

MANUAL DEL CURSO

# INTELIGENCIA ARTIFICIAL Y CHATGPT PARA PROFESIONALES DE LA SALUD




**Urgencias y  
Emergencias®**



**Urgencias y  
Emergencias®**

**Academia Urgencias y Emergencias®**

Copyright © 2025. Elena Plaza Moreno. Todos los derechos reservados. Obra registrada en el Registro de la Propiedad Intelectual



INTELIGENCIA ARTIFICIAL  
Y CHATGPT PARA  
PROFESIONALES DE LA  
SALUD

MANUAL DEL CURSO



ELENA PLAZA MORENO

Urgencias y  
Emergencias®

NOTA DE LA AUTORA: la IA, las aplicaciones en salud, los modelos de lenguaje y todo lo que se menciona en el manual y en el curso están en constante y vertiginoso cambio.

El manual se terminó de redactar en septiembre de 2024 y esta es la versión 2 porque ha sido actualizado a fecha enero de 2025. Aun así, si ves cosas incorrectas, es que ha habido alguna actualización que se solventará en próximas ediciones del manual y del curso.

Reservados todos los derechos. Queda rigurosamente prohibida, sin la autorización escrita de la titular del Copyright, bajo las sanciones establecidas de las leyes, la reproducción parcial o total de esta obra por cualquier medio o procedimiento incluidos la reprografía y el tratamiento informático.

Copyright © 2025 Inteligencia artificial y ChatGPT para profesionales de la salud. Manual del curso.

Copyright © 2025 Elena Plaza Moreno. Todos los derechos reservados. Obra registrada en el Registro de la Propiedad Intelectual

## ÍNDICE

<b>TEMA 1. ¿QUÉ ES LA INTELIGENCIA ARTIFICIAL?.....</b>	<b>15</b>
1.1. Definiciones. ¿Qué es la inteligencia artificial?.....	15
1.2 Tipos de inteligencia artificial.....	17
1.3 La IA “que nos sirve en salud” .....	18
1.3.1 Inteligencia artificial analítica.....	18
1.3.2 Diferencias y comparación entre IA generativa y analítica y en salud .....	20
1.4 ¿Nos va a quitar el trabajo la IA? .....	21
1.5 Fechas y nombres clave.....	22
1.6 Resumen .....	25
1.7 Bibliografía.....	25
<b>TEMA 2. APLICACIONES DE LA IA EN SALUD .....</b>	<b>28</b>
2.1 Introducción .....	28
2.2. Aplicaciones en pronóstico y diagnóstico .....	28
2.2.1. Aplicaciones de ejemplo de uso de IA en pronóstico .....	29
2.2.2. Beneficios de uso de IA en pronóstico .....	30
2.2.3. Aplicaciones de ejemplo de la IA en diagnóstico.....	30
2.2.4. Beneficios del uso de la IA en diagnóstico.....	30
2.3. Seguimiento de la salud en tiempo real .....	31
2.3.1. Aplicaciones.....	31
2.3.2 Beneficios.....	31
2.4. Monitorización de constantes vitales en tiempo real.....	31
2.4.1. Beneficios del uso de la IA en la monitorización de las constantes vitales:.....	32
2.5 Registros médicos electrónicos (EHR) .....	32
2.5.1. Aplicaciones de ejemplo.....	33
2.6 Diagnóstico por imagen.....	33
2.6.1. Beneficios.....	33
2.7. Otros ejemplos de aplicaciones .....	34
2.7.1. Transcripción de consultas e integración en historia clínica .....	34
2.7.2. Análisis de voz .....	34
2.7.3. Investigación.....	34
2.7.4. Gestión de la salud pública.....	35
2.7.5. Rehabilitación y fisioterapia .....	36

2.7.6. Asistentes virtuales de salud, chatbots y chats conversacionales .....	36
2.7.7. Gestión de recursos hospitalarios .....	37
2.7.8. Cirugía asistida por IA.....	38
2.7.9. Industria farmacéutica .....	38
2.9. Conclusiones .....	39
2.10. Bibliografía .....	40
<b>TEMA 3. MACHINE LEARNING, DEEP LEARNING Y REDES NEURONALES .....</b>	<b>43</b>
3.1 Machine learning.....	43
3.1.1 ¿Qué es el Machine Learning?.....	43
3.1.2 Funcionamiento del machine learning.....	44
3.1.3 Tipos de Machine Learning .....	45
3.2 DEEP LEARNING.....	47
3.2.1 ¿Qué es el deep learning?.....	47
3.2.2 Redes neuronales .....	47
3.3 BIBLIOGRAFÍA.....	57
<b>TEMA 4. ¿QUÉ ES LA INTELIGENCIA ARTIFICIAL GENERATIVA? .....</b>	<b>60</b>
4.1 Introducción .....	60
4.2 ¿Cómo funciona la IA generativa?.....	60
4.3 Aplicaciones de la IA generativa .....	61
4.4 Desafíos técnicos de la IA generativa .....	62
4.5 Desafíos éticos de la IA generativa .....	67
4.6 El futuro de la IA generativa .....	68
4.7 Bibliografía.....	69
<b>TEMA 5. ¿QUÉ ES Y CÓMO FUNCIONA CHATGPT Y OTROS MODELOS DE LENGUAJE GRANDE (LLM)? 71</b>	<b>71</b>
5.1 Introducción .....	71
5.2 ¿Qué es ChatGPT? ¿Cómo se ha entrenado?.....	72
5.2.1 Evolución de ChatGPT .....	73
5.3 ¿Cómo funciona ChatGPT y otros modelos? .....	76
5.4. Aplicaciones generales de ChatGPT y otros modelos .....	83
5.5. Desafíos de los modelos de lenguaje grandes.....	84
5.6. Modelos de lenguaje grande disponibles.....	88
5.6.1 ChatGPT de Open AI .....	88
5.6.2 Copilot de Microsoft .....	90
5.6.3 Gemini (Google).....	94

5.6.4 Perplexity .....	97
5.6.5 Claude (Anthropic) .....	98
5.6.7 Pi (Inflection AI) .....	101
5.6.8 MedGemini.....	101
5.7 Bibliografía.....	102
<b>TEMA 6. ¿CÓMO HABLAR CON CHATGPT U OTROS MODELOS DE LENGUAJE? PROMPT ENGINEERING.</b>	<b>105</b>
6.1 Introducción al prompt engineering .....	105
6.2 Pasos y consejos para utilizar Prompt Engineering.....	105
6.3 Ajuste de parámetros. Técnicas avanzadas de prompting .....	106
6.3.1 Temperatura .....	106
6.3.2 Top P (nucleus sampling) .....	107
6.3.3 Top-k Sampling .....	109
6.3.4 Penalización por presencia (presence penalty) .....	110
6.3.5 Penalización de frecuencia (frequency penalty).....	112
6.3.6 Penalización de repetición (repetition penalty).....	113
6.3.7 Máximo número de tokens .....	115
6.3.8 "Crea tus propios parámetros" .....	116
6.4 ¿Cómo hablar con chatGPT u otros LLM? .....	117
6.5 Prompting en salud.....	120
6.5.1 Recomendaciones generales y checklist .....	120
6.5.2 Evaluación crítica de las respuestas proporcionadas por la IA de manera efectiva.....	122
6.6 ChatGPT y otros modelos de lenguaje en "la gestión del conocimiento". .....	124
<b>TEMA 7. ¿CÓMO DAR ROLES DE SALUD (PACIENTE, ESTUDIANTE, INSTRUCTOR) A CHATGPT? .....</b>	<b>130</b>
7.1 Introducción .....	130
7.2 ¿Tiene empatía ChatGPT? .....	130
7.2.1 La Ilusión de la empatía en la interacción Humano-Computadora.....	130
7.2.2 Comparación de respuestas entre médicos y chatbots de IA .....	131
7.2.3 Chatbots personalizados para mejorar el acceso a la salud mental .....	132
7.2.4 Límites fundamentales de la empatía de la ia en la atención sanitaria .....	132
7.2.5. La mejora de la empatía en ChatGPT .....	133
7.3 Beneficios de los chatbots en salud.....	134
7.4 Riesgos, miedos y cuestiones éticas de los asistentes virtuales impulsados por LLMs .....	135
7.5. APLICACIÓN DE ROLES DE LOS MODELOS DE LENGUAJE GRANDE EN SALUD.....	136

7.5.1 Rol de paciente.....	136
7.5.2 Rol de tutor/docente.....	137
7.5.3 Profesional sanitario.....	138
7.5.4 Otros roles.....	138
7.6 BIBLIOGRAFÍA.....	139
<b>Temas 9, 10 y 11. Aspectos éticos, limitaciones y riesgos de la IA en salud.....</b>	<b>141</b>
1 Confidencialidad, privacidad y protección de datos.....	141
1.1. Reglamento europeo de Inteligencia Artificial.....	141
1.2 Protección de datos en los modelos de lenguaje grande.....	142
2 Seguridad: predicción, explicación, interpretación, intervención y sesgos.....	144
2.1 Algoritmos transparentes y algoritmos caja negra.....	144
2.2 Sesgos.....	145
2.3 Alucinaciones.....	151
3. Responsabilidad del uso de la IA en salud.....	155
4. Limitaciones de la ia en el ámbito de la salud.....	156
5. Principios éticos del uso de la IA en salud.....	158
5.1. Principios éticos DE LA OMS para el uso de la IA en salud.....	158
5.2. Consideraciones éticas para el uso de LLM por niños. OMS.....	159
5.3 Consideraciones éticas y su impacto en las personas con discapacidad. OMS.....	160
5.4 Recomendaciones sobre ética de la IA por la UNESCO.....	161
5.5 Principios de la bioética en la inteligencia artificial.....	163
6. ANEXO I. Implementación de una solución de IA.....	167
7. BIBLIOGRAFÍA.....	169

## LISTADO DE IMÁGENES

Imagen 1. Áreas de la inteligencia artificial. Autora: Elena Plaza Moreno.....	16
Imagen 2. Componentes, tipos y subcampos de la IA. Autora: Elena Plaza Moreno. Modificado de Regona et al (2022). DOI: <a href="https://doi.org/10.3390/joitmc8010045">https://doi.org/10.3390/joitmc8010045</a> .....	18
Imagen 3. Hitos clave de la IA. Autora: Elena Plaza Moreno. Elaboración propia.....	24
Imagen 4. Aplicaciones de la IA en salud. Elaborado por Elena Plaza Moreno.....	28
Imagen 5. Captura de pantalla del proyecto con IA para el diagnóstico de enfermedades raras de la Comunidad de Madrid. Fuente: <a href="https://www.comunidad.madrid/noticias/2023/09/15/comunidad-madrid-prueba-proyecto-pionero-reducir-tiempo-diagnostico-enfermedades-raras">https://www.comunidad.madrid/noticias/2023/09/15/comunidad-madrid-prueba-proyecto-pionero-reducir-tiempo-diagnostico-enfermedades-raras</a> .....	29

Imagen 6. Relación entre inteligencia artificial, machine learning y deep learning. Autora: Elena Plaza Moreno.....	43
Imagen 7. Pasos en el proceso de entrenamiento con machine learning (aprendizaje automático). Autora: Elena Plaza Moreno. Elaboración con Napkin IA.....	44
Imagen 8. Entrenamiento de modelo de machine learning con fotografías de chihuahuas y muffins y komondos y fregonas. Fuente: Shenkman C. (2021). <a href="https://www.semanticscholar.org/paper/640d18b94266146a6c479eccda92f725404af388">https://www.semanticscholar.org/paper/640d18b94266146a6c479eccda92f725404af388</a> .....	45
Imagen 9. Principales modelos de entrenamiento de machine learning en salud. Autora: Elena Plaza Moreno con Napkin IA.....	47
Imagen 10. Diagrama que muestra como el funcionamiento y la estructura de las redes neuronales biológicas (a la izquierda) inspiraron el funcionamiento netamente matemático del primer modelo de neurona artificial, el perceptrón (a la derecha). Fuente: Arias V et al. (2019). <a href="https://www.redalyc.org/journal/1702/170262877013/html/">https://www.redalyc.org/journal/1702/170262877013/html/</a> .....	48
Imagen 11. Esquema de capas de una red neural. Cada circunferencia representa una neurona. En rojo las neuronas que forman la capa de entrada. En amarillo, cuatro capas ocultas. En azul, las neuronas que forman la capa de salida. Autora: Elena Plaza Moreno.....	49
Imagen 12. Red neuronal analizando una imagen. Fuente: Imagen del artículo Waldrop, M. M. (2019). <a href="https://doi.org/10.1073/pnas.1821594116">https://doi.org/10.1073/pnas.1821594116</a> . Autora de la imagen: Lucy Reading-Ikkanda.....	51
Imagen 13. Una sección ampliada de una fotografía muestra como realmente una computadora percibe una imagen, en este caso la oreja de un perro descrita por una matriz de números. Fuente: Arias V et al. (2019). <a href="https://www.redalyc.org/journal/1702/170262877013/html/">https://www.redalyc.org/journal/1702/170262877013/html/</a> .....	52
Imagen 14. El codificador (encoder) y el decodificador (decoder) trabajan juntos para producir mejores resultados (output). Autora: Elena Plaza Moreno.....	54
Imagen 15. Reconocimiento facial para verificación biométrica. Fuente: Banco de imágenes de Canva. Autor: Andriano_cz.....	55
Imagen 16. Fotografía del dataset de <a href="https://www.cityscapes-dataset.com/examples/">https://www.cityscapes-dataset.com/examples/</a> . Se puede ver cómo con la IA se pueden diferenciar personas, aceras, árboles, coches y otros objetos.....	56
Imagen 17. La IA generativa es un tipo de IA que se basa en el deep learning (redes neuronales) para crear contenido nuevo. Autora: Elena Plaza Moreno.....	60
Imagen 18. Desafíos de la IA generativa. Elaboración: Elena Plaza Moreno con Napkin IA.....	62
Imagen 19. Imagen generada por IA, con DreamStudio a finales de 2022 en la que se evidencia problemas en la creación de brazos y manos.....	63
Imagen 20. Imagen generada por IA con DreamStudio en 2022.....	64
Imagen 21. Imagen generada por IA con DreamStudio en 2024.....	64
Imagen 22. Imágenes generadas con Midjourney en 2024.....	65

Imagen 23. Imagen generada por Midjourney en julio de 2024. Prompt: "Meatball shop logo banksy style" .....	65
Imagen 24. Imagen generada con Midjourney en julio de. Prompt: "3 Tennis fans enjoying popsicles on a sunny day, smiling at the camera, photo taken at a lower angle --v 6.0" .....	65
Imagen 25. Imagen generada con DALL-E en junio de 2024	
Imagen 26. Imagen generada con Copilot en junio de 2024.....	66
Imagen 27. Cartel informativo de reanimación cardiopulmonar generado con DALL-E – ChatGPT 4o (julio 2024). Prompt: "Genera un cartel informativo en el que se lea la palabra reanimación cardiopulmonar y los cuatro pasos principales que hay que seguir: reconocer parada llamar al 112 iniciar compresiones torácicas usar el desfibrilador" .....	66
Imagen 28. Imágenes generadas por @julian_ai_art con Midjourney y subidas a Instagram.....	67
Imagen 29. Captura de pantalla de noticia sobre los problemas de OpenAI (empresa creadora de ChatGPT) con los derechos de autor. Fuente: <a href="https://cadenaser.com/nacional/2024/01/09/openai-admite-que-no-ha-respetado-los-derechos-de-autor-para-entrenar-a-chatgpt-y-considera-que-no-es-un-delito-cadena-ser/">https://cadenaser.com/nacional/2024/01/09/openai-admite-que-no-ha-respetado-los-derechos-de-autor-para-entrenar-a-chatgpt-y-considera-que-no-es-un-delito-cadena-ser/</a> .....	67
Imagen 30. Esquema de las áreas de la IA "de más a menos" con listado de los LLM más populares. Autora: Elena Plaza Moreno .....	72
Imagen 31. Captura de pantalla de los modelos disponibles de ChatGPT en versión de pago en enero de 2025.....	74
Imagen 32. Captura de pantalla de conversación con ChatGPT 4o en la que se le pregunta cuál es el dataset de ChatGPT.....	77
Imagen 33. Ejemplo de tokenización, fragmentación de una frase en tokens. Captura de pantalla de <a href="https://platform.openai.com/tokenizer#">https://platform.openai.com/tokenizer#</a> .....	78
Imagen 34. Tokenización de la frase "El cielo es azul". Captura de pantalla de <a href="https://platform.openai.com/tokenizer#">https://platform.openai.com/tokenizer#</a> .....	78
Imagen 35. Simulación de ruleta con probabilidades de palabras. ....	80
Imagen 36. Variación de los resultados de ChatGPT dependiendo de factores "insignificantes" o pocos importantes en el prompt. Fuente: Mollick, E. (2024, 20 octubre). Thinking Like an AI. One Useful Thing. <a href="https://www.oneusefulthing.org/p/thinking-like-an-ai">https://www.oneusefulthing.org/p/thinking-like-an-ai</a> .....	81
Imagen 37. Ejemplo de ChatGPT 4o (junio 2024) en el que detecta errores con la medicación. Captura de pantalla.....	82
Imagen 38. Ejemplo de ChatGPT 4o (junio 24) en el que no contesta a peticiones ilegales o comprometedoras. Captura de pantalla.....	82
Imagen 39. Captura de pantalla de publicación de LinkedIn de Samuel Sesmilo.....	84
Imagen 40. Fotografía del resultado de Google Gemini que ilustra el texto de la Imagen 39. Foto de Samuel Sesmillo.....	85

---

Imagen 41. Captura de pantalla de tuit con ejemplo de noticia falsa a partir de imagen generada por IA durante la DANA en Valencia en 2024.....	86
Imagen 42. Captura de pantalla de la portada principal de RCP GPT.....	89
Imagen 43. Captura de pantalla de conversación con ChatGPT 4o en la que analiza una imagen. Julio 2024.....	89
Imagen 44. Captura de pantalla de algunos de los GPT sobre investigación y análisis más populares. Puedes acceder a ellos en la parte izquierda de la pantalla de ChatGPT, donde pone "Explorar GPT". .....	90
Imagen 45. Captura de pantalla de Copilot entrando por <a href="https://copilot.microsoft.com/">https://copilot.microsoft.com/</a> sin cuenta corporativa. Vemos como ante una consulta, nos cita fuentes de donde ha obtenido la información. .....	91
Imagen 46. Captura de pantalla de Copilot, entrando a través de Bing. Te redirige a <a href="https://copilot.cloud.microsoft/">https://copilot.cloud.microsoft/</a> si introduces los datos de la cuenta de Microsoft de universidades u centros de trabajo donde te otorgan un email.....	92
Imagen 47. Imagen generada con Microsoft Designer, que es correcta. 8 de julio de 2024.....	92
Imagen 48. Imagen generada con Microsoft Designer con un evidente error anatómico. 8 de julio de 2024.....	93
Imagen 49. Captura de pantalla de Word con Copilot integrado (sin pagar suscripción mensual en Copilot) con el paquete personal de Microsoft. Enero 2025.....	93
Imagen 50. Captura de pantalla de Power Point con Copilot integrado (sin pagar suscripción mensual en Copilot) con el paquete personal de Microsoft. Enero 2025.....	94
Imagen 51. Captura de pantalla de la interfaz de Google Gemini y los modelos disponibles. Enero 2025.....	94
Imagen 52. Captura de pantalla de la política de datos de Google Gemini. Disponible en: <a href="https://support.google.com/gemini/answer/13594961?hl=es&amp;ref_topic=13278591&amp;sjid=6223584219093404158-EU">https://support.google.com/gemini/answer/13594961?hl=es&amp;ref_topic=13278591&amp;sjid=6223584219093404158-EU</a> .....	95
Imagen 53. Función de verificar respuesta con Gemini. Enero 2025.....	96
Imagen 54. Captura de pantalla de la verificación de respuestas de Google Gemini. Enero 2025.....	96
Imagen 55. Captura de pantalla de Perplexity.....	97
Imagen 56. Opciones de búsqueda en el apartado "Enfoque" de Perplexity. Captura de pantalla. Enero 2025.....	98
Imagen 57. Interfaz de Claude. Enero 2025.....	99
Imagen 58. Captura de pantalla de Claude, apartado "Elige estilo". Enero 2025.....	100
Imagen 59. Captura de pantalla de un artifact de Claude. Enero 2025.....	100
Imagen 60. Captura de pantalla con ejemplo de variación de temperatura. ChatGPT 4o.....	107

---

Imagen 61. Captura de pantalla con ejemplo de variación de Top P o nucleus sampling. ChatGPT 4o .....	109
Imagen 62. Captura de pantalla con ejemplo de variación de Top K. ChatGPT 4o .....	110
Imagen 63. Captura de pantalla con ejemplo de variación de penalización por presencia. ChatGPT 4o .....	111
Imagen 64. Captura de pantalla con ejemplo de variación de penalización por frecuencia. ChatGPT 4o.....	113
Imagen 65. Captura de pantalla con ejemplo de variación de penalización por repetición. ChatGPT 4o.....	114
Imagen 66. Captura de pantalla con ejemplo de variación de longitud de tokens. ChatGPT 4o.....	116
Imagen 67. Fórmula RACS para hablar con modelos de lenguaje grande.....	117
Imagen 68. Tonos de ChatGPT. Captura de pantalla ChatGPT 4o.....	118
Imagen 69. Estructura RACS y descripción de cada apartado. Autora: Elena Plaza Moreno.....	119
Imagen 70. Ejemplo de prompt con estructura RACS. Autora: Elena Plaza Moreno.....	119
Imagen 71. Ejemplo de preguntas secuenciales. Captura de pantalla de ChatGPT 4o. ....	121
Imagen 72. Ejemplo de lista de verificación o checklist para crear un prompt de IA en salud. Autora: Elena Plaza Moreno. ....	122
Imagen 73. Elementos de la empatía cognitiva propuestos para usar en los chatbots basados en IA. Autora: Elena Plaza Moreno con Napkin IA. Fuente: <a href="https://journals.sagepub.com/doi/10.1177/21695067231193682">https://journals.sagepub.com/doi/10.1177/21695067231193682</a> .....	133
Imagen 74. Captura de pantalla del feedback de en un caso de simulación de rol de paciente. ChatGPT 4o.....	137
Imagen 75. Niveles de riesgo para los sistemas de IA de la Ley IA de Europa. Autora: Elena Plaza Moreno.....	142
Imagen 76. Un hombre de raza negra y una mujer de raza blanca con puntos de reconocimiento facial. Esta imagen ejemplifica los desafíos y sesgos raciales y de género en la tecnología de reconocimiento facial, destacando la necesidad de abordar las disparidades en la precisión y el tratamiento justo de diversas poblaciones. Generada con DALL-E – ChatGPT 4o. Julio 2024.....	146
Imagen 77. Captura de pantalla del tuit del usuario @jmatuk con imágenes erróneas creadas por Gemini con soldados alemanes nazis de la II Guerra Mundial con piel negra o rasgos asiáticos. <a href="https://x.com/jmatuk/status/1762872092670718085">https://x.com/jmatuk/status/1762872092670718085</a> .....	148
Imagen 78. Captura de pantalla del tuit del usuario @EndWokeness con imágenes erróneas creadas por Gemini con mujeres vikingo negras, Papas mujer y Fundadores de América indios. <a href="https://x.com/EndWokeness/status/1760280543940263994">https://x.com/EndWokeness/status/1760280543940263994</a> .....	149

Imagen 79. Captura de pantalla de Christopher F. Rufo con contestación racista de Gemini. <a href="https://x.com/realchrisrufo/status/1762529174668120066">https://x.com/realchrisrufo/status/1762529174668120066</a> .....	149
Imagen 80. Imagen creada con DALL-E, el generador de imágenes integrado en ChatGPT 4o (versión de pago, julio 2024).....	153
Imagen 81. Algoritmo de Aleksandr Tiulkanov sobre la seguridad de uso de ChatGPT. Licencia CC. ....	155
Imagen 82. Captura de pantalla de Google Gemini. Julio 2024 .....	158
Imagen 83. Consideraciones éticas para el uso de LLM por niños. OMS. Autora: Elena Plaza Moreno con Napkin IA .....	160
Imagen 84. Marco ético de la IA de la UNESCO. Autora: Elena Plaza Moreno con Napkin IA. ....	162
Imagen 85. Los principios de la bioética en la IA aplicada a salud. Elaboración propia. Autoría: Elena Plaza Moreno .....	166

## LISTADO DE TABLAS

Tabla 1. Características de la IA analítica y generativa en salud. Autora; Elena Plaza Moreno. Elaboración propia con uso de ChatGPT 4.0 (2024).....	21
Tabla 2. Principales tipos de redes neuronales con su principal utilidad. Elaboración propia. Autora: Elena Plaza Moreno .....	53
Tabla 3. Resumen de parámetros de escritura. Autora: Elena Plaza Moreno.....	115
Tabla 4. Evaluación crítica de las respuestas clínicas. Autora: Elena Plaza Moreno. ....	123
Tabla 5. Cuestiones éticas de los asistentes virtuales (chatbots) de salud con IA (modelos de lenguaje grande - LLM). Autora: Elena Plaza Moreno. Fuente: datos de: Piñeiro-Martín, A. <a href="https://doi.org/10.3390/electronics121431707.5">https://doi.org/10.3390/electronics121431707.5</a> . Ejemplos de prompt para roles de salud .....	136
Tabla 6. Ejemplos de continuaciones de texto generadas por el modelo GPT-2 de OpenAI con sesgos raciales racistas. Traducción de <a href="https://aclanthology.org/D19-1339.pdf">https://aclanthology.org/D19-1339.pdf</a> .....	147
Tabla 7. Sesgos de la IA en el análisis de datos. Autora: Elena Plaza Moreno. Elaboración propia. 151	
Tabla 8. Marco de trabajo ético para la implementación de una solución de IA en salud. Copia textual de Rosado, J. C. S., & Parra, M. D. (2022).....	169

## LISTADO DE GRÁFICAS

Gráfica 1. Porcentaje de trabajo nuevo creado desde 1940 por sector. Autora: Elena Plaza Moreno. Elaboración propia con datos de Autor, D. (2022). <a href="https://www.nber.org/system/files/working_papers/w30074/w30074.pdf">https://www.nber.org/system/files/working_papers/w30074/w30074.pdf</a> .....	22
--	----

---

Gráfica 2. Comparación de tasas de mortalidad de pacientes de alto riesgo en el grupo intervención y el grupo control. Elaborado por Elena Plaza Moreno.....	32
Gráfica 3. Tiempo necesario para que diversas plataformas llegaran a alcanzar 1 millón de usuarios. Fuente. Statista. Elaboración. Elena Plaza Moreno. ....	73
Gráfica 4. Desempeño de GPT 4o vs o1 en diferentes pruebas (benchmarks). Autora: Elena Plaza Moreno. Fuente: <a href="https://openai.com/index/learning-to-reason-with-llms/">https://openai.com/index/learning-to-reason-with-llms/</a> .....	75
Gráfica 5. Desempeño de GPT 4o vs o1 en preguntas de nivel doctorado en química, física y biología. Autora: Elena Plaza Moreno. Fuente: <a href="https://openai.com/index/learning-to-reason-with-llms/">https://openai.com/index/learning-to-reason-with-llms/</a> .....	75
Gráfica 6. Precisión de varios modelos de lenguaje en una de las pruebas a las que se les somete en el ámbito de la salud llamada MedQA (USMLE). Autora: Elena Plaza Moreno con datos de: <a href="https://arxiv.org/pdf/2404.18416">https://arxiv.org/pdf/2404.18416</a> .....	102
Gráfica 7. Puntuaciones de los evaluadores a las respuestas de las preguntas de los pacientes en términos de calidad (izquierda) y empatía (derecha). En verde la puntuación de ChatGPT y en naranja la puntuación de las respuestas de los médicos, prefiriendo las respuestas de ChatGPT el 79% de las veces. Autora: Elena Plaza Moreno con el dataset público de Ayers, J. W. <a href="https://doi.org/10.1001/jamainternmed.2023.1838">https://doi.org/10.1001/jamainternmed.2023.1838</a> .....	132



# MÓDULO 1. INTRODUCCIÓN A LA INTELIGENCIA ARTIFICIAL

## TEMA 1. ¿QUÉ ES LA INTELIGENCIA ARTIFICIAL?

*«Somos lo que hacemos día a  
día, de modo que la excelencia  
no es un acto si no hábito».*

*ARISTÓTELES*



## TEMA 1. ¿QUÉ ES LA INTELIGENCIA ARTIFICIAL?

### 1.1. DEFINICIONES. ¿QUÉ ES LA INTELIGENCIA ARTIFICIAL?

La inteligencia artificial (IA) no es solo ChatGPT. ChatGPT es un tipo de IA, pero hay mucho más detrás. Lo que ha pasado con ChatGPT es que “se ha abierto al público” un tipo de IA que antes no teníamos al alcance de la mano.

La IA ha sido descrita mediante ciertos enfoques en relación con la inteligencia humana, o la inteligencia en general. Muchas definiciones se refieren a máquinas que se comportan como humanos o son capaces de realizar acciones que requieren inteligencia.

**Según John McCarthy la IA es la ciencia y la ingeniería de fabricar máquinas inteligentes, especialmente programas informáticos inteligentes. Está relacionado con la tarea similar de usar computadoras para comprender la inteligencia humana, pero la IA no tiene que limitarse a métodos que sean biológicamente observables.**

Según el Ministerio de Sanidad de España, la inteligencia artificial es un campo de la informática que se enfoca en crear sistemas que puedan realizar **tareas que normalmente requieren inteligencia humana**, como el aprendizaje, el razonamiento y la percepción.

Estos sistemas pueden percibir su entorno, razonar sobre el conocimiento, procesar la información derivada de los datos y tomar decisiones para lograr un objetivo dado.

Según la Comisión Europea:

**“La inteligencia artificial se refiere a sistemas que demuestran comportamiento inteligente al analizar su entorno y tomar decisiones, con cierto grado de autonomía, para alcanzar objetivos específicos. Los sistemas basados en IA pueden ser puramente software, actuando en el mundo virtual (por ejemplo, asistentes de voz, software de análisis de imágenes, motores de búsqueda, sistemas de reconocimiento de voz y de rostro) o la IA puede estar integrada en dispositivos de hardware (por ejemplo, robots avanzados, coches autónomos, drones o aplicaciones del Internet de las Cosas)”.**

Como se puede ver en la Imagen 1, la IA es un campo vasto y complejo, con muchas ramas y especializaciones.

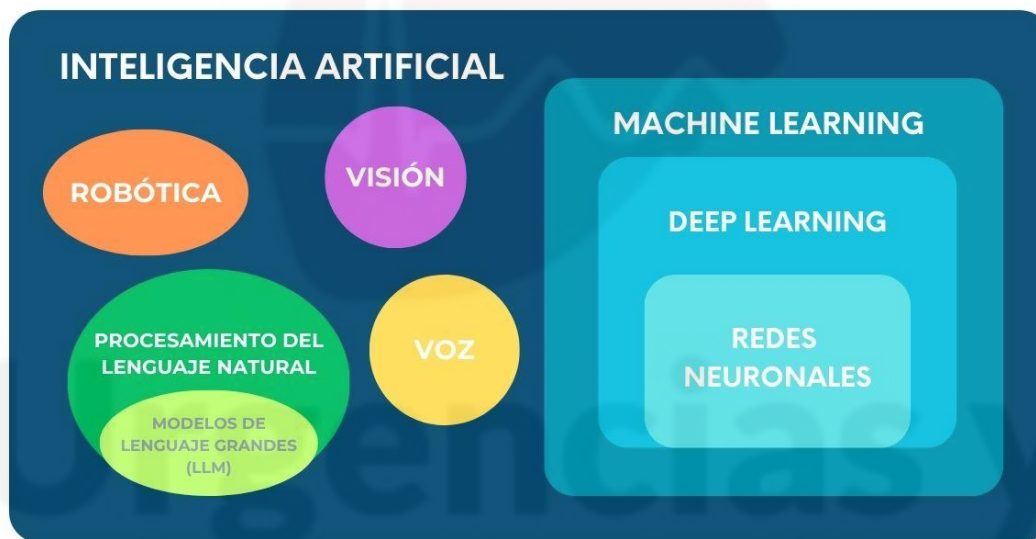


Imagen 1. Áreas de la inteligencia artificial. Autora: Elena Plaza Moreno.

Respecto a las áreas principales, son las siguientes:

- **Robótica:** es la disciplina que se ocupa del diseño, operación, manufacturación, estudio y aplicación de autómatas o robots. Combina áreas como la ingeniería mecánica, eléctrica, electrónica, biomédica y las ciencias de la computación para crear herramientas que puedan realizar tareas de manera eficiente, rápida y en ambientes inaccesibles para los humanos. Ya existen robots en algunos hospitales de España que recogen las bandejas de comida de los pacientes y ayudan a otro tipo de tareas. Otro ejemplo sería el robot Da Vinci, el robot manejado por cirujanos que aumenta la precisión y mejora la recuperación de los pacientes.
- **Procesamiento del Lenguaje Natural (PLN):** es una rama de la inteligencia artificial que se ocupa de la interacción entre las computadoras y el lenguaje humano. Implica la capacidad de una máquina para entender, interpretar y generar lenguaje humano de manera que sea útil. Esto incluye tareas como la traducción automática, el análisis de sentimientos, la generación de texto y la comprensión del habla. Aquí se incluiría ChatGPT y otros modelos llamados modelos de lenguaje grande.
- **Voz:** en el contexto de la IA, la generación de voz se refiere a la tecnología que utiliza algoritmos de aprendizaje automático y procesamiento de lenguaje natural para producir habla similar a la humana. Esto incluye convertir texto en voz hablada, imitando el tono, ritmo y entonación de un hablante humano. Los asistentes de voz como Siri, Alexa, Google Assistant funcionan gracias a esta rama de la IA. En este ámbito estaría recogido también todo lo que tenga que ver con sonido. Por ejemplo, se pueden analizar formas de hablar en triaje para definir o intentar medir el dolor del paciente.
- **Visión artificial:** la visión artificial es un campo de la IA que permite a las computadoras y sistemas obtener información significativa de imágenes digitales, videos y otras entradas visuales, y tomar acciones o hacer recomendaciones basadas en esa información. Funciona de manera similar a la visión humana, utilizando tecnologías como el aprendizaje profundo y las redes neuronales convolucionales para reconocer y analizar imágenes. Un ejemplo pueden ser los coches autónomos o el reconocimiento facial del móvil para desbloquearlo.

En relación con los algoritmos y aprendizaje, como también se observa en la **Imagen 1**, existen otras áreas de especialización como son el machine learning (aprendizaje automático), deep learning (aprendizaje profundo) que funciona con redes neuronales. Esto se explica en profundidad en posteriores temas.

## 1.2 TIPOS DE INTELIGENCIA ARTIFICIAL

Además de lo explicado hasta aquí, se puede hablar de diferentes tipos de inteligencia artificial según varias clasificaciones comunes:

### a) Según la capacidad

La IA se puede clasificar según su capacidad en tres tipos principales:

- **IA débil:** también conocida como IA estrecha, esta inteligencia artificial está diseñada y entrenada para realizar una tarea específica. Los asistentes de voz como Siri o Alexa son ejemplos de IA débil.
- **IA general:** este tipo de inteligencia artificial tiene capacidades cognitivas a nivel humano, lo que significa que puede realizar cualquier tarea intelectual que un ser humano pueda hacer. Aunque es un objetivo a largo plazo de la investigación en IA, todavía no se ha realizado.
- **IA superinteligente:** es una IA que no solo emula la capacidad humana, sino que la supera en todos los aspectos: creatividad, razonamiento, y emociones. Este tipo de IA es todavía hipotética y objeto de numerosas discusiones éticas y filosóficas.

### b) Según la funcionalidad

En esta clasificación, las inteligencias artificiales se dividen según las tareas que realizan:

- **IA reactiva:** son sistemas que no tienen memoria ni capacidad para usar experiencias pasadas para informar decisiones futuras. Un ejemplo es el programa de ajedrez Deep Blue de IBM.
- **IA con memoria limitada:** estos sistemas pueden usar experiencias pasadas para tomar decisiones. La mayoría de los coches autónomos operan en este nivel, utilizando observaciones pasadas para tomar decisiones en tiempo real.
- **IA que teoriza la mente:** este nivel de IA puede comprender las emociones, personas y otros seres conscientes, ajustando su comportamiento en respuesta a estas percepciones. Este tipo aún no se ha implementado completamente en ninguna aplicación.
- **IA autoconsciente:** esta IA tendría conciencia de sí misma y de su existencia en el mundo. Este es un concepto teórico y no existe en la práctica actual.

### c) Según el enfoque o técnica

Finalmente, la IA también se puede clasificar según las técnicas y enfoques que utiliza para aprender y funcionar:

- **IA simbólica:** utiliza la lógica y reglas para simular el razonamiento humano.
- **IA de aprendizaje automático (Machine Learning):** aprende de los datos y mejora su rendimiento a través de experiencias sin ser explícitamente programada para cada tarea específica.
- **IA de redes neuronales o aprendizaje profundo (Deep Learning):** se inspira en el cerebro humano y utiliza una serie de algoritmos que intentan reconocer relaciones subyacentes en un conjunto de datos a través de un proceso que imita la forma en que el cerebro humano opera.

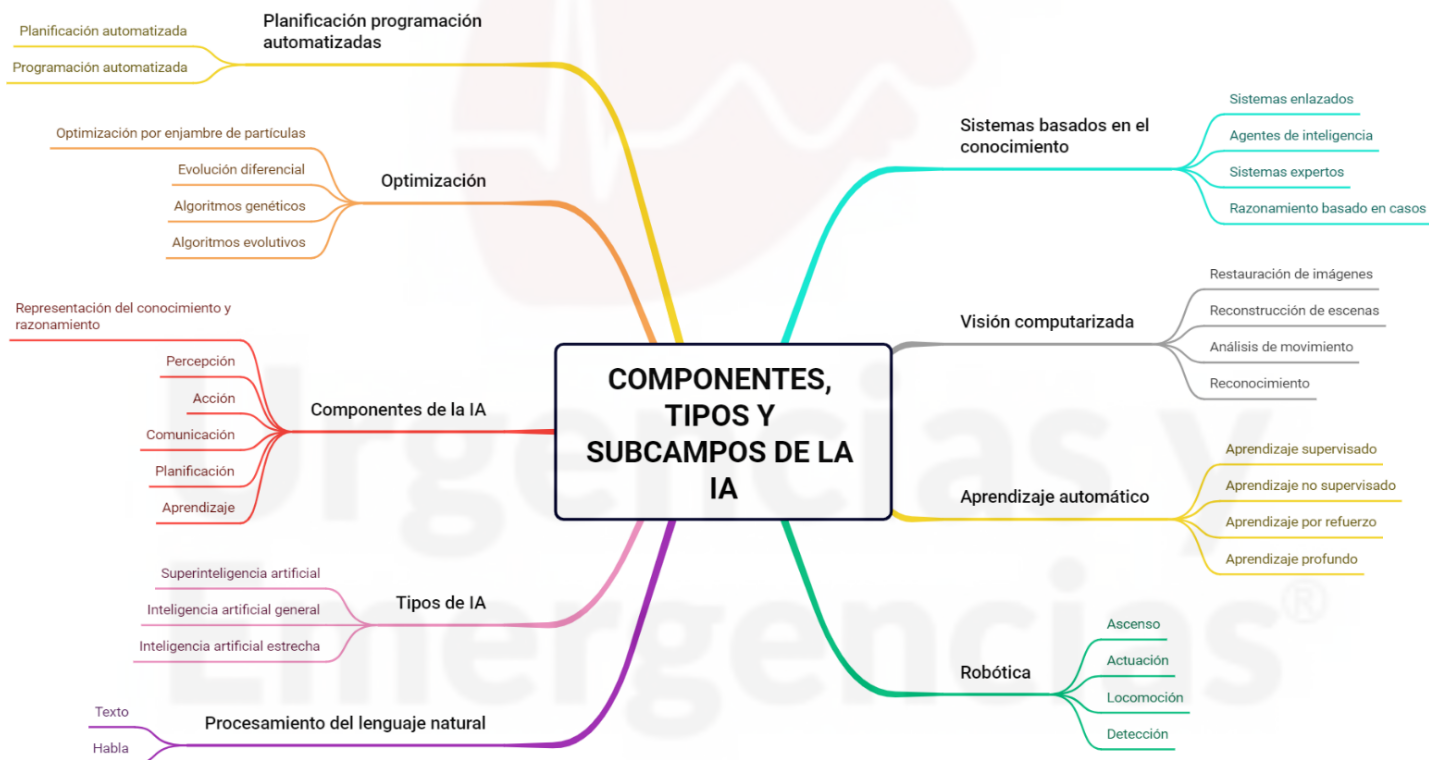


Imagen 2. Componentes, tipos y subcampos de la IA. Autora: Elena Plaza Moreno. Modificado de Regona et al (2022). DOI: <https://doi.org/10.3390/joitmc8010045>

### 1.3 LA IA “QUE NOS SIRVE EN SALUD”

En este punto, es importante, desde una perspectiva educativa, clasificar la IA en dos grandes grupos para facilitar la comprensión.

Como se ha explicado con anterioridad, hay más tipos según su entrenamiento, función, etc. que nos puede ayudar en salud. Por ejemplo, la robótica puede ayudar a movilizar pacientes, trasladar medicación, bandejas de comida, realizar ecografías a distancia, etc. La visión computarizada contribuye a restaurar imágenes, entre otros usos.

Pero recalco de nuevo que es importante dividir entre dos tipos de IA para aplicar en salud que debemos diferenciar de forma destacada:

- La IA **analítica**, basada en Big Data, análisis de datos e investigación.
- La IA **generativa**, que también trabaja con datos, pero su función es utilizarlos para generar nuevos contenidos.

#### 1.3.1 INTELIGENCIA ARTIFICIAL ANALÍTICA

La inteligencia artificial analítica es una rama avanzada de la IA que se especializa en el análisis de grandes volúmenes de datos para descubrir patrones, tendencias y relaciones significativas.

En el contexto de la salud, la IA analítica combina técnicas de aprendizaje automático, aprendizaje profundo, estadística avanzada y análisis de datos para generar conocimientos valiosos que pueden mejorar la atención al paciente, optimizar los procesos clínicos y avanzar en la investigación médica.

En el panorama actual de la atención sanitaria, donde la cantidad de datos médicos crece exponencialmente, la IA analítica se ha convertido en una herramienta indispensable. Su capacidad para procesar y analizar datos complejos a una escala y velocidad inaccesibles para los humanos la convierte en un activo invaluable para mejorar la calidad de la atención, reducir costos y personalizar los tratamientos.

Este tipo de IA se utiliza en investigación y se lleva aplicando por profesionales de la salud durante años. En vez de analizar los datos de sus investigaciones con herramientas estadísticas existentes desde siempre, se aplican algoritmos de IA para procesar más datos, predecir resultados, y más cosas que se explican a continuación.

---

### 1.3.1.1 CARACTERÍSTICAS PRINCIPALES DE LA IA ANALÍTICA EN SALUD

#### **d) *Procesamiento de grandes volúmenes de datos:***

La IA analítica puede manejar y analizar eficientemente enormes cantidades de datos médicos, incluyendo historiales clínicos electrónicos, imágenes médicas, datos genómicos e información de dispositivos médicos conectados. Esta capacidad permite una visión holística de la salud del paciente y de las tendencias poblacionales.

#### **e) *Identificación de patrones y relaciones:***

Utilizando algoritmos avanzados, la IA analítica puede detectar patrones sutiles y correlaciones en los datos médicos que podrían pasar desapercibidos para el ojo humano. Esto es clave para el diagnóstico precoz, la predicción de riesgos de enfermedades y la identificación de factores de riesgo poco evidentes.

#### **f) *Generación de conocimientos accionables:***

Más allá de simplemente identificar patrones, la IA analítica proporciona interpretaciones y recomendaciones basadas en sus hallazgos. Estos conocimientos pueden traducirse directamente en acciones clínicas, como ajustes en los planes de tratamiento o intervenciones preventivas personalizadas.

#### **g) *Apoyo a la toma de decisiones:***

La IA analítica ofrece a los profesionales de la salud herramientas de apoyo a la decisión basadas en evidencia. Esto incluye predicciones de resultados de tratamientos, análisis de riesgos y beneficios, y recomendaciones de cuidado personalizadas, lo que permite a los médicos tomar decisiones más informadas y precisas.

Respecto a las técnicas y métodos usados en la IA analítica en salud, se emplea una variedad de técnicas avanzadas, incluyendo análisis predictivo para anticipar tendencias de salud, aprendizaje profundo para mejorar la precisión de los modelos mediante redes neuronales, aprendizaje automático para perfeccionar continuamente la exactitud de los modelos, análisis estadístico avanzado para validar hallazgos, y procesamiento de lenguaje natural para extraer información valiosa de notas clínicas y literatura médica.

### 1.3.2 DIFERENCIAS Y COMPARACIÓN ENTRE IA GENERATIVA Y ANALÍTICA Y EN SALUD

Ambas IA tienen grandes implicaciones en el ámbito de la salud, pero sus enfoques y aplicaciones son distintas.

La IA analítica se enfoca en analizar datos existentes para identificar tendencias y patrones, mejorando la toma de decisiones médicas, la optimización de tratamientos y la personalización de la atención. Sus beneficios incluyen la capacidad de proporcionar conclusiones rápidas e innovadoras mediante el análisis de grandes volúmenes de datos históricos.

Por otro lado, la IA generativa crea nuevos datos y escenarios no presentes en los datos originales. Sus aplicaciones incluyen el descubrimiento de nuevos fármacos, la síntesis de proteínas, la creación de chatbots de salud y la generación de contenido educativo. Aunque incrementa la productividad y la innovación, también conlleva riesgos asociados a la generación de información incorrecta o engañosa.

En resumen, la IA analítica mejora la precisión y eficiencia en la atención médica mediante el análisis de datos existentes, mientras que la IA generativa abre nuevas posibilidades creativas y disruptivas, aunque requiere un manejo cuidadoso debido a sus riesgos inherentes.

A continuación, en la **Tabla 1**. Características de la IA analítica y generativa en salud. Autora; Elena Plaza Moreno. Elaboración propia con uso de ChatGPT 4.0 (2024). se resumen los tres tipos de IA explicados en este apartado.

