A. V., Capcha Huamaní, M. M. L., Quispe Olano, D. J., and Reza Condori, S. Z. (2021). Estrés académico en estudiantes universitarios en contexto de la pandemia por covid-19: una revisión sistemática. Ciencia Latina Revista Científica Multidisciplinar, 5(6):11279–11290.

Sanseviero, O., Tunstall, L., Schmid, P., Mangrulkar, S., Belkada, Y., and Cuenca, P. (2023). Mixture of experts explained.

Santillán, M. L. (2023). Cómo afecta la salud mental en el rendimiento escolar.

Saravia, E. (2022). Prompt Engineering Guide. https://github.com/dair-ai/Prompt-Engineering-Guide.

Saravia, E., Liu, H.-C. T., Huang, Y.-H., Wu, J., and Chen, Y.-S. (2018). CARER: Contextualized affect representations for emotion recognition. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 3687–3697, Brussels, Belgium. Association for Computational Linguistics.

Sarrió, C. (2021). ¿Qué es el contacto en Terapia Gestalt?

Science, B. O. C. and Science, B. O. C. (2023). Training and Validation Loss in Deep Learning | Baeldung on Computer Science.

Serrano, M. (2017). Acompañamiento emocional respetuoso: tres condiciones básicas para acompañar.

Services, H. C. P. (2018). Depression Anxiety and Stress Scale 21 (DASS-21) – Health-Focus Clinical Psychology Services.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., and Potts, C. (2013a). Recursive deep models for semantic compositionality over a sentiment treebank. In Yarowsky, D., Baldwin, T., Korhonen, A., Livescu, K., and Bethard, S., editors, Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., and Potts, C. (2013b). Recursive deep models for semantic compositionality over a sentiment treebank. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Solaiman, I., Brundage, M., Clark, J., Askell, A., Herbert-Voss, A., Wu, J., Radford, A., Krueger, G., Kim, J. W., Kreps, S., McCain, M., Newhouse, A., Blazakis, J., McGuffie, K., and Wang, J. (2019). Release strategies and the social impacts of language models.

Sprang, G. and Silman, M. (2013). Post-traumatic stress disorder in parents and youth after health-related disasters. Disaster Medicine and Public Health Preparedness, 7(1):105–110.

Suttles, J. and Ide, N. (2013). Distant Supervision for Emotion Classification with Discrete Binary Values. Springer Berlin Heidelberg.

Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., and Hashimoto, T. B. (2023). Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. (2023a). LLaMA: Open and Efficient Foundation Language Models. arXiv (Cornell University).

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Roziè-

re, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. (2023b). Llama: Open and efficient foundation language models.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P. S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X. E., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., and Scialom, T. (2023c). Llama 2: Open foundation and fine-tuned chat models.

UNESCO (2020). Education: from school closure to recovery. Online; accessed 19 August 2023.

Van Dyne, A. (unknow). NVIDIA H100 for AI Paperspace.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017a). Attention is all you need. CoRR, abs/1706.03762.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017b). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc.

Vera, J. (2022). El impacto de la salud mental en la economía global: 16.300 millones de euros de pérdidas entre 2011 y 2030.

von Werra, L., Belkada, Y., Tunstall, L., Beeching, E., Thrush, T., and Lambert, N. (2020). Trl: Transformer reinforcement learning.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. In Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Wang, Z., Hamza, W., and Florian, R. (2017). Bilateral multi-perspective matching for natural language sentences. In Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17, pages 4144–4150.

Warstadt, A., Singh, A., and Bowman, S. R. (2019). Neural network acceptability judgments.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., and Fedus, W. (2022). Emergent abilities of large language models. Transactions on Machine Learning Research. Survey Certification.

Welbl, J., Glaese, A., Uesato, J., Dathathri, S., Mellor, J. W., Hendricks, L. A., Anderson, K. M., Kohli, P., Coppin, B., and Huang, P.-S. (2021). Challenges in Detoxifying Language Models. arXiv (Cornell University).

Weng, L. (2021). Reducing Toxicity in Language Models.

Wenzek, G., Lachaux, M.-A., Conneau, A., Chaudhary, V., Guzmán, F., Joulin, A., and Grave, E. (2019). Ccnet: Extracting high quality monolingual datasets from web crawl data.

Werbos, P. J. (1974). Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences. PhD thesis, Harvard University.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 38–45, Online. Association for Computational Linguistics.

Xu, J., Zheng, Y., Wang, M., Zhao, J., Zhan, Q., Fu, M., et al. (2011). Predictors of symptoms of posttraumatic stress in chinese university students during the 2009 h1n1 influenza pandemic. Medical Science Monitor, 17(7):PH60.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., and Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, Advances in Neural Information Processing Systems, volume 32. Curran Associates, Inc.

Yasar, K. (2023). AI prompt.

Zhang, B. and Sennrich, R. (2019). Root mean square layer normalization. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors,

Advances in Neural Information Processing Systems, volume 32. Curran Associates, Inc.

Zhou, Hao, Huang, Minlie, Lin, Yi-Li, Zhu, Xiaoyan, Liu, and Bing (2017). Emotional Chatting Machine: Emotional Conversation Generation with Internal and External Memory. arXiv (Cornell University).

Zhu, C., Zhang, T., Li, Q., Chen, X., and Wang, K. (2022). Depression and Anxiety during the COVID-19 Pandemic: Epidemiology, mechanism, and treatment. Neuroscience Bulletin, 39(4):675–684.