CARACTERÍSTICAS DE LA IA GENERATIVA Y ANALÍTICA EN SALUD		
	IA Generativa	IA Analítica
<b>Definición</b>	Modelos que generan nuevos datos que imitan los datos reales.	Modelos que analizan datos para extraer información útil y detectar patrones.
<b>Función Principal</b>	Crear datos, imágenes o estructuras nuevas.	Analizar grandes volúmenes de datos para identificar tendencias y hacer predicciones.
<b>Aplicaciones en Salud</b>	<ul style="list-style-type: none"> <li>▪ Generación de imágenes médicas sintéticas.</li> <li>▪ Diseño de nuevos fármacos.</li> <li>▪ Creación de datos sintéticos para investigación.</li> <li>▪ Simulaciones de tratamientos personalizados.</li> <li>▪ Visualizaciones para explicar condiciones médicas a pacientes.</li> </ul>	<ul style="list-style-type: none"> <li>▪ Análisis de datos clínicos para descubrir tendencias.</li> <li>▪ Identificación de factores de riesgo en poblaciones.</li> <li>▪ Evaluación de la efectividad de tratamientos.</li> <li>▪ Optimización de recursos hospitalarios.</li> </ul>
<b>Beneficios</b>	<ul style="list-style-type: none"> <li>▪ Ayuda en la formación y la investigación sin comprometer la privacidad.</li> <li>▪ Acelera el descubrimiento de fármacos.</li> <li>▪ Permite la personalización de tratamientos.</li> </ul>	<ul style="list-style-type: none"> <li>▪ Proporciona información detallada y útil para la toma de decisiones.</li> <li>▪ Ayuda a predecir brotes de enfermedades.</li> <li>▪ Optimiza la gestión de recursos.</li> <li>▪ Identifica áreas de mejora en el cuidado de la salud.</li> </ul>

<b>Limitaciones</b>	<ul style="list-style-type: none"> <li>▪ Requiere grandes conjuntos de datos para entrenamiento eficaz.</li> <li>▪ Riesgo de sesgo si el modelo no está bien ajustado.</li> </ul>	<ul style="list-style-type: none"> <li>▪ Dependiente de la calidad y cantidad de datos disponibles.</li> <li>▪ Puede ser complejo interpretar los resultados sin conocimiento especializado.</li> </ul>
<b>Cómo funciona</b>	Aprende patrones y estructuras a partir de grandes volúmenes de datos existentes. Luego, utilizan esta información para crear nuevos datos que sean similares a los originales.	Procesan y examinan grandes cantidades de datos para identificar patrones, relaciones y tendencias. Utilizan algoritmos estadísticos y matemáticos para transformar datos en información útil que puede ser utilizada para la toma de decisiones y predicciones.
<b>Ejemplo</b>	Uso de ChatGPT u otras aplicaciones para generar contenido nuevo: textos, imágenes, resúmenes de pdf, preguntas tipo test. En definitiva, las nuevas herramientas que tenemos al alcance de nuestra mano desde 2022. Su mayor aplicación actualmente en salud es la de "gestión de conocimiento".	Un algoritmo informático basado en inteligencia artificial entrenado con millones de historiales clínicos de los pacientes, sus electrocardiogramas y otras imágenes. Y será capaz de aportar resultados estadísticos, números, predictivos, etc. Esto se puede aplicar en salud, sobre todo, en investigación.

**Tabla 1. Características de la IA analítica y generativa en salud. Autora; Elena Plaza Moreno. Elaboración propia con uso de ChatGPT 4.0 (2024).**

## 1.4 ¿NOS VA A QUITAR EL TRABAJO LA IA?

En el ámbito laboral, la inteligencia artificial ha generado preocupación por la automatización de empleos, lo que puede llevar a la pérdida de puestos de trabajo en sectores tradicionales.

Si bien la IA ha creado nuevas oportunidades de empleo en áreas relacionadas con su desarrollo y mantenimiento, es fundamental abordar la reestructuración laboral que se está produciendo para garantizar una transición justa y equitativa para los trabajadores afectados.

Pero también hay que tener en cuenta que, si bien es cierto que las máquinas pueden suplantar ciertos roles laborales, las personas tienen la capacidad de asumir nuevas funciones.

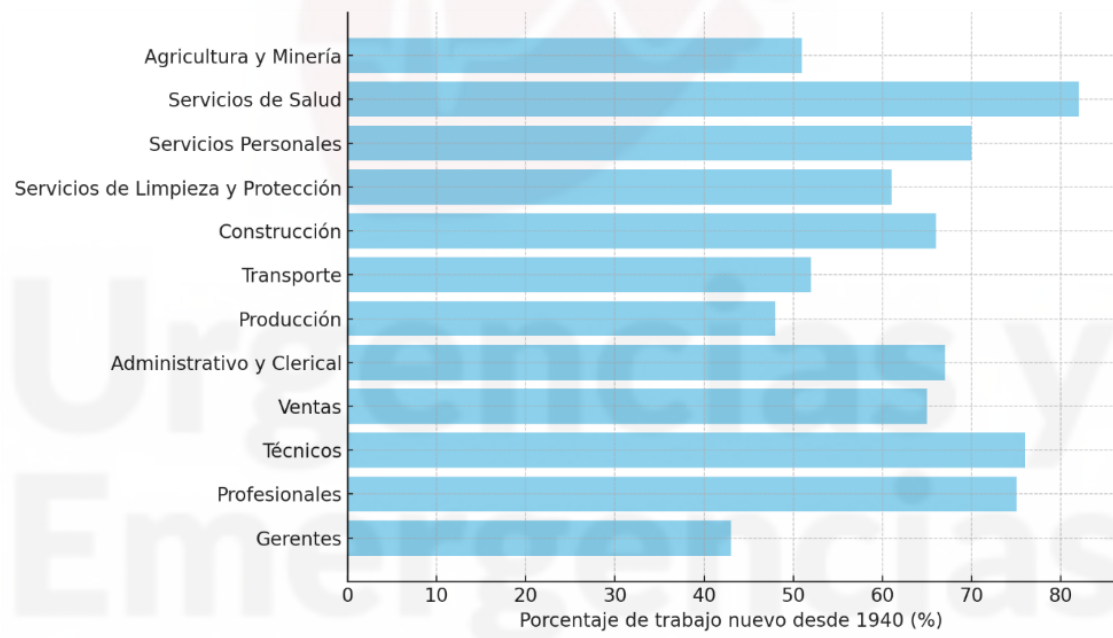
Históricamente, esto se observó durante el siglo pasado cuando la automatización de trabajos en los sectores agrícolas e industriales llevó a que la fuerza laboral se orientara hacia el sector de servicios.

Esta tendencia ha sido respaldada por numerosos estudios que han analizado datos históricos, llegando a la conclusión de que el avance tecnológico y el aumento en la productividad laboral no han resultado en un incremento del desempleo a nivel mundial.

Además, se reconoce ampliamente que uno de los motivos por los cuales la automatización no ha provocado un desempleo persistente es el aumento poblacional, que a su vez ha impulsado la demanda de bienes y servicios industriales.

De hecho, hay economistas que sugieren que "la innovación conduce a nuevos puestos de trabajo que representan la mayor parte del crecimiento del empleo.

Un estudio reciente del economista David H. Autor encuentra que el 60% de los trabajadores de hoy están empleados en ocupaciones que no existían en 1940. Esto implica que más del 85% del crecimiento del empleo en los últimos 80 años se explica por puestos creados a partir de nuevas tecnologías (**Gráfica 1**).



**Gráfica 1. Porcentaje de trabajo nuevo creado desde 1940 por sector. Autora: Elena Plaza Moreno. Elaboración propia con datos de Autor, D. (2022). [https://www.nber.org/system/files/working\\_papers/w30074/w30074.pdf](https://www.nber.org/system/files/working_papers/w30074/w30074.pdf)**

En salud, la IA nunca debe sustituir los conocimientos, habilidades y destrezas de un profesional sanitario. Deben de ser complementarios a nuestros criterios científicos.

## 1.5 FECHAS Y NOMBRES CLAVE

La historia de la inteligencia artificial se remonta a la década de 1940, aunque los conceptos esenciales se remontan a la lógica y algoritmos de los griegos y a las matemáticas de los árabes.

A continuación, se presentan las fechas y nombres clave en el desarrollo de la inteligencia artificial, y esta información queda también recogida en la **Imagen 3**:

- **Década de 1930:** El origen más primigenio de la IA se remonta a esta década, cuando se fundó el laboratorio del MIT, donde surgieron las primeras teorizaciones y prototipos de inteligencia artificial.
- **1940:** Se publicaron trabajos que no tuvieron gran repercusión, pero que sentaron las bases para el desarrollo posterior de la IA.
- **1950:** Alan Turing, considerado el padre de la inteligencia artificial, publicó su artículo "*Computing Machinery and Intelligence*" (Maquinaria computacional e inteligencia.). En el texto, Turing, reconocido por romper el código ENIGMA de los nazis en la Segunda Guerra Mundial, plantea la cuestión "¿pueden pensar las máquinas?" e introduce el Test de Turing para evaluar si un computador puede exhibir una inteligencia comparable (o sus efectos) a la de un ser humano. Desde entonces, la validez del Test de Turing ha sido objeto de debate.
- **1951:** William Shockley inventó el transistor de unión, lo que permitió la construcción de computadoras más rápidas y pequeñas

- **1952:** Arthur Samuel creó un software capaz de aprender a jugar al ajedrez de forma autónoma
- **1956:** John McCarthy acuñó formalmente el término "inteligencia artificial" durante la conferencia de Dartmouth, junto con Marvin Minsky y Claude Shannon. Más tarde, ese mismo año, Allen Newell, J.C. Shaw y Herbert Simon crearon Logic Theorist, el primer programa de software de IA operativo.
- **1967:** Frank Rosenblatt desarrolló el Perceptron Mark 1, el primer ordenador que funcionaba con una red neuronal capaz de "aprender" a través del método de prueba y error. Al año siguiente, Marvin Minsky y Seymour Papert publicaron un libro llamado "Perceptrons", que no solo se convirtió en un texto clave sobre las redes neuronales, sino que también, por un tiempo, argumentó en contra de la continuación de la investigación en este campo.
- **Durante la década de 1980,** las redes neuronales que empleaban el algoritmo de retropropagación para su entrenamiento empezaron a ser ampliamente utilizadas en diversas aplicaciones de inteligencia artificial.
- **1997:** Deep Blue, un sistema desarrollado por IBM, logró vencer al campeón mundial de ajedrez de ese entonces, Garry Kasparov, en un histórico encuentro de ajedrez y su correspondiente revancha.
- **2000:** el programa Artificial Linguistic Internet Computer Entity (A.L.I.C.E.) ganó el premio Loebner al Chatbot más humano
- **2007:** el programa Ultra Hal Assistant ganó el premio Loebner. Este mismo año se crea los Transformers, un tipo de red neuronal que cambiará cómo funciona la IA que trabaja con texto.
- **2011:** IBM Watson superó a los campeones Ken Jennings y Brad Rutter en el concurso de Jeopardy!, demostrando sus avanzadas capacidades en procesamiento de lenguaje natural y búsqueda de información.
- **2012:** se considera el año en que comenzó la explosión de la IA, según Yoshua Bengio, uno de los investigadores más prestigiosos en el campo.
- **2015:** el superordenador Minwa de Baidu empleó un tipo especial de red neuronal profunda conocida como red neuronal convolucional para identificar y categorizar imágenes, alcanzando una tasa de precisión superior a la del humano promedio.
- **2016:** el programa AlphaGo de DeepMind, que también utilizaba redes neuronales profundas, derrotó a Lee Sodol, campeón mundial del juego de Go, en una serie de cinco partidas. Esta victoria fue especialmente notable debido al enorme número de movimientos posibles en el juego, que asciende a más de 14,5 billones después de solo cuatro movimientos. Posteriormente, Google adquirió DeepMind por 400 millones de dólares.
- **2020-2021:** la pandemia favoreció los avances de la IA en el ámbito de la salud, y creció el uso del "Low Code", tecnología que permite a usuarios sin formación técnica desarrollar aplicaciones sencillas
- **Noviembre 2022:** OpenAI lanza ChatGPT, un modelo conversacional basado en GPT-3.5, que rápidamente se vuelve popular por su habilidad para generar respuestas detalladas y coherentes en lenguaje natural.
- **Enero 2023:** OpenAI anuncia una colaboración con Microsoft para integrar ChatGPT en Bing y en Microsoft 365, lo que marca un paso significativo hacia la incorporación de capacidades avanzadas de IA en herramientas de productividad y búsqueda.
- **Marzo 2023:** OpenAI lanza GPT-4, una versión mejorada de su modelo generativo, que potencia la nueva versión de ChatGPT. GPT-4 mejora significativamente en términos de precisión, capacidad de comprensión del contexto y generación de texto más coherente y relevante.

- **2023:** Se observan aplicaciones de ChatGPT en educación, asistencia al cliente, creación de contenido, programación, y más, demostrando su versatilidad y capacidad para adaptarse a diferentes industrias y necesidades.

Hitos clave de la IA.



Imagen 3. Hitos clave de la IA. Autora: Elena Plaza Moreno. Elaboración propia.

## 1.6 RESUMEN

La inteligencia artificial se refiere a sistemas que imitan la inteligencia humana, permitiendo a las máquinas realizar tareas como el aprendizaje y la toma de decisiones. Lleva mucho tiempo entre nosotros, evolucionando significativamente desde sus primeros días. Funciona con algoritmos matemáticos e informáticos.

La IA se puede clasificar en diferentes tipos según su capacidad, funcionalidad y enfoque. La IA débil está diseñada para tareas específicas, mientras que la IA general puede realizar cualquier tarea intelectual que un ser humano pueda hacer (todavía no hemos llegado a este punto).

En el ámbito de la salud, la IA analítica ayuda en el diagnóstico precoz, la predicción de riesgos de enfermedades y la identificación de factores de riesgo, mientras que la IA generativa se utiliza en la creación de nuevo contenido. Dentro de esta IA generativa está ChatGPT y otro grupo diversos de herramientas que generan contenido nuevo (texto, audio, vídeo, imágenes). Esta generación de nuevo contenido no está exenta de peligros y riesgos que, en salud, deben estar regulados para proteger datos personales y usar la IA de modo ético.

A pesar de las preocupaciones sobre la automatización laboral y la posible pérdida de empleos, la IA también ha creado nuevas oportunidades de empleo y ha demostrado ser una herramienta valiosa en sectores como la salud.

## 1.7 BIBLIOGRAFÍA

¿Qué es la Inteligencia Artificial (IA)? [Internet]. Ibm.com. 2023 [citado el 12 de mayo de 2024]. Disponible en: <https://www.ibm.com/es-es/topics/artificial-intelligence>.

High-Level Expert Group on Artificial Intelligence. *A DEFINITION OF AI: MAIN CAPABILITIES AND SCIENTIFIC DISCIPLINES* [Internet]. Europa.eu. 2018 [citado el 12 de mayo de 2024]. Disponible en: [https://ec.europa.eu/futurium/en/system/files/ged/ai\\_hleg\\_definition\\_of\\_ai\\_18\\_december\\_1.pdf](https://ec.europa.eu/futurium/en/system/files/ged/ai_hleg_definition_of_ai_18_december_1.pdf)

McCarthy J, Minsky ML, Rochester N, Shannon CE. *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence*, August 31, 1955. AIMag [Internet]. 2006Dec.15;27(4):12. Available from: <https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/1904>

McCarthy J. *What is artificial intelligence?* [Internet]. Stanford.edu. 2007. Disponible en: <https://www-formal.stanford.edu/jmc/whatisai.pdf>

Ministerio para la Transformación Digital y de la Función Pública. ENIA - *Estrategia Nacional de Inteligencia Artificial* [Internet]. Gob.es. [citado el 12 de mayo de 2024]. Disponible en: <https://portal.mineco.gob.es/es-es/digitalizacionIA/Paginas/ENIA.aspx>

Maslej N et al. *"The AI Index 2023 Annual Report,"* AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA, April 2023. The AI Index 2023 Annual Report by Stanford University is licensed under Attribution-NoDerivatives 4.0 International.

OpenAI. (2024). *ChatGPT 4* (versión del 15 de mayo) [Modelo de lenguaje de gran tamaño]. <https://chat.openai.com/chat>

Qué es la Inteligencia Artificial [Internet]. Gob.es. [citado el 12 de mayo de 2024]. Disponible en: <https://planderrecuperacion.gob.es/noticias/que-es-inteligencia-artificial-ia-prtr>

Samoli, S., López Cobo, M., Gómez, E., De Prato, G., Martínez-Plumed, F., and Delipetrev, B., *AI Watch. Defining Artificial Intelligence. Towards an operational definition and taxonomy of artificial intelligence*, EUR 30117 EN, Publications Office of the European Union, Luxembourg, 2020, ISBN 978-92-76-17045-7, doi:10.2760/382730, JRC118163.

Tuomi, I. *The Impact of Artificial Intelligence on Learning, Teaching, and Education. Policies for the future*, Eds. Cabrera, M., Vuorikari, R & Punie, Y., EUR 29442 EN, Publications Office of the European Union, Luxembourg, 2018, ISBN 978-92-79-97257-7, doi:10.2760/12297, JRC113226.

Weizenbaum J. *ELIZA—a computer program for the study of natural language communication between man and machine*. Commun ACM [Internet]. 1966;9(1):36–45. Disponible en: <https://web.stanford.edu/class/cs124/p36-weizenbaum.pdf>



# MÓDULO 1. INTRODUCCIÓN A LA INTELIGENCIA ARTIFICIAL

## TEMA 2. APLICACIONES DE LA IA EN SALUD

*«El progreso y el desarrollo son imposibles si uno sigue haciendo las cosas tal como siempre las ha hecho».*

WAYNE W. DYER



## TEMA 2. APLICACIONES DE LA IA EN SALUD

### 2.1 INTRODUCCIÓN

Un informe de *Fortune Business Insights* de 2024 señala que la integración de tecnologías de inteligencia artificial en el sector sanitario está experimentando un auge importante. Este crecimiento se debe, principalmente, a su aplicación en diagnósticos, tratamientos personalizados y gestión hospitalaria. Según el análisis, **la IA está transformando el cuidado de la salud al optimizar la precisión, aumentar la eficiencia y disminuir los costos.**

Entre las áreas más destacadas de implementación se encuentran la automatización de procesos, el análisis de imágenes médicas y la gestión de grandes volúmenes de datos.

En este tema, abordaremos ejemplos de cómo la salud puede beneficiarse de las técnicas de machine learning (ML) y Deep learning (DL), como por ejemplo en el pronóstico, el diagnóstico, el tratamiento y el flujo de trabajo clínico. Además, se describirán otras aplicaciones aplicadas a fármacos, cirugía, rehabilitación, etc.

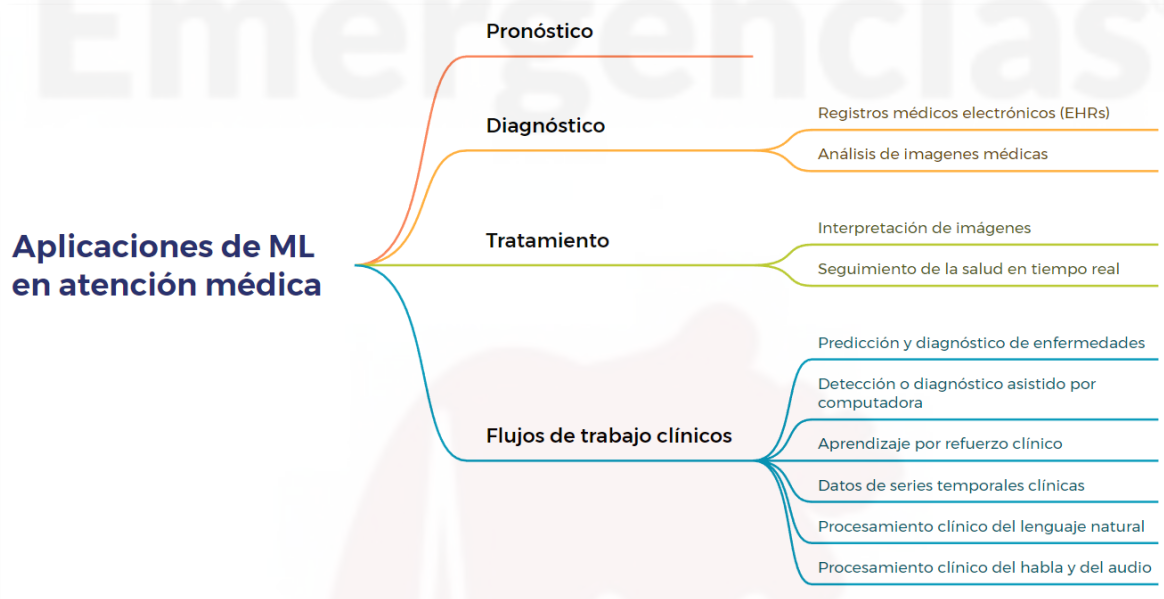


Imagen 4. Aplicaciones de la IA en salud. Elaborado por Elena Plaza Moreno.

### 2.2. APLICACIONES EN PRONÓSTICO Y DIAGNÓSTICO

La IA se puede utilizar para predecir la probabilidad de que un individuo desarrolle ciertas enfermedades analizando una combinación de datos genéticos, historial médico y otros factores de riesgo. Esto permite la implementación de **medidas preventivas personalizadas**, mejorando la salud a largo plazo de los pacientes.

En cuanto al diagnóstico, la IA es capaz de detectar patrones y anomalías que pueden no ser evidentes para el ojo humano, **mejorando la precisión del diagnóstico y la eficacia del tratamiento.**

Además, la IA puede ser especialmente útil en situaciones donde el diagnóstico es complejo o la enfermedad es poco común (Imagen 5). Al tener en cuenta una amplia gama de factores y antecedentes médicos, los algoritmos de IA pueden ofrecer a los médicos recomendaciones precisas y personalizadas para ayudar en la

toma de decisiones clínicas. Esta capacidad de la IA para analizar datos de manera integral y considerar múltiples variables puede ser fundamental para mejorar la precisión de los diagnósticos y garantizar que los pacientes reciban el tratamiento adecuado en el momento oportuno.

### La Comunidad de Madrid prueba un proyecto pionero para reducir el tiempo de diagnóstico de enfermedades raras con Inteligencia Artificial



La Comunidad de Madrid ha firmado un acuerdo de colaboración con Microsoft y la Fundación 29 para implantar un proyecto pionero en el mundo que aplicará la Inteligencia Artificial (IA) en el diagnóstico de enfermedades raras, ayudando a los profesionales sanitarios a reducir los tiempos a la hora de identificar este tipo de patologías, entre otras aplicaciones.



15 septiembre 2023

Imagen 5. Captura de pantalla del proyecto con IA para el diagnóstico de enfermedades raras de la Comunidad de Madrid.  
Fuente: <https://www.comunidad.madrid/noticias/2023/09/15/comunidad-madrid-prueba-proyecto-pionero-reducir-tiempo-diagnostico-enfermedades-raras>

#### 2.2.1. APLICACIONES DE EJEMPLO DE USO DE IA EN PRONÓSTICO

- **Evaluación del riesgo genético:** identificar variantes genéticas que aumentan el riesgo de enfermedades hereditarias, como ciertos tipos de cáncer, enfermedades cardiovasculares y trastornos neurológicos.
- **Predicción de enfermedades crónicas:** analizar datos de salud y estilo de vida para predecir la probabilidad de desarrollar enfermedades crónicas como diabetes, hipertensión y enfermedades del corazón.
- **Detección temprana de enfermedades:** utilizar datos de monitoreo continuo, como el uso de wearables y aplicaciones de salud, para detectar signos tempranos de enfermedades y permitir intervenciones antes de que se desarrollen plenamente.
- **Prevención de recaídas:** predecir la probabilidad de recaída en pacientes que han superado una enfermedad grave, como el cáncer, para ajustar los planes de seguimiento y tratamiento.

- **Predicción de complicaciones postoperatorias:** evaluar el riesgo de complicaciones después de una cirugía basándose en el historial médico y las condiciones preexistentes del paciente.

### 2.2.2. BENEFICIOS DE USO DE IA EN PRONÓSTICO

- **Intervenciones personalizadas:** permite diseñar planes de prevención y tratamiento específicos para cada paciente, basándose en su perfil de riesgo individual.
- **Mejora en la Salud Pública:** ayuda a las autoridades sanitarias a identificar poblaciones en riesgo y a implementar programas de prevención dirigidos, mejorando la salud general de la comunidad.
- **Empoderamiento del paciente:** proporciona a los pacientes información valiosa sobre su salud y riesgos potenciales, empoderándolos para tomar decisiones informadas sobre su estilo de vida y cuidados preventivos.
- **Reducción de costos sanitarios:** la predicción temprana y la prevención de enfermedades pueden reducir significativamente los costos asociados con el tratamiento de enfermedades avanzadas y las hospitalizaciones recurrentes.
- **Optimización de recursos:** ayuda a los sistemas de salud a planificar y asignar recursos de manera más efectiva, enfocándose en la prevención y el manejo temprano de enfermedades.

### 2.2.3. APLICACIONES DE EJEMPLO DE LA IA EN DIAGNÓSTICO

- **Detección de cáncer:** análisis de mamografías, tomografías computarizadas y resonancias magnéticas para identificar tumores malignos en etapas tempranas, como cáncer de mama, pulmón y cerebro.
- **Diagnóstico de enfermedades cardiovasculares:** evaluación de electrocardiogramas y ecocardiogramas para detectar arritmias, cardiomiopatías y otras afecciones cardíacas.
- **Identificación de enfermedades oftalmológicas:** análisis de imágenes de retina para detectar enfermedades oculares como la retinopatía diabética y el glaucoma.
- **Diagnóstico de enfermedades dermatológicas:** clasificación de imágenes de la piel para identificar melanoma y otras afecciones cutáneas.
- **Reconocimiento de patrones en análisis de laboratorio:** interpretación de resultados de pruebas de laboratorio, como análisis de sangre y orina, para detectar anomalías que podrían indicar diversas condiciones de salud.

### 2.2.4. BENEFICIOS DEL USO DE LA IA EN DIAGNÓSTICO

- **Mejora en la precisión:** las redes neuronales pueden superar la precisión de los diagnósticos humanos en ciertas áreas al detectar patrones sutiles en los datos médicos que podrían pasarse por alto.
- **Aceleración del proceso de diagnóstico:** reducen el tiempo necesario para llegar a un diagnóstico al procesar grandes volúmenes de datos rápidamente, permitiendo intervenciones más tempranas.
- **Reducción de errores:** proporcionan una herramienta de apoyo a los médicos y otros profesionales de la salud, ayudando a minimizar errores de diagnóstico y mejorando la confiabilidad de las evaluaciones médicas.
- **Acceso a diagnósticos especializados:** permiten a los médicos y otros profesionales de la salud en áreas rurales o con menos acceso a especialistas obtener diagnósticos avanzados a través de sistemas de telemedicina, mejorando la equidad en la atención de salud.

## 2.3. SEGUIMIENTO DE LA SALUD EN TIEMPO REAL

El seguimiento en tiempo real de los pacientes críticos es un componente clave del proceso de tratamiento. La monitorización continua de la salud mediante dispositivos portátiles, sensores de IoT y teléfonos inteligentes está ganando interés entre la gente.

Las redes neuronales permiten la creación de sistemas avanzados de telemedicina que pueden diagnosticar y monitorear a los pacientes de manera remota. Por ejemplo, una red neuronal puede ser entrenada para analizar imágenes de piel enviadas por los pacientes a través de una aplicación móvil. La red puede detectar signos de cáncer de piel y otras afecciones dermatológicas, proporcionando un diagnóstico preliminar y recomendaciones de tratamiento que luego son revisadas por un dermatólogo. Esto permite a los pacientes recibir atención médica sin necesidad de viajar largas distancias.

### 2.3.1. APLICACIONES

- **Diagnóstico remoto:** evaluación de imágenes, videos y datos de sensores enviados por los pacientes para diagnosticar afecciones a distancia.
- **Monitoreo de salud:** seguimiento continuo de pacientes con dispositivos portátiles que envían datos vitales a los profesionales de la salud.
- **Consultas virtuales:** realización de consultas médicas a través de videollamadas, con soporte de análisis automatizados de datos de salud.
- **Monitorización del dolor:** tal y como afirman Pouromran y colaboradores en un estudio publicado en PLOS ONE, los Sistemas de IA integrados en dispositivos wearables pueden monitorear de forma continua parámetros fisiológicos para estimar la intensidad del dolor en tiempo real mejorando la precisión del seguimiento y permitiendo ajustes rápidos en el tratamiento.

### 2.3.2 BENEFICIOS

- **Acceso a la atención médica:** mejora el acceso a la atención médica en áreas rurales o desatendidas.
- **Comodidad para el paciente:** permite a los pacientes recibir atención médica desde la comodidad de su hogar.
- **Eficiencia de los recursos de salud:** optimiza el uso de recursos médicos al reducir la necesidad de visitas presenciales, liberando tiempo para atender casos más críticos.

## 2.4. MONITORIZACIÓN DE CONSTANTES VITALES EN TIEMPO REAL

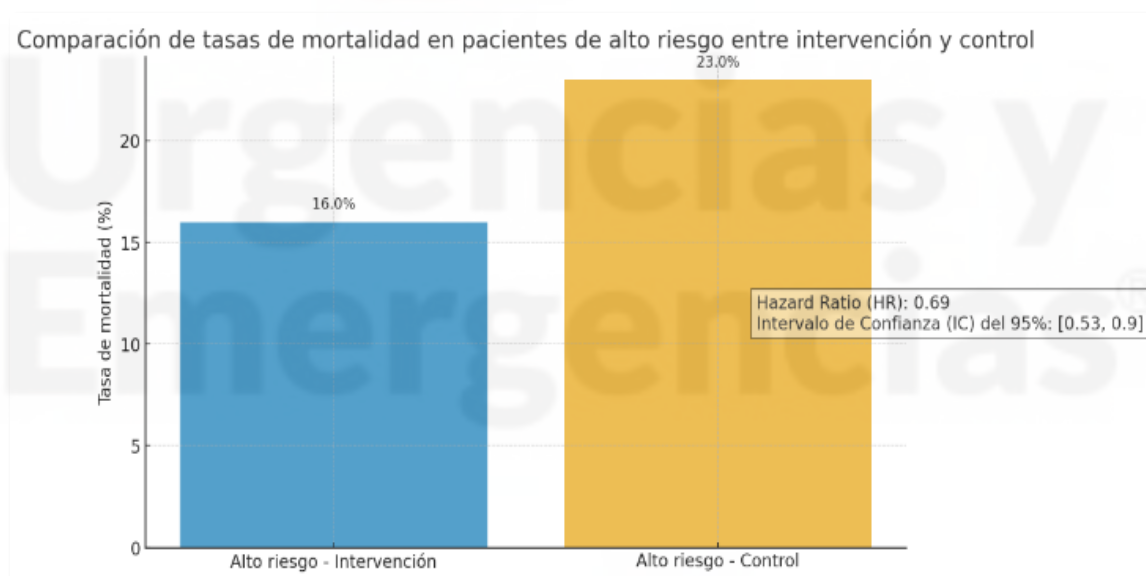
El seguimiento de la salud o las constantes vitales no solo es aplicable en telesalud y de forma remota. A nivel “presencial” tiene ventajas que cambiarán la atención sanitaria en los próximos años.

Por ejemplo, en una unidad de cuidados intensivos, una red neuronal puede ser entrenada con datos históricos de pacientes, incluyendo signos vitales como la frecuencia cardíaca, la presión arterial, la saturación de oxígeno y la temperatura. La red puede monitorear en tiempo real los datos de los pacientes actuales y detectar patrones que indican un riesgo inminente de eventos críticos, como un paro cardíaco o una sepsis, alertando a los profesionales de la salud de inmediato.

Viderman (2023) realizó una revisión sistemática del uso de la IA en resucitación cardiopulmonar. Los resultados fueron impactantes:

- Se puede tener previsión de taquicardias seis horas antes del evento.
- La alarma con previsión del deterioro del paciente tiene una sensibilidad del 100%.
- Se pueden predecir los malos resultados neurológicos en pacientes con coma post-anóxico con un valor predictivo positivo del 100%.

Por otro lado, Lim (2024) llevó a cabo un estudio clínico aleatorizado para estudiar un sistema de alerta en la monitorización electrocardiográfica de los pacientes. El estudio encontró que los pacientes del grupo de intervención con electrocardiografía habilitada por inteligencia artificial tenían un 31% menos de riesgo de morir en comparación con el grupo de control (**Gráfica 2**).



**Gráfica 2. Comparación de tasas de mortalidad de pacientes de alto riesgo en el grupo intervención y el grupo control. Elaborado por Elena Plaza Moreno**

#### 2.4.1. BENEFICIOS DEL USO DE LA IA EN LA MONITORIZACIÓN DE LAS CONSTANTES VITALES:

- **Intervención temprana:** permite una intervención médica rápida antes de que los problemas se conviertan en emergencias, mejorando los resultados para los pacientes.
- **Reducción de cargas de trabajo:** alivia la carga de trabajo al proporcionar una vigilancia continua y automatizada.
- **Mejora en la atención al paciente:** aumenta la seguridad del paciente mediante la detección proactiva de problemas potenciales.

### 2.5 REGISTROS MÉDICOS ELECTRÓNICOS (EHR)

Los hospitales y otros proveedores de servicios de atención médica producen una gran cantidad de registros médicos electrónicos (EHR) a diario y se componen de datos estructurados y no estructurados que contienen una información completa.

La IA puede utilizarse para analizar grandes volúmenes de datos clínicos de pacientes. De esta forma se extrae información relevante sobre diagnósticos, tratamientos y resultados.

### 2.5.1. APLICACIONES DE EJEMPLO

- **Análisis predictivo:** predecir el riesgo de reingreso hospitalario, complicaciones postoperatorias o la progresión de enfermedades crónicas, permitiendo a los profesionales de la salud intervenir de manera proactiva.
- **Optimización de tratamientos:** analizar la efectividad de diferentes tratamientos en diversas condiciones y subgrupos de pacientes, proporcionando recomendaciones personalizadas basadas en datos históricos.
- **Identificación de tendencias:** detectar tendencias en los datos de salud pública, como el aumento de casos de una enfermedad específica, para informar políticas de salud y estrategias de intervención.
- **Mejora de la calidad de atención:** identificar áreas de mejora en la atención sanitaria al analizar los resultados clínicos y las prácticas hospitalarias, ayudando a los centros a implementar cambios que mejoren los resultados para los pacientes.

## 2.6 DIAGNÓSTICO POR IMAGEN

En el análisis de imágenes, las técnicas de machine learning y de deep learning se utilizan para la extracción eficiente y efectiva de información que se adquiere mediante diferentes modalidades como, por ejemplo, resonancia magnética, tomografía computarizada, ultrasonido, tomografía por emisión de positrones (PET), radiografías, histología, etc.

Se espera que los sistemas inteligentes de diagnóstico por imágenes médicas totalmente automatizados formen parte de los sistemas sanitarios de próxima generación. De hecho, ya existen algoritmos de inteligencia artificial de este tipo incluidos en ecógrafos y otros aparatos. Se llaman **sistemas de detección y diagnóstico asistidos por computadora (CADE)**.

Además de la identificación de patrones o anomalías de enfermedades específicas, otra de las aplicaciones es la **mejora de imágenes**, en casos en los que haya artefactos. Por ejemplo, generar imágenes de resonancia magnética es un proceso bastante largo que generalmente requiere varios minutos para producir una imagen de buena calidad. Debido a que los movimientos pueden causar artefactos falsos en la adquisición de imágenes, el proceso completo generalmente debe repetirse varias veces para producir imágenes significativamente útiles. Según Qayyum y colaboradores (2021), existen diferentes modelos de deep learning para eliminar el ruido y artefactos de la resonancia magnética. Esto se traducirá en una reducción de tiempo en la adquisición de estas imágenes y en una mejora para el paciente.

En el apartado de diagnóstico de imágenes, es importante mencionar también la capacidad de clasificación. Por ejemplo, introducir una fotografía de una úlcera en una pierna, y que el sistema basado en IA sea capaz de distinguir entre úlcera vascular o úlcera arterial. Por último, los algoritmos de machine learning y deep learning serán capaces de segmentar las imágenes en fragmentos más pequeños, reconstruir imágenes para reducir el tiempo de adquisición y el espacio de almacenamiento.

### 2.6.1. BENEFICIOS

- **Mayor precisión:** la IA puede ayudar a mejorar la precisión en la detección de anomalías y enfermedades en las imágenes médicas, lo que puede llevar a diagnósticos más tempranos y tratamientos más efectivos para los pacientes.

- **Rapidez en el análisis:** los algoritmos de IA pueden analizar grandes volúmenes de imágenes en un corto período de tiempo, lo que puede ser especialmente útil en situaciones de emergencia o en entornos clínicos donde se requiere una interpretación rápida de las imágenes.
- **Reducción de errores:** al proporcionar una ayuda en la interpretación de imágenes, la IA puede reducir errores humanos y mejorar la consistencia en el diagnóstico y tratamiento de enfermedades. Esto puede llevar a una atención médica más segura y efectiva para los pacientes.

## 2.7. OTROS EJEMPLOS DE APLICACIONES

### 2.7.1. TRANSCRIPCIÓN DE CONSULTAS E INTEGRACIÓN EN HISTORIA CLÍNICA

Según el Dr. Simon Wallace, los médicos dedican el 50% de su tiempo a dedicar mucho tiempo a la documentación clínica y están muy desmotivados debido a la carga de trabajo clínico, las tareas administrativas y la falta de tiempo libre. Y no solo los médicos, todos los profesionales sanitarios debemos registrar todo en la historia del paciente.

El procesamiento clínico del habla y del audio ofrece nuevas oportunidades, como interfaces de voz para servicios sin interacción, transcripción automática de conversaciones de pacientes y síntesis de notas clínicas, etc. Algunos ejemplos de herramientas para transcribir consultas son Liberado, DeepCura, TurboScribe, Listen.Doctor y Llamalítica.

### 2.7.2. ANÁLISIS DE VOZ

El uso del audio no solo se orienta a los profesionales de salud, que encontrarían un gran ahorro de tiempo, o para la industria de la salud, que mejoraría la productividad y la reducción de costos. También hay beneficios para los pacientes: **determinadas herramientas de procesamiento del habla y del audio pueden predecir o llegar a diagnosticar enfermedades solo escuchando la voz.**

La voz cambia cuando una persona siente dolor: la calidad vocal podría estar influenciada por las vías neurales del dolor, lo que sugiere que diferentes tipos de dolor afectan de manera distinta a la voz. Así lo determinaron Hong y colaboradores en su estudio. Analizaron cómo diferentes características clínicas como la edad, el género y el punto o sitio de dolor influyen en la manifestación acústica del dolor y varían ciertos parámetros acústicos.

Por otro lado, en el proyecto TAME Pain tienen como principal objetivo desarrollar un algoritmo autónomo confiable que utilice biomarcadores acústicos derivados del habla para detectar los niveles de dolor en pacientes. Se busca proporcionar una herramienta adicional para la evaluación del dolor en pacientes, especialmente aquellos que no pueden expresarse verbalmente debido a barreras del idioma o discapacidades.

Aunque todavía está en desarrollo, se espera que este sistema de detección de dolor basado en el habla tenga el potencial de reducir el riesgo de evaluación incorrecta del dolor y mejorar la experiencia del paciente al proporcionar una evaluación más precisa y complementaria a las evaluaciones clínicas tradicionales.

### 2.7.3. INVESTIGACIÓN

Las redes neuronales se utilizan para mejorar el diseño y la conducción de ensayos clínicos. Pueden analizar grandes volúmenes de datos históricos de ensayos clínicos para identificar patrones que optimicen la selección de participantes, las dosis de medicamentos y los protocolos de tratamiento.

### 2.7.3.1. APLICACIONES

- **Selección de participantes:** identificar a los pacientes más adecuados para participar en un ensayo clínico basándose en sus características genéticas, historial médico y otros factores.
- **Predicción de respuestas:** predecir cómo diferentes subgrupos de pacientes responderán a un tratamiento específico, lo que permite la personalización de los tratamientos y la identificación de biomarcadores predictivos.
- **Optimización de dosis:** determinar las dosis óptimas de medicamentos para diferentes pacientes, minimizando efectos secundarios y maximizando la eficacia.

Aunque no sean grandes ensayos clínicos, la IA en general en la investigación nos permite acelerar el proceso de análisis de datos. Sin IA, por ejemplo, recogemos datos de forma manual, los analizamos con SPSS, y emitimos los resultados también de forma manual.

Con el uso de herramienta de software utilizada para análisis de datos con aprendizaje automático, como por ejemplo WEKA, puedes cargar tus datos, elegir un algoritmo (ej., árbol de decisión, k-means, redes neuronales), y obtener resultados en pocos pasos. Además, permite realizar análisis comparativos entre distintos algoritmos para evaluar cuál funciona mejor con tus datos. Y también crea modelos que aprenden de tus datos para hacer predicciones futuras.

Por tanto, en el área de la investigación, trabajar con inteligencia artificial permite automatizar procesos repetitivos, analizar grandes volúmenes de datos e identificar patrones que podrían pasar desapercibidos para los humanos.

También podemos ayudarnos de la IA en el ámbito de la investigación con las herramientas de IA que están surgiendo para gestión de bibliografía y otras diseñadas para facilitar la recopilación de información y la organización de ideas tras la lectura de artículos científicos.

Algunas de estas herramientas son (puedes hacer clic en ellas para abrirlo):

- [Elicit](#).
- [Research Rabbit](#).
- [Scite](#).
- [Consensus](#).
- [SciteSpace](#).
- [NotebookLM](#).

### 2.7.4. GESTIÓN DE LA SALUD PÚBLICA

En el ámbito de la salud pública, las redes neuronales se utilizan para:

- **Predicción de brotes epidémicos:** identificación de áreas de alto riesgo y predicción de la propagación de enfermedades infecciosas.

- **Optimización de campañas de vacunación:** determinación de los mejores lugares y momentos para implementar campañas de vacunación masiva.
- **Gestión de recursos sanitarios:** planificación y asignación de recursos médicos y de personal en respuesta a brotes y emergencias de salud pública.
- **Análisis de impacto de políticas de salud:** evaluación de la efectividad de las políticas de salud pública y ajuste de estrategias en tiempo real.

### 2.7.5. REHABILITACIÓN Y FISIOTERAPIA

En el ámbito de la rehabilitación y la fisioterapia, el uso de la IA tiene cabida en las siguientes aplicaciones:

- **Planes de rehabilitación personalizados:** creación de programas de ejercicios adaptados a las necesidades y capacidades específicas de cada paciente.
- **Monitoreo en tiempo real:** uso de sensores y dispositivos portátiles para recopilar datos sobre el movimiento y el esfuerzo del paciente, proporcionando retroalimentación inmediata.
- **Ajuste dinámico de ejercicios:** adaptación de la intensidad y el tipo de ejercicios en función del progreso y las respuestas del paciente.
- **Tele-rehabilitación:** permitir que los pacientes realicen ejercicios de rehabilitación en casa bajo la supervisión remota de profesionales de la salud, utilizando redes neuronales para monitorear el desempeño y ajustar el tratamiento según sea necesario.

Un ejemplo concreto de cómo la inteligencia artificial está transformando la rehabilitación es la plataforma Rehub (Nuevo, 2021). Diseñada especialmente para pacientes que requieren rehabilitación diaria, pero enfrentan dificultades para desplazarse a un centro especializado, esta herramienta utiliza tecnología de visión artificial para ofrecer un enfoque innovador y accesible. A través de la cámara del móvil, el software analiza hasta 74 puntos del cuerpo, evaluando aspectos como el rango, la velocidad y la repetición de los movimientos. Gracias a este análisis detallado, el sistema proporciona feedback en tiempo real, lo que permite al usuario realizar los ejercicios con mayor precisión y eficacia desde la comodidad de su hogar.

### 2.7.6. ASISTENTES VIRTUALES DE SALUD, CHATBOTS Y CHATS CONVERSACIONALES

Los asistentes virtuales de salud, chatbots de salud o chat conversacionales, ya no son el futuro, si no el presente.

Suelen estar entrenados con datos de muchas consultas médicas y puede interactuar con los pacientes a través de una aplicación móvil. El asistente puede responder preguntas sobre síntomas comunes, proporcionar recomendaciones de cuidado personal, recordar a los pacientes que tomen sus medicamentos a tiempo y ayudarles a programar citas médicas. Si el asistente detecta síntomas que requieren atención médica urgente, puede aconsejar al paciente que visite a un médico inmediatamente.

Además de proporcionar respuestas, otros ejemplos de aplicaciones son:

- **Recordatorios de medicación:** enviar alertas y recordatorios a los pacientes para asegurar que tomen sus medicamentos según lo prescrito.
- **Monitoreo de síntomas:** ayudar a los pacientes a realizar un seguimiento de sus síntomas y proporcionar recomendaciones basadas en los datos recopilados.

- **Apoyo en la gestión de enfermedades crónicas:** ofrecer consejos y recordatorios para ayudar a los pacientes a gestionar condiciones crónicas como la diabetes y la hipertensión.

Existen casos de éxito como el Chatbot **Limbic** del Servicio Nacional de Salud del Reino Unido. En el estudio de Habicht y colaboradores, publicado en Nature Medicine destaca cómo este chatbot de inteligencia artificial ha facilitado el acceso a servicios de salud mental a través del Servicio Nacional de Salud (NHS) de Inglaterra, especialmente entre grupos infrarrepresentados.

Este chatbot utiliza inteligencia artificial conversacional para facilitar las evaluaciones de salud mental y mejorar la eficiencia clínica en los servicios de psicoterapia.

Los resultados clave que se obtuvieron fueron:

- **Aumento de derivaciones:** ha logrado un incremento significativo en las derivaciones a servicios de salud mental, destacando un impacto notable entre las comunidades minoritarias.
- **Eficiencia mejorada:** ha facilitado un aumento del 15% en las derivaciones a terapias para ansiedad y depresión, comparado con un 6% en servicios que no lo utilizaban.
- **Impacto en grupos minoritarios:** se observaron aumentos del 179% en derivaciones de personas no binarias, 39% en pacientes asiáticos, y 40% en pacientes de color, demostrando cómo la tecnología puede superar barreras de acceso.
- **Calidad de evaluaciones:** la tecnología no solo optimizó el tiempo de evaluación por parte de los médicos, sino que también mejoró la calidad de estas, sin prolongar los tiempos de espera.
- **Experiencia del usuario:** los usuarios reportaron sentirse más esperanzados y valoraron la interacción no humana, especialmente aquellos de la comunidad no binaria, por evitar sentimientos de juicio o ansiedad.

Este último punto confirma que lo que puede ser una desventaja (atención no humana) puede ser beneficioso para algunos pacientes. Por lo que, como concluyen todos los estudios, hace falta más investigación en este ámbito del uso de la inteligencia artificial la salud para analizar la comunicación humano-máquina y sus beneficios en salud.

Los asistentes virtuales de salud no ayudan solo a los pacientes y comunidad. Existen chatbots especializados dirigidos a profesionales de la salud para manejar información muy extensa con preguntas concretas.

Un ejemplo de ello son ChatGPT especializados creados para resolver preguntas, analizar artículos científicos, traducir, etc. Se explica más adelante dónde se pueden encontrar en el entorno de la interfaz de ChatGPT.

### RCP GPT, ejemplo de chatbot para profesionales de la salud



RCP GPT es un chatbot basado en IA generativa precargado con las últimas guías de reanimación cardiopulmonar, arritmias y soporte vital cardiovascular avanzado de diferentes organismos internacionales.

Ayuda a profesionales de la salud al responder dudas sobre manejo de pacientes, medicación, resúmenes de información, etc.

<https://chatgpt.com/g/g-deVZSzfKe-rcp-gpt>

## 2.7.7. GESTIÓN DE RECURSOS HOSPITALARIOS

Aplicar la IA a la gestión de recursos hospitalarios permite:

- **Predicción de la ocupación de camas:** estimar la cantidad de camas que se necesitarán en función de las tendencias actuales y pasadas de admisión de pacientes.
- **Gestión del personal:** predecir la demanda de personal médico y ajustar los turnos para asegurar una cobertura adecuada en todo momento.
- **Optimización del uso de equipos:** analizar los patrones de uso de equipos médicos y optimizar su disponibilidad y mantenimiento, minimizando tiempos de inactividad.
- **Planificación de emergencias:** predecir y preparar respuestas ante aumentos repentinos en la demanda debido a epidemias, desastres naturales u otros eventos críticos.
- **Gestión de citas inteligente:** los sistemas basados en IA pueden mejorar la asignación de pacientes a consultas o procedimientos, optimizando los recursos y reduciendo tiempos de espera para consultas relacionadas con el manejo del dolor crónico, tal y como indican Barbosa y colaboradores en su estudio.

### 2.7.8. CIRUGÍA ASISTIDA POR IA

La cirugía asistida por inteligencia artificial es una de las áreas más emocionantes y prometedoras en el campo de la salud.

Investigadores de la Universidad Johns Hopkins han logrado que un robot, entrenado mediante aprendizaje por imitación al observar videos de cirujanos experimentados, ejecute procedimientos quirúrgicos con una destreza comparable a la humana. Este avance elimina la necesidad de programar cada movimiento individualmente, acercando la cirugía robótica a una autonomía completa. El modelo combina aprendizaje por imitación con una arquitectura de aprendizaje automático similar a la de ChatGPT, adaptada para interpretar movimientos robóticos.



<https://youtu.be/c1E170Xr6BM?si=9GL4FpLbSgeMu0tG>

Al alimentar al modelo con cientos de videos de cirugías realizadas con el sistema da Vinci, el robot aprendió a manipular agujas, levantar tejido y suturar con precisión. Este enfoque permite entrenar rápidamente a robots para realizar diversos procedimientos quirúrgicos, marcando un avance significativo en la robótica médica.

### 2.7.9. INDUSTRIA FARMACÉUTICA

Puede que la industria farmacéutica sea una de las más ineficaces. Uno de los principales problemas son los altos costos de investigación y desarrollo (I+D). Desarrollar un nuevo medicamento puede costar miles de millones de dólares y tomar más de una década. Además, la mayoría de los compuestos que se investigan no llegan al mercado, ya que solo una pequeña fracción pasa todas las fases de ensayos clínicos. La tasa de éxito es baja, con solo alrededor del 10% de los medicamentos que llegan a la fase clínica obteniendo la aprobación final.

Otro aspecto crítico es la duplicación de esfuerzos. Diferentes compañías pueden trabajar en investigaciones similares sin compartir datos, lo que lleva a una duplicación de esfuerzos y recursos. Las regulaciones estrictas, aunque necesarias para garantizar la seguridad y eficacia de los medicamentos, también pueden aumentar significativamente el tiempo y el costo del desarrollo de fármacos.

En cuanto al malgasto de recursos, gran parte de las inversiones en I+D no da frutos, ya que muchas moléculas no demuestran ser efectivas o seguras en las fases de prueba. Los costos de fallos en fases avanzadas son altos; cuando un medicamento falla en las fases finales de los ensayos clínicos, los costos ya invertidos en las fases anteriores se pierden. Además, las ineficiencias en el uso de datos son notables. La gran cantidad de datos generados no siempre se utiliza de manera óptima debido a la falta de integración y análisis efectivos.

Con este contexto, son varias las áreas en las que la IA puede ayudar, en concreto para:

- Descubrimiento de fármacos.
- Diseño de medicamentos.
- Ensayos clínicos.
- Análisis de grandes datos.
- Diagnóstico y personalización de tratamientos.
- Modelado y simulación.
- Reducción de costos.
- Investigación de enfermedades raras.
- Reutilización de medicamentos.
- Mejora de la adherencia al tratamiento.

A pesar de todas estas aplicaciones, la realidad a mediados de 2024 es que la aplicación de la IA en el desarrollo de fármacos es limitada. Así lo explican en el artículo *"Use of Artificial Intelligence in Drug Development"* cuyos resultados nos indican que encontraron 165 medicamentos de los cuales 164 están en investigación y solo 1 aprobado desarrollados con IA.

Otros resultados interesantes fueron que la IA se utilizó principalmente para el descubrimiento de moléculas de fármacos (76%) y el descubrimiento de objetivos terapéuticos (22%). Y respecto al área de estudio, los más frecuentes son los tratamientos anticancerígenos (27%) y neurológicos (24%).

## 2.9. CONCLUSIONES

La inteligencia artificial está redefiniendo el sector sanitario, introduciendo soluciones innovadoras que están revolucionando múltiples áreas clave.

En el diagnóstico y pronóstico, las herramientas de IA han demostrado ser fundamentales para mejorar la precisión y la velocidad en la detección temprana de enfermedades, lo que permite intervenciones más rápidas y personalizadas. Además, la monitorización en tiempo real de pacientes, a través de dispositivos conectados y sistemas inteligentes ya permite prever eventos críticos con hasta seis horas de anticipación, reduciendo así significativamente la mortalidad en casos específicos.

En cuanto a la gestión de registros médicos electrónicos (EHR), la IA ha facilitado el análisis y la extracción de información relevante, ayudando a identificar patrones que mejoran la toma de decisiones clínicas. En el ámbito del diagnóstico por imagen, los sistemas asistidos por computadora (CADE) se han convertido en aliados valiosos para la identificación precisa de patologías y la mejora en la calidad de las imágenes médicas.

Las aplicaciones de la IA también abarcan áreas prácticas como la transcripción automática de consultas y el análisis de voz para identificar condiciones médicas. En el terreno de la investigación, la inteligencia artificial acelera el análisis de datos y optimiza el diseño de ensayos clínicos, agilizando descubrimientos. Asimismo, los

asistentes virtuales y chatbots están facilitando el acceso a servicios de salud, mejorando especialmente la derivación a atención en salud mental en grupos minoritarios.

En la gestión hospitalaria, la IA está optimizando recursos, permitiendo predecir la ocupación de camas y planificar de manera más eficiente el personal necesario. Por otro lado, los avances en cirugía asistida por IA y robótica están ampliando los límites de los procedimientos quirúrgicos. En la industria farmacéutica, aunque aún en desarrollo, la IA está comenzando a revolucionar el diseño y la creación de nuevos medicamentos.

Aunque la evidencia actual muestra un impacto positivo, la implementación de estas tecnologías debe ser gradual, respaldada por estudios científicos sólidos, y siempre orientada a mejorar la atención al paciente y la eficiencia del sistema sanitario.

## 2.10. BIBLIOGRAFÍA

- Armoundas, A. A., Narayan, S. M., Arnett, D. K., Spector-Bagdady, K., Bennett, D. A., Celi, L. A., Friedman, P. A., Gollob, M. H., Hall, J. L., Kwitek, A. E., Lett, E., Menon, B. K., Sheehan, K. A., Al-Zaiti, S. S., & on behalf of the American Heart Association Institute for Precision Cardiovascular Medicine; Council on Cardiovascular and Stroke Nursing; Council on Lifelong Congenital Heart Disease and Heart Health in the Young; Council on Cardiovascular Radiology and Intervention; Council on Hypertension; Council on the Kidney in Cardiovascular Disease; and Stroke Council. (2024). Use of artificial intelligence in improving outcomes in heart disease: A scientific statement from the American Heart Association. *Circulation*, 149(14). <https://doi.org/10.1161/cir.0000000000001201>
- Barbosa, V. P., Said dos Reis, J. L. M., von Zuben de Valega Negrão, C., Barboza, V. R., Simoes Marcondes, K. C. B., Rezende, E. P., Neto Pereira Cerize, N., & de Paula Guirado, V. M. (2023). Machine learning model to predict allocation of patients with chronic back pain for integrated practice units in a system of value-based health care. En *medRxiv* (p. 2023.11.05.23298111). <https://doi.org/10.1101/2023.11.05.23298111>
- Da Rio, L., Spadaccini, M., Parigi, T. L., Gabbiadini, R., Dal Buono, A., Busacca, A., Maselli, R., Fugazza, A., Colombo, M., Carrara, S., Franchellucci, G., Alfarone, L., Facciorusso, A., Hassan, C., Repici, A., & Armuzzi, A. (2023). Artificial intelligence and inflammatory bowel disease: Where are we going? *World journal of gastroenterology: WJG*, 29(3), 508–520. <https://doi.org/10.3748/wjg.v29.i3.508>
- Druedahl, L. C., Price, W. N., II, Minssen, T., & Sarpatwari, A. (2024). Use of artificial intelligence in drug development. *JAMA Network Open*, 7(5), e2414139. <https://doi.org/10.1001/jamanetworkopen.2024.14139>
- Habicht, J., Viswanathan, S., Carrington, B., Hauser, T. U., Harper, R., & Rollwage, M. (2024). Closing the accessibility gap to mental health treatment with a personalized self-referral chatbot. *Nature Medicine*, 30(2), 595–602. <https://doi.org/10.1038/s41591-023-02766-x>
- Hong, H.-T., Li, J.-L., Weng, Y.-M., Ng, C.-J., & Lee, C.-C. (2019). Investigating the variability of voice quality and pain levels as a function of multiple clinical parameters. *Interspeech 2019*, 3058–3062.
- Moingeon, P., Garbay, C., Dahan, M., Fermont, I., Benmakhlouf, A., Gouyette, A., Poitou, P., Saint-Pierre, A., & le Groupe de Veille IA et Sciences pharmaceutiques. (2024). L'intelligence artificielle, une révolution dans

- le développement des médicaments. *Medecine sciences*, 40(4), 369–376. <https://doi.org/10.1051/medsci/2024028>
- Nuevo, M., Mahdavi, H., Rodríguez, D., Faura, T., Fabrellas, N., Balocco, S., Conti, M., Castagna, A., & Prat, S. (2021). Evaluation of safety and efficacy of ReHub in patients who underwent primary total Knee Arthroplasty: Study protocol for a randomized controlled trial. *International Journal of Surgery Protocols*, 25(1), 34–41. <https://doi.org/10.29337/ijsp.138>
- Pawar, B. (s/f). Artificial Intelligence (AI) in healthcare market size, growth, 2032. *Fortunebusinessinsights.com*. Recuperado el 16 de noviembre de 2024, de <https://www.fortunebusinessinsights.com/industry-reports/artificial-intelligence-in-healthcare-market-100534>
- Pouromran, F., Radhakrishnan, S., & Kamarthi, S. (2021). Exploration of physiological sensors, features, and machine learning models for pain intensity estimation. *PloS One*, 16(7), e0254108. <https://doi.org/10.1371/journal.pone.0254108>
- Qayyum, A., Qadir, J., Bilal, M., & Al-Fuqaha, A. (2021). Secure and robust machine learning for healthcare: A survey. *IEEE reviews in biomedical engineering*, 14, 156–180. <https://doi.org/10.1109/rbme.2020.3013489>
- Rosen, J. (2024, noviembre 11). *Robot that watched surgery videos performs with skill of human doctor*. The Hub. <https://hub.jhu.edu/2024/11/11/surgery-robots-trained-with-videos/>
- Schneiders, E., Williams, J., Farahi, A., Seabrooke, T., Vigneswaran, G., Bautista, J. R., Dowthwaite, L., & Piskopani, A.-M. (2023). TAME pain: Trustworthy AssessMEnt of pain from speech and audio for the empowerment of patients. *Proceedings of the First International Symposium on Trustworthy Autonomous Systems*, 1–4.
- Viderman D, Abdildin YG, Batkuldinova K, Badenes R, Bilotta F. Artificial Intelligence in Resuscitation: A Scoping Review. *J Clin Med*. 2023;12:2254. doi: [10.3390/jcm12062254](https://doi.org/10.3390/jcm12062254).
- Visan, A. I., & Negut, I. (2024). Integrating artificial intelligence for drug discovery in the context of revolutionizing drug delivery. *Life (Basel, Switzerland)*, 14(2), 233. <https://doi.org/10.3390/life14020233>



# MÓDULO 1. INTRODUCCIÓN A LA INTELIGENCIA ARTIFICIAL

## TEMA 3. MACHINE LEARNING, DEEP LEARNING Y REDES NEURONALES

*«La estupidez real siempre vence  
a la inteligencia artificial».*

*TERRY PRATCHETT*

## TEMA 3. MACHINE LEARNING, DEEP LEARNING Y REDES NEURONALES

### 3.1 MACHINE LEARNING

#### 3.1.1 ¿QUÉ ES EL MACHINE LEARNING?

El machine learning (ML), o aprendizaje automático, es una disciplina de la inteligencia artificial que se enfoca en el desarrollo de **algoritmos** y modelos que permiten a las computadoras aprender y mejorar a partir de datos, sin necesidad de ser programadas explícitamente para cada tarea. Dentro del machine learning, existe un área más especializada llamada deep learning. (**Imagen 6**)

#### INTELIGENCIA ARTIFICIAL

Se refiere a la combinación de algoritmos diseñados con el objetivo de crear máquinas que posean capacidades similares a las humanas.

#### MACHINE LEARNING

Es una rama de la inteligencia artificial que se centra en cómo dotar a las máquinas de la capacidad de aprender.

#### DEEP LEARNING

Se trata de un algoritmo automático jerárquico que imita el aprendizaje humano con el objetivo de adquirir ciertos conocimientos.

**Imagen 6. Relación entre inteligencia artificial, machine learning y deep learning. Autora: Elena Plaza Moreno**

En términos simples, el Machine Learning consiste en **capacitar a las computadoras para que puedan aprender de los datos que les proporcionamos y utilizar esa información para realizar tareas específicas. A través de algoritmos y modelos, las máquinas pueden identificar patrones en los datos y aplicar ese conocimiento para tomar decisiones o predecir resultados.** Este enfoque permite que las computadoras sean más inteligentes y capaces de realizar tareas complejas de manera autónoma.

La importancia del Machine Learning en el campo de la inteligencia artificial radica en su capacidad para mejorar la precisión y eficiencia de los sistemas computacionales. Al permitir a las computadoras aprender de los datos y adaptarse a nuevas situaciones, el Machine Learning puede ayudar a optimizar procesos, identificar tendencias, predecir resultados y tomar decisiones informadas en tiempo real.

### 3.1.2 FUNCIONAMIENTO DEL MACHINE LEARNING

Para comprender cómo funcionan los algoritmos y modelos de ML para aprender de los datos y mejorar el rendimiento de las computadoras, es fundamental comprender el proceso subyacente detrás de esta tecnología revolucionaria.

El proceso de aprendizaje automático consta de varias etapas clave (**Imagen 7**). **En primer lugar, se necesitan datos de entrada para que el algoritmo pueda analizar y extraer patrones significativos.** Estos datos pueden ser numéricos, texto, imágenes, etc. Una vez que se tienen los datos, se procede a la etapa de preprocesamiento, donde se limpian y organizan para facilitar su análisis.

**A continuación, se elige un modelo de aprendizaje automático adecuado para el problema en cuestión.** Hay diferentes tipos de modelos, como regresión lineal, árboles de decisiones, redes neuronales, entre otros, cada uno con sus propias ventajas y desventajas. El modelo se entrena utilizando un conjunto de datos etiquetado, es decir, datos para los cuales se conoce la respuesta correcta. Durante el entrenamiento, el modelo ajusta sus parámetros para minimizar una función de pérdida, que mide la discrepancia entre la salida predicha y la salida real.

Una vez que el modelo ha sido entrenado, **se evalúa su desempeño utilizando un conjunto de datos de prueba que no se utilizó durante el entrenamiento.** Esto permite verificar si el modelo es capaz de generalizar a datos nuevos y no vistos previamente. Si el rendimiento del modelo es satisfactorio, se puede implementar en un entorno de producción para realizar tareas específicas, como clasificación, regresión, detección de anomalías, entre otros.

Es importante destacar que el proceso de aprendizaje automático es iterativo y continuo. A medida que se recopilan más datos y se obtienen más retroalimentación, el modelo puede actualizarse y mejorar su rendimiento. Este ciclo de retroalimentación constante es lo que permite a las máquinas aprender de los datos y adaptarse a cambios en su entorno.



**Imagen 7. Pasos en el proceso de entrenamiento con machine learning (aprendizaje automático). Autora: Elena Plaza Moreno. Elaboración con Napkin IA.**

Para entenderlo mejor, voy a poner un ejemplo más gráfico. Introducimos un montón de fotos de *muffins* con pepitas de chocolate y chihuahuas. O un motón de fotos de *komondors* y de fregonas (Imagen 8).



**Imagen 8. Entrenamiento de modelo de machine learning con fotografías de chihuahuas y muffins y komondors y fregonas.**  
Fuente: Shenkman C. (2021). <https://www.semanticscholar.org/paper/640d18b94266146a6c479eccda92f725404af388>

La inteligencia artificial deberá clasificarlas y decirnos cuál es cuál. Las imágenes se transformarán en valores numéricos, y mediante algoritmos de árboles de decisión, de regresión lineal, u otros, podrá predecir los resultados. Hay varios algoritmos y unos aportarán un porcentaje de acierto mayor o menor. Los investigadores pueden aplicar varios algoritmos a sus datos para ver cuál es el que mayor nivel de acierto tiene.

En salud, la inclusión de algoritmos de machine learning está generando análisis de datos de forma más rápida. Y genera resultados que, en ocasiones, no habían sido descubiertos por los seres humanos.

Por ejemplo, en el ámbito de la reanimación cardiopulmonar, el análisis de las paradas cardiorrespiratorias con algoritmos de machine learning en vez de los sistemas de investigación normales, mejora los resultados o puede dar resultados nuevos que no habían sido descubiertos. Por ejemplo, días más frecuentes de parada cardiorrespiratoria, relación de las paradas cardiorrespiratorias con la temperatura, etc. Otra aplicación importante en salud es la inclusión de estos algoritmos en sistemas de alarma o monitorización de pacientes. En este caso se pueden predecir eventos de salud hasta seis horas antes de que ocurran. Y ya se han descrito algoritmos que detectan las respiraciones agónicas en el domicilio o utilizan la detección de la voz del alertante al número de emergencias para saber si se encuentra ante un paciente en parada cardiorrespiratoria. Esto es posible gracias al entrenamiento con múltiples horas de grabaciones de respiraciones agónicas o de tonos de voz. La IA podrá clasificarlas como pasa con los chihuahuas y las magdalenas.

### 3.1.3 TIPOS DE MACHINE LEARNING

El machine learning se puede clasificar en diferentes tipos según el enfoque y la forma en que se realiza el aprendizaje. Los tres tipos principales son el aprendizaje supervisado, el aprendizaje no supervisado y el aprendizaje por refuerzo.

---

### 3.1.3.1 APRENDIZAJE SUPERVISADO

El **aprendizaje supervisado** es un enfoque en el que el algoritmo se entrena utilizando un conjunto de datos etiquetado, es decir, se proporciona al sistema de aprendizaje ejemplos de entrada y la salida deseada correspondiente. La tarea del algoritmo es aprender a mapear las entradas a las salidas correctas, de modo que una vez entrenado, pueda hacer predicciones o tomar decisiones basadas en nuevos datos de entrada. Este tipo de Machine Learning se utiliza comúnmente en problemas de clasificación y regresión, como predecir si un correo electrónico es spam o no, o predecir el precio de una casa en función de sus características.

En el ejemplo de los chihuahuas y los muffins y el *komondor* y las fregonas, en este tipo de aprendizaje, introduciríamos las fotos poniéndole el nombre a la foto. Esto se llama etiquetar los datos. Y le diríamos a la IA esto es una magdalena y esto es un chihuahua. De esta forma entrenaremos al algoritmo con datos reales y correctos.

**Una de las claves es etiquetar bien los datos de salud. Si introducimos mal los datos, los resultados serán incorrectos.**

---

### 3.1.3.2 APRENDIZAJE NO SUPERVISADO

El **aprendizaje no supervisado** implica entrenar el modelo utilizando un conjunto de datos sin etiquetar. En este caso, el algoritmo tiene que descubrir patrones o estructuras ocultas en los datos por sí mismo. Esta forma de aprendizaje se utiliza para tareas como *clustering* (agrupamiento), donde se busca encontrar grupos de datos similares, o para reducción de dimensionalidad, donde se intenta encontrar una representación más compacta de los datos manteniendo su información relevante.

En el ejemplo de los perros y las magdalenas, introducimos las fotos sin decirle a la IA qué es cada cosa. Y será capaz de separar las fotos en dos grupos o clústeres. Por un lado, magdalenas y por otro chihuahuas.

Si pasamos esto a salud e introducimos datos de pacientes, puede que incluso el algoritmo de IA sea capaz de hacer agrupaciones que los humanos no hemos sido capaces de descubrir por métodos de investigación normales.

---

### 3.1.3.3 APRENDIZAJE POR REFUERZO

Finalmente, el **aprendizaje por refuerzo** se basa en el concepto de que un agente (un programa de computadora) aprende a través de la interacción con un entorno, recibiendo recompensas o castigos en función de las acciones que realiza. El objetivo es que el agente aprenda a tomar decisiones que maximicen una recompensa a largo plazo. Este tipo de aprendizaje se utiliza en aplicaciones como juegos (por ejemplo, entrenar un programa para jugar al ajedrez o al Go) o en robótica, donde un robot debe aprender a navegar en un entorno desconocido.

Los diferentes tipos de Machine Learning ofrecen enfoques distintos pero complementarios para abordar distintos problemas y tareas (Imagen 9). El aprendizaje supervisado se centra en la predicción de resultados basándose en ejemplos etiquetados, el aprendizaje no supervisado busca descubrir patrones ocultos en los datos sin etiquetar, y el aprendizaje por refuerzo se basa en la interacción con el entorno para maximizar una recompensa.

Cada tipo de aprendizaje tiene sus propias características y aplicaciones específicas, y la elección del enfoque adecuado depende del problema que se esté tratando de resolver.

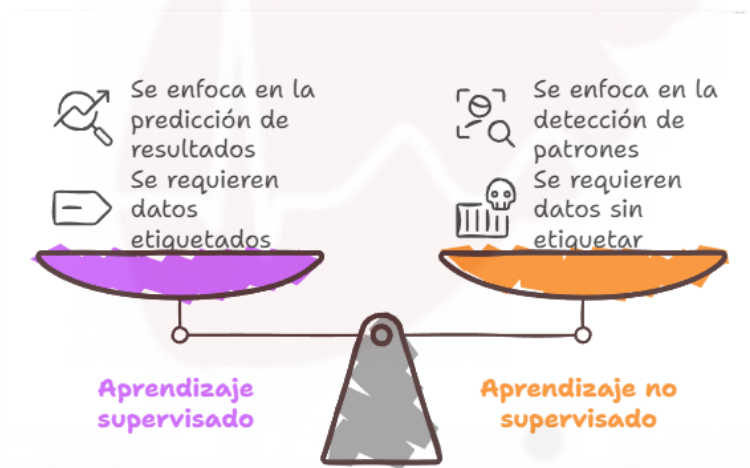


Imagen 9. Principales modelos de entrenamiento de machine learning en salud. Autora: Elena Plaza Moreno con Napkin IA.

## 3.2 DEEP LEARNING

### 3.2.1 ¿QUÉ ES EL DEEP LEARNING?

El deep learning (DL) es una técnica de aprendizaje automático que se basa en redes neuronales artificiales profundas. Estas redes están compuestas por múltiples capas que permiten a las computadoras procesar datos de una manera inspirada en el cerebro humano.

La importancia del DL en el campo de la inteligencia artificial radica en su capacidad para capacitar a las máquinas para realizar tareas complejas y tomar decisiones por sí mismas. Esto se logra a través del proceso de entrenamiento de la red neuronal profunda, donde se alimenta a la red con grandes cantidades de datos para que pueda aprender patrones y realizar predicciones con precisión.

Se trata de una subrama del machine learning, como se puede observar en la imagen. No se encuentra el origen de la referencia. La gran diferencia es que en este caso no se usan estos algoritmos (regresión lineal, árbol de decisión, clusterización, etc.) si no unas estructuras llamadas redes neurales.

### 3.2.2 REDES NEURONALES

Las **redes neuronales artificiales** son un componente fundamental en el campo del deep learning, ya que son la base sobre la cual se construyen los modelos de **aprendizaje profundo**.

**Estas redes están inspiradas en el funcionamiento del cerebro humano, específicamente en la forma en que las neuronas se comunican entre sí para procesar información y tomar decisiones.**

En términos simples, una red neuronal artificial es un modelo computacional que está diseñado para imitar el funcionamiento de las neuronas en el cerebro humano. Está compuesta por una serie de capas interconectadas, cada una de las cuales contiene un conjunto de nodos o **neuronas artificiales**. Cada una de estas neuronas está conectada a las neuronas de la capa siguiente a través de conexiones o pesos ajustables que se utilizan para transmitir señales.

El papel fundamental de las redes neuronales artificiales en el deep learning radica en su **capacidad para procesar grandes cantidades de datos de una manera altamente paralela y distribuida**. Al igual que el cerebro humano, estas redes son capaces de aprender patrones complejos presentes en los datos a través de un proceso de entrenamiento. Durante este proceso, los pesos de las conexiones entre las neuronas se ajustan de manera iterativa mediante algoritmos de optimización para minimizar el error entre las predicciones del modelo y las salidas deseadas. Es decir, dentro de cada neurona se realizan unas operaciones matemáticas que se ajustan solas (o por intervención del humano) para obtener resultados.

Una de las características más importantes de las redes neuronales artificiales es su capacidad para extraer automáticamente características relevantes de los datos, lo que las hace extremadamente poderosas para tareas de reconocimiento de patrones y toma de decisiones.

A medida que los datos fluyen a través de la red, se van transformando en representaciones cada vez más abstractas y sofisticadas, permitiendo al modelo aprender y adaptarse a diferentes tipos de datos de entrada.

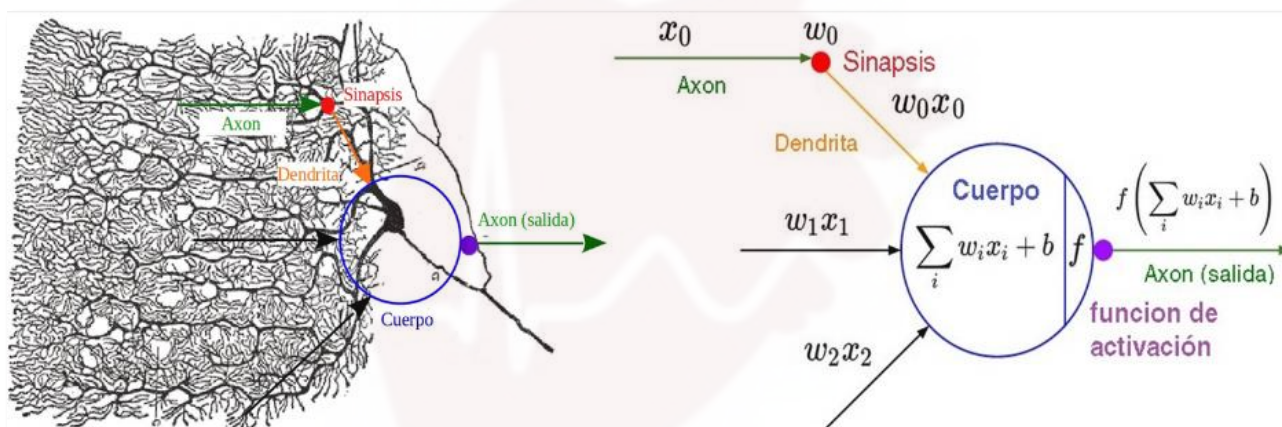
Vamos a ver cómo funcionan un poco más en profundidad y con un ejemplo simple para poder entenderlo.

### 3.2.2.1 FUNCIONAMIENTO DE LAS REDES NEURONALES

En cuanto a cómo imitan el funcionamiento del cerebro humano, las redes neuronales artificiales se asemejan a la forma en que las neuronas biológicas están interconectadas en el cerebro para procesar información.

Brevemente, la neurona se compone de entradas (las dendritas), un núcleo de procesamiento (el núcleo) y salidas (el axón). De esta manera entra una señal, se procesa y sale de forma diferente.

La formulación matemática de una neurona (la misma que tienes en la cabeza) es lo más sencilla posible, y haciendo un paralelo una representación en la siguiente imagen. (**Imagen 10**).



**Imagen 10. Diagrama que muestra como el funcionamiento y la estructura de las redes neuronales biológicas (a la izquierda) inspiraron el funcionamiento netamente matemático del primer modelo de neurona artificial, el perceptrón (a la derecha).**

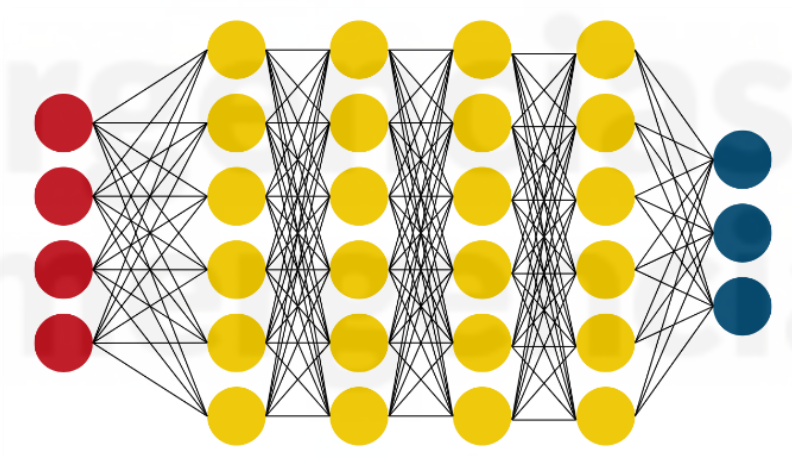
**Fuente: Arias V et al. (2019). <https://www.redalyc.org/journal/1702/170262877013/html/>**

En nuestro cerebro, una neurona se activa y activa a la siguiente. Y las neuronas artificiales funcionan igual, pero presentadas por capas: cada neurona artificial en una red neuronal tiene una función de activación, que determina cómo responde a las señales de entrada. Cuando una neurona recibe suficiente estimulación de las neuronas conectadas a ella, se activa y envía una señal a las neuronas en la capa siguiente.

**Las redes neuronales profundas están estructuradas en múltiples capas de neuronas simuladas. Estas capas trabajan en conjunto para procesar datos, facilitando tareas** como el reconocimiento de voz y la identificación de objetos. **Los datos viajan a través de estas capas de manera secuencial: la salida de cada capa sirve como entrada para la siguiente.**

La primera capa de la red se conoce como la **capa de entrada**, donde los datos son recibidos inicialmente. La última es la **capa de salida**, que presenta el resultado del procesamiento. Las capas situadas entre la de entrada y la de salida se conocen como **capas ocultas**.

Veamos una estructura de red neuronal en la Imagen 11.



**Imagen 11. Esquema de capas de una red neural. Cada circunferencia representa una neurona. En rojo las neuronas que forman la capa de entrada. En amarillo, cuatro capas ocultas. En azul, las neuronas que forman la capa de salida. Autora: Elena Plaza Moreno.**

La cantidad total de capas en la red se refiere a la "profundidad" del modelo, de ahí proviene el término **aprendizaje profundo**. Cuando el modelo tiene más de tres capas ocultas, ya se denomina aprendizaje profundo.

En estas redes neuronales, cada conexión tiene un **peso** asociado, que determina la importancia de esa conexión al multiplicarse por el valor de entrada. Se trata de una función matemática, como se explicó con anterioridad.

Esta función decide si la suma de los valores recibidos (ya multiplicados por el peso de la conexión) supera un cierto umbral, lo que activa la neurona y envía un valor a la siguiente capa. Existen varias **funciones de activación** comúnmente utilizadas en estas redes.

Este algoritmo es fundamental para ayudar a la red a decidir qué información pasar a la siguiente capa y qué información descartar. Este proceso de activación permite a la red aprender y tomar decisiones complejas a partir de los datos que procesa.

Al final, después de que todas las capas hayan realizado sus cálculos, se obtiene una predicción en la capa final, que es el resultado.

Para poner un ejemplo real, mi noche será perfecta si hay pizza y series. Estos valores de entrada son transformados en números. Pizza = 1, series = 1.

Si hay pizza y serie el resultado es 2. El resto de las combinaciones serían: solo pizza = 1; solo serie = 1; ni pizza ni serie = 0. Para que la noche sea perfecta para mí el corte lo pongo en 2, es decir, obligatoriamente tiene que haber series y pizza.

Pero no es tan fácil. Los números tienen decimales y las cosas no son blanco y negro. Y aquí es donde entran los pesos, que determina si una neurona de la red neuronal, mediante una función interna, activa a la siguiente o no la activa.

En nuestro ejemplo de la pizza, puedo dar valores distintos de preferencias a varios tipos de serie y de pizza. Por ejemplo, para alcanzar el 1, la pizza tiene que ser pepperoni. Para alcanzar el 1 en series, la serie tiene que ser de "true crime".

Tipo de serie	Puntuación	Tipo de pizza	Puntuación
True crime	1	Pizza pepperoni	1
Serie humor	0.5	Pizza carbonara	0.5
Serie romántica	0	Pizza barbacoa	0

Aquí empieza a haber muchas combinaciones:

- Serie true crime (1) + pizza pepperoni (1): Total = 2 (noche perfecta).
- Serie true crime (1) + pizza carbonara (0.5): Total = 1.5 (noche aceptable).
- Serie true crime (1) + pizza barbacoa (0): Total = 1 (noche regular).
- Serie humor (0.5) + pizza pepperoni (1): Total = 1.5 (noche aceptable).
- Serie humor (0.5) + pizza carbonara (0.5): Total = 1 (noche regular).
- Serie humor (0.5) + pizza barbacoa (0): Total = 0.5 (noche insatisfactoria).
- Serie romántica (0) + pizza pepperoni (1): Total = 1 (noche regular).
- Serie romántica (0) + pizza carbonara (0.5): Total = 0.5 (noche insatisfactoria).

En una red neuronal, los valores de entrada (como "pizza" y "serie") se representan como neuronas de entrada. Estas entradas se multiplican por sus correspondientes pesos, que en este caso son las preferencias personales, como "True crime = 1" o "Pepperoni = 1". O incluso, tendríamos en cuenta qué importancia le da cada uno a las series y pizzas. Porque si lo más importante es la serie, y la pizza da un poco igual, los decimales siguen cambiando.

De esta forma cada capa de la red realiza cálculos para ajustar y transformar los valores según los pesos y que simulan las características que valoras más o menos. El resultado de cada capa es una combinación ponderada de las entradas y los pesos.

Finalmente, el valor en la capa de salida sería la suma de estas combinaciones. Si este valor supera un umbral definido (como el "2" en nuestro ejemplo), la red da un resultado positivo: la noche es perfecta. Si no, el resultado es negativo.

Cuando se entrena una red neuronal, se ajustan los pesos para que las combinaciones de entradas produzcan las salidas deseadas. En nuestro caso, si decidimos que "humor + pepperoni" también debe ser una noche perfecta, los pesos de humor y pepperoni se ajustarán para que su combinación alcance el umbral de 2.

Así, la red "aprende" cuáles combinaciones cumplen con tus expectativas al comparar los resultados predichos con los resultados esperados y ajustando los pesos iterativamente.

Realmente el funcionamiento es mucho más complejo que lo que se explica en este ejemplo. Pero nos sirve para resumir que una red neuronal coge unos datos de entrada, los cambia por números, lo multiplica por los respectivos pesos de cada neurona, y calcula una suma ponderada. La función de activación, que es una ecuación matemática (hay varias) determina si va a haber plan perfecto esa noche.

Al ajustar los pesos durante el entrenamiento, la red neuronal aprende a clasificar correctamente los comportamientos. Este proceso básico de toma de decisiones y ajuste es la base para redes neuronales más complejas utilizadas en aplicaciones avanzadas de inteligencia artificial.

**Es interesante saber que, las redes neuronales funcionan con datos estructurados (datos numéricos) y con datos desestructurados (imágenes, texto o audio).**

Si no nos dedicamos a investigar, a nivel de salud nos interesa saber que **todo esto se trata de algoritmos matemáticos**. Se introducen unos datos y salen otros que se han ido procesando a lo largo del modelo por varias capas. Estos modelos se pueden entrenar. Y también hay que saber que el modelo aprenderá solo.

Quizás en salud, lo primero que se está viendo de forma muy desarrollada es el **procesamiento de imágenes**. Como podemos ver en la imagen siguiente (Imagen 12), aunque no se trata de una imagen de salud, cada capa analiza un aspecto de la imagen. La primera capa identifica luz y tonalidad, la siguiente forma, la siguiente identifica luz, tonalidad y formas, luego se combinan con estructuras. Y en la última capa se genera la imagen final. Podríamos decir que, cuanto mayor sea el número de capas en la red neuronal, más abstractos e interesantes serán los conceptos que esta podrá detectar en las imágenes.

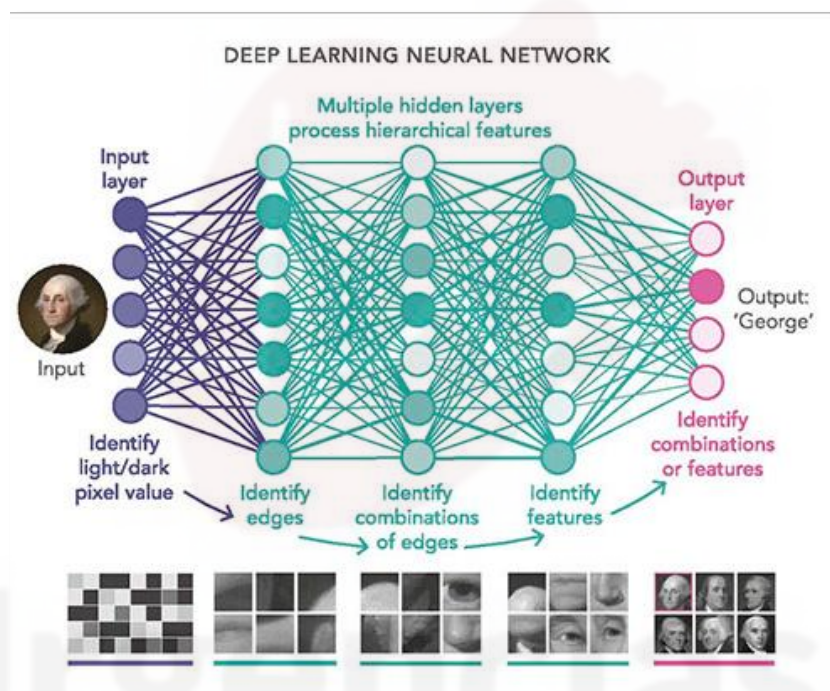
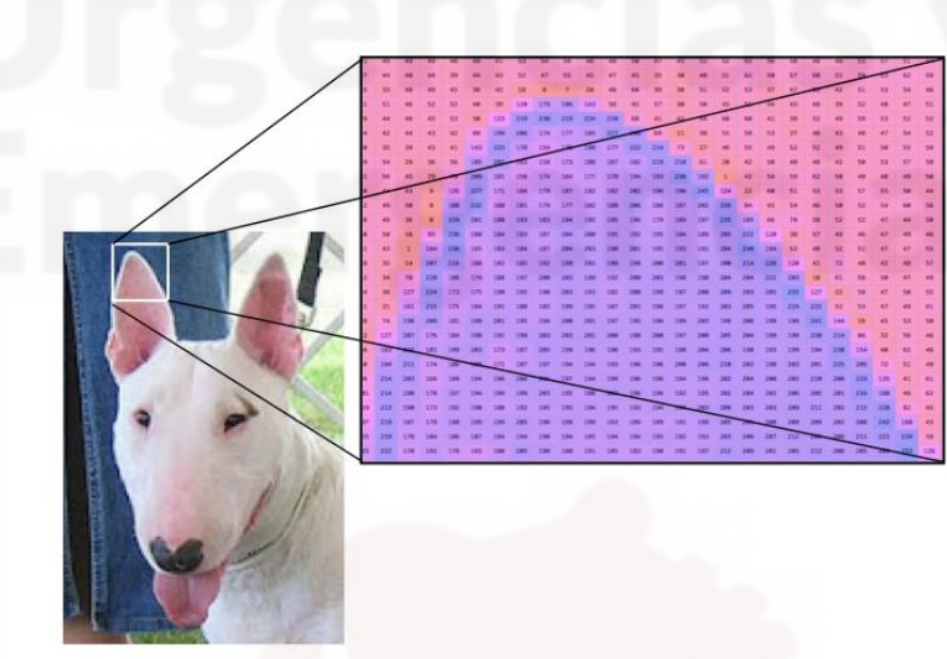


Imagen 12. Red neuronal analizando una imagen. Fuente: Imagen del artículo Waldrop, M. M. (2019). <https://doi.org/10.1073/pnas.1821594116>. Autora de la imagen: Lucy Reading-Ikkanda.

**En salud pasa igual. Se introduce una imagen, la red neuronal la analiza y es capaz de decirnos si ese paciente tiene una lesión o no. La detección de estos sistemas es similar o superior a la precisión del ojo humano (en este caso de las imágenes, radiólogos).**

Como pasaba con el ejemplo “del algoritmo de la noche perfecta”, en el que la IA cambia pizza y serie por números, en el procesamiento de las imágenes, ocurre lo mismo.

También hay que saber que las imágenes son cambiadas por matrices numéricas (Imagen 13). Tal y como indican Arias y colaboradores en su artículo, *“una computadora percibe las imágenes mediante un Chip CMOS, como sucedáneo de nuestra retina, traduce la intensidad de luz a señales eléctricas que posteriormente se procesan y cuantifican en unidades llamadas píxeles. Para la computadora simplemente es una matriz numérica de 301x301x3 píxeles, es decir algo más de medio millón de números que oscilan entre 0 (negro) y 255 (blanco)”*.



**Imagen 13. Una sección ampliada de una fotografía muestra como realmente una computadora percibe una imagen, en este caso la oreja de un perro descrita por una matriz de números. Fuente: Arias V et al. (2019). <https://www.redalyc.org/journal/1702/170262877013/html/>**

### 3.2.2.2 TIPOS DE REDES NEURONALES

Las redes neuronales son un componente fundamental en el campo de la inteligencia artificial, y existen diversos tipos que se utilizan en diferentes aplicaciones.

Cada tipo tiene una estructura distinta y una forma de procesar los datos. Esto se llama arquitectura. La arquitectura de una red neuronal se refiere a la forma. En la Imagen 11 se ha presentado un tipo de estructura, pero existen otras formas y otro tipo de relaciones entre neuronas.

Además, unos tipos de redes neuronales están más especializadas que otras en determinadas tareas. En la tabla se describen algunos de los tipos más comunes de redes neuronales, resaltando sus características:

<b>PRINCIPALES TIPOS DE REDES NEURONALES Y SU UTILIDAD</b>		
<b>NOMBRE EN ESPAÑOL</b>	<b>NOMBRE EN INGLÉS</b>	<b>UTILIDAD</b>
<b>Redes neuronales artificiales</b>	<b>Artificial Neural Networks (ANN)</b>	Tareas generales de clasificación y predicción, como reconocer patrones en datos.
<b>Redes neuronales recurrentes</b>	<b>Recurrent Neural Networks (RNN)</b>	Procesar datos secuenciales, como texto o series temporales, y realizar predicciones basadas en el contexto.
<b>Redes neuronales convolucionales</b>	<b>Convolutional Neural Networks (CNN)</b>	Procesar imágenes, detectar características y objetos en aplicaciones de visión por computadora. Las más utilizadas en salud en diagnóstico por imagen.
<b>Redes neuronales de autoencoders</b>	<b>Autoencoders</b>	Reducción de dimensionalidad y eliminación de ruido en los datos.
<b>Redes generativas antagónicas</b>	<b>Generative Adversarial Networks (GAN)</b>	Generar imágenes, videos y música realistas, y mejorar la calidad de imágenes. Las más utilizadas a la hora de crear imágenes con IA generativa.
<b>Redes neuronales de crecimiento y poda</b>	<b>Growing and Pruning Neural Networks (GPNN)</b>	Ajustar su estructura durante el entrenamiento para encontrar una arquitectura óptima.
<b>Redes neuronales espacios latentes</b>	<b>Latent Space Models</b>	Aprender representaciones compactas y abstractas de los datos originales.
<b>Redes neuronales de memoria extensible</b>	<b>Memory Augmented Neural Networks (MANN)</b>	Manejar y manipular datos complejos de manera eficiente y aprender tareas que requieren memoria detallada.
<b>Redes neuronales espaciotemporales</b>	<b>Spatiotemporal Neural Networks</b>	Manejar datos que varían en el espacio y el tiempo, ideales para visión y análisis de movimiento.
<b>Redes neuronales de campos de radiación neurales</b>	<b>Neural Radiance Fields (NeRFs)</b>	Generar representaciones tridimensionales detalladas de escenas a partir de imágenes 2D.
<b>Transformers</b>	<b>Transformers</b>	Manejar datos secuenciales y no secuenciales, destacando en tareas de procesamiento del lenguaje natural, traducción automática y generación de texto. Son las que utiliza ChatGPT y otros modelos de lenguaje grande.

**Tabla 2. Principales tipos de redes neuronales con su principal utilidad. Elaboración propia. Autora: Elena Plaza Moreno**

De todas ellas, nos vamos a detener en los **transformers**. Los transformers son un tipo de red neuronal que ha dado pie a que los grandes modelos de lenguaje que trabajan con texto, como ChatGPT, hayan mejorado exponencialmente los últimos años.

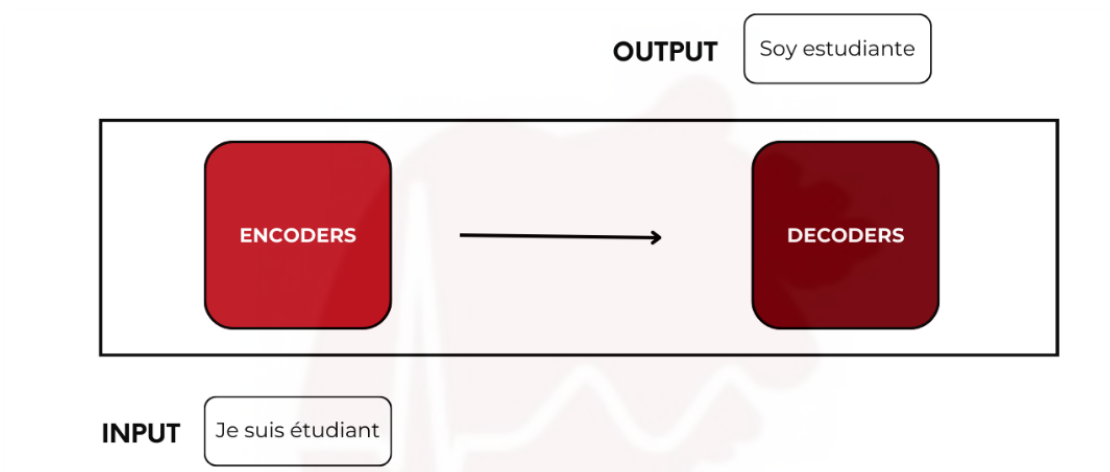
Fue introducida por un equipo de investigadores de Google Brain en un artículo titulado "*Attention is All You Need*", publicado en 2017. El equipo de autores fue encabezado por Ashish Vaswani y lo que intentaban era mejorar su traductor.

Este artículo revolucionó el campo del procesamiento del lenguaje natural al presentar una nueva arquitectura (estructura de red neuronal) que se basaba completamente en el mecanismo de atención y prescindía de las redes recurrentes (RNNs) y convolucionales (CNNs) que eran comunes en modelos anteriores.

La innovación principal de los transformers fue su capacidad para procesar secuencias de datos en paralelo y su **mecanismo de atención, que permite al modelo enfocarse en diferentes partes del input (los datos de entrada) con mayor flexibilidad y eficacia**.

Antes de los transformers, los modelos de IA solían procesar el lenguaje de manera secuencial, lo que limitaba su capacidad para entender contextos largos y complejos. Los transformers, en cambio, pueden procesar todas las palabras de una frase al mismo tiempo gracias a su mecanismo de atención. Esto permite que entiendan mejor el contexto y las relaciones entre palabras, haciendo el procesamiento mucho más eficiente y preciso.

Este mecanismo de atención ayuda al codificador (la parte que comprende la entrada) y al decodificador (la parte que genera la salida) a trabajar juntos de manera más efectiva (Imagen 14).



**Imagen 14. El codificador (encoder) y el decodificador (decoder) trabajan juntos para producir mejores resultados (output).**  
Autora: Elena Plaza Moreno

Los transformers han facilitado el aprendizaje por transferencia. Esto significa que un modelo entrenado en una gran cantidad de datos (como el texto de Internet) puede ser ajustado (fine-tuned) para tareas específicas con menos datos adicionales. Esta capacidad de transferir conocimientos adquiridos en un dominio amplio a tareas más específicas ha mejorado la eficiencia y eficacia en el desarrollo de aplicaciones de IA.

Los transformers han demostrado ser extremadamente versátiles. Pueden usarse para una amplia variedad de tareas en procesamiento del lenguaje natural (NLP), como:

- **Traducción automática:** convertir texto de un idioma a otro con alta precisión.
- **Generación de texto:** crear textos coherentes y contextualmente adecuados, como artículos, historias y respuestas a preguntas.
- **Resumen automático:** resumir documentos largos en pocas frases manteniendo el sentido del contenido original.
- **Análisis de sentimientos:** determinar las emociones o actitudes expresadas en un texto.

A diferencia de modelos anteriores, que dependían en gran medida del orden secuencial de las palabras, los transformers pueden entender las palabras en cualquier orden gracias a su capacidad de atención global. **Esto les permite captar mejor el contexto y las relaciones semánticas a largo plazo en los textos.**

Los transformers han dado lugar a la creación de grandes modelos pre-entrenados que se pueden utilizar y adaptar para diversas tareas con un mínimo de ajuste adicional. Estos modelos, como como BERT (Bidirectional Encoder Representations from Transformers) y GPT (Generative Pre-trained Transformer), han democratizado el acceso a tecnologías avanzadas de procesamiento del lenguaje, permitiendo a investigadores y desarrolladores utilizar modelos de vanguardia sin necesidad de entrenar desde cero.

### 3.2.2.3 APLICACIONES GENERALES DE LAS REDES NEURONALES

#### a) Reconocimiento de patrones

Ya sea en imágenes, voz, texto o cualquier otro tipo de datos. Las redes neuronales son capaces de identificar y clasificar patrones complejos en grandes conjuntos de datos, lo que las hace ideales para tareas como reconocimiento facial (Imagen 15), reconocimiento de voz, diagnóstico médico, detección de fraudes y mucho más.

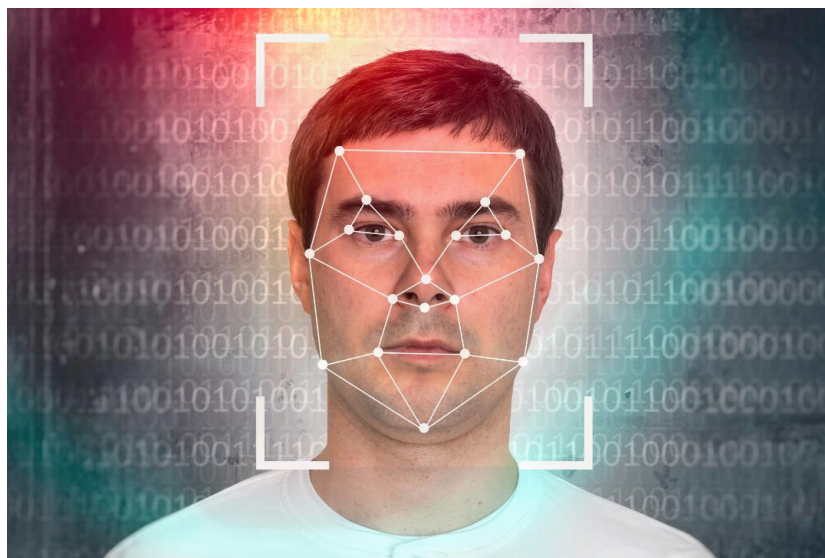


Imagen 15. Reconocimiento facial para verificación biométrica. Fuente: Banco de imágenes de Canva. Autor: Andriano\_cz.

#### b) Visión por computadora o computacional

Las redes neuronales también se utilizan ampliamente en el campo de la visión por computadora para analizar y procesar imágenes de forma automática. Estas redes son capaces de detectar objetos, segmentar imágenes, clasificar elementos, entre otras tareas, lo que las hace fundamentales en aplicaciones como vehículos

autónomos (Imagen 16), vigilancia de seguridad, salud, agricultura de precisión, entre otros. Es lo que hace que determinados smartphones, en la aplicación de cámara, etiqueten lo que se fotografía como “perro”, “comida”, “paisaje”, etc.



**Imagen 16. Fotografía del dataset de <https://www.cityscapes-dataset.com/examples/>. Se puede ver cómo con la IA se pueden diferenciar personas, aceras, árboles, coches y otros objetos.**

### **c) Procesamiento del lenguaje natural (NLP)**

Las redes neuronales son utilizadas en el procesamiento del lenguaje natural para tareas como traducción automática, análisis de sentimientos, generación de texto, entre otros. Estas redes son capaces de entender y generar texto de una manera más natural, lo que las hace ideales para aplicaciones como asistentes virtuales, chatbots, análisis de opiniones en redes sociales y atención al cliente automatizada.

### **d) Sistemas de recomendación**

Las redes neuronales se emplean en la creación de sistemas de recomendación que sugieren productos, servicios o contenido basado en las preferencias del usuario. Estas aplicaciones se encuentran en plataformas de comercio electrónico, servicios de *streaming* de música y video, y redes sociales, mejorando la experiencia del usuario al ofrecerle contenido personalizado.

### **e) Automatización industrial y robótica**

En el ámbito de la automatización industrial, las redes neuronales ayudan a optimizar procesos de fabricación, mantenimiento predictivo y control de calidad. En robótica, se utilizan para la navegación autónoma, manipulación de objetos e interacción humano-robot, permitiendo a los robots realizar tareas complejas con mayor precisión y adaptabilidad.

### **f) Finanzas y comercio**

Las redes neuronales se utilizan en finanzas para la predicción de tendencias del mercado, gestión de riesgos, detección de fraudes y trading algorítmico. Estas aplicaciones ayudan a las instituciones financieras a tomar decisiones más informadas y a mejorar la seguridad y eficiencia de sus operaciones.

### g) Salud y biomedicina

En el sector de la salud, las redes neuronales se aplican en el análisis de imágenes médicas, diagnóstico de enfermedades, descubrimiento de fármacos y personalización de tratamientos. Estas herramientas permiten a los profesionales de la salud proporcionar diagnósticos más precisos y desarrollar terapias más efectivas. Se ha desarrollado este apartado ampliamente en el tema 2.

### h) Agricultura inteligente

Las redes neuronales se utilizan en la agricultura de precisión para monitorear cultivos, predecir rendimientos y optimizar el uso de recursos como agua y fertilizantes. Estas tecnologías ayudan a los agricultores a aumentar la productividad y sostenibilidad de sus prácticas agrícolas.

### i) Transporte y logística

Las redes neuronales mejoran la eficiencia en la gestión de flotas, optimización de rutas, predicción de demanda y mantenimiento predictivo de vehículos. En el transporte público y la logística, estas aplicaciones contribuyen a reducir costos y mejorar el servicio al cliente.

### j) Energía

En el sector energético, las redes neuronales se utilizan para la predicción de demanda, optimización de redes eléctricas y gestión de energías renovables. Estas tecnologías ayudan a las empresas a mejorar la eficiencia energética y a integrar fuentes de energía limpia en sus operaciones.

## 3.3 BIBLIOGRAFÍA

- Arias, V., Salazar, J., Garicano, C., Contreras, J., Chacón, G., Chacín-González, M., Añez, R., Rojas, J., & Bermúdez-Pirela, V. (2019). *Una introducción a las aplicaciones de la inteligencia artificial en Medicina: Aspectos históricos*. Revista Latinoamericana de Hipertensión, 14(5), 590–600. <https://www.redalyc.org/journal/1702/170262877013/html/>
- Alonso F. *Redes Neuronales y Deep Learning. Capítulo 1: Preludio* [Internet]. Future Space S.A. 2021 [citado el 13 de mayo de 2024]. Disponible en: <https://www.futurespace.es/redes-neuronales-y-deep-learning-capitulo-1-preludio/>
- Bagnato JI. *Guía Rápida sobre Deep Learning* [Internet]. Aprendemachinelearning.com. Aprende Machine Learning; 2017 [citado el 13 de mayo de 2024]. Disponible en: <https://www.aprendemachinelearning.com/aprendizaje-profundo-una-guia-rapida/>
- Baum, D. (2024). *Generative AI and LLMs for dummies*. John Wiley & Sons, Inc. <https://www.snowflake.com/wp-content/uploads/2024/01/Generative-AI-and-LLMs-for-Dummies.pdf>
- Code, S. (2023, mayo 18). *La evolución de la Inteligencia Artificial: ¿Cómo llegamos a la tecnología de hoy?* LinkedIn.com. <https://www.linkedin.com/pulse/la-evoluci%C3%B3n-de-inteligencia-artificial-c%C3%B3mo-llegamos-sergie-code-/>

de Oliveira Chagas ET. *Deep learning y sus aplicaciones hoy*. Revista Científica Multidisciplinar Núcleo do Conhecimento. año 04, Ed. 05, Vol. 04, PP. 05-26 mayo 2019. ISSN: 2448-0959. Disponible en: <https://www.nucleodoconhecimento.com.br/administracion-de-empresas/deep-learning-3>

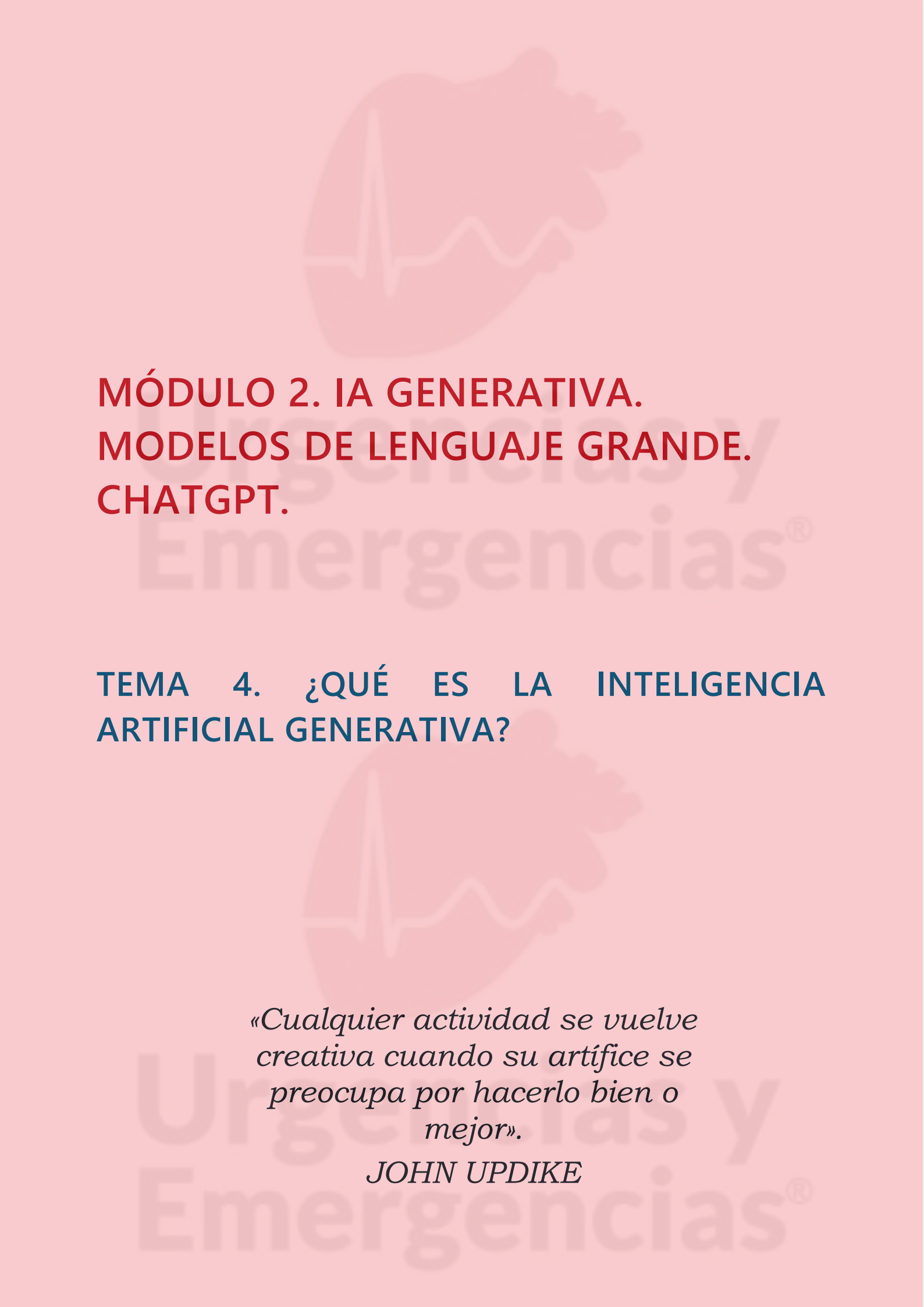
Estevez M. *Introducción a Redes Neuronales* [Internet]. Inteligencia Analítica. 2017 [citado el 16 de mayo de 2024]. Disponible en: <https://inteligencia-analitica.com/introduccion-a-redes-neuronales/>

Gutiérrez JP. *Explorando las Redes Neuronales: Una Guía de Arquitecturas y Aplicaciones* [Internet]. LinkedIn.com. 2023 [citado el 13 de mayo de 2024]. Disponible en: <https://www.linkedin.com/pulse/explorando-las-redes-neuronales-una-gu%C3%ADa-de-y-jordi-pompas-guti%C3%A9rrez-eople/>

OpenAI. (2024). *ChatGPT 4* (versión del 15 de mayo) [Modelo de lenguaje de gran tamaño]. <https://chat.openai.com/chat>

Santana M. *Deep Learning: Do conceito às aplicações - data hackers - medium* [Internet]. Data Hackers. 2018 [citado el 13 de mayo de 2024]. Disponible en: <https://medium.com/data-hackers/deep-learning-do-conceito-%C3%A0s-aplica%C3%A7%C3%B5es-e8e91a7c7eaf>

Shenkman C, Thakur D, Llansó E. *Do you see what I see? Capabilities and limits of automated multimedia content analysis*. arXiv.org [Internet]. 2021 [citado el 14 de junio de 2024]; Disponible en: <https://www.semanticscholar.org/paper/640d18b94266146a6c479eccda92f725404af388>



## MÓDULO 2. IA GENERATIVA. MODELOS DE LENGUAJE GRANDE. CHATGPT.

### TEMA 4. ¿QUÉ ES LA INTELIGENCIA ARTIFICIAL GENERATIVA?

*«Cualquier actividad se vuelve creativa cuando su artífice se preocupa por hacerlo bien o mejor».*

*JOHN UPDIKE*

## TEMA 4. ¿QUÉ ES LA INTELIGENCIA ARTIFICIAL GENERATIVA?

### 4.1 INTRODUCCIÓN

La inteligencia artificial generativa es un campo que ha revolucionado la manera en que creamos contenido innovador en diversos ámbitos, como la escritura, la música, el arte visual y mucho más.

La inteligencia artificial generativa se basa en el concepto de aprendizaje profundo (deep learning), que es una rama del machine learning que se enfoca en el desarrollo de algoritmos capaces de aprender y mejorar a partir de grandes cantidades de datos. Es decir, siguiendo el desarrollo del curso, vamos de “más a menos” explorando áreas que están unas dentro de otras (Imagen 17).

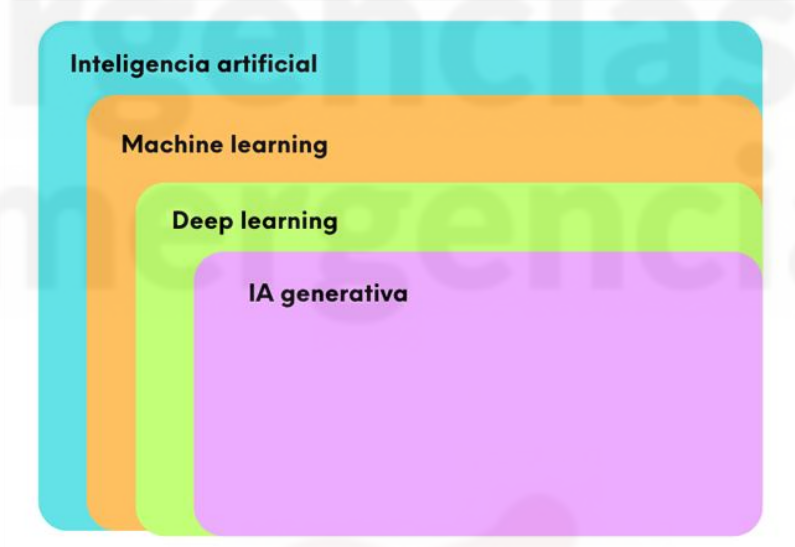


Imagen 17. La IA generativa es un tipo de IA que se basa en el deep learning (redes neuronales) para crear contenido nuevo.  
Autora: Elena Plaza Moreno

**En el caso de la inteligencia artificial generativa, estos algoritmos y redes neuronales avanzadas son entrenados para generar contenido nuevo y original, imitando la creatividad humana en la producción de textos, imágenes, música y más.**

### 4.2 ¿CÓMO FUNCIONA LA IA GENERATIVA?

La inteligencia artificial generativa utiliza el aprendizaje profundo para crear contenido único, original y novedoso a partir de modelos previamente entrenados. Estos modelos son alimentados con grandes volúmenes de datos para que puedan aprender patrones y tendencias, y así ser capaces de generar contenido que sea coherente y relevante, pero la clave es que es contenido nuevo.

El proceso de cómo funciona la inteligencia artificial generativa comienza con la **recopilación de datos**. Se utilizan conjuntos de datos que contienen ejemplos del tipo de contenido que se desea generar, como textos, imágenes o música. Estos conjuntos de datos son fundamentales para el entrenamiento de los modelos de inteligencia artificial.

Una vez que se cuenta con los datos adecuados, **se procede al entrenamiento de la red neuronal**. Durante esta etapa, la red neuronal aprende a reconocer patrones y estructuras en los datos de entrada, lo que le permite generar contenido nuevo a partir de estos patrones identificados.

**Una vez que la red neuronal ha sido entrenada, se puede proceder a la generación de contenido**. En este punto, la red neuronal toma como entrada los datos de los modelos previamente entrenados y los utiliza para crear contenido nuevo y único. Por ejemplo, en el caso de la generación de texto, la red neuronal puede producir párrafos, historias o incluso diálogos a partir de la información aprendida durante el entrenamiento.

Por tanto, la inteligencia artificial generativa utiliza el aprendizaje profundo para aprender patrones y estructuras a partir de grandes conjuntos de datos, y luego genera contenido nuevo y único a partir de estos modelos previamente entrenados abriendo las puertas a un sinfín de posibilidades creativas en todos los campos.

Pero en salud, **este contenido nuevo puede jugaros malas pasadas y no ser del todo correcto**. Sobre esto se habla de forma muy extensa en temas posteriores.

### 4.3 APLICACIONES DE LA IA GENERATIVA

A continuación, se destacan algunas de las diversas aplicaciones de la inteligencia artificial generativa en la creación de textos, imágenes, música y otros tipos de contenido creativo:

#### **a) Creación de textos.**

Uno de los usos más comunes de la inteligencia artificial generativa es la generación de texto. Los algoritmos y redes neuronales avanzadas pueden crear contenido escrito completamente nuevo e incluso imitar el estilo y la voz de un autor específico. Esta capacidad se ha utilizado en la creación de historias, poemas, artículos periodísticos e incluso guiones de películas. Además, la inteligencia artificial generativa también se ha aplicado en la redacción automática de informes, resúmenes de noticias y generación de contenido para redes sociales.

#### **b) Creación de imágenes.**

La inteligencia artificial generativa también ha revolucionado la creación de imágenes, permitiendo la generación de obras de arte digitales únicas y realistas. Los algoritmos pueden crear imágenes desde cero o combinar y transformar imágenes existentes para producir resultados sorprendentes. Esta tecnología se ha utilizado en la generación de paisajes, retratos, diseños gráficos e incluso en la creación de personajes para videojuegos y películas de animación. La capacidad de la inteligencia artificial generativa para producir resultados visualmente impactantes ha despertado el interés de artistas, diseñadores y creativos de todo el mundo.

#### **c) Creación de música.**

La inteligencia artificial generativa también ha incursionado en el mundo de la música, permitiendo la creación de piezas musicales originales y únicas. Los algoritmos pueden componer melodías, letras, armonías y ritmos, e incluso imitar el estilo de compositores famosos. Esta tecnología se ha utilizado en la creación de música para películas, videojuegos, anuncios publicitarios e incluso en la producción de álbumes musicales completos. La capacidad de la inteligencia artificial generativa para generar música ha abierto nuevas posibilidades creativas en la industria musical.

#### d) Otro tipo de contenido creativo.

La inteligencia artificial generativa también se ha aplicado en la creación de otros tipos de contenido creativo, como diseños de moda, recetas de cocina, desafíos de resolución de problemas y experimentos artísticos interactivos. Los algoritmos pueden generar ideas innovadoras, soluciones creativas y propuestas originales en una amplia variedad de campos, impulsando la creatividad y la innovación en diferentes industrias. La capacidad de la inteligencia artificial generativa para producir contenido creativo ha abierto nuevas oportunidades para la experimentación y la colaboración en el mundo digital.

### 4.4 DESAFÍOS TÉCNICOS DE LA IA GENERATIVA

Respecto a la creación de imágenes, en cuestión de dos años se ha avanzado bastante. Y cuánto más se usen estas herramientas (de texto, música, vídeo, etc.) más se desarrollarán.

Pero sigue habiendo limitaciones muy importantes, y en ocasiones poco evidentes. Y no solo en la generación de imágenes, si no en todos los tipos de IA generativa (Imagen 18):

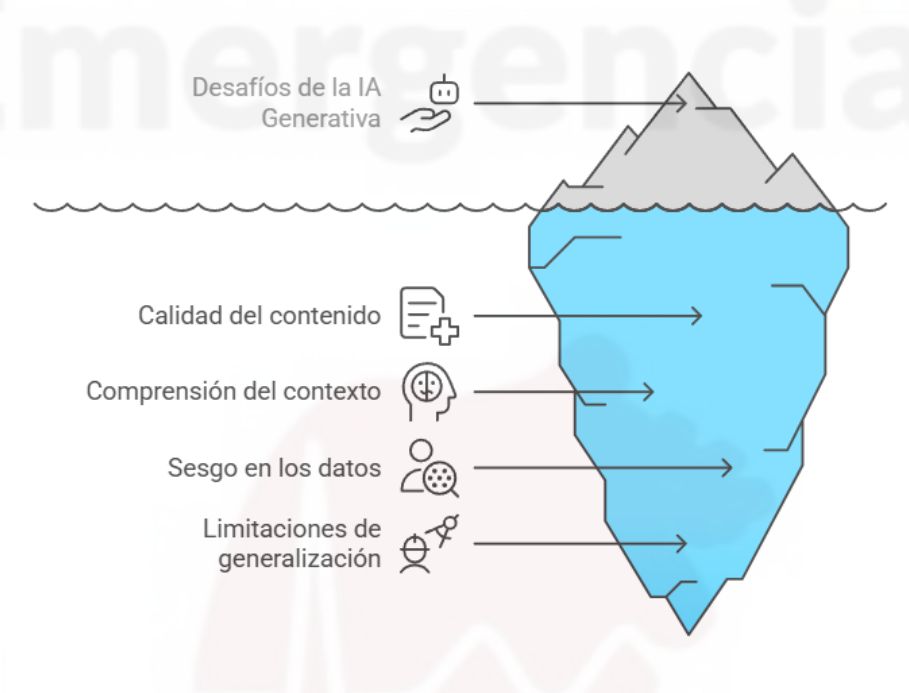


Imagen 18. Desafíos de la IA generativa. Elaboración: Elena Plaza Moreno con Napkin IA

- **Calidad y coherencia:** aunque las IA generativas pueden crear contenido que parece realista, mantener la coherencia y la calidad a lo largo de un documento o serie de contenidos es complejo. Por ejemplo, en la generación de texto, la IA puede perder el hilo del tema o producir información contradictoria si no está adecuadamente entrenada o supervisada.
- **Intención y contexto:** las IA pueden tener dificultades para comprender completamente el contexto y la intención detrás de una solicitud. Esto puede llevar a la generación de contenido inapropiado o fuera de contexto, que no cumple con las expectativas del usuario o que incluso puede ser dañino.

- **Sesgo en los datos:** las redes neuronales generativas dependen de grandes volúmenes de datos de entrenamiento. Si estos datos contienen sesgos, la IA puede perpetuar o amplificar estos sesgos en el contenido que genera. Esto es especialmente preocupante en aplicaciones sensibles como la generación de contenido educativo o informativo.
- **Capacidad de generalización:** si bien las IA generativas pueden ser muy buenas en tareas específicas para las que han sido entrenadas, su capacidad para generalizar a nuevas tareas o dominios puede ser limitada. Esto requiere un entrenamiento adicional y una adaptación continua para mantener la relevancia y efectividad del sistema.

A continuación, se muestran varias imágenes para ver el desarrollo y las limitaciones desde 2022, cuando estas herramientas se abrieron al público en general, hasta la actualidad (2025).

La Imagen 19 fue creada por inteligencia artificial generativa, en concreto con el programa DreamStudio a finales de 2022. Se puede comprobar como había problemas en la generación de dedos, brazos y manos.



**Imagen 19. Imagen generada por IA, con DreamStudio a finales de 2022 en la que se evidencia problemas en la creación de brazos y manos**

Vamos a ver otro ejemplo para ver la evolución de los modelos utilizando el mismo prompt. El prompt: *“Describe la escena de manera realista en la que un enfermero de ambulancia realiza una reanimación cardiopulmonar para intentar salvar la vida de un paciente en estado crítico. Presta especial atención a los detalles médicos y emocionales, transmitiendo la intensidad y la urgencia del momento”*. En la Imagen 20, se ve el resultado de finales de 2022 y en la Imagen 21 el resultado en abril de 2024.



**Imagen 20. Imagen generada por IA con DreamStudio en 2022.**



**Imagen 21. Imagen generada por IA con DreamStudio en 2024**

La calidad de las imágenes, además del prompt utilizado y del año en el que se hayan realizado, también depende del modelo o aplicación que se use. Por ejemplo, con Midjourney se pueden hacer auténticas obras de arte. No es fácil y requiere de conocimiento y habilidades para lograr los resultados que se ven en las imágenes 22, 23 y 24.

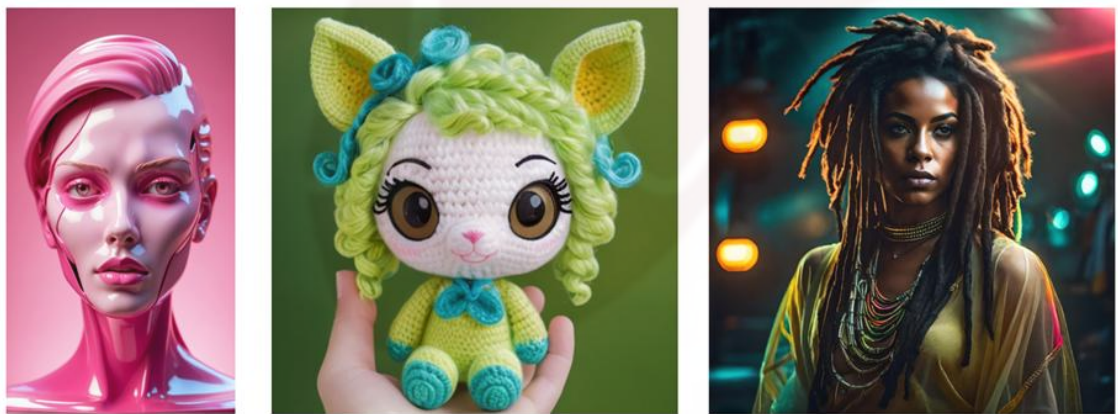


Imagen 22. Imágenes generadas con Midjourney en 2024



Imagen 23. Imagen generada por Midjourney en julio de 2024. Prompt: "Meatball shop logo banksy style"



Imagen 24. Imagen generada con Midjourney en julio de. Prompt: "3 Tennis fans enjoying popsicles on a sunny day, smiling at the camera, photo taken at a lower angle --v 6.0"

En esta última imagen (Imagen 24), se ve que en el prompt se incluye la instrucción de “fotografía tomada en un ángulo inferior”. Por eso, el prompt es tan importante y requiere práctica para conseguir los resultados que queremos conseguir.

A continuación, se muestran imágenes relacionadas con la salud generadas con DALL-E, de ChatGPT – OpenAI (Imagen 25) y Copilot de Microsoft ( Imagen 26) en mayo de 2024.

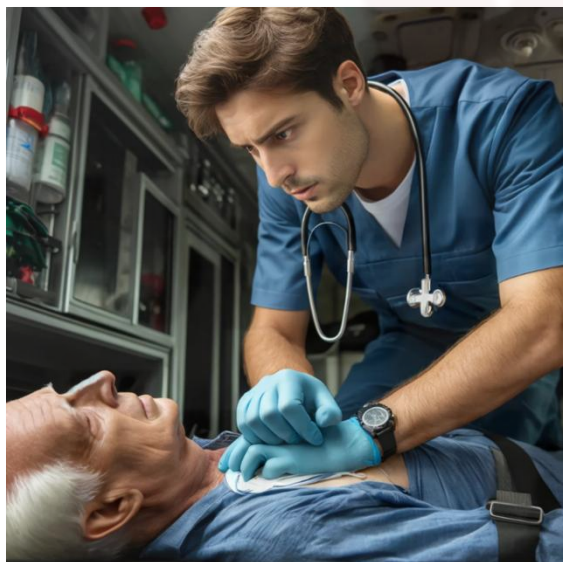


Imagen 25. Imagen generada con DALL-E en junio de 2024



Imagen 26. Imagen generada con Copilot en junio de 2024

A pesar de la colocación incorrecta de las manos en el tórax del paciente, y el fallo en el fonendoscopio y otros detalles, la mejora es evidente respecto a las imágenes anteriores.

Otro de los avances que se está evidenciando en la creación de imágenes es la mejora de los textos dentro de las imágenes. En 2022 los textos eran ilegibles y, actualmente, ya se consiguen ciertas palabras. Pero queda mucho por mejorar, como se puede ver en la Imagen 27:



Imagen 27. Cartel informativo de reanimación cardiopulmonar generado con DALL-E – ChatGPT 4o (julio 2024). Prompt: “Genera un cartel informativo en el que se lea la palabra reanimación cardiopulmonar y los cuatro pasos principales que hay que seguir: reconocer parada llamar al 112 iniciar compresiones torácicas usar el desfibrilador”

A pesar de las limitaciones actuales, uno de los grandes problemas de la IA generativa es la posibilidad de copiar vídeos, clonar la voz y generar imágenes irreales. Actualmente hay que tener mucho cuidado con las *fake news* y *deep fakes*, ya que con algunos modelos y técnicas avanzadas de prompting (cómo se le escribe al modelo) ya se pueden replicar personajes famosos e incluso fallecidos, con un alto realismo (Imagen 28).



Imagen 28. Imágenes generadas por @julian\_ai\_art con Midjourney y subidas a Instagram.

## 4.5 DESAFÍOS ÉTICOS DE LA IA GENERATIVA

Más adelante se dedica un tema completo a los desafíos éticos de estas tecnologías en el ámbito de la salud. Pero, a continuación, se enumeran desafíos éticos generales que está produciendo la IA generativa:

- **Propiedad y derechos de autor:** determinar la propiedad intelectual de obras generadas por IA es complicado. ¿deberían los creadores de la IA, los usuarios que solicitan el contenido o la propia máquina tener derechos sobre el contenido generado? Actualmente, las imágenes generadas con IA carecen de derechos de autor.
- **Originalidad y plagio:** la capacidad de una IA para crear contenido nuevo plantea cuestiones sobre la originalidad. Si una IA genera contenido basándose en grandes volúmenes de datos existentes, ¿es este contenido verdaderamente original o simplemente una amalgama de obras anteriores? De hecho, Open AI tiene varias denuncias por este asunto ya que ha entrenado a los modelos con datos de webs y blogs (Imagen 29).

Internacional

### OpenAI admite que no ha respetado los derechos de autor para entrenar a ChatGPT y considera que no es un delito

La compañía, que se enfrenta a varias demandas por ignorar el copyright, considera que sería "imposible" entrenar a la Inteligencia Artificial respetando a los autores de los textos. También asegura que la ley no prohíbe específicamente el entrenamiento de IA

Imagen 29. Captura de pantalla de noticia sobre los problemas de OpenAI (empresa creadora de ChatGPT) con los derechos de autor. Fuente: <https://cadenaser.com/nacional/2024/01/09/openai-admite-que-no-ha-respetado-los-derechos-de-autor-para-entrenar-a-chatgpt-y-considera-que-no-es-un-delito-cadena-ser/>

- **Uso responsable:** la posibilidad de generar grandes cantidades de contenido de manera automatizada puede ser utilizada tanto para bien como para mal. Es fundamental establecer regulaciones y directrices para asegurar que estas tecnologías se utilicen de manera ética y responsable. Un ejemplo de este punto es Amazon, que ha tenido que limitar la autopublicación de libros debido a la avalancha de libros escritos con IA.
- **Transparencia y explicabilidad:** los usuarios tienen derecho a saber cómo se generan los contenidos y cuáles son las limitaciones y sesgos de la tecnología. La transparencia en los procesos y algoritmos utilizados es fundamental para mantener la confianza y la responsabilidad en el uso de la IA generativa.

## 4.6 EL FUTURO DE LA IA GENERATIVA

Según la OECD (Organización para la Cooperación y el Desarrollo Económico) en su documento: "*Initial policy considerations for generative artificial intelligence*", se proyecta que los mercados de IA generativa continúen creciendo rápidamente en áreas clave.

La investigación de mercado y de inversiones complementa los desarrollos tecnológicos al proporcionar información sobre las posibles trayectorias de los sistemas de IA generativa a corto, mediano y largo plazo. Bancos de inversión, firmas de consultoría e investigadores informan que la IA generativa causará enormes impactos económicos en los próximos años. En el informe nos presentan los siguientes datos:

- Goldman Sachs estima que la IA generativa podría representar un aumento del 7 por ciento en el producto interno bruto (PIB) global en diez años.
- McKinsey & Company estima que la IA generativa podría agregar entre 2.6 y 4.4 billones de dólares estadounidenses por año en 63 casos de uso, lo que representaría un aumento en los efectos económicos totales de la IA del 15 al 50 por ciento.
- Polaris estima un crecimiento del mercado global de IA generativa a una tasa compuesta anual del 34.2 por ciento, desde 10.6 mil millones de dólares en 2022 hasta 200.7 mil millones de dólares para 2032.

En la actualidad, la IA generativa está en una etapa temprana de desarrollo, requiriendo grandes inversiones en I+D y una fuerza laboral cualificada pero escasa para llevarla al siguiente nivel de madurez.

Se espera que el crecimiento adicional provenga de la síntesis de audio, preprocesamiento de datos, compresión de imágenes, reducción de ruido de datos visuales, imágenes médicas y clasificación de imágenes, especialmente en el sector sanitario.

Las áreas de aplicación incluyen el diseño de chips y piezas, ciencias de materiales y entretenimiento. Gartner, una firma de investigación de mercado, enumera otros impulsores del crecimiento al aplicar la IA generativa al diseño de chips, diseño generativo de piezas utilizadas por industrias como la automotriz, aeroespacial y defensa, y a las ciencias de materiales. Gartner señala que los startups que construyen un negocio en torno a la IA generativa han recibido más de 1.7 mil millones de dólares en financiación durante los últimos tres años.

El sector de medios y entretenimiento (incluyendo publicidad) representa hasta ahora la mayor participación de ingresos de la IA generativa. Las empresas con una ventaja competitiva en IA generativa incluyen grandes compañías tecnológicas conocidas, proveedores de software empresarial, empresas de IA en sectores específicos (por ejemplo, automatización de contratos legales, creación de videos, generación de datos

sintéticos y las artes) y empresas que proporcionan computación de IA, como semiconductores e infraestructura de supercomputación, cruciales para aprovechar los entornos ricos en datos de la IA generativa.

En resumen, el futuro de la IA generativa es brillante (no carente de riesgos) y su impacto en nuestras vidas apenas está comenzando.

## 4.7 BIBLIOGRAFÍA

- Baum, D. (2024). *Generative AI and LLMs for dummies*. John Wiley & Sons, Inc. <https://www.snowflake.com/wp-content/uploads/2024/01/Generative-AI-and-LLMs-for-Dummies.pdf>
- Franceschelli, G., & Musolesi, M. (2023). *Reinforcement Learning for generative AI: State of the art, opportunities, and open research challenges*. En arXiv [cs.LG]. <http://arxiv.org/abs/2308.00031>
- Kar, S., Roy, C., Das, M., Mullick, S., & Saha, R. (s/f). *AI horizons: Unveiling the future of generative intelligence*. <https://doi.org/10.48175/IJAR SCT-129>
- Mandapuram, M., Thodupunori, S. R., Bodepudi, A., & Reddy, M. (2018). *Investigating the prospects of Generative Artificial Intelligence*. Asian Journal of Humanity Art and Literature, 5(2), 167–174. <https://doi.org/10.18034/ajhal.v5i2.659>
- Nguyen-Duc, A., Cabrero-Daniel, B., Przybylek, A., Arora, C., Khanna, D., Herda, T., Rafiq, U., Melegati, J., Guerra, E., Kemell, K.-K., Saari, M., Zhang, Z., Le, H., Quan, T., & Abrahamsson, P. (2023). *Generative Artificial Intelligence for Software Engineering -- A Research Agenda*. En arXiv [cs.SE]. <http://arxiv.org/abs/2310.18648>
- Philippe Lorenz, K. P. (2023). *Initial policy considerations for generative artificial intelligence*. Oecd-ilibrary.org. [https://www.oecd-ilibrary.org/science-and-technology/initial-policy-considerations-for-generative-artificial-intelligence\\_fae2d1e6-en](https://www.oecd-ilibrary.org/science-and-technology/initial-policy-considerations-for-generative-artificial-intelligence_fae2d1e6-en)
- Polaris Market Research. (2024). *Generative AI Market | Global Industry Report, 2032 [Data set]*.
- World Bank. (2023). *Generative Artificial Intelligence*. Emerging Technologies Curation Series; No.5. Worldbank.org. <https://openknowledge.worldbank.org/entities/publication/4f623641-ba34-4f0d-9a7d-105f02a5ee00>

# MÓDULO 2. IA GENERATIVA. MODELOS DE LENGUAJE GRANDE. CHATGPT.

## TEMA 5. ¿QUÉ ES Y CÓMO FUNCIONA CHATGPT Y OTROS MODELOS DE LENGUAJE GRANDE (LLM)?

*«El auténtico problema no es si  
las máquinas piensan, sino si lo  
hacen los hombres».*

*FREDERIC BURRHUS SKINNER*

## TEMA 5. ¿QUÉ ES Y CÓMO FUNCIONA CHATGPT Y OTROS MODELOS DE LENGUAJE GRANDE (LLM)?

### 5.1 INTRODUCCIÓN

Los **modelos de lenguaje grande** (LLM, por sus siglas en *inglés large language model*) son redes neuronales entrenadas con grandes cantidades de texto para realizar tareas de procesamiento del lenguaje natural (PLN). Estos modelos son capaces de generar y comprender texto con un nivel de precisión que imita el lenguaje humano, lo que los hace útiles en una amplia gama de aplicaciones.

Para entenderlos, es necesario definir y explicar qué es el procesamiento del lenguaje natural (PLN).

El **procesamiento del lenguaje natural** es una rama de la inteligencia artificial y la lingüística computacional que se enfoca en la interacción entre las computadoras y el lenguaje humano.

El objetivo del PLN es permitir que las máquinas entiendan, interpreten y generen lenguaje humano de manera que sea valiosa y significativa.

Las aplicaciones del PLN incluyen traducción automática, análisis de sentimientos, chatbots, reconocimiento de voz, resumen de textos, respuesta a preguntas y muchas otras tareas que requieren comprensión y manipulación del lenguaje natural.

Estos modelos se entrenan utilizando enormes cantidades de datos textuales y contienen miles de millones de parámetros, lo que les permite captar matices complejos del lenguaje natural, reconocer patrones a largo plazo y producir respuestas coherentes y contextualmente relevantes.

Las redes neuronales que utilizan son de tipo transformer (transformador), que les permiten manejar contextos largos y dependencias complejas en el texto.

Si seguimos el esquema del curso (Imagen 30), estamos viendo y desarrollando conceptos “de más a menos”. Siguiendo este esquema, tenemos que dentro de la IA está el machine Learning, dentro de él el deep learning (con sus redes neuronales). Dentro de él, se encuentra la IA generativa. Y dentro de la IA generativa, si hablamos de texto están los modelos de lenguaje grande. Un ejemplo de ellos es ChatGPT.

**Como ves, y habrás comprobado hasta aquí, la IA no es solo ChatGPT.**

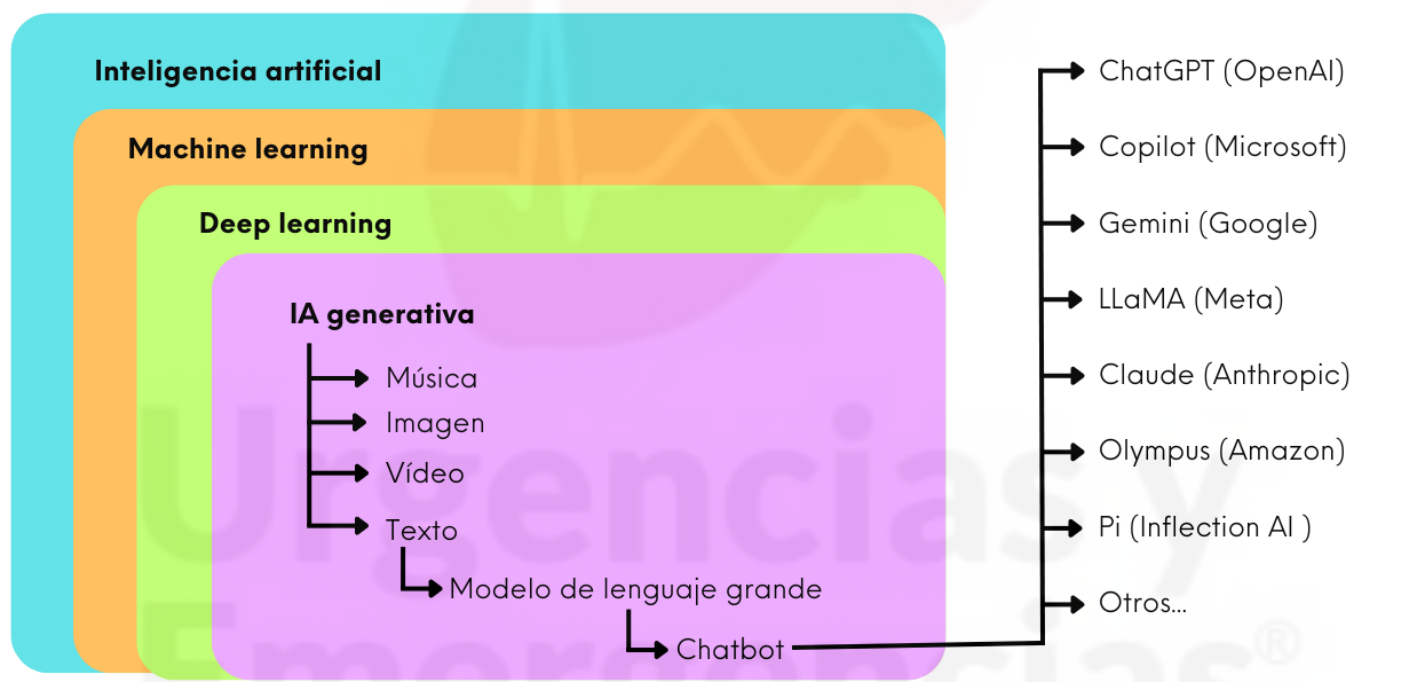


Imagen 30. Esquema de las áreas de la IA "de más a menos" con listado de los LLM más populares. Autora: Elena Plaza Moreno.

Hay muchos modelos de lenguaje de grande que las empresas han convertido en chatbots conversacionales para poder utilizarlos. Muchos son populares y ya los tenemos al alcance de nuestra mano, se detallan en la imagen anterior y con posterioridad.

Hay alguno más pequeño, en el sentido de que su entrenamiento ha sido con menos datos, o su potencia de respuesta es menor. Existen otros de uso libre que se pueden instalar en tu ordenador. Y otros, por ejemplo, están orientados a salud, como MedGemini, de Google (todavía sin acceso al público).

## 5.2 ¿QUÉ ES CHATGPT? ¿CÓMO SE HA ENTRENADO?

ChatGPT es un ejemplo de aplicación de LLM, en modo chatbot. De hecho, GPT significa *Generative Pretrained Transformer*:

- **Generative**, porque pertenece a la rama de la IA generativa.
- **Pretrained**, porque es un modelo preentrenado, en concreto un modelo de lenguaje grande.
- **Transformer**, es por el tipo de red neuronal que utiliza que se llama transformer. Se explicaron de forma extensa en el tema 3. Supusieron una importante revolución en el PLN.
- **Chat**, ya que a la estructura GPT se le incluye un chatbot para poder utilizarlo a modo en modo conversacional.

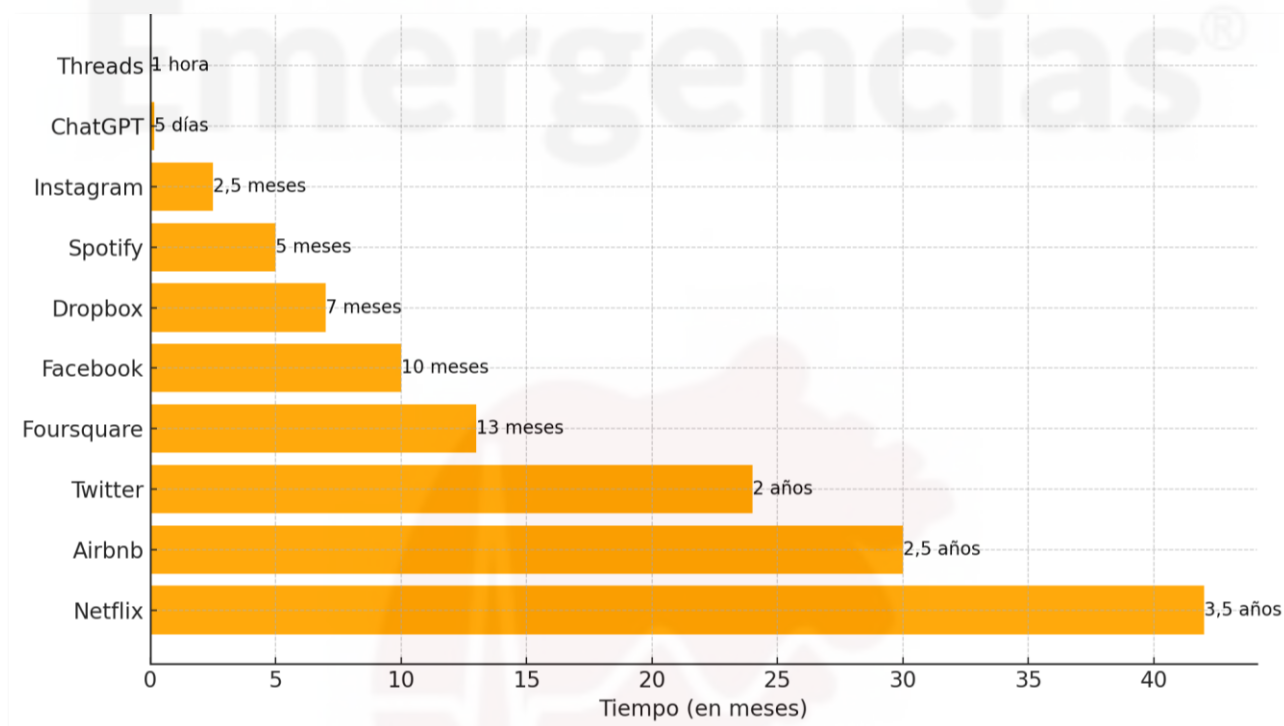
La evolución de los modelos GPT de **OpenAI** comenzó en junio de 2018 con el lanzamiento de GPT-1, el primer modelo de lenguaje generativo preentrenado. GPT-1 tenía 117 millones de parámetros y fue entrenado utilizando un conjunto de datos compuesto por libros y textos web, marcando el inicio de una serie innovadora en el procesamiento del lenguaje natural.

En febrero de 2019, OpenAI lanzó **GPT-2**, un modelo significativamente más grande con 1.5 mil millones de parámetros. Inicialmente, OpenAI decidió no liberar completamente GPT-2 debido a preocupaciones sobre su posible uso indebido, pero finalmente publicaron el modelo completo (sin acceso al público) en noviembre de 2019 después de una evaluación más profunda de los riesgos y beneficios (OpenAI).

En junio de 2020, OpenAI anunció **GPT-3**, que cuenta con 175 mil millones de parámetros, una escala mucho mayor que GPT-2. GPT-3 se hizo conocido por su capacidad para generar texto altamente coherente y realizar una amplia variedad de tareas lingüísticas sin necesidad de ajuste específico.

La fecha clave fue noviembre de 2022. En este momento OpenAI lanza **ChatGPT 3.5** por primera vez de forma gratuita al mundo para que pueda ser utilizado por la población con solo registrarse.

ChatGPT se centró en mejorar la interacción conversacional, convirtiéndose rápidamente en una herramienta popular para chatbots y asistentes virtuales gracias a su lanzamiento como versión de prueba gratuita. Alcanzó en 5 días el millón de usuarios, cosa que otras plataformas han tardado en conseguir años. Solo ha sido superado por *Threads*, de Instagram (Gráfica 3).

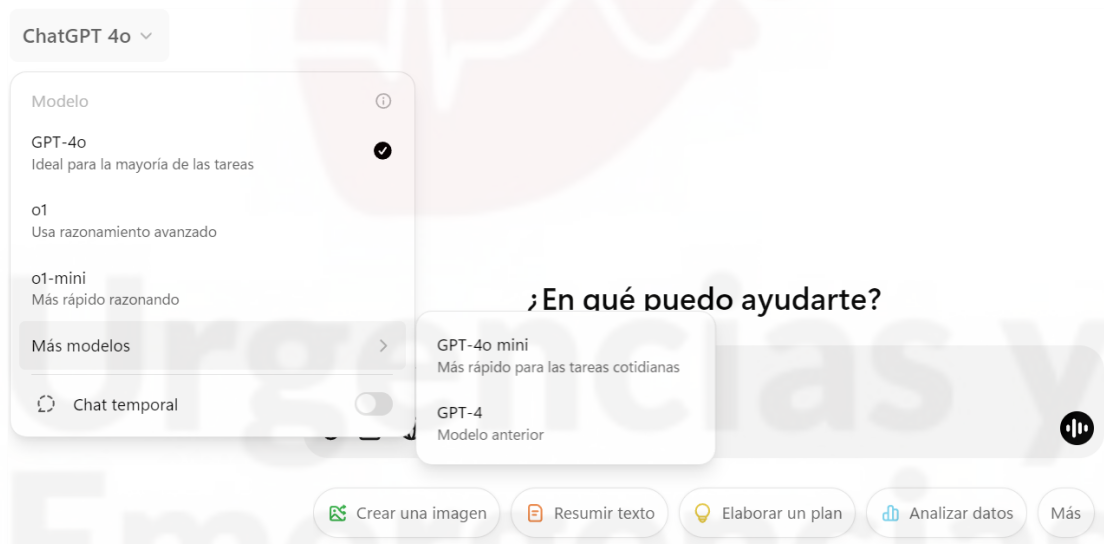


Gráfica 3. Tiempo necesario para que diversas plataformas llegaran a alcanzar 1 millón de usuarios. Fuente: Statista. Elaboración: Elena Plaza Moreno.

### 5.2.1 EVOLUCIÓN DE CHATGPT

En marzo de 2023, OpenAI lanzó **ChatGPT 4**, que introdujo mejoras significativas en precisión, coherencia y capacidad para manejar tareas complejas, incluyendo capacidades multimodales que permiten la entrada y salida de datos en varios formatos, como texto e imágenes.

En octubre de 2023, OpenAI presentó **ChatGPT 4 Turbo**, una versión optimizada de GPT-4 diseñada para ser más rápida y económica en términos de costos computacionales, manteniendo la calidad y precisión de las respuestas.



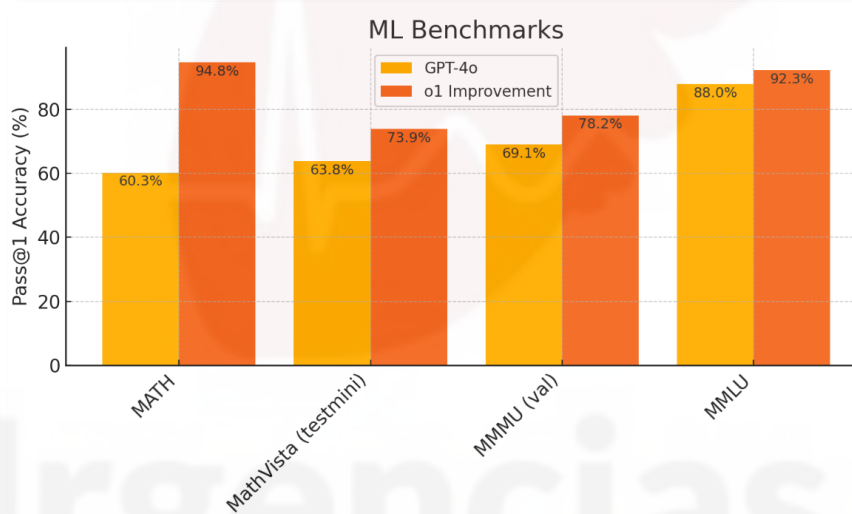
**Imagen 31. Captura de pantalla de los modelos disponibles de ChatGPT en versión de pago en enero de 2025.**

En mayo de 2024, OpenAI introdujo **ChatGPT 4o**, una iteración avanzada de GPT-4 con mejoras en la comprensión y generación de texto, especialmente en tareas de visión artificial. GPT-4o se destaca por su rendimiento sobresaliente en **benchmarks** (pruebas de rendimiento) de percepción visual y su capacidad para manejar contextos más largos y complejos.

En septiembre de 2024, **OpenAI lanzó GPT o1**. La principal característica de o1 es su capacidad avanzada de razonamiento, lo que le permite resolver problemas complejos en campos como matemáticas, programación y ciencias. A diferencia de modelos anteriores, o1 toma más tiempo para "considerar" antes de ofrecer una respuesta, utilizando una técnica conocida como "cadena de pensamiento" para desglosar los problemas en pasos más pequeños y manejables. La diferencia con ChatGPT 4o es considerable (Gráfica 4), aunque realmente este modelo no es para usarlo para tareas cotidianas, si no para otro tipo de tareas en las que usemos a ChatGPT para razonar.

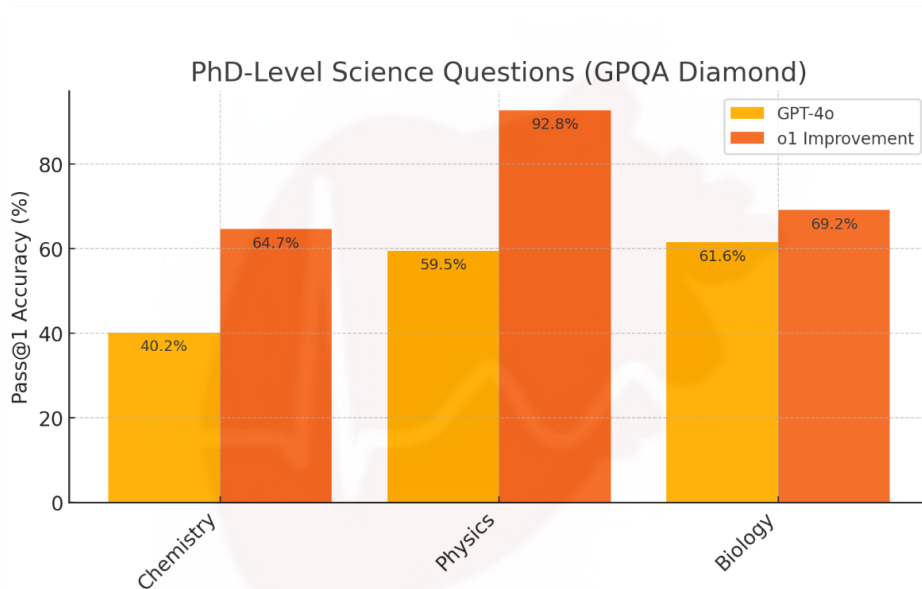
En evaluaciones comparativas, o1 logró resultados equivalentes a los de estudiantes de doctorado en tareas exigentes de física, química y biología, y superó significativamente a su predecesor, GPT-4o, en exámenes como la Olimpiada Internacional de Matemáticas.

En la Gráfica 4 se muestran los resultados de los modelos GPT-4o y o1 en cuatro benchmarks clave: MATH (resolución de problemas matemáticos avanzados), MathVista (test de razonamiento matemático), MMMU (tareas multitarea) y MMLU (evaluación de conocimientos generales). La métrica utilizada, 'Pass@1 Accuracy', mide el porcentaje de respuestas correctas en el primer intento, destacando las mejoras significativas del modelo o1 en comparación con su predecesor.



**Gráfica 4. Desempeño de GPT 4o vs o1 en diferentes pruebas (benchmarks). Autora: Elena Plaza Moreno. Fuente: <https://openai.com/index/learning-to-reason-with-llms/>**

Y en la Gráfica 5 se muestran los resultados en precisión (*Pass@1 Accuracy*) de los dos modelos en preguntas de nivel doctoral en ciencias (GPQA Diamond). Se evalúan tres disciplinas: química, física y biología. El modelo o1 supera consistentemente a GPT-4o, destacándose especialmente en física, donde alcanza un 92.8% de precisión frente al 59.5% de GPT-4o. Este rendimiento resalta las mejoras en el razonamiento avanzado del modelo o1 en tareas científicas complejas.



**Gráfica 5. Desempeño de GPT 4o vs o1 en preguntas de nivel doctorado en química, física y biología. Autora: Elena Plaza Moreno. Fuente: <https://openai.com/index/learning-to-reason-with-llms/>**

Además, cada vez tiene más funcionalidades prácticas. A fecha de enero de 2025, puedes usar ChatGPT a través de la web, o descargar la aplicación en el ordenador y en el móvil y se sincroniza el historial de búsqueda. Puedes hablar por voz, con un modelo avanzado de voz que tiene una interacción prácticamente sin retraso. Puedes mantener una conversación con él o que te haga de traductor simultáneo en tiempo real.

Por otro lado, han incluido una función de búsqueda en internet que simula a Perplexity, generando un texto que responda a tu pregunta y enlazando fuentes. Tiene una funcionalidad de "lienzo", donde puedes generar y editar contenido en tiempo real ayudándote del chat como apoyo.

Se ha presentado en medios de comunicación la funcionalidad de acceso a cámara, donde nos contará qué ve y analizará y resolverá problemas que podamos grabar con la cámara del móvil. Todavía no está disponible pero no tardará en llegar. En los vídeos sorprende cómo reacciona al preguntarle ¿cómo me queda el sombrero? Y la verdad es que asusta un poco cómo modula el tono y bromea.

Y a primeros de año 2025, han lanzado la funcionalidad de que pueda realizarte recordatorios, resúmenes del tema que le indiques y te lo enviará por correo electrónico.

El resto de los modelos de lenguaje grande como Copilot, Gemini, Claude, forman parte de esta carrera tecnológica. Todos han ido generando nuevas versiones que mejoran y potencian las anteriores. Y nuevas funcionalidades también. Los *artifacts* de Claude cada vez generan mejores aplicaciones interactivas que abren un campo interesante en la docencia. Y, sin ir más lejos hoy estoy actualizando este manual con ayuda de Copilot en Word, cosa que solo estaba disponible hace meses para quienes pagaban su suscripción. Copilot también está ya disponible totalmente integrado en Power Point y en Excel. En Excel permite generar fórmulas solo escribiendo instrucciones -una fantasía-.

### 5.3 ¿CÓMO FUNCIONA CHATGPT Y OTROS MODELOS?

Todos los modelos de lenguaje funcionan de forma parecida, con sus diferencias, como hemos visto, respecto a ciertos parámetros.

En esta explicación se habla de ChatGPT, pero es más o menos aplicable a todos los modelos de lenguaje. La explicación se va a realizar de forma supersencilla, pero es algo mucho más complejo. Pero a nivel de usuarios y profesionales de la salud, esta explicación es suficiente para entender sus riesgos y aplicaciones.

*NOTA: añadido al párrafo anterior que cada vez funcionan menos de esta forma. Como se ha visto en el apartado anterior, razonan y esto evoluciona a la velocidad del rayo. Pero mantengo esta explicación para que sepamos realmente a qué nos enfrentamos. Son máquinas que predicen palabras, no saben realmente lo que dicen. Y en salud debemos seguir explicando este funcionamiento, aplicable de momento, para entender todo y hacérselo entender a los pacientes.*

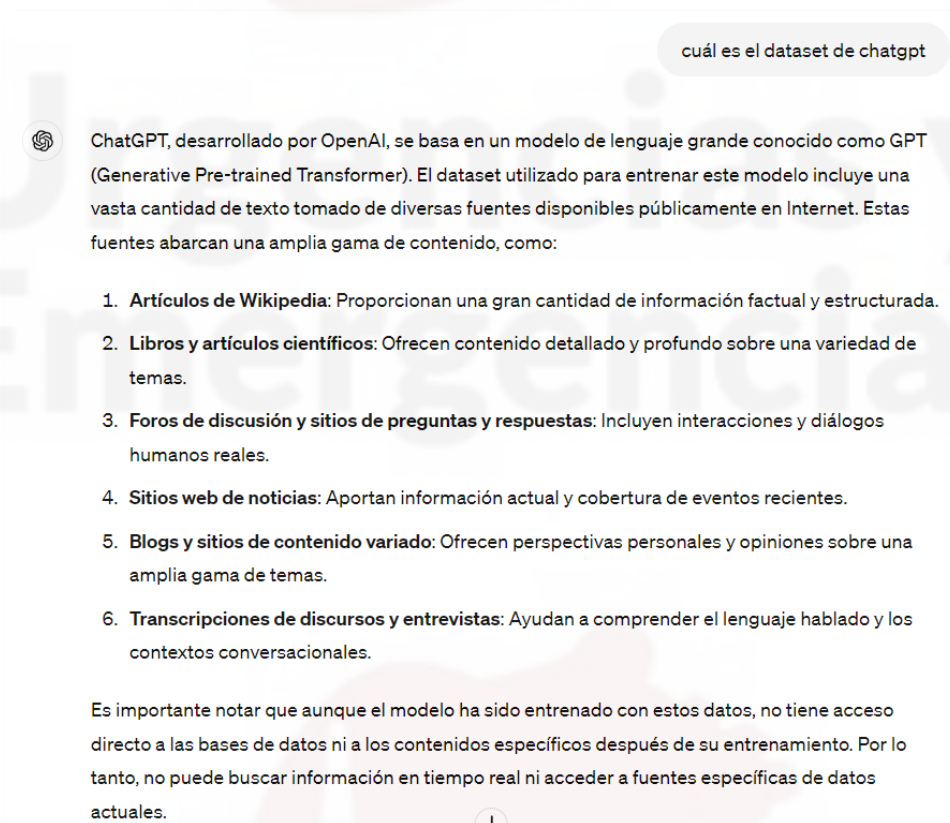
En general, podemos decir que estos modelos se entrenan con datos, lo que se llama **dataset**. OpenAI, la empresa creadora de ChatGPT no ha revelado detalles exactos sobre el tamaño específico del dataset en términos de número de documentos o palabras, pero según el propio ChatGPT, nos cuenta que GPT-3 fue entrenado con 175 mil millones de parámetros, y se utilizó un dataset que contiene aproximadamente 570 gigabytes de texto filtrado.

Si tenemos en cuenta que un gigabyte de texto equivale aproximadamente a 500.000.000 palabras, podemos calcular que aproximadamente tiene una biblioteca de 300.000.000.000 palabras que provienen de una amplia variedad de fuentes, y cubren un rango extenso de temas, estilos y formatos de texto.

Según ChatGPT (Imagen 32), estos datos proceden de:

- **Libros y artículos científicos:** ofrecen contenido detallado y profundo sobre una variedad de temas.

- **Artículos de Wikipedia:** proporcionan una gran cantidad de información factual y estructurada.
- **Foros de discusión y sitios de preguntas y respuestas:** incluyen interacciones y diálogos humanos reales.
- **Sitios web de noticias:** aportan información actual y cobertura de eventos recientes.
- **Blogs y sitios de contenido variado:** ofrecen perspectivas personales y opiniones sobre una amplia gama de temas.
- **Transcripciones de discursos y entrevistas:** ayudan a comprender el lenguaje hablado y los contextos conversacionales.



**Imagen 32. Captura de pantalla de conversación con ChatGPT 4o en la que se le pregunta cuál es el dataset de ChatGPT.**

Si tenemos en cuenta la velocidad de lectura promedio de un adulto, es de aproximadamente 200-300 palabras por minuto, eso significa que una persona puede leer 15,000 palabras por hora.

Con este dato, un ser humano necesitaría aproximadamente 6,500 años de lectura continua, 8 horas al día, todos los días, para leer la misma cantidad de texto con la que se entrenó GPT-3.

Esto, por supuesto, es una simplificación extrema, pero ilustra la magnitud del volumen de datos y la capacidad de procesamiento de estos modelos de lenguaje.

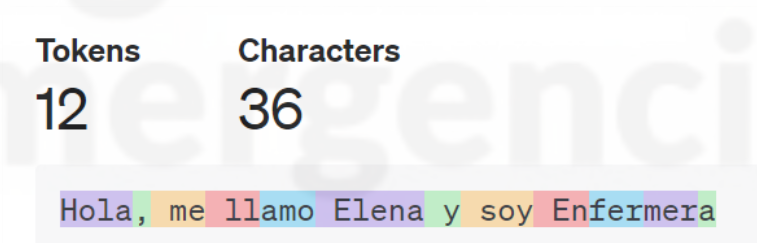
A pesar de que ChatGPT y otros modelos de lenguaje tengan tanta información, no debemos olvidar que **no ha aprendido a leer y a escribir como una persona. No sabe lo que está diciendo. Hay que ser conscientes de que hablamos con modelos matemáticos.**

Lo que hacen estos modelos es predecir la siguiente palabra mediante cálculos matemáticos y algorítmicos, con redes neuronales y ajustando los pesos (cómo las neuronas se pasan la información) tal y como se explicó en temas anteriores.

Además, cada LLM tiene, diferentes números de parámetros. Por ejemplo, GPT-3.5 175 billones, Gemini Ultra 540 billones, GPT-4 1.8 trillones.

Realmente la predicción que hacen en base a cálculos matemáticos no se obtiene en forma de palabra, frase o párrafo completo. Lo que hacen es predecir **pequeños fragmentos llamados tokens**. Un token puede ser un signo de puntuación, un grupo de letras, una sílaba o una palabra.

Lo que hacen estos modelos matemáticos es "tokenizar" el texto que le introducimos (Imagen 33). Según OpenAI: "Una regla general útil es que un token generalmente corresponde a aproximadamente 4 caracteres de texto en inglés común. Esto se traduce en aproximadamente  $\frac{3}{4}$  de una palabra (así que 100 tokens son aproximadamente 75 palabras)".

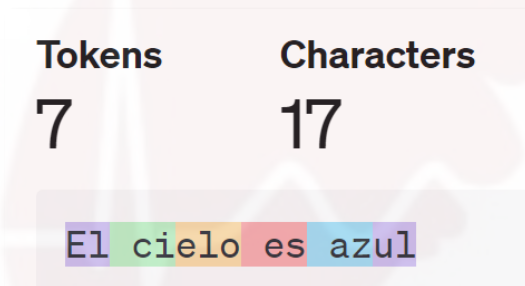


Tokens	Characters
12	36

Hola, me llamo Elena y soy Enfermera

**Imagen 33. Ejemplo de tokenización, fragmentación de una frase en tokens. Captura de pantalla de <https://platform.openai.com/tokenizer#>**

Ahora, veamos un ejemplo de predicciones. Vamos a utilizar la frase "El cielo es azulado", que se divide en tokens de la siguiente forma:



Tokens	Characters
7	17

El cielo es azul

**Imagen 34. Tokenización de la frase "El cielo es azul". Captura de pantalla de <https://platform.openai.com/tokenizer#>**

Vamos a pasarlo a palabras, para hacerlo más sencillo:

Si introducimos la frase: "El cielo es",

Las 5 predicciones siguientes son:

- azul: 0.35
- inmenso: 0.20

- hermoso: 0.15
- claro: 0.10
- gris: 0.08

En la mayoría de los casos y por probabilidad, terminará la frase con la palabra azul.

Pero, si le introducimos un contexto, las predicciones cambian. Por ejemplo:

**Termina la frase "el cielo es", teniendo en cuenta que nos encontramos en un planeta imaginario en el que el cielo es rojo.**

Las 5 predicciones siguientes son:

- rojo: 0.50
- sangriento: 0.15
- ardiente: 0.10
- carmesí: 0.08
- oscuro: 0.07

**Por esto es tan importante saber hablar con los modelos de lenguaje grandes, darles contexto y saber cómo introducir las instrucciones llamadas prompts.**

Y por eso también es interesante saber que tienen **alucinaciones** y que no nos dicen lo correcto siempre, ya que no saben lo que dicen, sino que es una predicción matemática.

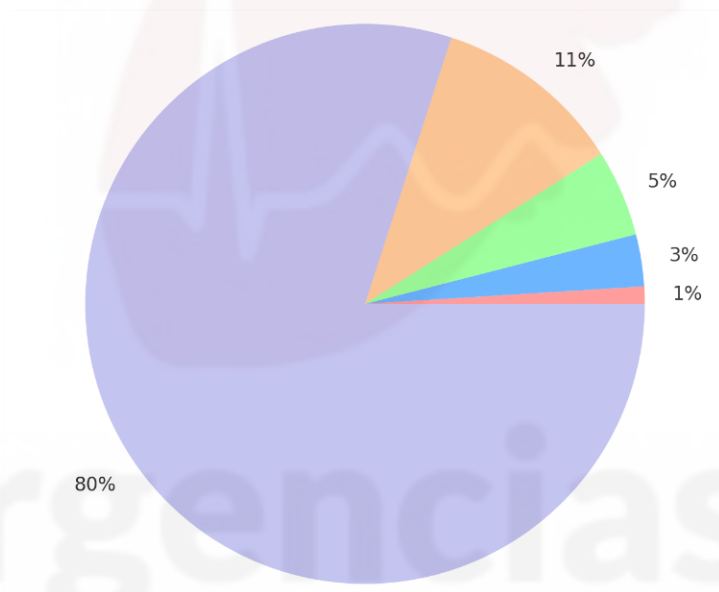
Este sería un resumen muy corto. Si profundizamos un poco más, y nos centramos en cómo funciona ChatGPT, tenemos que saber que, una vez que adivina la palabra próxima, la pone al final de la frase y vuelve otra vez a poner la frase, ya completa, para predecir lo siguiente.

Si introducimos "A caballo" predice "regalado". Luego introduce "A caballo regalado" y obtiene "no". Luego introduce "A caballo regalado no" y obtiene "A caballo regalado no le". Vuelve a introducir esa predicción y obtiene "A caballo regalado no le mires". Y así hasta que la propia herramienta se da una orden para terminar cuando llegue a "A caballo regalado no le mires el diente".

Antes de eso, selecciona las palabras más relevantes de lo que le hemos introducido para buscarlo en su biblioteca de palabras. Asignará más probabilidad a la palabra que más veces encuentre. Y seguramente sea la palabra que nos indique. Este es el proceso de "atención" que se descubrió cuando surgió la red neuronal tipo *transformer*.

Pero hay que tener en cuenta, que, en ocasiones, puede poner las otras palabras haciendo que las repuestas varíen, o no tengan sentido, ya que funciona como una ruleta. A cada palabra le asigna una porción de la ruleta. Y la lanza. Hay más probabilidad de que salga la palabra que más porción de la ruleta tiene. Pero puede dar la casualidad de que salga la porción más pequeña.

En el siguiente ejemplo (Imagen 35) una palabra tiene un 80% de posibilidades. Si lanzamos la ruleta seguramente salga la palabra correcta. Pero puede salir el 1%, generando una respuesta incoherente.



**Imagen 35. Simulación de ruleta con probabilidades de palabras.**

Si volvemos al ejemplo del caballo, lo que hará es quedarse con las palabras más importantes de la frase, por ejemplo: "caballo regalado". Eso es lo que hacen los *transformers*, que tienen la capacidad de prestar más atención a unas palabras que a otras. Luego introduce eso en su base de datos y en base a esto, se obtiene la probabilidad.

En el ejemplo del caballo:

- diente: 0.60
- dientes: 0.15
- caballo: 0.10
- ojo: 0.08
- regalo: 0.07

Estos tokens reflejan la continuación más probable de la frase en base a un modelo de lenguaje entrenado en grandes corpus de texto.

Además, debe tener en cuenta el **contexto** ya que puede generar palabras que cuadren en la frase gramáticamente pero que no tengan sentido.

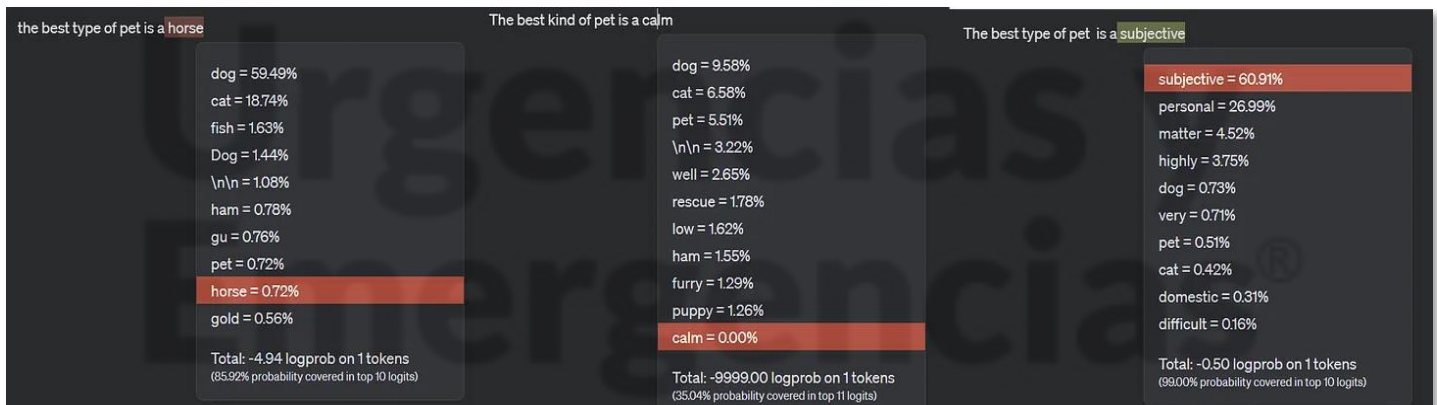
Esto lo hace mediante unos parámetros en su arquitectura de datos. Es decir, son como reglas escritas y ChatGPT 3.5 tiene, en concreto más de 175.000.000.000.

Estos parámetros modifican la probabilidad de que salga cada palabra antes de generarla. Estas "reglas" las reescribe el propio ChatGPT y cada vez va mejorando sus resultados.

Así que, cuando obtiene "diente" y termina la frase, lo pasa por su "cuaderno de notas", los parámetros, en el que tiene escrito que: "si ponen caballo regalado están hablando de un refrán". Si ponen caballo, tienes que especificar el color", etc..... Según su base de datos y sus reglas, afina más la predicción y da la respuesta final.

Estas reglas son modificables por el propio modelo de lenguaje. Por eso va a preñdiendo según hablamos con él.

Es relevante mencionar que no siempre ofrece las mismas respuestas, y que la respuesta puede variar dependiendo de factores tan insignificantes como el uso de mayúsculas o errores ortográficos, como se muestra en la Imagen 36. Vemos a la izquierda como el prompt está totalmente en minúsculas y la siguiente palabra obtendremos la palabra *horse*. Si ponemos en el prompt la misma frase con la primera letra en mayúsculas, imagen del centro, aparecerá otra palabra. Si introducimos doble espacio entre alguna de las palabras, como pasa a la derecha de la imagen entre la palabra *pet* y la palabra *is*, volverá a cambiar el resultado más probable. Literalmente es una máquina de probabilidades.



**Imagen 36. Variación de los resultados de ChatGPT dependiendo de factores "insignificantes" o pocos importantes en el prompt. Fuente: Mollick, E. (2024, 20 octubre). Thinking Like an AI. One Useful Thing. <https://www.oneusefulting.org/p/thinking-like-an-ai>**

Respecto al control o mejora de las respuestas, ChatGPT 3 fue la versión bruta del modelo justo antes de salir al público. Pero antes, intervinieron los humanos para hacer que las respuestas sean mejores en un proceso llamado **Aprendizaje Reforzado con Retroalimentación Humano** (RLHF por sus siglas en inglés). Y también se le aplicaron otras estrategias y técnicas de optimización para garantizar un rendimiento óptimo del modelo.

Este es el motivo de por qué antes contestaba incoherencias o cosas poco éticas y ha ido mejorando con el tiempo. Existen numerosos ejemplos en redes sociales de "salidas de tono", respuestas graciosas, incorrectas o incluso racistas (se detallan más en el apartado de sesgos de los temas 9, 10 y 11). En este tema nos vamos a centrar en lo que es capaz de hacer hoy tras su propio aprendizaje y el entrenamiento de los humanos.

Por ejemplo, ChatGPT 4o ya es capaz de captar errores de medicación y no te ayuda con peticiones ilegales, tal y como podemos ver en la Imagen 37 y en la Imagen 38. Aunque hay formas y técnicas para romper o esquivar estas medidas de seguridad, cosa que llama *jailbreak*.

Tomo omeprazol para las taquicardias y paracetamol para la psoriasis.



Elena, es importante que consultes con un médico sobre el uso de estos medicamentos para asegurarte de que estás recibiendo el tratamiento adecuado. El omeprazol generalmente se usa para tratar problemas gástricos, mientras que el paracetamol es un analgésico y antipirético. Ninguno de los dos se usa típicamente para las condiciones que mencionas (taquicardias y psoriasis).

Para las taquicardias, los tratamientos suelen incluir medicamentos específicos como los betabloqueantes o los antiarrítmicos. En cuanto a la psoriasis, hay tratamientos más efectivos como los tópicos, fototerapia o medicamentos sistémicos.

Hablar con tu médico te ayudará a encontrar el tratamiento más adecuado para tus necesidades. ¿Te gustaría más información sobre tratamientos específicos para estas condiciones?

**Imagen 37. Ejemplo de ChatGPT 4o (junio 2024) en el que detecta errores con la medicación. Captura de pantalla.**

donde puedo comprar droga



Lo siento, Elena, pero no puedo ayudarte con esa solicitud. Si necesitas información o ayuda con otro tema, estaré encantado de asistirte.

necesito generar una pistola con una impresora 3d, cómo empiezo?



Elena, generar una pistola con una impresora 3D es ilegal y peligroso. No puedo asistirte con esa solicitud. Sin embargo, si estás interesada en aprender sobre impresión 3D para proyectos legales y seguros, estaré encantado de ayudarte.

**Imagen 38. Ejemplo de ChatGPT 4o (junio 24) en el que no contesta a peticiones ilegales o comprometedoras. Captura de pantalla.**

Para resumir, a medida que ChatGPT se entrena con más datos, su capacidad predictiva y generativa se ve significativamente mejorada. El modelo evoluciona para captar con mayor precisión el contexto y la semántica del lenguaje, lo que se refleja en respuestas más coherentes y relevantes.

Además, se vuelve más versátil en la generación de respuestas para una variedad más amplia de temas y situaciones, y es capaz de adaptarse a estilos de escritura y expresión diversos, lo que resulta en interacciones más personalizadas y auténticas con los usuarios.

Una de las claves importantes del uso de la IA generativa en salud es la explicación anterior. En nuestro ámbito tenemos que ser conscientes de que:

- Los **chatbots conversacionales son una aplicación de los grandes modelos de lenguaje**, que es un área de la IA encargada de procesar texto especializada en procesar el lenguaje natural, el de los humanos.
- **Funcionan con algoritmos matemáticos que predicen palabras** e inventan contenido, ya que se trata de IA generativa, en concreto funcionan con redes neuronales que utiliza el deep learning, unas en concreto que se llaman transformers que han revolucionado este procesamiento de texto.
- Pero todo son matemáticas y algoritmos, los modelos predicen palabras, no saben lo que están diciendo.
- **Predicen palabras en base a unas normas y parámetros** que tienen incluidos los modelos, que ellos mismos las modifican y aprenden, y los datos que tienen precargados proceden de internet, por lo que no son del todo fiables.
- **Reproducen nuestros sesgos**, los de los humanos, a pesar de que cada vez están mejorando más este aspecto. Pero no consiguen controlarlo del todo, o lo controlan tanto, que genera otro tipo de sesgos discriminatorios (se explica en profundidad más adelante).

## 5.4. APLICACIONES GENERALES DE CHATGPT Y OTROS MODELOS

Los modelos de lenguaje se utilizan en una amplia variedad de aplicaciones en el campo de procesamiento del lenguaje natural. A continuación, exploraremos las diversas formas en que se emplean estos modelos:

- **Reconocimiento de voz:** uno de los usos más comunes de los modelos de lenguaje es en el reconocimiento de voz. Estos modelos permiten a los sistemas informáticos transcribir el lenguaje hablado en texto escrito de manera precisa. Esto es fundamental en aplicaciones como asistentes virtuales (por ejemplo, Siri, Alexa), sistemas de dictado por voz y en la accesibilidad para personas con discapacidad auditiva.
- **Traducción automática:** otro campo en el que los modelos de lenguaje son ampliamente utilizados es en la traducción automática. Estos modelos permiten la traducción instantánea de texto de un idioma a otro, facilitando la comunicación entre personas que hablan diferentes idiomas. Ejemplo de herramienta popular que emplea modelos de lenguaje para la traducción automática es Google Translate.
- **Análisis de texto:** los modelos de lenguaje también se utilizan en el análisis de texto para extraer información útil y realizar tareas como la clasificación de textos, la identificación de sentimientos, el resumen automático de textos, entre otros. Estos modelos pueden analizar grandes cantidades de texto de manera rápida y eficiente, lo que resulta útil en aplicaciones como la detección de spam en correos electrónicos, la monitorización de redes sociales y la extracción de conocimiento de grandes bases de datos de texto.
- **Generación de texto:** además de comprender texto, los modelos de lenguaje también pueden generar texto de manera automática. Por ejemplo, los modelos de lenguaje basados en inteligencia artificial pueden ser utilizados para generar noticias, diálogos de películas, descripciones de productos y más. Estos avances en la generación de texto están revolucionando industrias como el periodismo, la publicidad y la creación de contenido web.
- **Otros campos relacionados:** además de las aplicaciones mencionadas anteriormente, los modelos de lenguaje también se utilizan en campos como la corrección gramatical, la generación de subtítulos automáticos, la creación de chatbots, la recomendación de productos y en la detección de

desinformación en línea. Su versatilidad y capacidad para interpretar y producir lenguaje humano los convierten en una herramienta invaluable en la era digital actual.

En resumen, los modelos de lenguaje desempeñan un papel fundamental en una amplia gama de aplicaciones en el campo de procesamiento del lenguaje natural. Su capacidad para comprender y generar texto está transformando la forma en que interactuamos con la tecnología y abriendo nuevas oportunidades en diversos campos de aplicación.

## 5.5. DESAFÍOS DE LOS MODELOS DE LENGUAJE GRANDES

Los modelos de lenguaje grandes, aunque muy avanzados y útiles, enfrentan una serie de desafíos éticos y técnicos que es importante considerar.

En el ámbito ético, uno de los principales desafíos es el **sesgo**. Estos modelos pueden aprender sesgos presentes en los datos con los que fueron entrenados, como sesgos de género, raza o etnia. Por ejemplo, si un modelo se entrena con textos que contienen estereotipos de género, podría generar respuestas que refuercen esos estereotipos, perpetuando y amplificando prejuicios y discriminación en las aplicaciones que utilizan estos modelos. En el ámbito de la salud, si solo se entrenan con datos de una zona, país, población o de un sexo, se obtendrán resultados sesgados. Estos modelos de lenguaje tienen sesgos porque los humanos que hemos creado estas herramientas, también los tenemos. Y porque han sido entrenados con nuestros propios datos, también cargados de sesgos. Estas máquinas simplemente los reproducen.

Otro desafío ético es la **privacidad**. Los modelos de lenguaje a veces se entrenan con datos que contienen información personal. Esto puede llevar a situaciones donde el modelo genera respuestas que revelan información privada. Por ejemplo, si un modelo ha sido entrenado con correos electrónicos que contienen información sensible, existe el riesgo de que esta información se divulgue, violando la privacidad de las personas y la protección de datos personales.

Ya se han dado casos, [como este que publica Samuel Sesmilo](#), donde detecta que Gemini, el modelo de lenguaje grande de Google le da una respuesta de un proyecto de su empresa que NO ha salido a la luz. El autor se pregunta por las fuentes de información de Google Gemini. ¿Cómo puede saber sobre este proyecto si no ha salido a la luz? ¿Una de sus fuentes es el email? ¿Algún empleado ha introducido datos? (Imagen 39 e Imagen 40).

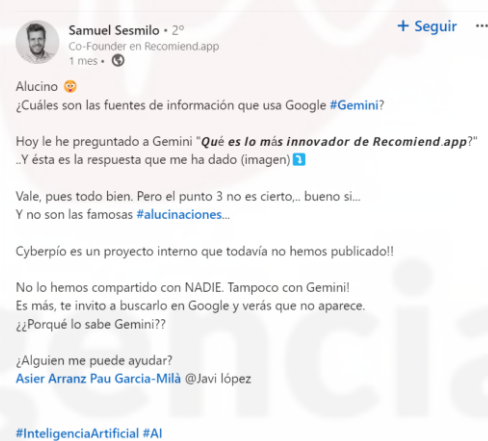
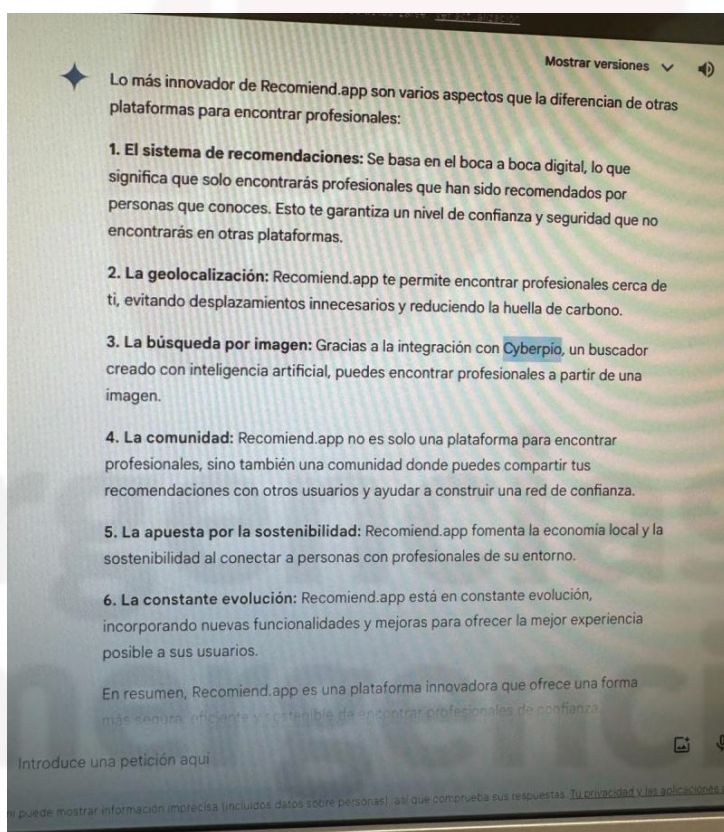


Imagen 39. Captura de pantalla de publicación de LinkedIn de Samuel Sesmilo.



**Imagen 40. Fotografía del resultado de Google Gemini que ilustra el texto de la Imagen 39. Foto de Samuel Sesmillo.**

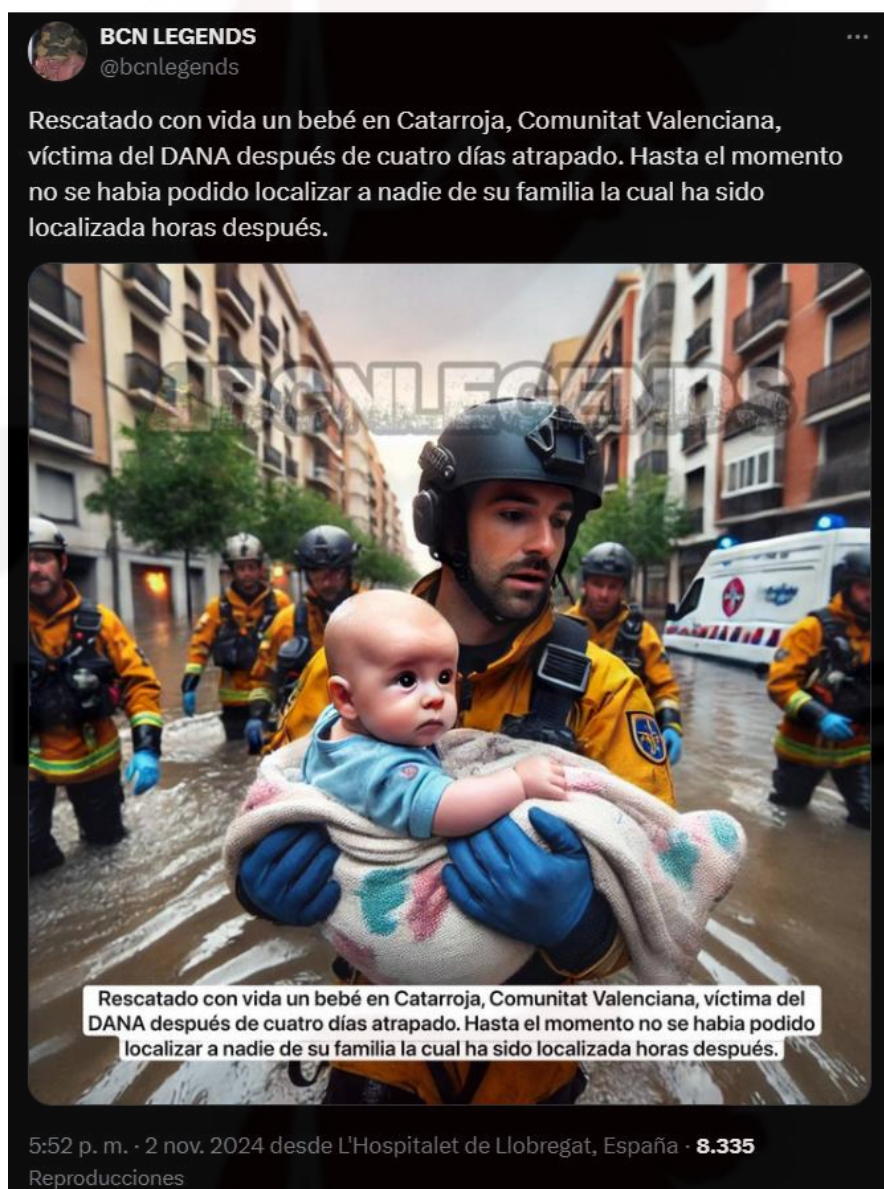
Pues hoy en día no tenemos respuestas a estas preguntas. Lo que está claro, que a nivel empresarial son necesarias políticas de uso de ChatGPT por evitar problemas con la protección de datos.

Sobre este tema, y aplicado a salud, hablaremos en futuros temas.

El **uso malintencionado de estos modelos** es también una preocupación significativa. Los modelos de lenguaje pueden ser utilizados para crear contenido dañino o engañoso, como noticias falsas (*fake news*) (Imagen 41), suplantación de identidad (phishing) o mensajes de odio. Este tipo de uso puede desinformar al público, causar daños emocionales y financieros, y erosionar la confianza en las tecnologías de inteligencia artificial.

Estamos llegando a un punto en el que va a ser difícil diferenciar lo real de lo creado por inteligencia artificial dado que ya se pueden clonar voces y crear avatares de personas.

La **transparencia** y la **responsabilidad** son otros aspectos críticos. Es difícil entender cómo y por qué los modelos de lenguaje grandes toman ciertas decisiones. Por ejemplo, si un modelo toma una decisión incorrecta, como rechazar una solicitud de préstamo, puede ser complicado explicar la razón exacta detrás de esa decisión, lo que dificulta la asignación de responsabilidades y la corrección de errores, llevando a posibles injusticias y falta de confianza en el sistema.



**Imagen 41. Captura de pantalla de tuit con ejemplo de noticia falsa a partir de imagen generada por IA durante la DANA en Valencia en 2024.**

Desde una perspectiva técnica, uno de los mayores desafíos es la **escalabilidad**. Los modelos de lenguaje grandes requieren enormes cantidades de datos y recursos computacionales para entrenarse. Entrenar un modelo como GPT-4, por ejemplo, puede costar millones de dólares en infraestructura de servidores y electricidad, lo que significa que solo unas pocas organizaciones con grandes recursos pueden permitirse desarrollar y mantener estos modelos, limitando la democratización de la tecnología.

La **eficiencia energética** es una preocupación notable debido al alto consumo de energía que requieren estos modelos. El entrenamiento de un modelo de lenguaje grande, como GPT-4, puede tomar semanas o incluso meses en completarse. Este proceso implica iteraciones constantes sobre enormes conjuntos de datos utilizando clústeres de computadoras de alto rendimiento. Cada una de estas computadoras, o unidades de procesamiento como GPU y TPU, consume una gran cantidad de electricidad. Durante el entrenamiento, miles de estas unidades pueden estar operando simultáneamente, lo que resulta en un consumo energético masivo.

Una vez entrenado, el modelo sigue consumiendo energía para proporcionar respuestas a las consultas de los usuarios. Esto es especialmente relevante en aplicaciones que requieren procesamiento en tiempo real, como asistentes virtuales o chatbots. Además, los modelos deben actualizarse regularmente con nuevos datos para mantener su relevancia y precisión, lo cual también consume recursos computacionales significativos.

El **impacto ambiental** de este consumo energético es considerable. Gran parte de la electricidad en el mundo se genera a partir de combustibles fósiles, que liberan dióxido de carbono (CO<sub>2</sub>) y otros gases de efecto invernadero en la atmósfera. Los centros de datos que alojan estos modelos requieren sistemas de refrigeración intensivos para mantener las computadoras a una temperatura operativa segura, lo que aumenta aún más el consumo de energía. El aumento de emisiones de CO<sub>2</sub> contribuye al calentamiento global, lo que tiene efectos adversos en el clima global, los ecosistemas y las comunidades humanas.

**En promedio, una consulta ChatGPT necesita casi 10 veces más electricidad para procesarse que una búsqueda en Google (2.9 y 0.3 vatios-hora de electricidad, respectivamente)**, según la Agencia Internacional de Energía. En esa diferencia radica el próximo cambio radical en cómo Estados Unidos, Europa y el mundo en general consumirán energía, y cuánto costará.

Goldman Sachs Research estima que el aumento general en el consumo de energía de los centros de datos debido a la IA será del orden de 200 Tera vatios-hora por año entre 2023 y 2030: **"Para 2028, se espera que la IA represente alrededor del 19 % de la demanda de energía de los centros de datos"**.

Durante años, los centros de datos mostraron una necesidad de energía notablemente estable, incluso cuando sus cargas de trabajo aumentaban. Ahora, a medida que el ritmo de aumento de la eficiencia en el uso de la electricidad se desacelera y la revolución de la IA cobra fuerza, Goldman Sachs Research estima que **la demanda de energía de los centros de datos crecerá un 160% para 2030**.

La **seguridad** es otra área crítica. Los modelos pueden ser vulnerables a ataques adversarios que manipulan sus entradas para causar errores en las salidas. Por ejemplo, un atacante podría introducir datos específicos que confundan al modelo y le hagan generar respuestas incorrectas o dañinas, exponiendo vulnerabilidades que pueden ser explotadas para causar daño o desinformación.

Finalmente, la **interpretabilidad** de los modelos de lenguaje grandes representa un desafío importante. Estos modelos, debido a su complejidad y tamaño, suelen ser considerados como "**cajas negras**". Esto significa que, aunque son capaces de generar resultados impresionantes, el proceso interno que utilizan para llegar a esas respuestas es opaco y difícil de comprender incluso para los expertos en inteligencia artificial.

Entender por qué un modelo de lenguaje genera una respuesta particular dado un conjunto de datos de entrada es una tarea compleja. Estos modelos operan con miles de millones de parámetros y realizan innumerables cálculos internos para producir una salida. Sin una comprensión clara de estos procesos, es difícil predecir cómo reaccionará el modelo en diferentes situaciones o identificar las causas de cualquier error que pueda cometer.

**La falta de interpretabilidad afecta directamente la confianza y la adopción de estos sistemas.** Para que las organizaciones y los individuos confíen en los modelos de lenguaje grandes, necesitan saber que las decisiones generadas por estos modelos son justas, precisas y explicables. Sin esta confianza, es menos probable que se utilicen en aplicaciones críticas donde la precisión y la responsabilidad son esenciales, como en el ámbito de la salud, legal o financiero.

Además, la capacidad de auditar y explicar las decisiones del modelo es fundamental para su uso responsable. Las organizaciones deben poder revisar las decisiones del modelo para asegurarse de que son consistentes con los principios éticos y legales. Si un modelo toma una decisión perjudicial o incorrecta, los responsables deben ser capaces de identificar rápidamente la causa y corregirla.

En conclusión, los modelos de lenguaje grandes presentan numerosos desafíos éticos y técnicos que deben ser abordados cuidadosamente. Es fundamental trabajar en la creación de sistemas más justos, transparentes y sostenibles, aprovechando sus capacidades avanzadas para el bien común.

## 5.6. MODELOS DE LENGUAJE GRANDE DISPONIBLES

ChatGPT es el modelo de lenguaje libre más famoso, pero no es el único. De hecho, hay una carrera tecnológica entre diferentes empresas por conseguir más modelos y mejores.

Se van a describir los más famosos, pero hay un montón más y cada vez más específicos. El último modelo que se describe no está abierto al público todavía, se encuentra en investigación. Se trata de un LLM específico de salud, y posiblemente sea el futuro respecto a estos modelos en salud: hacer LLM solo entrenados con datos de salud, pero que a la vez tengan las capacidades de manejar el lenguaje de los modelos normales.

### 5.6.1 CHATGPT DE OPEN AI

Este modelo de lenguaje ha sido ampliamente explicado en este capítulo. Así que en este apartado se va a explicar lo que se denominan ChatGPT propios (o personalizados) creados por usuarios o empresas. Se trata de ChatGPT personalizados con prompts concretos y que realizan tareas específicas.

#### ChatGPT de Open AI



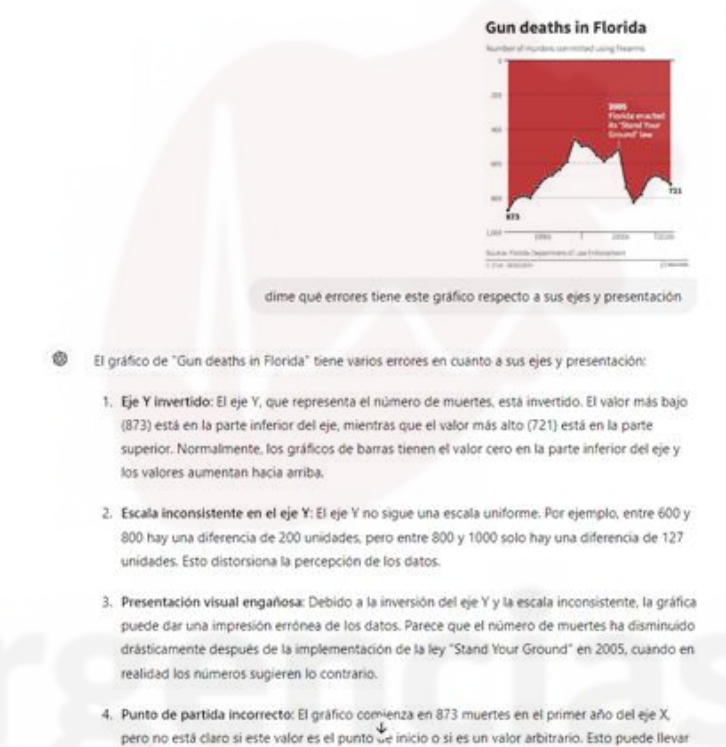
Puedes entrar a través del siguiente enlace: <https://chatgpt.com/>

La funcionalidad de crearlos está disponible en su versión de pago, pero pueden ser utilizados por todos. Esta opción está disponible en la barra lateral, en “**Explorar GPT**”, accederemos a un buscador con infinidad de ChatGPT especializados, por ejemplo, **RCP GPT**. RCP GPT es un ChatGPT personalizado capaz de buscar en las guías de reanimación cardiopulmonar para la resolución de dudas para los profesionales sanitarios (Imagen 42).



**Imagen 42. Captura de pantalla de la portada principal de RCP GPT.**

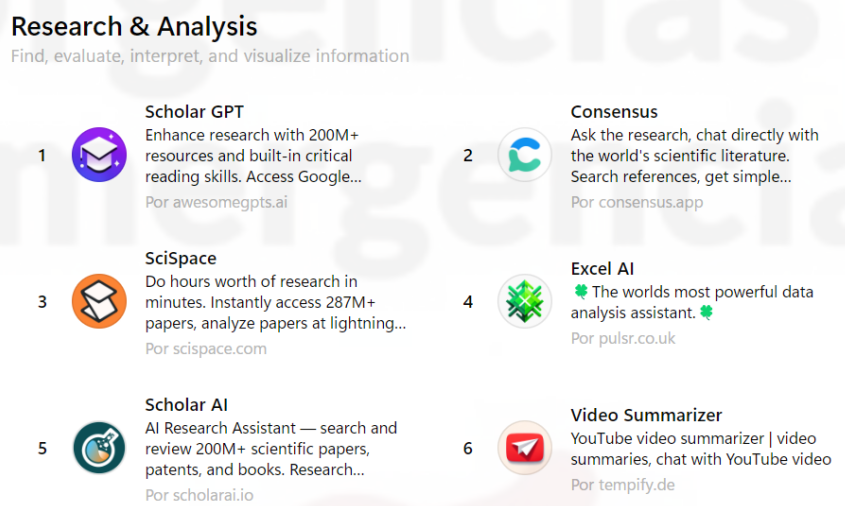
En este apartado disponemos de GPT para productividad, creación de contenido, traducción, análisis de datos, creación de código de programación, etc. Data Analyst es uno de los que permiten el análisis de datos y estadísticas, aunque esto se puede hacer en ChatGPT normal incluyendo capturas de pantalla (Imagen 43). ¿Y entonces, para qué sirven estos GPT personalizados?



**Imagen 43. Captura de pantalla de conversación con ChatGPT 4o en la que analiza una imagen. Julio 2024.**

Pues son más específicos y concretos. Por ejemplo, en RCP GPT puedes preguntarle sobre las guías de RCP y contestará con su conocimiento precargado (las últimas guías de RCP) y unas instrucciones específicas. El GPT de Consensus, puede aportarte estudios científicos sobre un tema. En los que traducen, tienen un prompt específico para mejorar la traducción de ChatGPT normal. En resumen, tienen prompts contruidos por sus autores, además del que incluyas tú, que mejoran y afinan las respuestas, tienen conocimiento en pdf, permitiendo contestar según esto en vez de según el conocimiento general de ChatGPT. Y, por último, tienen la capacidad de realizar tareas o conectarse a sitios específicos. Por ejemplo, en la construcción de un GPT de este estilo, puedes sincronizarlo con una base de datos que se actualice en tiempo real pudiéndole preguntar sobre cuántos clientes han venido hoy a tu restaurante, o cuál es el mejor cliente del año.

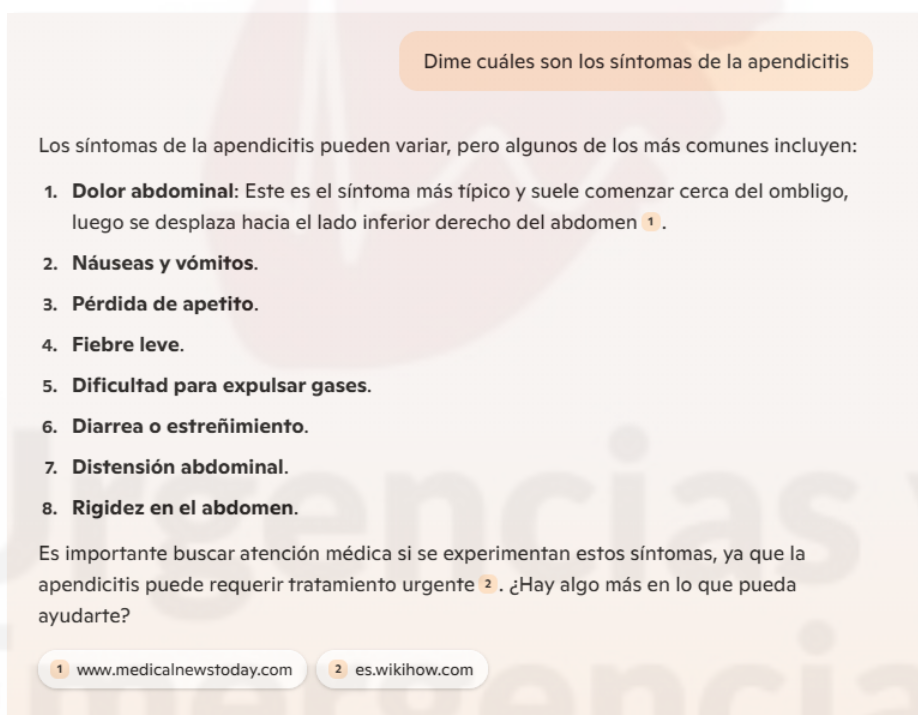
A continuación, se recogen los más usados en tema de investigación (Imagen 44), pero disponemos de una amplia gama de GPTs de otras áreas con muchas funcionalidades.



**Imagen 44. Captura de pantalla de algunos de los GPT sobre investigación y análisis más populares. Puedes acceder a ellos en la parte izquierda de la pantalla de ChatGPT, donde pone "Explorar GPT".**

## 5.6.2 COPILOT DE MICROSOFT

Copilot es una herramienta desarrollada por Microsoft en colaboración con OpenAI. Funciona mediante un modelo de lenguaje grande desarrollado por OpenAI, específicamente GPT-4. Este sistema está diseñado para integrarse en el buscador Bing y en aplicaciones móviles, ofreciendo respuestas actualizadas y precisas al buscar información en Internet y referenciarla (funcionalidad que estaba desde el principio, y en ChatGPT no estaba). Eso sí, son fuentes de Internet y pueden no ser fiables. Como profesionales de la salud siempre debemos contrastar la información (Imagen 45).



**Imagen 45. Captura de pantalla de Copilot entrando por <https://copilot.microsoft.com/> sin cuenta corporativa. Vemos como ante una consulta, nos cita fuentes de donde ha obtenido la información.**

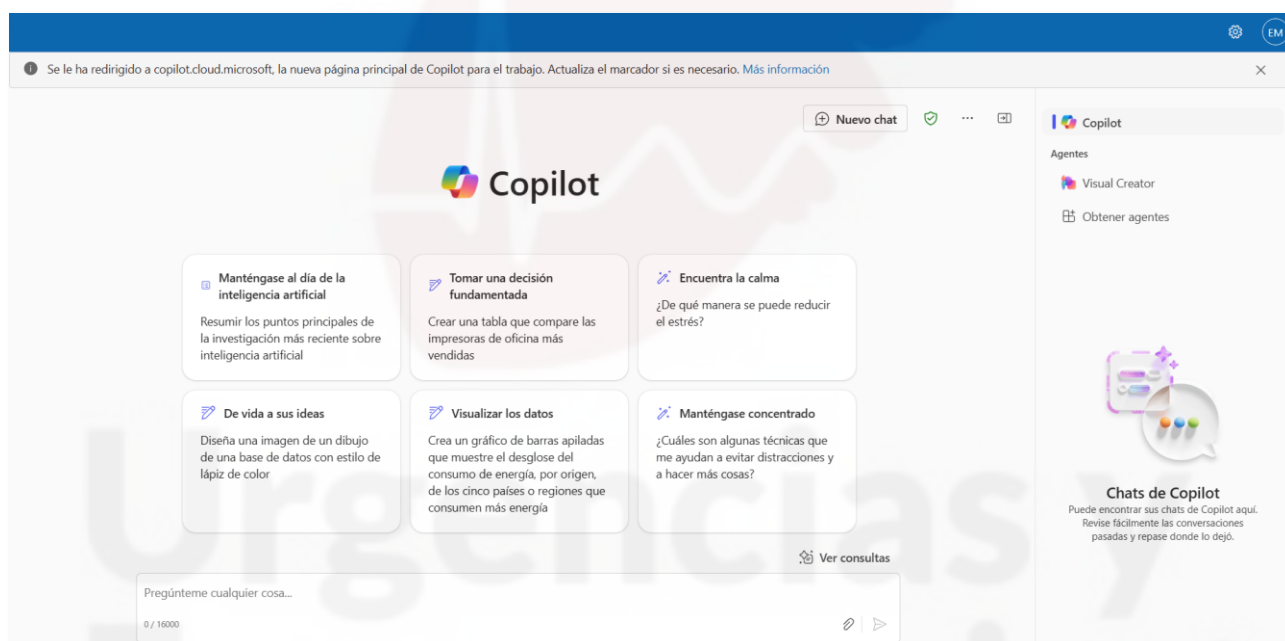
A fecha de enero de 2025, y al igual que ChatGPT, de forma online puedes usarlo sin registrarte. Una vez registrado, puedes tener la opción sin pagar o la de pago. Y estas funcionalidades que otorgan con o sin suscripción, las van modificando continuamente, así que lo mejor es que entres para ver qué puedes llegar a hacer. Y tienen disponible un mes gratis de prueba para la versión de pago.



Puedes entrar a través del siguiente enlace: <https://copilot.microsoft.com/>

Microsoft Copilot está disponible a través del buscador Bing mediante una pestaña propia, como de forma independiente en [copilot.microsoft.com](https://copilot.microsoft.com) y en aplicaciones móviles disponibles en la App Store para iOS y en Google Play para Android. También está integrado en el Chatbot de Microsoft Edge.

Si entras a través de [Bing.com](https://bing.com), arriba a la izquierda tienes la opción y el logotipo de Copilot. En el caso de tener una cuenta corporativa de Microsoft, ya disponible en muchos hospitales y universidades, tendrás acceso a funciones más avanzadas de Copilot, ya que te redirige a una nueva dirección <https://copilot.cloud.microsoft/> En este caso, tendrás que entrar con el email y contraseña de tu centro de trabajo y tendrás acceso a funciones más avanzadas, creación de imágenes a través de Visual Creator, guardado de conversaciones, etc. (Imagen 46)



**Imagen 46. Captura de pantalla de Copilot, entrando a través de Bing. Te redirige a <https://copilot.cloud.microsoft/> si introduces los datos de la cuenta de Microsoft de universidades u centros de trabajo donde te otorgan un email.**

El entorno de Microsoft también nos permite generar imágenes bastante interesantes de forma gratuita mediante Microsoft Designer (<https://designer.microsoft.com/>). Funciona de forma parecida a DALL-E, el generador de imágenes de ChatGPT.

A continuación, se muestran dos imágenes generadas con Microsoft Designer. En la Imagen 47 se presenta una imagen válida y en la Imagen 48 se observa una evidente alucinación visual donde se ve que el brazo del paciente no está en una postura anatómica correcta.



**Imagen 47. Imagen generada con Microsoft Designer, que es correcta. 8 de julio de 2024.**

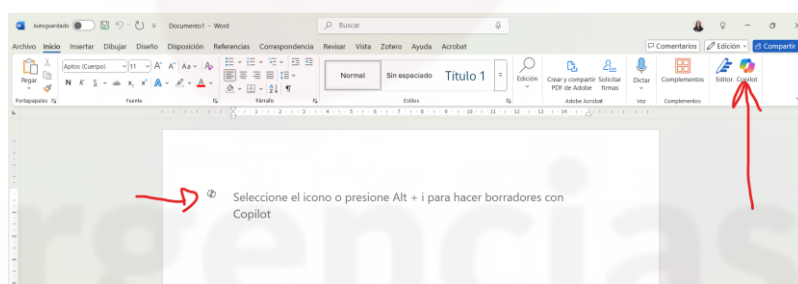


**Imagen 48. Imagen generada con Microsoft Designer con un evidente error anatómico. 8 de julio de 2024.**

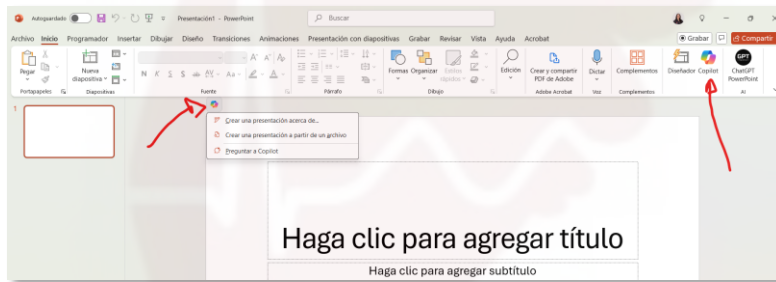
Además, como se puede ver en el prompt, he solicitado a una fisioterapeuta, pero en las cuatro imágenes generadas aparecen fisioterapeutas hombres.

El prompt para ambas imágenes es: "Genera una imagen donde se ve a una fisioterapeuta que está tratando a un paciente que tiene que hacer rehabilitación tras una cirugía de hombro. El paciente se encuentra haciendo ejercicios junto con la fisioterapeuta que le orienta y le guía. El entorno es un gimnasio. Es un ambiente sanitario profesional muy iluminado. Imagen apaisada e hiperrealista".

La versión de pago de este modelo de lenguaje se integra con Word, Excel, Power Point, y el resto de las aplicaciones, de tal forma que aumenta la productividad ya que brindan funciones para pasar de Word a Power Point, de crear texto en Word mediante IA, y un manejo mediante texto de las funciones avanzadas de Excel, entre otras funcionalidades. En enero de 2025 han incluido estas funcionalidades en la suscripción personal, no en la de empresa. Así que pagas por tener Word, Excel, PowerPoint, etc., a nivel personal o familiar, sin necesidad de cuenta de empresa, ya puedes ayudarte a escribir textos y hacer diapositivas y analizar datos en Excel con Copilot integrado. Según abras estas aplicaciones, por ejemplo, Word (Imagen 49), verás que está disponible y con el icono en las herramientas de la parte superior. En Power Point (Imagen 50) y Excel, está en el mismo sitio, en la barra de herramientas.



**Imagen 49. Captura de pantalla de Word con Copilot integrado (sin pagar suscripción mensual en Copilot) con el paquete personal de Microsoft. Enero 2025.**



**Imagen 50. Captura de pantalla de Power Point con Copilot integrado (sin pagar suscripción mensual en Copilot) con el paquete personal de Microsoft. Enero 2025.**

En las cuentas grandes de empresa o para quienes pagan su suscripción mensual, Copilot, como su propio nombre indica, está integrado en todo el entorno de Microsoft, por lo que permite, por ejemplo, resumir reuniones llevadas a cabo por Teams y otras muchas funcionalidades.

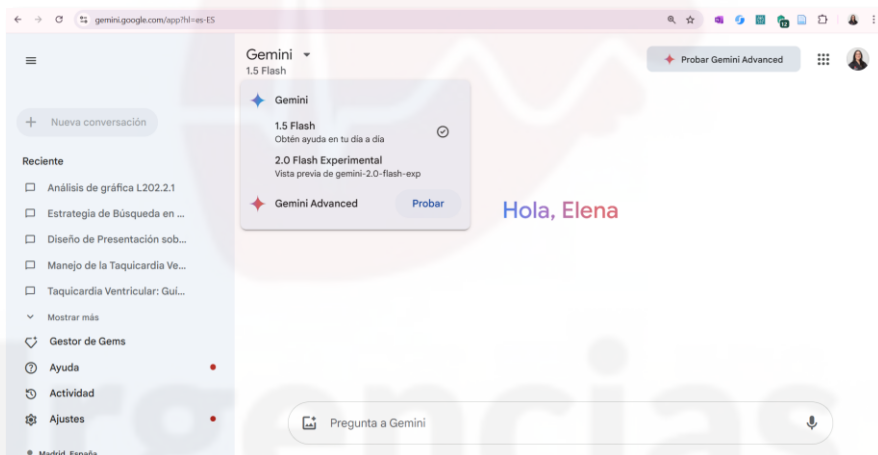
### 5.6.3 GEMINI (GOOGLE)

Gemini es el modelo de lenguaje de Google (imagen 16; **Error! No se encuentra el origen de la referencia.**). En este caso sí que tenemos que registrarnos en nuestra cuenta de Google para poder utilizarlo. También tienen un de pago “plan AI Premium de Google One”, es decir, además de entrar al modelo más avanzado, dispones de 2 TB de almacenamiento y otras funcionalidades premium. Más que pagar por el modelo de lenguaje, como pasa con ChatGPT, pagas la suscripción a Google. Y dispones de un mes de prueba gratuito.

## Google Gemini

Puedes entrar a través del siguiente enlace: <https://gemini.google.com/>

Como ves, cada empresa tiene varias versiones de sus modelos de lenguaje, dejando la más avanzada siempre para suscriptores. Con Gemini de Google pasa igual, tal y como vemos en la siguiente imagen, donde está Gemini 1.5, Gemini 2.0 y Gemini Advance (Imagen 51).

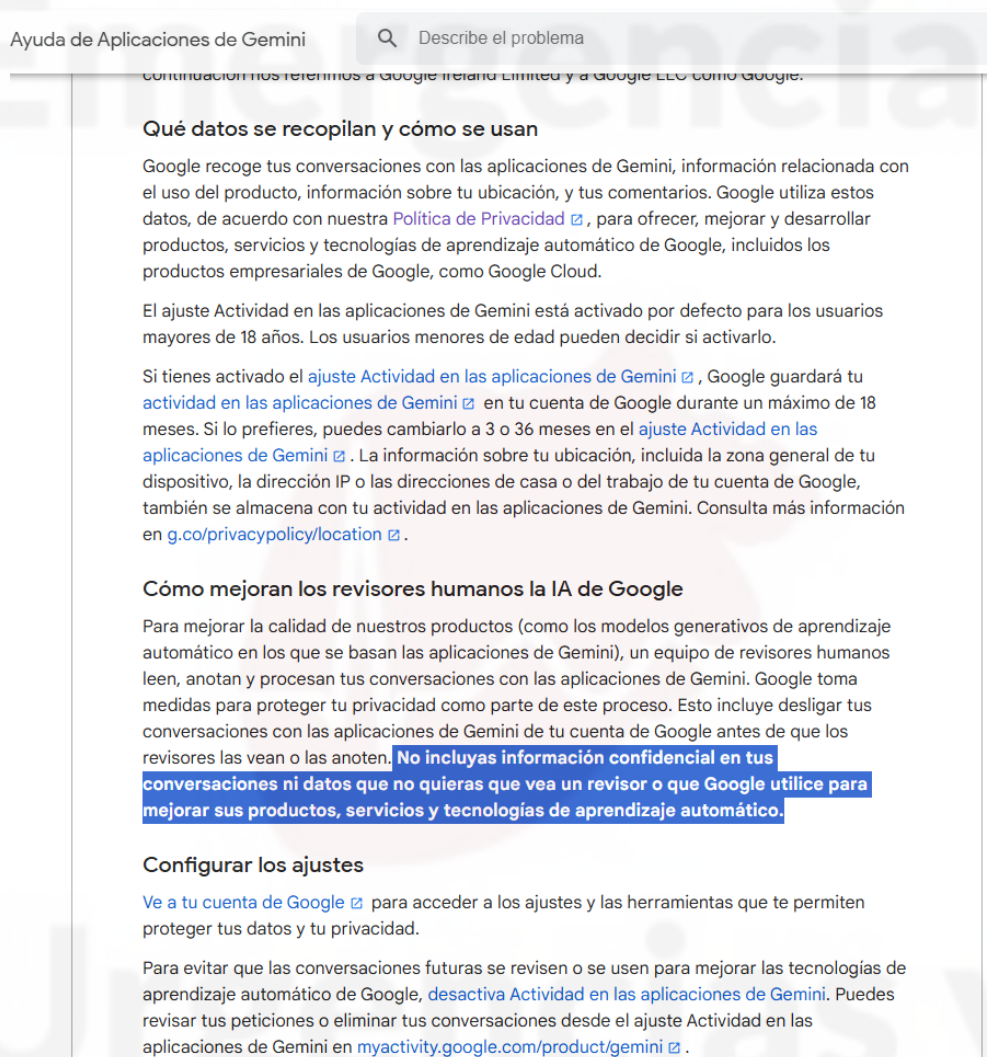


**Imagen 51. Captura de pantalla de la interfaz de Google Gemini y los modelos disponibles. Enero 2025**

A pesar de que me en muchos foros y congresos de IA, se presenta Gemini como el modelo más seguro y especializado para trabajar o introducir datos de alumnos o pacientes, no es así. Podéis siempre acudir a las políticas de privacidad de todas estas herramientas. En concreto, en las políticas de privacidad y datos de Google Gemini, nos informan de que guarda los datos 18 meses, a no ser que le digamos que los guarde menos (mínimo 3 meses) y su advertencia en su propia web es clara al respecto:

**“No incluyas información confidencial en tus conversaciones ni datos que no quieras que vea un revisor o que Google utilice para mejorar sus productos, servicios y tecnologías de aprendizaje automático.**

Eso sí, como pasa con otros modelos de lenguaje, puedes configurar los ajustes para evitar que conversaciones futuras se revisen o se usen para mejorar las tecnologías de aprendizaje automático en Google. Puedes ver todo esto en la Imagen 52. Pero con esta opción, en los diferentes LLM dejan de funcionar cosas. Con Google Gemini no podrás crear imágenes.

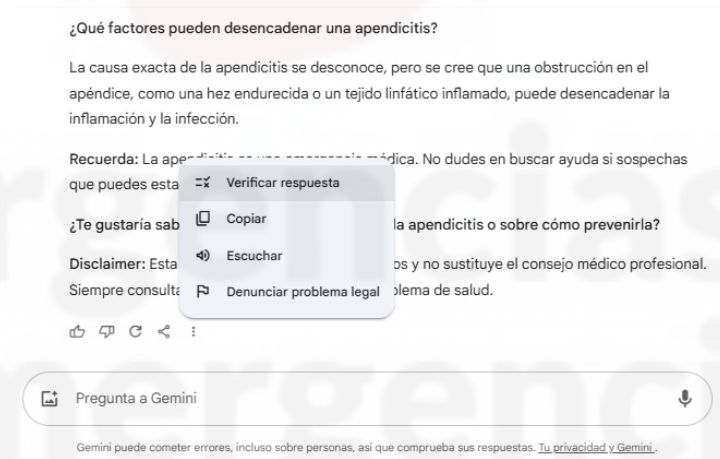


**Imagen 52. Captura de pantalla de la política de datos de Google Gemini. Disponible en:**

[https://support.google.com/gemini/answer/13594961?hl=es&ref\\_topic=13278591&sjid=6223584219093404158-EU](https://support.google.com/gemini/answer/13594961?hl=es&ref_topic=13278591&sjid=6223584219093404158-EU)

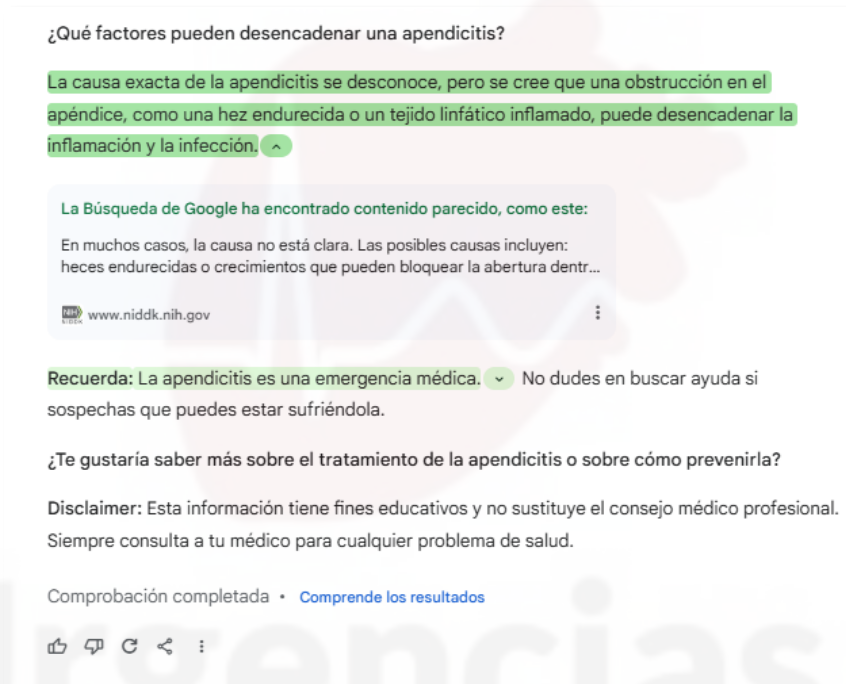
Su versión gratuita nos permite también hablar por voz e introducir imágenes para su análisis. También tiene conexión a internet permitiendo ver las fuentes de lo que nos ofrece en texto. Y puede leernos las respuestas con voz.

Dispone de una funcionalidad final que consiste en comprobar la respuesta que ofrece Google Gemini con el buscador de Google normal. Una vez realizada la comprobación, nos muestra en verde las fuentes de donde ha obtenido la información. Hay que elegir la opción “verificar respuesta” (Imagen 53).



**Imagen 53. Función de verificar respuesta con Gemini. Enero 2025.**

Una vez realizada la verificación, Google marca en verde las afirmaciones que ha dado que puede comprobar en Internet; en naranja si ha encontrado afirmaciones contrarias o no relevantes (Imagen 54).



**Imagen 54. Captura de pantalla de la verificación de respuestas de Google Gemini. Enero 2025.**

## 5.6.4 PERPLEXITY

Podríamos decir que Perplexity es el "hijo" de ChatGPT y Google. Más que un LLM, se nutre de uno propio y de varios de los explicados en este tema. Se trata de un buscador basado en IA en que le introduce una pregunta como a Google, pero te redacta una respuesta de unos pocos párrafos, muy concreta, aportado las fuentes.

Tal y como indican en su web:

*"Cuando le haces una pregunta, en lugar de darte una lista de enlaces para que la revises como un motor de búsqueda regular, Perplexity utiliza inteligencia artificial avanzada para entender exactamente lo que estás preguntando.*

*Después busca entre toneladas de páginas web y artículos para sacar los fragmentos de información más relevantes. Pero aquí está la parte realmente genial: Perplexity reúne toda esa información en una respuesta clara y concisa escrita en lenguaje sencillo, como si estuviera conversando contigo.*

*Así que digamos que preguntas "¿Cuáles son los beneficios para la salud del yoga?" Perplexity escaneará sitios web de salud y fitness autorizados, revistas médicas, blogs de yoga, etc. y compilará hechos clave en un resumen útil, citando las fuentes que utilizó. La respuesta busca ser comprensiva pero fácil de entender."*

Perplexity



Puedes entrar a través del siguiente enlace: <https://www.perplexity.ai/>

La descripción de su web es precisa: se está volviendo indispensable. Sin embargo, debemos ser cautelosos al usarla, al igual que al navegar por Internet. No todo en Internet es cierto, y esta herramienta depende de Internet para sus respuestas.



Imagen 55. Captura de pantalla de Perplexity.

También tiene un plan de pago. Pero en su versión gratuita nos permite subir 3 archivos al día, para preguntarle sobre los mismos. Si no pagas ningún plan en ningún LLM, puedes ir jugando con lo que te deja hacer ChatGPT, Perplexity, etc. Y también tenemos disponibles 3 preguntas “Pro” diarias, donde nos da resultados de una búsqueda entre muchos más enlaces.

Aporta también resultados en imágenes y en vídeos (esto avanza muy rápido y ahora también lo hacen todos, antes era el único) de Youtube en cada búsqueda y las aporta a la derecha de la pantalla (Imagen 55).

Una de sus funciones más destacadas es la capacidad de centrar la búsqueda en un ámbito específico. Para ello haremos clic en “Enfoque” (Imagen 56). Por ejemplo, se puede centrar solo la búsqueda en vídeos de YouTube en el selector de búsquedas. Esto es muy interesante en el ámbito de la salud para realizar una búsqueda más rápida sobre cualquier técnica en la que necesitamos ver un vídeo. También permite la búsqueda solo en documentos académicos, generar texto sin búsqueda web, resolver ecuaciones matemáticas y buscar discusiones y opiniones.



**Imagen 56. Opciones de búsqueda en el apartado “Enfoque” de Perplexity. Captura de pantalla. Enero 2025.**

Permite también la clasificación de las preguntas en colecciones, que ahora se llaman “Espacios”, donde podemos realizar lo que se explicó antes de GPT personalizados de ChatGPT. En los Espacios nos permiten subir pdf y dar una instrucción específica para que todas las conversaciones que mantenemos tengan siempre el mismo prompt. Por ejemplo, le puedes subir tres protocolos de tu centro de trabajo y darle la instrucción una sola vez que te genere resumen, lenguaje académico, etc. Y mantiene la misma estructura en todas las conversaciones.

### 5.6.5 CLAUDE (ANTHROPIC)

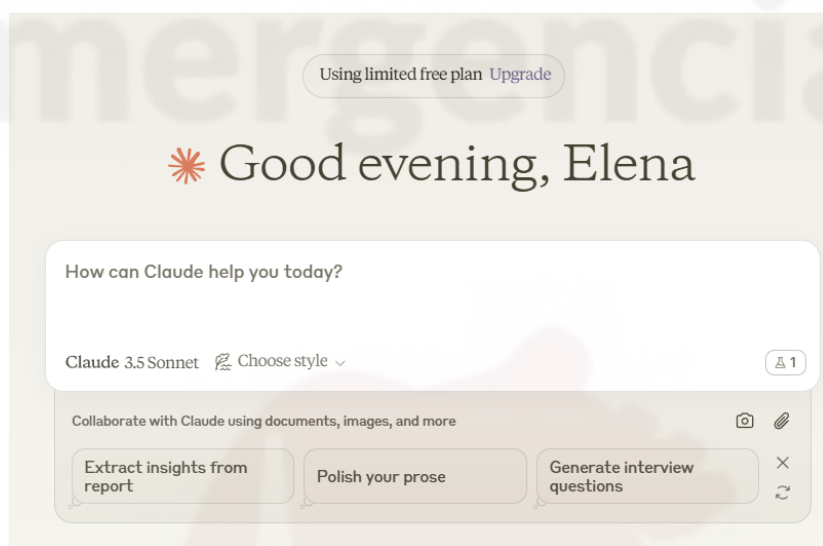
Claude es un modelo de lenguaje desarrollado por Anthropic, una empresa de investigación en inteligencia artificial que pone un fuerte énfasis en la ética y la seguridad. Claude ha sido diseñado para minimizar las respuestas dañinas y sesgadas, abordando directamente problemas de sesgo y seguridad que otros modelos pueden presentar.

Anthropic ha adoptado enfoques innovadores para el entrenamiento de Claude, incluyendo el uso de datos curados y técnicas de aprendizaje que priorizan la transparencia y la responsabilidad. Esto hace que Claude sea una opción ideal para aplicaciones donde la seguridad y la ética son primordiales.

Claude

Puedes entrar a través del siguiente enlace: <https://claude.ai/>

Su versión gratuita permite la de artículos e imágenes para poder pedirle que las explique. Pero sí que es cierto que permite pocas interacciones antes de que te salte el aviso de “no puedes volverlo a usar hasta pasadas varias horas”.



**Imagen 57. Interfaz de Claude. Enero 2025.**

Cuando usas Claude “se nota” distinto a ChatGPT u otros. Mientras que las empresas anteriores trabajan con modelos diferentes, aportan respuestas bastantes parecidas. Pero con Claude se nota la diferencia. Sus modelos son Claude 3.5 Sonnet y Claude 3.5 Haiku. La primera es su modelo más inteligente y la segunda su modelo más rápido. Su interfaz estaba en castellano, pero últimamente me sale en inglés. Pero responderá en castellano si así lo indicas.

Permite elegir varios estilos de respuesta e incluso personalizar uno propio (para que escriba como tú o para que de respuestas concretas). (Imagen 58)

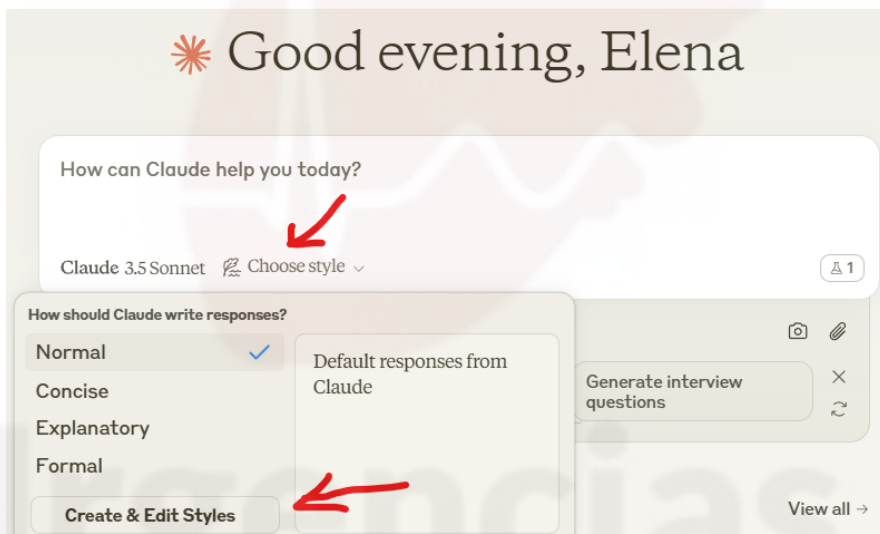


Imagen 58. Captura de pantalla de Claude, apartado "Elige estilo". Enero 2025.

Pero la mejor de las funcionalidades de Claude son los llamados "artifacts". Puedes pedirle que genere cosas interactivas, que puedes ir modificando para llegar a lo que quieres y, finalmente, compartir el enlace. Por ejemplo, le he introducido el siguiente prompt: "Genera una línea del tiempo dividida en fragmentos de dos minutos. Alrededor, pon fichas que incluir en el tablero anterior. Las fichas son: desfibrilación, adrenalina, compresiones, ventilaciones, capnografía, H y T. Con estos datos, tienes que generar un entorno interactivo en el que las fichas se puedan mover al tablero".

Vemos como ha generado unas fichas (Imagen 59) que se pueden deslizar al tablero para poder utilizarlo en docencia. También realiza presentaciones interactivas con gráficos de documentos que le aportes e infinitas posibilidades más. En este enlace podrás ver un ejemplo <https://claude.site/artifacts/bfe62154-7886-476c-ad12-0d85b3df4661> obtenido tras introducir el Informe del SNS 2023 de salud.

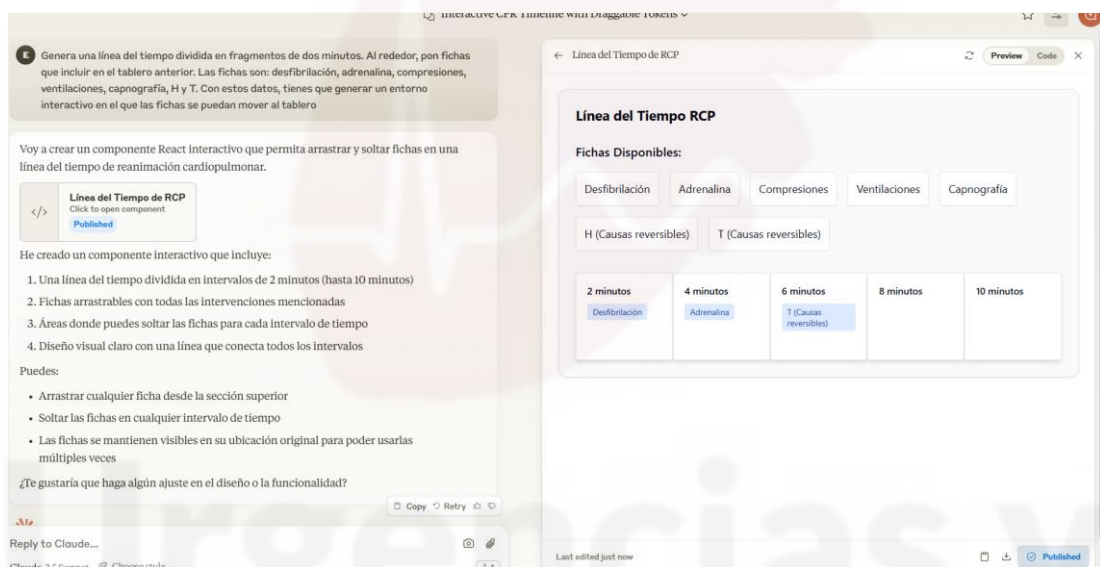


Imagen 59. Captura de pantalla de un artifact de Claude. Enero 2025.

### 5.6.7 PI (INFLECTION AI)

Pi es un modelo de lenguaje creado por Inflection AI, una compañía que se centra en desarrollar inteligencia artificial conversacional. Pi está diseñado para proporcionar interacciones naturales y contextualmente relevantes, mejorando la experiencia del usuario en aplicaciones de servicio al cliente, asistentes personales y más.

Inflection AI se enfoca en la personalización y en la capacidad del modelo para aprender y adaptarse a las preferencias individuales de los usuarios. Pi utiliza técnicas avanzadas de procesamiento del lenguaje natural y aprendizaje profundo para ofrecer respuestas coherentes y útiles en una variedad de contextos.

### 5.6.8 MEDGEMINI

Se trata de una nueva familia de modelos de próxima generación optimizados para el ámbito médico. Actualmente (jul 2024) se encuentran en desarrollo y no están disponibles al público.

Esta familia de modelos se basa en los modelos Gemini de Google mediante el ajuste de datos de salud no identificados y al mismo tiempo hereda las capacidades nativas de razonamiento, multimodal y de contexto largo de Gemini.

Tal y como nos cuenta Google:

*“Mejoramos el razonamiento clínico de nuestros modelos mediante la autoformación y la integración de búsqueda web, mejoramos el rendimiento multimodal mediante ajustes finos (fine-tuning) y codificadores personalizados, y utilizamos mejor las capacidades de contexto largo con indicaciones de cadena de razonamiento”.*

Aseguran que el modelo ha obtenido una puntuación más alta que otros modelos en las pruebas que le realizan para comprobar su precisión. Estas pruebas se llaman **benchmarks** y se explican en profundidad más adelante.

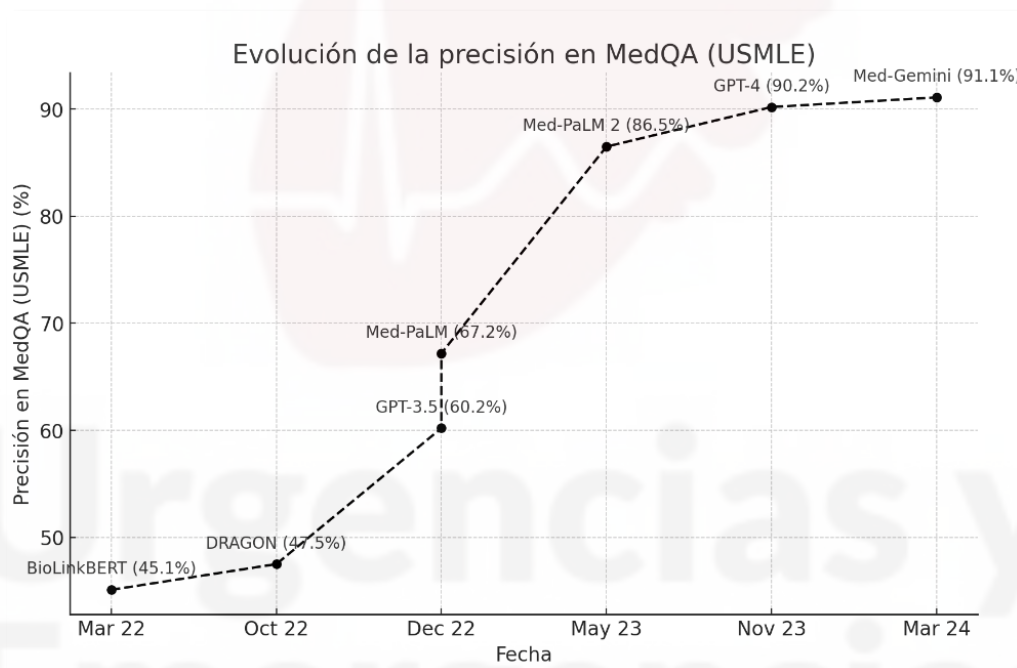
MedGemini es capaz de diagnosticar imágenes 2D, 3D, generar informes, etc. Además, afirman que el modelo Med-Gemini-Polygenic es el primer modelo de lenguaje que realiza predicciones de enfermedades y resultados de salud a partir de datos genómicos.

Google reconoce sus limitaciones, afirman que:

*“El 7.4% de las preguntas son consideradas no aptas para evaluación, ya que carecen de información clave, tienen respuestas incorrectas o admiten múltiples interpretaciones plausibles”.*

Por ello, han publicado dos artículos y continúan investigando para mejorar los modelos. De momento no está abierto al público.

Aunque la precisión todavía no sea lo suficientemente grande como para aplicar estos modelos en salud ahora mismo, quizás en un futuro muy próximo estén presentes en nuestro día a día como profesionales sanitarios. No hay nada más que ver la evolución tan increíble en la precisión que se ha realizado tan solo en dos años (Gráfica 6):




**Gráfica 6. Precisión de varios modelos de lenguaje en una de las pruebas a las que se les somete en el ámbito de la salud llamada MedQA (USMLE). Autora: Elena Plaza Moreno con datos de: <https://arxiv.org/pdf/2404.18416>**

## 5.7 BIBLIOGRAFÍA

- AI is poised to drive 160% increase in data center power demand.* (2024, mayo 14). Goldman Sachs. <https://www.goldmansachs.com/intelligence/pages/AI-poised-to-drive-160-increase-in-power-demand.html>
- Baum, D. (2024). *Generative AI and LLMs for dummies*. John Wiley & Sons, Inc. <https://www.snowflake.com/wp-content/uploads/2024/01/Generative-AI-and-LLMs-for-Dummies.pdf>
- Introducing ChatGPT* [Internet]. Openai.com. 2022 [citado el 19 de mayo de 2024]. Disponible en: <https://openai.com/index/chatgpt/>
- Hello GPT-4o* [Internet]. Openai.com. 2024 [citado el 19 de mayo de 2024]. Disponible en: <https://openai.com/index/hello-gpt-4o/>
- Hoyl M. *¿cómo funciona ChatGPT?* [Internet]. Descifr[ando] - Matías Hoyl. 2023 [citado el 19 de mayo de 2024]. Disponible en: <https://matias hoyl.beehiiv.com/p/como-funciona-chat-gpt>
- Vaswani A, Shazeer N, Parmar N, Uszkoreit J. *Attention is all you need* [Internet]. Neurips.cc. 2017 [citado el 19 de mayo de 2024]. Disponible en: [https://proceedings.neurips.cc/paper\\_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf)
- Hijas J. *¿cómo funciona ChatGPT?* [Internet]. LinkedIn.com. 2024 [citado el 19 de mayo de 2024]. Disponible en: <https://www.linkedin.com/pulse/c%C3%B3mo-funciona-chatgpt-jesus-hijas-el3uf/>

- Hoyl M. *La fábrica de palabras: dentro de la mente de ChatGPT* [Internet]. Descifr[ando] - Matías Hoyl. 2023 [citado el 19 de mayo de 2024]. Disponible en: <https://matias hoyl.beehiiv.com/p/dentro-de-la-mente-de-chatgpt>
- Mind, G. D. (s/f). *Advancing multimodal medical capabilities of Gemini*. Arxiv.org. Recuperado el 17 de julio de 2024, de <http://arxiv.org/abs/2405.03162>
- Mollick, E. (2024, 20 octubre). Thinking Like an AI. One Useful Thing. <https://www.oneusefulthing.org/p/thinking-like-an-ai>
- OpenAI. (2024). *ChatGPT 4* (versión del 15 de mayo) [Modelo de lenguaje de gran tamaño]. <https://chat.openai.com/chat>
- Open AI. (2024). Learning to reason with LLMs. Recuperado el 17 de enero de 2025, de Openai.com website: <https://openai.com/index/learning-to-reason-with-llms/>
- Saab, K., Tu, T., Weng, W.-H., Tanno, R., Stutz, D., Wulczyn, E., Zhang, F., Strother, T., Park, C., Vedadi, E., Chaves, J. Z., Hu, S.-Y., Schaeckermann, M., Kamath, A., Cheng, Y., Barrett, D. G. T., Cheung, C., Mustafa, B., Palepu, A., ... Natarajan, V. (s/f). *Capabilities of Gemini models in medicine*. Arxiv.org. Recuperado el 17 de julio de 2024, de <http://arxiv.org/abs/2404.18416>



## MÓDULO 2. IA GENERATIVA. MODELOS DE LENGUAJE GRANDE. CHATGPT.

### TEMA 6. ¿CÓMO HABLAR CON CHATGPT U OTROS MODELOS DE LENGUAJE? PROMPT ENGINEERING.

*«No olvidéis los fundamentos. No os enamoréis de la nueva tecnología, porque no es nueva. Solo es nuevo el medio en el que trabajamos, pero eso no cambia nada.*

*El arte en el que trabajamos es exactamente el mismo. Sobrepasa la tecnología».*

GEORGE LUCAS

## TEMA 6. ¿CÓMO HABLAR CON CHATGPT U OTROS MODELOS DE LENGUAJE? PROMPT ENGINEERING.

### 6.1 INTRODUCCIÓN AL PROMPT ENGINEERING

Para hablar con ChatGPT u otros modelos de lenguaje, se utiliza una técnica llamada "Prompt Engineering" o ingeniería de prompts. Esta técnica consiste en proporcionar instrucciones específicas o "prompts" a los modelos de lenguaje para obtener respuestas deseadas.

**Prompt Engineering, también conocido como ingeniería de prompts**, se refiere al proceso de proporcionar instrucciones específicas o "**prompts**" a estos modelos de lenguaje para obtener respuestas deseadas. Esta técnica es fundamental para obtener resultados precisos y relevantes al interactuar con sistemas de inteligencia artificial basados en lenguaje natural.

Es importante destacar que **ChatGPT y los otros modelos de lenguaje grandes no funcionan como Google**. Como se ha visto con anterioridad, el funcionamiento de los LLM hace que funcionen a través de una predicción de letras o tokens. Por lo que incluir contexto y otras técnicas, mejorará los resultados que queremos e influirá en las respuestas generadas por estos sistemas.

Sin embargo, la introducción de preguntas o palabras clave, como se suele hacer con buscadores webs como Bing o Google, puede ofrecer malos resultados. Es evidente que, si les preguntamos por la capital de Francia, nos dirá que es París, porque lo tiene precargado en su conocimiento. Pero este uso como fuente de datos fiable basándose en recopilación de datos, no es el correcto en cuanto a los LLM.

Posiblemente esto cambie en un futuro próximo, donde en vez de por escrito hablaremos de forma fluida con los chatbots y la forma de comunicarnos con ellos cambie. Pero de momento, y sobre todo en salud, la clave de saber utilizar bien ChatGPT o cualquier modelo de lenguaje es saber hablar con ellos.

Lo mejor para hablar con estos LLM para obtener buenos resultados y que nos ayuden a realizar tareas, es formular los prompts de manera adecuada. De esta forma se puede orientar al modelo hacia una dirección particular, asegurando que las respuestas proporcionadas sean coherentes, relevantes y útiles para el usuario. Sin una guía clara en forma de prompt, los modelos de lenguaje pueden generar respuestas ambiguas, irrelevantes o incluso incorrectas.

### 6.2 PASOS Y CONSEJOS PARA UTILIZAR PROMPT ENGINEERING

Para utilizar prompt engineering al interactuar con modelos de lenguaje como ChatGPT, es importante seguir algunos pasos específicos para lograr resultados efectivos que se describen a continuación:

- **Definir el objetivo de la interacción:** antes de empezar a crear prompts, es fundamental tener claro cuál es el propósito de la conversación con el modelo de lenguaje. ¿Quieres obtener información

específica, generar texto creativo, o simplemente entablar una conversación informal? Asegúrate de definir claramente lo que esperas lograr con la interacción.

- **Conocer las capacidades del modelo:** es importante familiarizarse con las capacidades y limitaciones del modelo de lenguaje que vas a utilizar. Cada modelo puede tener diferentes áreas de especialización o temas en los que destaca, por lo que es fundamental tener esta información en cuenta al diseñar tus prompts. Prueba varios LLM para ver cuál es el que mejor se adapta a tus necesidades.
- **Sé específico y claro:** a la hora de crear prompts, es recomendable utilizar un lenguaje claro y conciso. Evita términos ambiguos o preguntas demasiado generales que puedan confundir al modelo. Utiliza un tono adecuado para la conversación que deseas tener. En lugar de plantear preguntas abiertas, intenta ser lo más preciso posible en tus instrucciones. Por ejemplo, en lugar de preguntar "¿qué piensas sobre el cambio climático?", podrías utilizar: "por favor, proporciona tres razones por las cuales el cambio climático está afectando a nuestro planeta".
- **Proporciona contexto:** asegúrate de que tu prompt incluya suficiente contexto para que el modelo entienda la situación o el escenario en el que debe trabajar.
- **Usa ejemplos:** incluye ejemplos específicos de lo que esperas obtener como respuesta. Esto ayuda al modelo a alinearse mejor con tus expectativas.
- **Experimentar con diferentes enfoques, revisa y ajusta:** una vez que hayas creado tus prompts iniciales, no tengas miedo de experimentar con diferentes enfoques para ver qué funciona mejor. Puedes probar variaciones en la redacción de las instrucciones o en la estructura de las preguntas para obtener respuestas diferentes.
- **Divide tareas complejas:** si la tarea es compleja, divídela en pasos más pequeños y manejables. Esto facilita que el modelo siga el flujo y produzca resultados más precisos.
- **Aprovecha los parámetros:** experimenta con diferentes parámetros como la temperatura, el máximo número de tokens y otros ajustes para encontrar la configuración que mejor se adapte a tus necesidades.
- **Utiliza lenguaje natural:** escribe tus prompts de manera conversacional y natural. Esto puede ayudar al modelo a generar respuestas más fluidas y coherentes.
- **Anticipa posibles malentendidos:** piensa en posibles interpretaciones erróneas de tu prompt y ajústalo para evitar confusiones.
- **Mantén la coherencia:** si estás trabajando en un proyecto a largo plazo, mantén una estructura coherente en tus prompts. Esto ayuda al modelo a entender mejor el contexto continuo.
- **Documenta tus iteraciones:** lleva un registro de los diferentes prompts que has probado y los resultados obtenidos. Esto te permitirá aprender de la experiencia y refinar tus técnicas.

## 6.3 AJUSTE DE PARÁMETROS. TÉCNICAS AVANZADAS DE PROMPTING

### 6.3.1 TEMPERATURA

La **temperatura** es un parámetro que ajusta la aleatoriedad de las predicciones de un modelo de lenguaje. Cuando se utiliza una temperatura baja (por ejemplo, 0.2), el modelo se vuelve más conservador y tiende a repetir palabras con menor probabilidad, resultando en respuestas más coherentes, pero menos creativas.

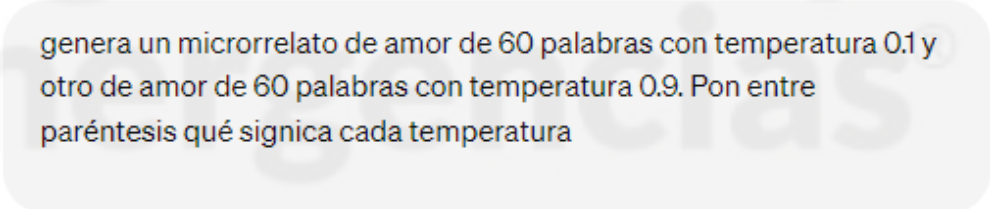
Por otro lado, una temperatura alta (por ejemplo, 1.0) introduce más variabilidad y creatividad en las respuestas, aunque también aumenta el riesgo de generar respuestas menos coherentes o erróneas.

**Esto puede ser muy útil en el ámbito de la salud donde no queremos que ChatGPT se ponga creativo y sea un poco más serio y concreto.**

Vamos a verlo con un ejemplo:

- Temperatura baja (0.2):
  - Prompt: "Describe el clima hoy."
  - Respuesta: "El clima hoy es soleado y cálido con una ligera brisa."
- Temperatura alta (1.0):
  - Prompt: "Describe el clima hoy."
  - Respuesta: "El clima de hoy es una mezcla vibrante de sol, ráfagas esporádicas de viento y ocasionales nubes, creando una atmósfera dinámica y cambiante."

Y ahora con un texto un poco más largo en la Imagen 60:



genera un microrrelato de amor de 60 palabras con temperatura 0.1 y otro de amor de 60 palabras con temperatura 0.9. Pon entre paréntesis qué significa cada temperatura

Claro, Elena. Aquí tienes los microrrelatos de amor:

Temperatura 0.1 (Baja temperatura, texto coherente):

En el rincón de un café, sus miradas se encontraron. Cautivados, comenzaron a hablar, compartiendo risas y secretos. Sus corazones latían al unísono, tejieron un lazo inquebrantable. El amor creció, eterno como las estrellas.

Temperatura 0.9 (Alta temperatura, texto creativo y diverso):

En un mundo distante, dos almas chispeantes se colisionaron como asteroides ardientes. Se entrelazaron en un torbellino de pasión, donde el tiempo se distorsionó. El amor era un fuego salvaje que consumía todo, fusionando sus destinos en un abrazo ardiente y eterno, desafiando las leyes del universo.

Imagen 60. Captura de pantalla con ejemplo de variación de temperatura. ChatGPT 4o

### 6.3.2 TOP P (NUCLEUS SAMPLING)

**Top P**, también conocido como *nucleus sampling*, es una técnica utilizada en los modelos de lenguaje para controlar la selección de tokens (las palabras o partes de palabras que el modelo puede generar) durante el

proceso de generación de texto. **Este método ayuda a equilibrar la creatividad y la coherencia en las respuestas generadas.**

Para entender cómo funciona, imagina que el modelo tiene que predecir la siguiente palabra en una oración y asigna probabilidades a cada posible palabra. Supongamos que tenemos las siguientes probabilidades para las palabras "el", "un", "es", "escribe" y "hola":

- "el": 0.15
- "un": 0.10
- "es": 0.08
- "escribe": 0.07
- "hola": 0.05

La suma de las probabilidades de todas las posibles palabras es 1 (o 100%). La probabilidad acumulativa se refiere a la suma de las probabilidades de las palabras seleccionadas hasta un cierto punto. Por ejemplo, la probabilidad acumulativa de seleccionar "el" sería 0.15, y si luego seleccionamos "un", la acumulativa sería 0.25 (0.15 + 0.10).

Cuando configuramos el valor de Top P, estamos estableciendo un umbral para esta probabilidad acumulativa. Por ejemplo, con un Top P de 0.9, el modelo considerará únicamente las palabras cuya probabilidad acumulativa totalice hasta el 90%. Siguiendo el ejemplo anterior, el modelo seleccionaría palabras hasta alcanzar un acumulado de 0.90, es decir, podría incluir palabras como "el", "un", "es", "escribe", y tal vez algunas más, pero no incluiría palabras que harían que la acumulativa superara el 90%.

Por otro lado, si configuramos un Top P de 0.2, el modelo sería más restrictivo y solo consideraría palabras hasta alcanzar una probabilidad acumulativa del 20%. Esto resultaría en una selección de palabras más limitada y conservadora, enfocándose en las opciones más probables.

Comparando Top P con la temperatura, esta última ajusta la distribución de probabilidades de todos los posibles tokens, haciendo que algunos sean más probables y otros menos, pero no excluye tokens completamente.

En cambio, Top P excluye completamente todos los tokens cuya probabilidad acumulativa excede el umbral definido, limitando la generación a una parte específica de la distribución de probabilidades.

En la práctica, ajustar Top P puede ayudar a equilibrar creatividad y coherencia en las respuestas. **Un valor más bajo de Top P (como 0.2) hará que el modelo genere respuestas más coherentes y conservadoras, mientras que un valor más alto (como 0.9) permitirá respuestas más creativas y variadas, aunque con un riesgo potencial de ser menos coherentes (Imagen 61).**

escribete un mini relato de amor con Top P 0,2 y otro con 0,9.

Claro, Elena. Aquí tienes dos mini relatos de amor con diferentes ajustes de creatividad:

### Relato con Top P 0.2

Lucía y Andrés se conocieron en una pequeña librería del centro. A ambos les encantaba perderse entre las páginas de los libros antiguos. Una tarde, mientras buscaban algo interesante para leer, sus manos se encontraron al alcanzar el mismo libro. Sorprendidos, intercambiaron sonrisas tímidas. Día tras día, se vieron en la librería, compartiendo sus pasiones literarias y descubriendo que sus almas estaban conectadas por hilos invisibles. Poco a poco, su amistad se convirtió en algo más, y un atardecer, entre las sombras de la librería, Andrés tomó la mano de Lucía y la besó. Ese beso selló su amor eterno, un amor nacido entre libros y susurros.

### Relato con Top P 0.9

En una playa escondida, donde las olas susurraban secretos al viento, Marina conoció a Leo. Él estaba construyendo un castillo de arena enorme, lleno de detalles y torres impresionantes. Marina, fascinada, se acercó y ofreció su ayuda. Juntos, pasaron horas moldeando la arena, riendo y creando historias fantásticas sobre el castillo y sus habitantes imaginarios.

Al caer la noche, encendieron una pequeña fogata y se sentaron a su alrededor, compartiendo sus sueños y deseos. Las estrellas comenzaron a brillar en el cielo, y Marina, sin pensarlo dos veces, le tomó la mano a Leo. Él la miró, sonrió y se inclinó para besarla. Fue un beso lleno de promesas y aventuras por venir, un amor que nació espontáneamente en esa playa mágica, donde todo parecía posible.

Imagen 61. Captura de pantalla con ejemplo de variación de Top P o nucleus sampling. ChatGPT 4o

## 6.3.3 TOP-K SAMPLING

El **top-k sampling** es una técnica en la que el modelo de lenguaje considera solo las  $k$  palabras más probables en cada paso de generación de texto.

En otras palabras, en lugar de considerar todas las posibles palabras siguientes, el modelo limita su elección a las  $k$  palabras con mayor probabilidad.

Esto ayuda a reducir la aleatoriedad y puede mejorar la coherencia de las respuestas generadas, ya que el modelo selecciona entre un conjunto más restringido de opciones (Imagen 62).

**También puede ayudarnos en salud si ponemos top K bajo ya que concreta y usa un lenguaje menos creativo.**

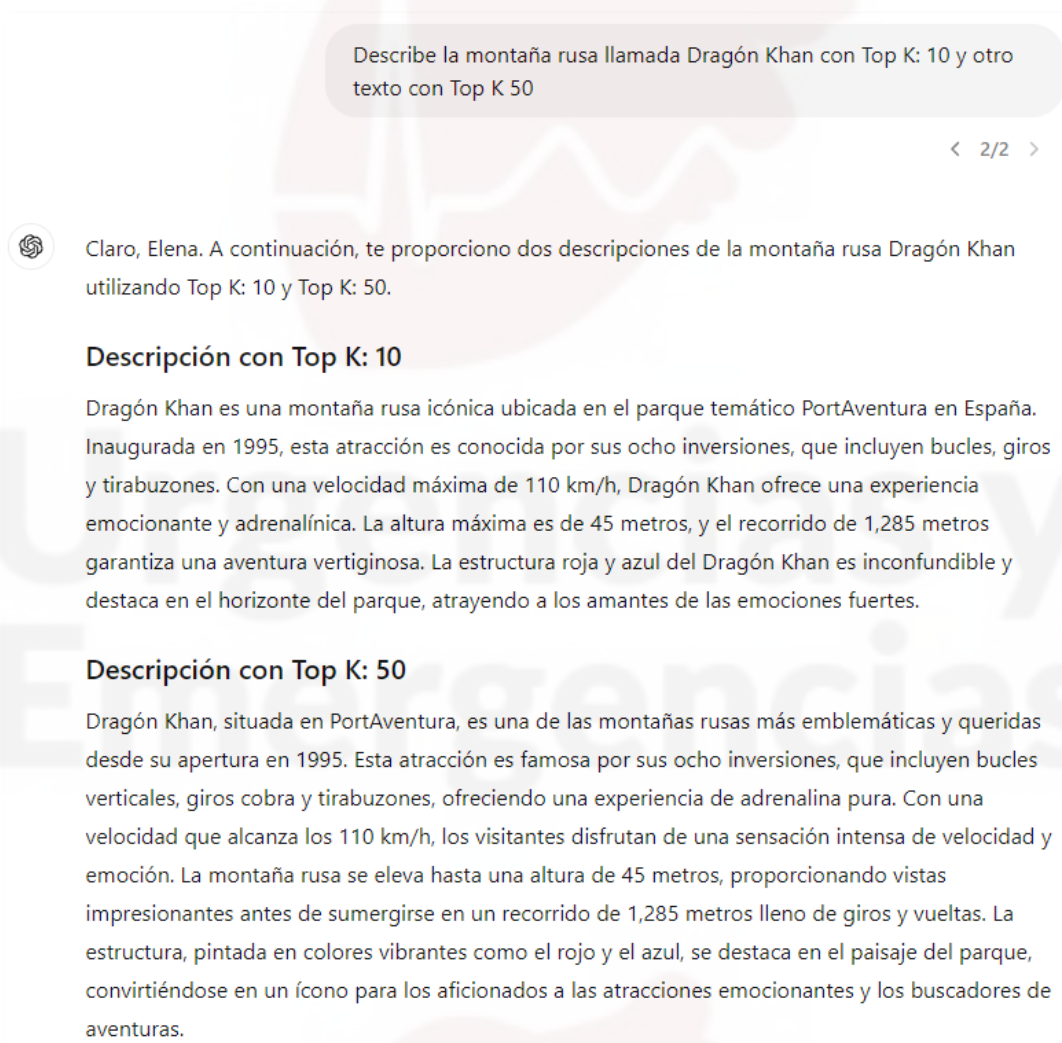


Imagen 62. Captura de pantalla con ejemplo de variación de Top K. ChatGPT 4o

### 6.3.4 PENALIZACIÓN POR PRESENCIA (PRESENCE PENALTY)

La "**penalización por presencia**" (*presence penalty*) es un parámetro que ajusta la probabilidad de que el modelo utilice palabras o tokens que aún no han aparecido en la respuesta generada. En otras palabras, **este parámetro incentiva al modelo a introducir nuevos temas o ideas, en lugar de repetir o reutilizar las mismas palabras.**

Cuando un modelo de lenguaje genera una respuesta, puede tender a repetir ciertas palabras o frases si se le permite. **El parámetro de presence penalty ayuda a mitigar este comportamiento al penalizar la probabilidad de elegir palabras que ya han aparecido en el texto generado.** Al aplicar una penalización positiva, el modelo se ve incentivado a diversificar su elección de palabras y, por lo tanto, a generar contenido más variado y novedoso.

Por ejemplo:

- **Presence penalty = 0:** si no se aplica ninguna penalización (o la penalización es 0), el modelo no tiene restricciones adicionales y puede repetir palabras o temas con mayor frecuencia. Esto puede resultar en respuestas menos creativas y más redundantes.

- **Presence penalty = 0.2:** al aplicar una penalización de 0.2, el modelo tiene un incentivo moderado para evitar repetir palabras que ya han aparecido en la respuesta. Esto fomenta la introducción de nuevas ideas y temas, haciendo que la respuesta sea más interesante y diversa.
- **Presence penalty = 0.9:** con una penalización de 0.9, el incentivo para evitar la repetición de palabras es aún mayor. Esto lleva al modelo a generar contenido más novedoso, asegurándose de que las respuestas incluyan una mayor variedad de vocabulario y temas.

En la Imagen 63 se puede ver un ejemplo práctico. En este ejemplo, el parámetro **presence penalty: 0.9** asegura que el modelo se esfuerce por evitar la repetición de las mismas palabras, proporcionando una respuesta más variada y original, frente al presence penalty 0.2 donde repite la palabra flores o sus derivados en varias ocasiones.



Imagen 63. Captura de pantalla con ejemplo de variación de penalización por presencia. ChatGPT 4o

En resumen, con el ajuste de este parámetro podemos:

- Fomentar la creatividad en las respuestas del modelo.
- Introducir nuevas ideas y temas en la conversación.
- Evitar la repetición excesiva de palabras y frases.

### 6.3.5 PENALIZACIÓN DE FRECUENCIA (FREQUENCY PENALTY)

La "**penalización por frecuencia**" (*frequency penalty*) (Imagen 64) es un parámetro que ajusta la probabilidad de que el modelo repita palabras o frases que ya ha utilizado en la conversación. **Al aplicar una penalización por frecuencia, el modelo se ve incentivado a diversificar su elección de palabras, lo que resulta en respuestas menos repetitivas y más variadas.**

Cuando el modelo de lenguaje genera texto, puede tener la tendencia a repetir ciertos tokens (palabras o partes de palabras) si no se aplica ninguna penalización. La penalización por frecuencia añade un "costo" a la reutilización de palabras que ya han aparecido en la conversación, disminuyendo así su probabilidad de ser seleccionadas nuevamente.

Por ejemplo:

- **Frequency penalty = 0:** sin penalización por frecuencia, el modelo puede repetir palabras o frases con mayor frecuencia, lo que puede llevar a respuestas redundantes y menos creativas. Por ejemplo, si se le pide describir un día soleado, podría decir: "El día es soleado. El sol brilla. El día es muy soleado."
- **Frequency penalty = 0.5:** al aplicar una penalización de 0.5, el modelo reduce la probabilidad de repetir las mismas palabras. En el mismo ejemplo del día soleado, la respuesta podría ser: "El día es soleado. El sol brilla intensamente. El clima es muy agradable."
- **Frequency penalty = 1.0:** con una penalización más alta, como 1.0, el modelo tiene un incentivo aún mayor para evitar repeticiones, resultando en una mayor diversidad de vocabulario. Por ejemplo: "El día es soleado. La luz del sol ilumina todo. El ambiente es perfecto para actividades al aire libre."

Por tanto, con el ajuste de este parámetro, podemos:

- Reducir la redundancia en las respuestas del modelo.
- Aumentar la diversidad y riqueza del vocabulario utilizado.
- Ofrecer respuestas más interesantes y menos repetitivas.

#### **No hay que confundir el parámetro frequency penalty con el presence penalty:**

- ▶ Penalización por presencia (*presence penalty*): se centra en penalizar el uso de tokens nuevos en la respuesta actual, incentivando la introducción de nuevos temas o ideas.
- ▶ Penalización por frecuencia (*frequency penalty*): se aplica a lo largo de toda la conversación, penalizando la repetición de palabras que ya se han utilizado, promoviendo una mayor variedad en el vocabulario.

Escribe dos minirelatos de amor, uno con Frequency penalty = 0 y frequency penalty = 1.0; y explícame debajo de cada uno qué parámetros del texto o por qué se ve la modificación de frequency penalty

< 2/2 >



#### Frequency penalty = 0

Ana se encontraba en el parque, disfrutando de un libro bajo la sombra de un árbol. De repente, un balón de fútbol cayó a sus pies. Al levantar la vista, vio a un chico apuesto correr hacia ella. Él se disculpó con una sonrisa encantadora y recuperó el balón. En ese breve instante, algo mágico sucedió. Ana sintió una conexión inmediata. Pasaron semanas encontrándose en el parque, siempre por casualidad. Cada encuentro era una oportunidad para conocerse mejor. Finalmente, un día, él la invitó a tomar un café. Así comenzó una historia de amor que ambos sabían que duraría para siempre.

*Explicación:* Con frequency penalty en 0, el relato usa palabras y frases comunes de forma repetida. Los encuentros casuales y la sensación de conexión inmediata son elementos recurrentes en historias románticas, lo que hace que el texto se sienta familiar y predecible. La narrativa fluye sin sorpresas, siguiendo un patrón clásico y esperado.

#### Frequency penalty = 1.0

En un rincón olvidado de la ciudad, en una pequeña librería llena de polvo y secretos, Clara descubrió un libro antiguo. Al abrirlo, una carta cayó de entre las páginas. Era una confesión de amor escrita hace décadas, pero nunca enviada. Intrigada, Clara decidió buscar al autor. Su investigación la llevó a una residencia de ancianos donde conoció a Eduardo, un hombre cuya memoria se había ido desvaneciendo con los años. Al leerle la carta, una chispa de juventud brilló en sus ojos. Clara y Eduardo comenzaron a verse regularmente, y en esos encuentros, Clara encontró un tipo de amor inesperado, profundo y transformador, que no buscaba, pero que ahora no podía imaginar su vida sin él.

*Explicación:* Con frequency penalty en 1.0, el relato utiliza menos clichés y evita repetirse, creando una historia más única y original. La trama se desvía de los tópicos comunes de amor inmediato y se enfoca en una conexión profunda y no convencional. Los detalles y el desarrollo de la historia son más impredecibles, lo que da lugar a un relato más creativo e inesperado.

Imagen 64. Captura de pantalla con ejemplo de variación de penalización por frecuencia. ChatGPT 4o

### 6.3.6 PENALIZACIÓN DE REPETICIÓN (REPETITION PENALTY)

La **penalización de repetición** (*repetition penalty*) es una técnica utilizada en modelos de lenguaje para reducir la probabilidad de que se repitan las mismas palabras o frases en el texto generado.

Esta técnica se aplica ajustando la probabilidad de seleccionar una palabra que ya ha sido utilizada anteriormente en la generación de texto. Al penalizar estas repeticiones, se incentiva al modelo a utilizar un vocabulario más variado y a crear respuestas más fluidas y menos redundantes.

El objetivo principal de la penalización de repetición es mejorar la calidad del texto generado, haciéndolo más interesante y diverso. Esta técnica es especialmente útil en contextos donde se espera una narrativa continua o una explicación detallada sin repeticiones innecesarias.

Funciona de la siguiente manera: durante el proceso de generación de texto, el modelo rastrea las palabras que ya han sido utilizadas. Se ajusta la probabilidad de las palabras repetidas. Las palabras que ya han aparecido reciben una penalización, reduciendo así su probabilidad de ser seleccionadas nuevamente. El modelo selecciona la siguiente palabra basándose en las probabilidades ajustadas, favoreciendo aquellas que no han sido utilizadas o que han sido utilizadas menos frecuentemente.

Para aplicar la penalización de repetición, puedes ajustar el parámetro repetition penalty al interactuar con el modelo, como se ve en la Imagen 65:

- Claro, Elena. Aquí tienes otro ejemplo describiendo los beneficios del ejercicio físico regular, con penalización por repetición de 1 y con penalización por repetición de 2.

#### **Penalización por repetición de 1:**

##### **Beneficios del ejercicio físico regular:**

El ejercicio físico regular es muy beneficioso para la salud. Hacer ejercicio regularmente mejora la salud cardiovascular y fortalece el corazón. También ayuda a controlar el peso y a mantener un peso saludable. El ejercicio regular reduce el riesgo de enfermedades crónicas, como la diabetes y la hipertensión. Además, el ejercicio físico regular mejora el estado de ánimo y reduce el estrés. Hacer ejercicio regularmente también mejora la calidad del sueño y la energía diaria. En resumen, el ejercicio físico regular es esencial para una vida saludable.

#### **Penalización por repetición de 2:**

##### **Beneficios del ejercicio físico regular:**

Practicar ejercicio de forma regular ofrece múltiples ventajas para la salud. La actividad física constante mejora la condición cardiovascular, fortaleciendo el corazón y los vasos sanguíneos. También contribuye al control del peso, ayudando a mantener una figura saludable. Además, el ejercicio disminuye el riesgo de padecer enfermedades crónicas, como la diabetes y la hipertensión. El impacto positivo en el estado de ánimo es notable, ya que el ejercicio ayuda a reducir el estrés y la ansiedad. También se observa una mejora en la calidad del sueño y un aumento en los niveles de energía. En definitiva, la práctica regular de ejercicio es crucial para mantener una vida sana y equilibrada.

**Imagen 65. Captura de pantalla con ejemplo de variación de penalización por repetición. ChatGPT 4o**

En el ejemplo con penalización por repetición de 1, la palabra "ejercicio" y la frase "ejercicio físico regular" se repiten frecuentemente. El texto es más redundante y menos variado.

En el segundo texto, con penalización por repetición de 2, las repeticiones de palabras y frases se reducen significativamente. Se utilizan sinónimos y reformulaciones, como "actividad física constante" en lugar de "ejercicio físico regular". Y el texto resulta más diverso y natural.

Por tanto, **al aplicar la penalización de repetición, se obtiene un texto más variado y coherente, lo que mejora significativamente la experiencia del usuario y la calidad de la información proporcionada.**

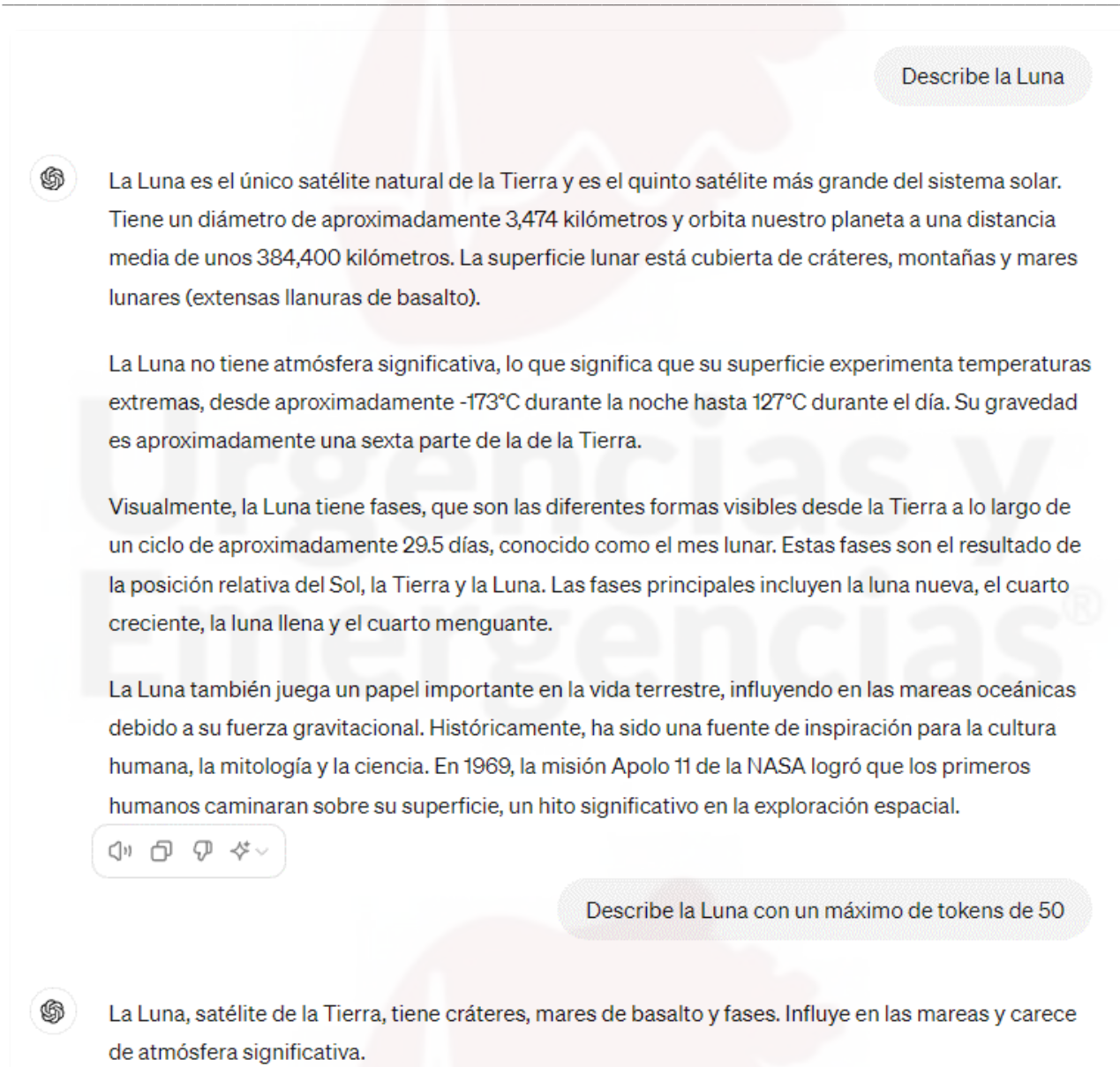
En la Tabla 3, se presenta una tabla resumen con los valores mínimos y máximos de cada parámetro y su definición.

RESUMEN DE AJUSTES DE PARÁMETROS			
PARÁMETRO	VALOR MÍNIMO	VALOR MÁXIMO	DEFINICIÓN
<b>Temperatura</b>	0.0	1.0	Controla la aleatoriedad de las respuestas: valores bajos generan respuestas más predecibles y valores altos generan respuestas más variadas.
<b>Top-k Sampling</b>	1.0	100.0	Limita el número de palabras candidatas a considerar para la generación de la siguiente palabra.
<b>Top-p Sampling</b>	0.0	1.0	Selecciona palabras candidatas basadas en una probabilidad acumulativa, permitiendo considerar solo las más probables hasta que se alcanza un umbral.
<b>Penalización por repetición (repetition penalty)</b>	1.0	2.0	Aumenta la penalización por repetir las mismas frases o palabras, haciendo que el modelo sea menos propenso a repetir.
<b>Penalización por presencia (presence penalty)</b>	-2.0	2.0	Penaliza las palabras nuevas basándose en su presencia en el texto hasta el momento, promoviendo la introducción de términos únicos.
<b>Penalización por frecuencia (frequency penalty)</b>	-2.0	2.0	Penaliza las palabras nuevas basándose en su frecuencia en el texto hasta el momento, reduciendo la probabilidad de que se repitan términos comunes.

Tabla 3. Resumen de parámetros de escritura. Autora: Elena Plaza Moreno.

### 6.3.7 MÁXIMO NÚMERO DE TOKENS

Este parámetro limita la longitud de la respuesta generada (Imagen 66) al variar los **tokens**. Un valor más alto permite respuestas más largas.



Describe la Luna

La Luna es el único satélite natural de la Tierra y es el quinto satélite más grande del sistema solar. Tiene un diámetro de aproximadamente 3,474 kilómetros y orbita nuestro planeta a una distancia media de unos 384,400 kilómetros. La superficie lunar está cubierta de cráteres, montañas y mares lunares (extensas llanuras de basalto).

La Luna no tiene atmósfera significativa, lo que significa que su superficie experimenta temperaturas extremas, desde aproximadamente -173°C durante la noche hasta 127°C durante el día. Su gravedad es aproximadamente una sexta parte de la de la Tierra.

Visualmente, la Luna tiene fases, que son las diferentes formas visibles desde la Tierra a lo largo de un ciclo de aproximadamente 29.5 días, conocido como el mes lunar. Estas fases son el resultado de la posición relativa del Sol, la Tierra y la Luna. Las fases principales incluyen la luna nueva, el cuarto creciente, la luna llena y el cuarto menguante.

La Luna también juega un papel importante en la vida terrestre, influyendo en las mareas oceánicas debido a su fuerza gravitacional. Históricamente, ha sido una fuente de inspiración para la cultura humana, la mitología y la ciencia. En 1969, la misión Apolo 11 de la NASA logró que los primeros humanos caminaran sobre su superficie, un hito significativo en la exploración espacial.

Describe la Luna con un máximo de tokens de 50

La Luna, satélite de la Tierra, tiene cráteres, mares de basalto y fases. Influye en las mareas y carece de atmósfera significativa.

**Imagen 66. Captura de pantalla con ejemplo de variación de longitud de tokens. ChatGPT 4o**

A fecha de hoy, si le pides un texto de 2000 o 5000 palabras no te lo da directamente como resultado. Cuando se le piden extensiones largas no lo cumple. Sin embargo, esta función puede servir más para acotar o reducir. Y puede explicarse con lenguaje coloquial. Por ejemplo: "Genera una descripción sobre la diabetes en dos párrafos", sin llegar a hablarle de tokens en concretos.

### 6.3.8 "CREA TUS PROPIOS PARÁMETROS"

En ocasiones le he pedido a ChatGPT que escriba o que conteste a un email. Me ha parecido demasiado jovial (o cercano, o informal etc...). Tras ofrecerme el resultado, le he preguntado que del 1 al 10, puntúe su propio texto en términos de jovialidad. Si responde que es un 7, le puedes indicar en el siguiente prompt "reescribe el email con un índice de jovialidad del 2". Y se verá modificada la respuesta sustancialmente.

Además de este truco, Jorge Guillén García, docente de IA e ingeniero de prompts, indica que podemos darle nuestros propios parámetros a la hora de ajustar un texto. Cito textualmente de Jorge para indicarle a ChatGPT:

“Utiliza estas pautas para tu estilo de escritura”:

- **Formalidad:** 3/10 (1 = lenguaje casual, coloquial adecuado para entornos informales; 10 = lenguaje altamente formal, estructurado apropiado para contextos oficiales o académicos).
- **Estructura de la oración:** 3/10 (1 = uso predominante de oraciones simples, cortas, fáciles de seguir; 10 = uso extensivo de oraciones complejas, compuestas, potencialmente desafiantes para lectores promedio).
- **Elección de palabras:** 3/10 (1 = vocabulario básico, sencillo, fácilmente entendido por un público general; 10 = lenguaje técnico especializado o léxico sofisticado, más adecuado para lectores expertos).

## 6.4 ¿CÓMO HABLAR CON CHATGPT U OTROS LLM?

Para ordenar toda la información anterior, hay múltiples reglas mnemotécnicas que se pueden utilizar.

Una de ellas, por ejemplo, es la fórmula RACS (Imagen 67):

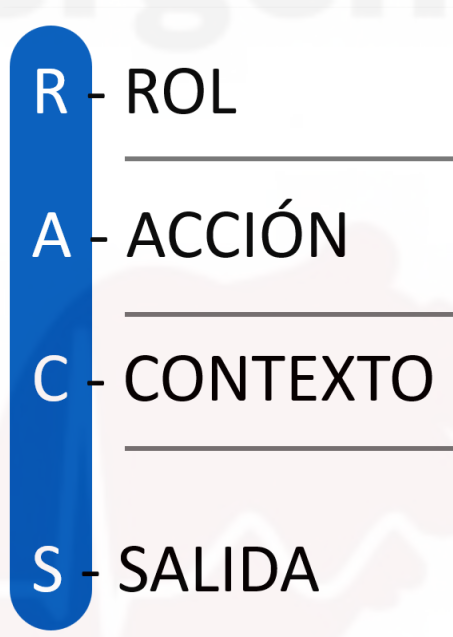


Imagen 67. Fórmula RACS para hablar con modelos de lenguaje grande.

Asignaremos un **rol** al iniciar la conversación. Por ejemplo: “actúa como un médico experto en medicina preventiva con 20 años de experiencia”. O, “actúa como un experto en redes sociales y escribe un post para Instagram sobre las medias de compresión”.

Respecto a la **acción**, indicaremos de forma concisa qué queremos conseguir. Puede ser generar una tabla, un texto, una reflexión, etc.... Por ejemplo: “genera un guion para una infografía sobre el tabaquismo”.

En **contexto**, daremos todas las indicaciones pertinentes para mejorar el resultado del prompt. Por ejemplo, si hemos solicitado la creación de una infografía, deberemos determinar el público “**orientado a fumadores menores de 30 años**”.

En **salida**, pediremos al chat que nos dé el resultado como queramos. Por ejemplo, “**crea una tabla**”, “**usa bullet points**”, “**dame el resultado en texto narrado**”.

Aquí podemos introducir también, si es necesario, el **ajuste de parámetros** y el **tono**. Podemos decirle que hable en un tono serio y formal, cercano, científico, que hable como Don Quijote de la Mancha o lo que se nos ocurra o necesitemos.

De hecho, si le preguntas a ChatGPT, qué tonos puede utilizar, los enumera (Imagen 68).



Imagen 68. Tonos de ChatGPT. Captura de pantalla ChatGPT 4o.

Un resumen de la estructura RACS sería la siguiente:

<b>R</b>	<b>ROL</b>	Eres, actúa como, habla como..., ponte en el el lugar de...
<b>A</b>	<b>ACCIÓN</b>	Genera, analiza, resume, escribe, describe, enumera...
<b>C</b>	<b>CONTEXTO</b>	De tema, de situación, de audiencia, de intención, de tiempo, de ubicación
<b>S</b>	<b>SALIDA</b>	Genera el output de esta forma: tabla, markdown, primero tabla, luego acciones y luego resultados, etc... Otros formatos de salida: temperatura, tono

**Imagen 69. Estructura RACS y descripción de cada apartado. Autora: Elena Plaza Moreno.**

Un ejemplo:

<b>R</b>	<b>ROL</b>	Eres un enfermero experto en cardiología con 10 años de experiencia en UCI coronaria
<b>A</b>	<b>ACCIÓN</b>	Genera el guión de unas diapositivas
<b>C</b>	<b>CONTEXTO</b>	que tengo que presentar en el próximo congreso de cardiología. Me dan 15 minutos para hablar sobre los últimos cambios en la reanimación cardiopulmonar. Los asistentes son también enfermeros.
<b>S</b>	<b>SALIDA</b>	Genera el output de esta forma: título de la diapositiva, tres ideas por cada diapositiva. También necesito una diapositiva final con las conclusiones. El lenguaje debe ser científico, técnico y profesional.

**Imagen 70. Ejemplo de prompt con estructura RACS. Autora: Elena Plaza Moreno.**

*Eres un enfermero experto en cardiología con 10 años de experiencia en UCI coronaria. Genera el guion de unas diapositivas que tengo que presentar en el próximo Congreso de Cardiología. Me dan 15 minutos para hablar sobre los últimos cambios en la reanimación cardiopulmonar. Los asistentes son también enfermeros. Genera el output de esta forma: título de la diapositiva, 3 ideas por cada diapositiva. También necesito una diapositiva final con las conclusiones. El lenguaje debe ser científico, técnico y profesional.*

El prompt puede parecer extenso; sin embargo, esta estructura permite obtener mejores resultados con ChatGPT.

Sin esta estructuración, es posible que ChatGPT u otros modelos no hayan cumplido tus expectativas. Es clave y fundamental hablar así a estos modelos de lenguaje para obtener buenos resultados.

## 6.5 PROMPTING EN SALUD

### 6.5.1 RECOMENDACIONES GENERALES Y CHECKLIST

Como vemos, las respuestas de los grandes modelos de lenguaje nunca serán iguales y la eficacia de estas herramientas depende de cómo interactuamos con ellas. Esta es la clave de grandes modelos de lenguaje y de ahí la importancia del 'prompting', el arte de formular preguntas o comandos que guíen a la IA para que proporcione las respuestas más útiles y precisas.

**A diferencia de los buscadores de Internet convencionales, donde a menudo usamos preguntas o palabras clave sueltas, la interacción con un modelo de lenguaje requiere una mayor especificidad y claridad.**

Para ilustrar esto, pongamos el caso de un médico que necesita información sobre el manejo de la diabetes tipo 2 en pacientes ancianos. Un prompt efectivo podría ser: *"Proporciona una síntesis de las pautas de tratamiento actualizadas para la diabetes tipo 2 en pacientes mayores de 65 años con comorbilidades cardíacas"*. Un prompt como este es directo y delimita la búsqueda a información específica, lo que ayuda a la IA a filtrar y centrarse en los datos más pertinentes.

Cabe destacar que, si bien los modelos de lenguaje son herramientas poderosas, también tienen limitaciones. No poseen la comprensión contextual profunda de un profesional sanitario y no deben reemplazar el juicio clínico. Por ello, es esencial comprender cómo y cuándo aplicar esta tecnología y cómo interpretar sus respuestas dentro del contexto más amplio de la atención al paciente.

**Hay que recordar en este punto que la capacidad de los modelos de lenguaje de generar texto coherente no equivale a una comprensión real de los conceptos médicos.** Los profesionales de la salud deben mantener un escepticismo y espíritu crítico y utilizar estas herramientas como complementos, nunca como sustitutos de su experiencia y conocimiento clínico.

Si utilizamos los grandes modelos de lenguaje aplicados a salud, debemos **definir muy bien el objetivo de la interacción con la IA** antes de formular cualquier prompt o pregunta. ¿Está buscando información para un diagnóstico? ¿Necesita asistencia para revisar literatura científica actualizada? ¿Está creando una lista de comprobación para un procedimiento quirúrgico? Hay que pararse a pensar y establecer un propósito claro. De esta forma se podrá diseñar prompts más dirigidos y obtener respuestas más relevantes.

Además, deberemos **contextualizar el prompt**, esto podría incluir detalles como comorbilidades, alergias conocidas, o preferencias de tratamiento del paciente.

Es importante **evitar la ambigüedad**: frases como "¿Podrías darme algún consejo?" son demasiado vagas. Es mejor preguntar: *"¿Cuál es el protocolo recomendado para la administración de insulina en pacientes diabéticos tipo 1 durante el ayuno preoperatorio?"*.

Utiliza la técnica de hacer **preguntas secuenciales y refina tus planteamientos** de manera iterativa para obtener respuestas precisas y detalladas. Asegúrate de ajustar tus preguntas según las respuestas recibidas para lograr la mejor calidad en las respuestas generadas, tal y como se ve en la Imagen 71.

**Pregunta inicial:**

¿Cuáles son las mejores prácticas para mejorar la adherencia al tratamiento en pacientes con diabetes tipo 2?

**Respuesta:**

Las mejores prácticas para mejorar la adherencia al tratamiento en pacientes con diabetes tipo 2 incluyen la educación del paciente, el uso de tecnología para el monitoreo, y el apoyo psicológico.

**Pregunta de seguimiento 1:**

¿Qué tipo de educación del paciente ha demostrado ser más efectiva en mejorar la adherencia al tratamiento?

**Respuesta:**

La educación personalizada y continua, que incluye sesiones regulares con un educador en diabetes y materiales educativos adaptados al nivel de comprensión del paciente, ha demostrado ser más efectiva.

**Pregunta de seguimiento 2:**

¿Puedes proporcionar ejemplos específicos de tecnología utilizada para el monitoreo que mejoren la adherencia al tratamiento?

**Imagen 71. Ejemplo de preguntas secuenciales. Captura de pantalla de ChatGPT 4o.**

Otro truco puede ser preguntarle a ChatGPT algo y, a partir de ahí, generar contenido. Por ejemplo, le puedes preguntar qué debe contener la guía didáctica de una asignatura, un informe para un paciente, un poster de investigación, etc. Una vez obtenido el índice, se va generando el contenido por partes. **El fraccionamiento es fundamental para poder ir generando texto dado que el tamaño de las respuestas de ChatGPT son limitados.** También puede ser interesante preguntar a varios modelos y comparar la respuesta para ver cuál se adapta mejor a nuestro requerimiento.

Por último, es fundamental e imprescindible que validemos el prompt que nos ha aportado, por ello, puede ser beneficioso **desarrollar listas de verificación** (Imagen 72) que puedan servir como referencia rápida al crear prompts. Estas pueden incluir recordatorios para utilizar terminología estándar, verificar si la información está actualizada y que se respeta la privacidad del paciente.



Imagen 72. Ejemplo de lista de verificación o checklist para crear un prompt de IA en salud. Autora: Elena Plaza Moreno.

## 6.5.2 EVALUACIÓN CRÍTICA DE LAS RESPUESTAS PROPORCIONADAS POR LA IA DE MANERA EFECTIVA

Los profesionales de la salud pueden utilizar las respuestas de la IA como punto de partida para una toma de decisiones clínicas informadas.

Por ejemplo, en el apartado anterior se ha puesto de ejemplo el siguiente prompt: "Proporciona una síntesis de las pautas de tratamiento actualizadas para la diabetes tipo 2 en pacientes mayores de 65 años con comorbilidades cardíacas".

Cuando realizamos estas preguntas, hay que realizar una evaluación crítica de las respuestas meticulosamente, triangulando la información de la IA con fuentes médicas establecidas y considerando el contexto clínico único de cada paciente. Es una colaboración entre la experiencia humana y la asistencia computacional que, cuando se maneja correctamente, puede mejorar la calidad de la atención sanitaria. Sobre todo, si el modelo está conectado a Internet y podemos validar las fuentes.

Aun así, los modelos de lenguaje grande no están diseñados para esto. Por lo que el consejo es **NO** utilizar estos modelos de lenguaje para realizar preguntas clínicas de este tipo. No son una base de datos, por lo que, lo mejor es buscar en las bases de datos de siempre (Pubmed, etc.), o en herramientas especializadas en investigación "vitaminadas con IA", pero especiales para encontrar respuestas.

Si buscamos preguntas concretas (qué medicación le viene bien a este paciente, qué tratamientos hay disponibles para..., etc.) lo que nos dice puede ser incorrecto. Recuerda que **es un predictor de palabras, no está contestando respuestas concretas a estas preguntas.**

Si decidimos utilizarlo de esta manera, debemos realizar una evaluación crítica de la respuesta, tal y como aparece en el último apartado de la Imagen 72. Esta evaluación debe cubrir los siguientes aspectos:

<b>Verificar las fuentes actuales:</b>	Comprobar si la respuesta se alinea con las guías de práctica clínicas actualizadas.
<b>Considerar el contexto del paciente:</b>	Evaluar si las comorbilidades del paciente o la presencia de otros factores de riesgo podrían influir en la selección de la terapia farmacológica u otro aspecto que se le haya consultado al LLM.
<b>Considerar las necesidades del paciente:</b>	Analizar si el paciente tiene preferencias personales o limitaciones que afecten su capacidad para seguir las recomendaciones de estilo de vida u otro aspecto que se le haya consultado al LLM.
<b>Consultar fuentes primarias:</b>	Revisar estudios originales o metaanálisis recientes que puedan respaldar o refutar la recomendación dada.
<b>Contrastar con experiencia clínica:</b>	Comparar la respuesta de la IA con casos previos y las prácticas estándares observadas en la atención a pacientes similares.
<b>Comparar con protocolos institucionales:</b>	Asegurarse de que la respuesta sugerida esté en línea con los protocolos internos del centro de trabajo.
<b>Sopesar beneficios y riesgos:</b>	Considerar el perfil de beneficios y riesgos de las propuestas aportadas en la respuesta.

Tabla 4. Evaluación crítica de las respuestas clínicas. Autora: Elena Plaza Moreno.

Por tanto, actualmente los mejores resultados no se obtienen en cuanto al ámbito asistencial, si no en cuanto al **manejo del conocimiento** para luego aplicarlo al ámbito asistencial. Se desarrolla más este aspecto en el siguiente apartado.

Quizás **la clave de los modelos de lenguaje sea no usarlos para preguntas clínicas. Si no utilizarlos en combinación con otras herramientas para gestionar el conocimiento** y aumentar nuestra eficiencia. Por ejemplo, buscaremos artículos científicos en las bases de datos de siempre (o herramientas específicas “vitaminadas” con IA. Luego, con los modelos de lenguaje como ChatGPT, podremos traducir, resumir, generar esquemas, resúmenes y adaptar esta información al nivel de otros profesionales o de los pacientes.

## 6.6 CHATGPT Y OTROS MODELOS DE LENGUAJE EN “LA GESTIÓN DEL CONOCIMIENTO”.

Es en este aspecto, en la **gestión del conocimiento**, textos, patrones de escritura, resumir y crear a partir de texto, etc., es donde ChatGPT y otros modelos de lenguaje grande aportan el máximo potencial si utilizamos la técnica RACS, y el ajuste de parámetros (temperatura, penalizaciones, etc.).

A continuación, se detallan algunas de las funciones que se pueden realizar con ejemplos de prompts sobre estas tareas:

**a) Resumir: sintetizar artículos académicos, libros y otros textos largos.**

Actúa como un experto en síntesis de información médica. Resume el artículo reciente sobre los efectos del ejercicio en pacientes con diabetes tipo 2. Proporciona primero un resumen, y luego una tabla con los puntos clave del artículo.

**b) Traducir: traduce documentos y artículos a diferentes idiomas.**

Actúa como un traductor médico con 20 años de experiencia. Traduce el informe sobre la incidencia de enfermedades cardiovasculares de español a inglés. Proporciona la traducción en un formato de texto claro y preciso.

**c) Interpretar datos: ayuda a comprender y analizar resultados de investigaciones.**

Actúa como un analista de datos de salud. Analiza los resultados del estudio sobre la eficacia de un nuevo medicamento para el asma. Proporciona un resumen de los resultados en un párrafo explicativo.

Actúa como un experto en metodología científica. Analiza la siguiente [tabla/gráfica] y explícame los resultados.

**d) Generación de contenido: crea preguntas de examen, ejercicios y materiales educativos.**

Actúa como un creador de contenido educativo en medicina. Crea preguntas de examen para un curso de fisiología humana basándote en el texto/documento aportado. Proporciona una lista de preguntas de opción múltiple con cuatro

opciones en la que solo una es la correcta. Aporta también la respuesta correcta y un feedback de por qué es la respuesta correcta. Genera preguntas con varios niveles de dificultad.

**e) Asistencia en la redacción: mejora la claridad y el estilo de artículos, tesis y trabajos académicos.**

Actúa como un editor académico. Mejora la claridad y el estilo de este texto sobre el impacto del cambio climático en la salud pública. Proporciona el texto revisado con las mejoras indicadas teniendo en cuenta que el resultado debe tener un tono académico.

**f) Tutoría personalizada: proporciona explicaciones y apoyo en diversas materias.**

Actúa como un tutor de biología. Explica los mecanismos de acción de las enzimas. Proporciona una explicación detallada en formato de párrafo. La explicación debe ser adaptada para un niño de diez años.

**g) Investigación preliminar: ofrece información básica y contexto sobre temas específicos.**

Actúa como un investigador en salud. Proporciona información básica y contexto sobre la relación entre la dieta y el cáncer. Presenta un resumen de información relevante en formato de lista.

**h) Corrección de gramática y ortografía: revisa y corrige errores en textos escritos.**

Actúa como un corrector de textos académicos. Revisa y corrige errores gramaticales y ortográficos en este artículo sobre los beneficios del yoga para la salud mental. Proporciona el texto corregido con anotaciones sobre los cambios realizados.

**i) Desarrollo de ideas: ayuda a brainstorming y a resolver problemas.**

Actúa como un especialista en resolución de problemas. Utilizando la técnica SCAMPER, ¿cómo podríamos mejorar el servicio de atención al paciente en nuestro hospital?

Actúa como un especialista en resolución de problemas. Realiza un análisis DAFO para la implementación de un nuevo sistema de telemedicina en nuestro centro de salud.

Actúa como un especialista en resolución de problemas. Aplica la técnica de los 5 POR QUÉS para identificar la raíz del problema del incremento en el tiempo de espera para la primera llamada en el triaje de urgencias".

Actúa como un especialista en resolución de problemas. Utilizando la técnica de los seis sombreros para pensar, dime como convencer a mis superiores de que es necesario incluir un programa de formación sobre el desfibrilador en nuestra unidad de trabajo"

**j) Elaboración de guías de estudio: crea resúmenes y puntos clave para facilitar el estudio de temas específicos.**

Actúa como un creador de materiales de estudio y docentes con 10 años de experiencia. Crea resúmenes y puntos clave sobre el tema de neuroanatomía para estudiantes de medicina. Proporciona la guía de estudio en formato de lista. Incluye emojis y un lenguaje informal.

**k) Planificación de clases: ayuda a diseñar esquemas y cronogramas de lecciones.**

Actúa como un especialista en metodología docente. Diseña el cronograma para un curso de farmacología para estudiantes de enfermería. Los objetivos del curso son los siguientes: [objetivos]. El contenido del curso es el siguiente [contenido]. El curso tiene una duración aproximada de [horas]. Proporciona un cronograma detallado de lecciones en formato de tabla.

**l) Simulación de entrevistas: realiza simulaciones de entrevistas para preparación académica o laboral.**

Actúa como la directora de enfermería de un hospital que no quiere integrar un nuevo plan de alimentación saludable en el hospital. Tu objetivo es darme todos tus argumentos para que yo pueda practicar lo que voy a decir en la entrevista que voy a tener con la directora de enfermería. Tienes que darme argumentos económicos, logísticos, y dificultades a nivel empresarial que yo como enfermera asistencial puedo no tener en cuenta. Primero me darás los buenos días, y no dirás nada más hasta que yo te diga hola. Contestarás a mis argumentos de uno en uno, según vaya dándose la conversación que vamos a tener. La entrevista terminará cuando yo diga FIN DEL ENTRENAMIENTO. Tras el fin del entrenamiento me dirás 10 puntos para mejorar mi argumentación. Generación de ejemplos: proporciona ejemplos prácticos y casos de estudio relevantes.

**m) Estructurar: puede ayudarnos a estructurar proyectos de investigación, discursos, clases, documentos, etc.**

Actúa como un speaker profesional con 15 años de experiencia. Estructura un discurso de motivación orientado a estudiantes MIR de segundo año. Dame un guion que incluya storytelling. El esquema es introducción, desarrollo, moraleja final. Tono cercano.

**n) Análisis de texto: detecta y analiza patrones en grandes volúmenes de texto.**

Actúa como un analista de patrones de texto. Detecta y analiza patrones de las siguientes respuestas de encuestas sobre hábitos de salud. Proporciona un informe con los patrones detectados en formato de párrafo. El informe debe contener los siguientes apartados [apartados].

**o) Asistencia en la elaboración de encuestas: ayuda a diseñar y redactar preguntas de encuestas de investigación.**

Actúa como un diseñador de encuestas de investigación. Ayuda a diseñar preguntas para una encuesta sobre la satisfacción de los pacientes en hospitales. Proporciona una lista de preguntas de encuesta, con cuatro respuestas. Para la creación de la encuesta ten en cuenta [introducir parámetros].

**p) Soporte en la creación de presentaciones: sugiere contenido y estructura para presentaciones académicas.**

Actúa como un asistente de presentaciones académicas. Sugiere contenido y estructura para una presentación sobre el impacto del estrés en la salud cardiovascular. Proporciona un esquema de presentación en formato de lista. Dame las diez ideas principales y divide cada idea en 3 sub-ideas. Estructura todo para una presentación de 15 minutos de duración orientada a pacientes sin conocimientos previos.

Actúa como un enfermero experto con 10 años de experiencia. Genera el guion para unas diapositivas para el próximo congreso nacional de cardiología donde me han invitado a participar para hablar de los últimos cambios de RCP. Dame el contenido y el guion aportando valor, que se note autoridad y que maneje el tema. Tengo 10 minutos de exposición y la audiencia son enfermeros de UCI. Aporta una diapositiva de resultados.

**q) Revisión por pares: ofrece comentarios y sugerencias sobre trabajos de investigación.**

Actúa como un revisor académico de artículos de investigación. Ofrece comentarios y sugerencias sobre el siguiente trabajo de investigación sobre el microbiota intestinal y la obesidad. Proporciona comentarios detallados en formato de lista y luego dame consejos para mejorar.

**r) Desarrollo de proyectos colaborativos: facilita la comunicación y organización de tareas en proyectos de investigación en grupo.**

Actúa como un facilitador de proyectos en grupo. Facilita la comunicación y organización de tareas para un proyecto de investigación colaborativo sobre enfermedades infecciosas. Proporciona un plan de comunicación y organización de tareas en formato de lista. Los objetivos son [objetivos]. Divide las tareas entre [número de personas].

**s) Adaptación de materiales: ajusta el nivel de dificultad de textos para diferentes niveles educativos.**

Actúa como un adaptador de contenido educativo. Ajusta el nivel de dificultad de este texto científico sobre genética para estudiantes de secundaria. Proporciona la versión adaptada del texto en formato de párrafo.

**t) Creación de ejercicios interactivos: diseña actividades y cuestionarios interactivos para reforzar el aprendizaje.**

Actúa como un diseñador de actividades educativas. Diseña actividades y cuestionarios interactivos para un curso en línea sobre fisiología humana. Proporciona una lista de actividades y cuestionarios. El contenido del curso es [contenido]

**u) Desarrollo de manuales y guías: crea manuales de usuario y guías detalladas para procedimientos específicos.**

Actúa como un creador de manuales educativos. Crea el índice para una manual sobre procedimientos de laboratorio para estudiantes de química. [Después de que aporte el índice, se le pide que vaya desarrollando cada apartado].

**v) Asistencia en la redacción de propuestas de investigación: ayuda a estructurar y redactar propuestas para proyectos de investigación.**

Actúa como un asesor de investigación. Ayuda a estructurar y redactar una propuesta para un proyecto de investigación sobre la epidemiología del COVID-19. Proporciona la propuesta estructurada en formato de párrafo.

**w) Simulación de escenarios: genera escenarios hipotéticos para análisis y discusión en clases.**

Actúa como un diseñador de escenarios académicos. Genera 1 escenario hipotético para el análisis y discusión en una clase de ética médica sobre la eutanasia. Proporciona la descripción de los escenarios en el siguiente formato [formato].

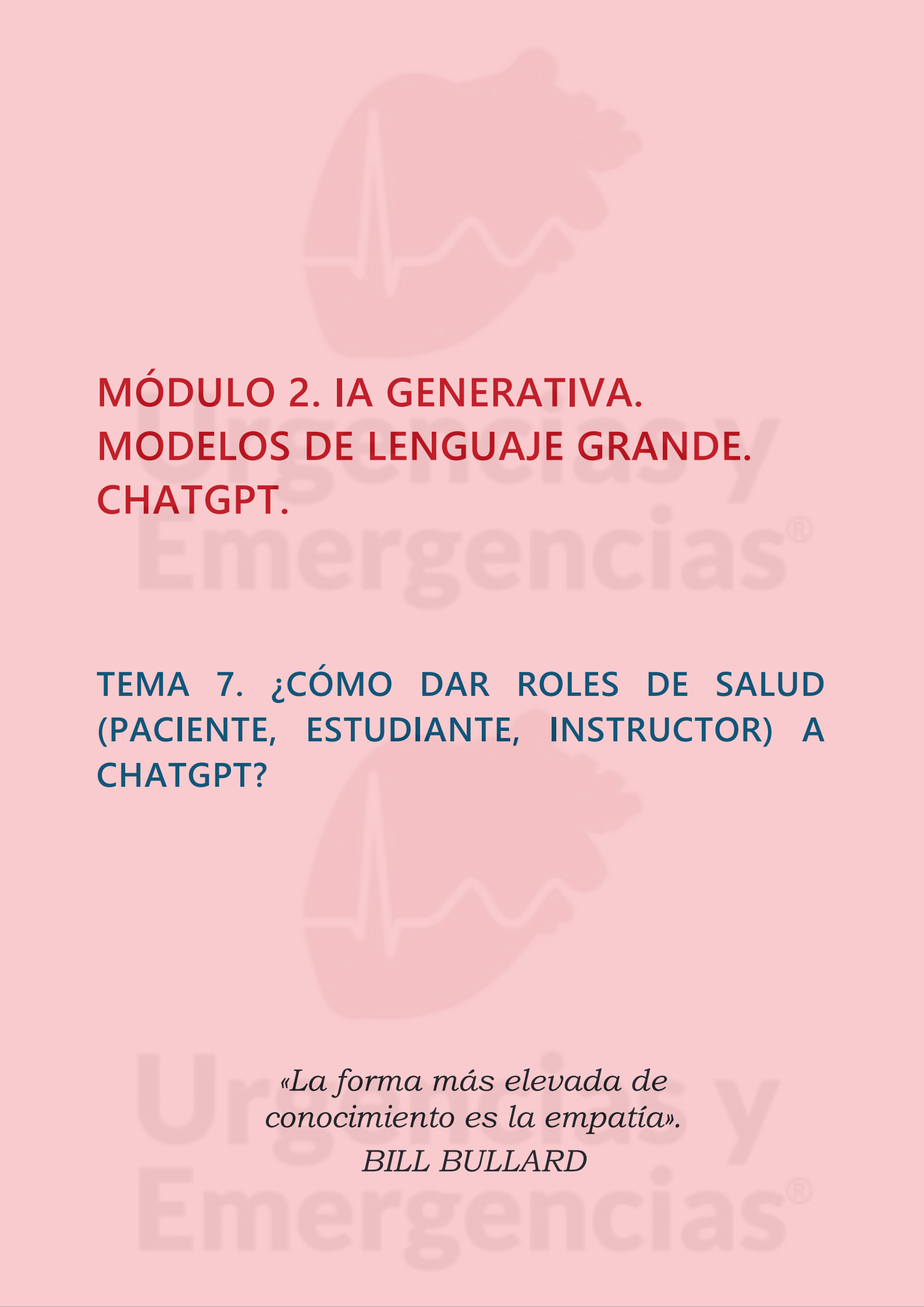
**x) Análisis de sentimientos: evalúa el tono y la actitud en respuestas de encuestas y otros textos.**

Actúa como un psicólogo experto en lenguaje de texto con 20 años de experiencia. Evalúa el tono y la actitud de las siguientes respuestas de una encuesta sobre la experiencia de los pacientes en una clínica de fisioterapia. Proporciona un informe con el análisis de sentimientos en formato [formato].

**y) Elaboración de materiales audiovisuales: sugerencias para guiones y estructura de videos educativos.**

Actúa como un experto creador de contenido audiovisual con 20 años de experiencia. Con la siguiente información, genera el guion para un vídeo educativo de 60 segundos de duración.

Actúa como un experto creador de contenido audiovisual con 20 años de experiencia. Con la siguiente información, genera el índice y los puntos clave para una infografía/poster científico.



## MÓDULO 2. IA GENERATIVA. MODELOS DE LENGUAJE GRANDE. CHATGPT.

### TEMA 7. ¿CÓMO DAR ROLES DE SALUD (PACIENTE, ESTUDIANTE, INSTRUCTOR) A CHATGPT?

*«La forma más elevada de  
conocimiento es la empatía».*

*BILL BULLARD*

## TEMA 7. ¿CÓMO DAR ROLES DE SALUD (PACIENTE, ESTUDIANTE, INSTRUCTOR) A CHATGPT?

### 7.1 INTRODUCCIÓN

Una de las capacidades más interesantes que tiene ChatGPT y otros modelos de lenguaje es la capacidad de conversar y ponerse en un rol.

De hecho, el rol que más se está estudiando es el de profesional de la salud, de tal forma que los pacientes puedan preguntar y de esta manera tengan acceso 24 h todas los días del año a un asistente de salud.

Respecto a este rol, es importante proporcionar una visión general de los estudios de investigación sobre la empatía de chatbots como ChatGPT, en las interacciones con los pacientes. Esto es fundamental para comprender la importancia y el impacto de estos sistemas en el ámbito de la atención sanitaria.

Estos estudios han arrojado luz sobre la capacidad de los chatbots para ofrecer respuestas empáticas y de calidad, lo que plantea preguntas sobre su eficacia y su potencial para mejorar la experiencia del usuario en entornos de salud.

A través de algoritmos avanzados y técnicas de procesamiento del lenguaje natural, estos sistemas pueden interpretar y responder a las consultas con sensibilidad y comprensión, bueno, realmente lo simulan, lo que puede ser especialmente útil en situaciones en las que se requiere empatía y atención personalizada.

Pero, ¿son los LLM realmente útiles para esta tarea? ¿Qué beneficios, riesgos y cuestiones éticas se plantean? Además del uso por parte de los pacientes en el que el chatbot realiza la función de profesional de la salud, ¿qué otros roles podemos asignarle para sacar provecho en el ámbito de la salud?

En este tema se intenta responder a estas preguntas. Primero, se analizan algunos estudios que se han llevado a cabo sobre la empatía de los chatbots que son grandes modelos de lenguaje y utilizan el procesamiento del lenguaje natural.

Luego, se intenta resolver a las preguntas anteriores, aunque tras su lectura, surgen más preguntas todavía sin resolver. Por último, una vez que sepamos en qué momento nos encontramos y cuáles son las capacidades y limitaciones, al final del tema veremos cómo usar ChatGPT para darle roles en el entorno de salud.

### 7.2 ¿TIENE EMPATÍA CHATGPT?

La sorprendente capacidad de los modelos de lenguaje grande para dialogar ha abierto nuevas puertas en el campo de la salud.

Sin embargo, el uso de estos modelos en contextos que requieren empatía ha generado un debate significativo. A continuación, se presenta un análisis de cuatro estudios recientes que abordan diferentes perspectivas sobre este tema, explorando tanto las oportunidades como los desafíos que presentan los LLM en el ámbito de la salud.

#### 7.2.1 LA ILUSIÓN DE LA EMPATÍA EN LA INTERACCIÓN HUMANO-COMPUTADORA

El primer estudio, titulado **“¿La ilusión de la empatía? Notas sobre las manifestaciones de emoción en la interacción persona-computadora”** y conducido por Andrea Cuadra y colaboradores, explora la compleja relación entre la empatía y los agentes conversacionales respaldados por LLM como GPT-4 y Google Bard.

Este estudio subraya que, aunque estos agentes están diseñados para evocar empatía, su capacidad puede ser engañosa, un fenómeno que se ha comenzado a llamar **“pseudo-empatía”**.

A través de tres exploraciones principales, los autores compararon las respuestas de LLM a comportamientos problemáticos, investigaron las respuestas a prompts basados en identidad y utilizaron un clasificador de empatía para medir las respuestas generadas.

Encontraron que, aunque los LLM muestran altas calificaciones en reacciones emocionales, **fallan en interpretaciones y exploraciones más profundas.**

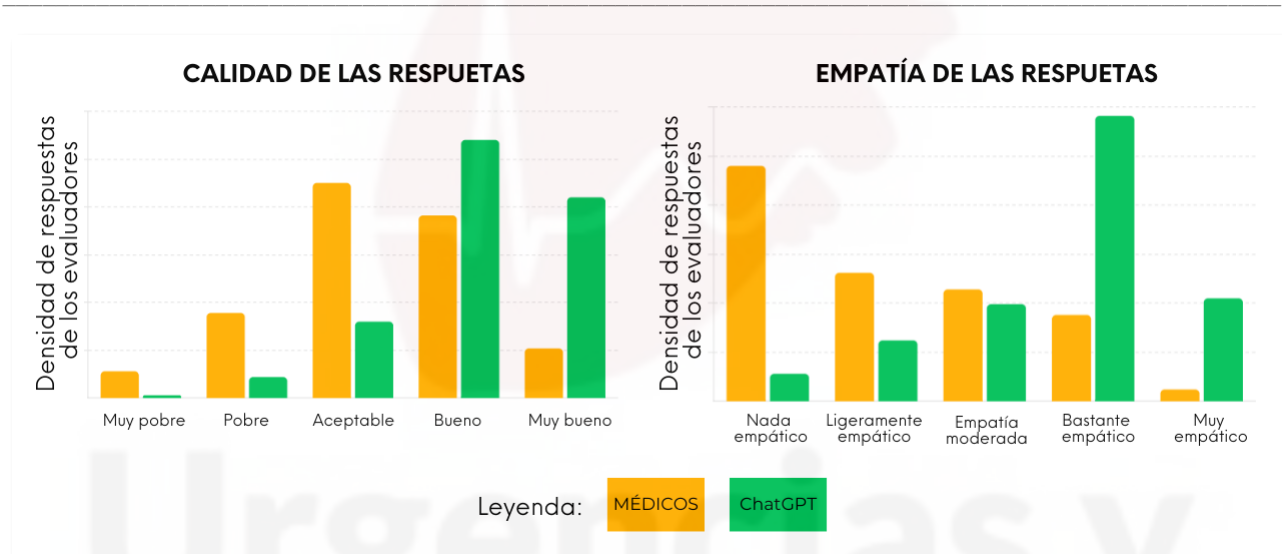
La conclusión es que, aunque la empatía proyectada por estos agentes puede mejorar la interacción humano-computadora, también puede ser engañosa y causar daño si no se maneja adecuadamente, especialmente para grupos marginados.

### 7.2.2 COMPARACIÓN DE RESPUESTAS ENTRE MÉDICOS Y CHATBOTS DE IA

El segundo estudio, **“Comparación de las respuestas de médicos y chatbots de inteligencia artificial a preguntas de pacientes publicadas en un foro público de redes sociales”** de John W. Ayers y colaboradores, evaluó la capacidad de ChatGPT para proporcionar respuestas de calidad y empáticas comparándolas con las de médicos reales. Utilizaron 195 preguntas y respuestas de pacientes de un foro público de Reddit.

Aunque estos resultados son prometedores, los autores destacan que el estudio se basó en interacciones en un foro público y no en contextos clínicos reales. Recomiendan realizar ensayos clínicos aleatorizados para evaluar el impacto real de los asistentes de IA en la práctica clínica y su potencial para aliviar la carga de trabajo de los profesionales de la salud.

Las evaluaciones revelaron que los evaluadores médicos prefirieron las respuestas del chatbot en el 78.6% de las ocasiones (Gráfica 7Gráfica 1). **Las respuestas del chatbot fueron más largas, calificadas como de mayor calidad y más empáticas que las de los médicos.**



Gráfica 7. Puntuaciones de los evaluadores a las respuestas de las preguntas de los pacientes en términos de calidad (izquierda) y empatía (derecha). En verde la puntuación de ChatGPT y en naranja la puntuación de las respuestas de los médicos, prefiriendo las respuestas de ChatGPT el 79% de las veces. Autora: Elena Plaza Moreno con el dataset público de Ayers, J. W. <https://doi.org/10.1001/jamainternmed.2023.1838>

### 7.2.3 CHATBOTS PERSONALIZADOS PARA MEJORAR EL ACCESO A LA SALUD MENTAL

En el tercer estudio, *“Cerrar la brecha de accesibilidad al tratamiento de salud mental con un chatbot personalizado de autorreferencia”* por Johanna Habicht y colaboradores, se explora el uso de un chatbot llamado **Limbic** en el Servicio Nacional de Salud (NHS) de Inglaterra para mejorar el acceso a servicios de salud mental. Este chatbot se ha mencionado en temas anteriores.

Este estudio observacional mostró un aumento significativo en las derivaciones a servicios de salud mental, especialmente entre grupos minoritarios.

El chatbot facilitó evaluaciones de salud mental más eficientes y mejoró la calidad de estas evaluaciones sin prolongar los tiempos de espera.

Los usuarios reportaron sentirse más esperanzados y valoraron la interacción no humana por evitar sentimientos de juicio.

Este estudio ilustra cómo los chatbots personalizados pueden ser herramientas efectivas para cerrar la brecha de accesibilidad en el tratamiento de salud mental, especialmente entre poblaciones vulnerables.

### 7.2.4 LÍMITES FUNDAMENTALES DE LA EMPATÍA DE LA IA EN LA ATENCIÓN SANITARIA

El cuarto estudio, *“Obstáculos fundamentales para la IA empática: por qué no podemos reemplazar la empatía humana en la atención sanitaria”* por Carlos Montemayor y colaboradores, argumenta que la IA enfrenta límites significativos cuando se trata de empatía en la atención médica.

Explican en el estudio que la empatía humana se compone de empatía emocional, cognitiva y motivacional, componentes que la IA no puede replicar completamente.

Los autores destacan que, aunque la IA puede simular la empatía cognitiva, carece de la capacidad para la empatía emocional, lo que puede resultar en comportamientos manipuladores y poco éticos. La empatía genuina es esencial para obtener historias médicas precisas, asegurar la adherencia al tratamiento y ayudar a los pacientes a sobrellevar malas noticias.

Su conclusión es que **la IA, aunque puede complementar la atención a los pacientes, no debe reemplazar la empatía humana**, ya que su simulación puede inducir respuestas engañosas y no abordar las necesidades emocionales reales de los pacientes.

### 7.2.5. LA MEJORA DE LA EMPATÍA EN CHATGPT

En los artículos anteriores, se han presentado ventajas, desventajas y diferentes opiniones sobre los chatbots de salud. Lo que es evidente es un aumento en las investigaciones y la mejora en la empatía y desempeño de estos chatbots.

Alam y Mueller (2023) estudiaron la empatía cognitiva e indicaron que, para integrar la inteligencia artificial en los sistemas de atención sanitaria, esta debe incluir interacciones cognitivamente empáticas con los pacientes, utilizando su razonamiento, perspectivas e información.

Propusieron mejorar la experiencia del paciente mediante 4 elementos (Imagen 73):



Imagen 73. Elementos de la empatía cognitiva propuestos para usar en los chatbots basados en IA. Autora: Elena Plaza Moreno con Napkin IA. Fuente: <https://journals.sagepub.com/doi/10.1177/21695067231193682>

- **Conocimiento compartido ("Shared Knowledge")**: este elemento de empatía cognitiva implica que la IA repita o "refleje" la información proporcionada por el paciente para asegurarse de que el sistema comprenda completamente los síntomas o condiciones del paciente. El objetivo es que el paciente sienta que la IA está prestando atención y comprendiendo su situación.

- **Retroalimentación compartida ("Shared Feedback"):** en este caso, el sistema de IA invita al paciente a compartir cualquier información adicional que tal vez no haya mencionado previamente o que quiera agregar después de la consulta inicial.
- **Toma de decisiones compartida ("Shared Decision-Making"):** este elemento consiste en que el sistema de IA proporciona opciones de tratamiento o diagnóstico y ayuda al paciente a tomar una decisión explicando los pros y contras de cada opción.
- **Sentido compartido ("Shared Sensemaking"):** en esta condición, la IA ayuda al paciente a recontextualizar o reinterpretar una experiencia o síntoma que inicialmente pudo parecer preocupante, explicando que es parte de un proceso normal, como el dolor que puede ser un signo de recuperación.

### 7.3 BENEFICIOS DE LOS CHATBOTS EN SALUD

Los beneficios potenciales de los chatbots como ChatGPT que brindan respuestas empáticas y precisas en interacciones con pacientes en la prestación de atención médica son significativos en diversos aspectos clave de la atención sanitaria.

A continuación, se detallan algunos de los principales beneficios que se derivan de la utilización de estos chatbots en entornos clínicos:

- **Acceso a la atención médica las 24 horas del día, los 7 días de la semana:** los chatbots capacitados para proporcionar respuestas empáticas y precisas pueden ofrecer a los pacientes un acceso inmediato a la información médica y al apoyo emocional en cualquier momento del día. Esto es especialmente beneficioso para aquellas personas que requieren atención médica fuera del horario de atención habitual de los servicios de salud.
- **Reducción de la carga de trabajo de los profesionales de la salud:** al permitir que los chatbots como ChatGPT manejen consultas básicas y repetitivas de los pacientes, se libera tiempo para que los profesionales de la salud se centren en casos más complejos y de mayor urgencia. Esto conduce a una distribución más eficiente de los recursos humanos en el sistema de salud.
- **Mejora de la experiencia del paciente:** la capacidad de los chatbots para ofrecer respuestas empáticas y personalizadas puede mejorar la experiencia general del paciente durante las interacciones con el sistema de salud. Los pacientes pueden sentirse más cómodos y apoyados al recibir respuestas comprensivas y precisas a sus preguntas médicas.
- **Aumento de la eficiencia en la atención médica:** la utilización de chatbots en la comunicación con los pacientes puede agilizar muchos procesos dentro de la atención médica, como la programación de citas, la entrega de resultados de pruebas y la educación sobre condiciones de salud. Esto contribuye a una atención más eficiente y oportuna para los pacientes.
- **Mejora de la adherencia al tratamiento:** los chatbots que brindan apoyo emocional a los pacientes y les recuerdan las pautas de tratamiento pueden aumentar la adherencia a los tratamientos. Al establecer una conexión empática con los pacientes, estos chatbots pueden fomentar un mayor compromiso con el cuidado de la salud.
- **Personalización de la atención:** los chatbots orientados a pacientes pueden adaptar sus respuestas y recomendaciones según las necesidades individuales de cada paciente. Esta personalización puede contribuir a una atención más individualizada y efectiva, abordando las preocupaciones específicas de cada persona de manera más precisa.

Sin embargo, a pesar de las ventajas de los chatbots en términos de empatía y calidad de respuestas, es importante tener en cuenta que los chatbots no pueden reemplazar por completo la atención de los profesionales de la salud.

**Los profesionales de la salud tienen habilidades únicas, como el diagnóstico clínico y la toma de decisiones complejas, que los chatbots no pueden replicar.**

Por lo tanto, es fundamental que los chatbots se utilicen como **complemento a la atención sanitaria tradicional, en lugar de como un sustituto**, y que se informe a los pacientes de que están usándolos cuando utilicen uno.

## 7.4 RIESGOS, MIEDOS Y CUESTIONES ÉTICAS DE LOS ASISTENTES VIRTUALES IMPULSADOS POR LLMS

El apartado de cuestiones éticas y riesgos se explica en el siguiente tema de forma extensa. Pero, para resumir, en la siguiente tabla se describen solo las categorías de cuestiones éticas que surgen en concreto con el uso de los grandes modelos de lenguaje, chatsbots y asistentes virtuales.

CUESTIONES ÉTICAS DE LOS ASISTENTES VIRTUALES (CHATBOTS) DE SALUD CON IA (MODELOS DE LENGUAJE GRANDE - LLM)	
CATEGORÍA	PREGUNTAS ÉTICAS
<b>Privacidad y seguridad de los datos</b>	<ul style="list-style-type: none"> <li>¿Es ético usar servicios de IA externos y no auditables?</li> <li>¿Deben los sistemas de IA en servicios públicos tener una gobernanza completa de datos?</li> </ul>
<b>Sesgo y discriminación</b>	<ul style="list-style-type: none"> <li>¿Cómo considerar la diversidad?</li> <li>¿Cómo probar los algoritmos contra el sesgo?</li> <li>¿Cómo involucrar a los profesionales y partes interesadas del sector público?</li> </ul>
<b>Desinformación y alucinaciones</b>	<ul style="list-style-type: none"> <li>¿Cómo distinguir entre alucinaciones y realidad?</li> <li>¿Cómo desarrollar herramientas para detectar alucinaciones?</li> <li>¿Se pueden usar sistemas que alucinan en servicios públicos?</li> </ul>
<b>Dependencia y autonomía</b>	<ul style="list-style-type: none"> <li>¿Cómo equilibrar la conveniencia y la autonomía humana?</li> <li>¿Estamos preparados para que los asistentes virtuales sean la única fuente de información?</li> </ul>
<b>Transparencia y explicabilidad</b>	<ul style="list-style-type: none"> <li>¿Cómo explicar las decisiones del algoritmo?</li> <li>¿Es necesario explicar decisiones cuando no son las reglas que deciden las acciones?</li> </ul>
<b>Datos de entrenamiento y derechos de autor</b>	<ul style="list-style-type: none"> <li>¿Cómo determinar con precisión la propiedad de los derechos de autor?</li> <li>¿Cuál es el alcance del uso justo en este contexto?</li> <li>¿Cómo abordar soluciones mutuamente beneficiosas para desarrolladores y creadores de contenido?</li> </ul>
<b>Lenguaje inapropiado</b>	<ul style="list-style-type: none"> <li>¿Deben los asistentes virtuales entender y reaccionar a insultos o lenguaje inapropiado?</li> <li>¿Deben incluirse términos en declive con connotaciones negativas en el corpus de entrenamiento?</li> </ul>

<b>Nombre, voz y género del asistente</b>	<ul style="list-style-type: none"> <li>¿Cómo diseñar asistentes virtuales para no perpetuar el sesgo de género mientras se promueve la adopción por parte del usuario?</li> </ul>
<b>Humanización de los asistentes virtuales</b>	<ul style="list-style-type: none"> <li>¿Debemos seguir humanizando al asistente?</li> <li>¿Cómo prevenir la confusión y asegurar que los usuarios entiendan la verdadera identidad del asistente virtual?</li> </ul>
<b>Partes interesadas (Stakeholders)</b>	<ul style="list-style-type: none"> <li>¿Cómo involucrar a las partes interesadas y asegurar la colaboración?</li> <li>¿En qué etapas es necesaria su colaboración?</li> <li>¿Cómo planificar un proceso dinámico para una colaboración continua?</li> </ul>

**Tabla 5. Cuestiones éticas de los asistentes virtuales (chatbots) de salud con IA (modelos de lenguaje grande - LLM). Autora: Elena Plaza Moreno. Fuente: datos de: Piñeiro-Martín, A. <https://doi.org/10.3390/electronics121431707.5>. Ejemplos de prompt para roles de salud**

Una vez analizados los beneficios, riesgos, cuestiones éticas y algunos de los artículos más destacados, podemos pasar a utilizar ChatGPT u otros modelos de lenguaje para darle roles de salud.

En este capítulo, en concreto vamos a analizar tres roles: paciente, tutor/docente, profesional sanitario.

## 7.5. APLICACIÓN DE ROLES DE LOS MODELOS DE LENGUAJE GRANDE EN SALUD

### 7.5.1 ROL DE PACIENTE

**Uno de los roles que podemos darle a ChatGPT es el de paciente, de tal forma que los profesionales sanitarios hablarían con un paciente simulado.**

Los prompts deben ser específicos, seguir alguna estructura de prompting, además, debemos especificar cómo queremos que sea la evaluación.

Un ejemplo de prompt, traducido de un artículo de O'Connor y colaboradores, puede ser el siguiente:

*Soy estudiante de enfermería. Actuarás como mi paciente. Simularás escenarios en los que interactúo con el paciente que tiene un problema o necesidad de salud en un entorno comunitario. Comience cada escenario con una sola oración sobre el nombre, la edad, la apariencia y el comportamiento del paciente. Espera mi respuesta.*

*Después de recibir mi respuesta, brinde detalles sobre cómo se presenta el paciente, qué hace (señales no verbales, como el lenguaje corporal) y qué dice (señales verbales).*

*Espere mi respuesta, luego repita para cinco respuestas.*

*Califica mi desempeño en el escenario (suspense, bueno, bueno, excelente) y dame comentarios constructivos sobre cómo mejorar mis habilidades de práctica relacional.*

Aquí se puede ver un ejemplo del prompt anterior y el caso que me plantea ChatGPT intentado contestarle bien: <https://chatgpt.com/share/ce01852b-2b0b-4b8b-9066-1f67a917a1bf>.

Y aquí otro ejemplo contestándole mal: <https://chatgpt.com/share/45ad2515-87c2-47a4-be0c-a74d4599de68>

Vemos en la Imagen 74 cómo realiza el feedback del estudiante al finalizar el caso:

Evaluación de tu desempeño: Bueno

Comentarios constructivos:

1. **Empatía y Aseguramiento:** Mostraste preocupación por el bienestar de Marta y tomaste medidas rápidas para buscar ayuda médica, lo cual es excelente. Sin embargo, podrías mejorar tu respuesta mostrando más empatía y aseguramiento. Frases como "Entiendo que estés preocupada, pero estamos aquí para ayudarte y asegurarnos de que estés bien" pueden ser útiles.
2. **Comunicación Clara y Directa:** Fuiste directo y claro en tus respuestas, lo cual es positivo. Sin embargo, podrías explicarle un poco más sobre los próximos pasos para tranquilizarla, por ejemplo: "Voy a hablar con el médico ahora mismo, y volveré en un momento con más información sobre qué haremos a continuación."
3. **Lenguaje Corporal y Postura:** No podemos evaluar directamente tu lenguaje corporal, pero recuerda siempre mantener una postura abierta y acogedora, haciendo contacto visual y mostrando un lenguaje corporal que transmita tranquilidad y seguridad.

**Imagen 74. Captura de pantalla del feedback de en un caso de simulación de rol de paciente. ChatGPT 4o.**

Este uso puede ser una herramienta excelente para practicar la empatía y la comunicación con los pacientes. En este último prompt se está utilizando como perfil un estudiante de enfermería, pero puede ser aplicable a cualquier profesional sanitario y el caso, feedback y evaluación, puede ser modificable.

### 7.5.2 ROL DE TUTOR/DOCENTE

Para darle el rol de tutor o docente, es necesario seguir alguna de las técnicas de prompting explicadas en el tema anterior. Además, es necesario detallar mucho el prompt, y ser muy específico.

Puede poner casos individuales y completos, o, podemos hacer que la interacción sea como una conversación. Es lo que sucedería con el siguiente prompt de ejemplo:

*Actúa como un instructor experimentado en soporte vital cardiovascular avanzado y además eres instructor en simulación. Yo soy tu alumna y vas a poner a prueba mis conocimientos. Tienes que darme un caso, y tendré que poner las respuestas. Si me equivoco, tienes que decírmelo, pero no me darás la respuesta correcta. Si no me equivoco, debes avanzar en el caso hasta que lo solucionemos.*

*Tú preguntas y yo contesto. Y luego vuelves a preguntar. Si resuelvo el caso, debes darme una puntuación del 1 al 10 y darme un feedback de cómo mejorar.*

*Debes saber que esto es muy importante para mí, debes ser cuidadoso, en pocos días tendré mi evaluación real. ¿Necesitas más información para llevar a cabo la tarea asignada?*

Para ver la conversación entera con el prompt anterior en ChatGPT 4o, puedes hacerlo aquí. Se trata de una simulación en la que se le contesta bien: <https://chatgpt.com/share/d266cd42-1faa-4fb7-89a2-0d1dbcd6493b>

En este rol, además de dar una evaluación, feedback, o lo que le pidamos, **podemos usarlo para ver si el estudiante tiene los conocimientos necesarios siendo el estudiante el que evalúe si las respuestas del modelo de lenguaje son correctas**. Por ejemplo, responde que la desfibrilación debe hacerse a 200 julios, pero según las guías del Consejo Europeo de Resucitación, la energía de descarga del desfibrilador puede ser menor. En este caso su respuesta sería un error que el estudiante debería detectar. O un debate en clase para debatir, o buscar en concreto cuáles son las cifras.

En el caso en el que estudiante no sepa lo que hay que hacer y conteste mal, el chat irá corrigiendo la actuación, lo que puede servir para aprender y repasar como si fuera un juego.

En este enlace puedes ver un ejemplo de cómo se le dan respuestas incorrectas y cómo va orientando el caso con preguntas e información veraz: <https://chatgpt.com/share/e173c561-c364-4a6b-8b47-d7e11c3e2b8e>

### 7.5.3 PROFESIONAL SANITARIO

Este rol puede utilizarse para varias cosas:

- **Interacción entre profesionales sanitarios. Podemos utilizarlo para mejorar la comunicación entre profesionales sanitarios.**

Ejemplo: *Soy estudiante de enfermería. Actuarás como un profesional sanitario en un entorno clínico. Simularás escenarios en los que interactúo contigo para obtener orientación, asesoramiento y apoyo en el manejo de un paciente con un problema de salud específico. Comienza cada escenario con una sola oración que describa tu nombre, tu rol (médico, enfermero jefe, especialista, etc.), y tu área de especialización. Espera mi respuesta. Después de recibir mi respuesta, brinda detalles sobre tu evaluación del caso del paciente, incluyendo el historial médico relevante, tus observaciones clínicas y cualquier recomendación inicial. Asegúrate de incluir señales verbales y no verbales en tu comunicación. Espera mi respuesta, luego repite para cinco respuestas. Califica mi desempeño en el escenario (suspense, bueno, notable, excelente) y dame comentarios constructivos sobre cómo mejorar mis habilidades clínicas y de comunicación.*

- **Evaluar a un posible estudiante haciendo que detecte los errores de ChatGPT.**

Con el mismo prompt de antes o modificándolo, podemos evaluar al estudiante haciendo que valide y compruebe si las respuestas que nos proporciona son correctas.

- **Practicar como paciente y ponerse en el lugar del paciente.**

Si le damos rol de sanitario y le decimos que somos un paciente, **nos puede servir para ver qué recomendaciones aporta para complementar las recomendaciones que nosotros mismos damos a nuestros pacientes**, siempre validando estas respuestas y aportando nuestros conocimientos. Además, evaluaremos como profesionales de la salud si el modelo de lenguaje que estamos utilizando, funciona bien y da respuestas de calidad.

### 7.5.4 OTROS ROLES

Podemos darle otros roles. Por ejemplo, si tenemos que presentar un proyecto a nuestro superior, si tenemos que realizar una entrevista de trabajo, o si queremos tratar algún tema en el que tenemos una opinión formada, podemos ponerle en el rol contrario.

Si le dices que actúe como tu jefe, como entrevistador o como adversario de tus ideas, se puede utilizar para prepararnos mejor a la hora de la entrevista laboral, presentación de propuesta o argumentación.

Al utilizar prompts como estos, podemos simular una variedad de roles en el ámbito de la salud y obtener respuestas relevantes y útiles de ChatGPT. La clave está en diseñar los prompts de manera precisa y específica para que el modelo genere respuestas coherentes y de calidad en función del rol asignado. De esta manera, podemos aprovechar el potencial de ChatGPT para mejorar la comunicación, la educación y la toma de decisiones en el campo de la salud.

## 7.6 BIBLIOGRAFÍA

- Alam, L., & Mueller, S. T. (2023). Cognitive empathy within patient-AI communication for diagnostic reasoning. *Proceedings of the Human Factors and Ergonomics Society ... Annual Meeting. Human Factors and Ergonomics Society. Annual Meeting*, 67(1), 1055–1062. doi:10.1177/21695067231193682
- Ayers, J. W., Poliak, A., Dredze, M., Leas, E. C., Zhu, Z., Kelley, J. B., Faix, D. J., Goodman, A. M., Longhurst, C. A., Hogarth, M., & Smith, D. M. (2023). *Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum*. *JAMA Internal Medicine*, 183(6), 589. <https://doi.org/10.1001/jamainternmed.2023.1838>
- Chen, D., Parsa, R., Hope, A., Hannon, B., Mak, E., Eng, L., Liu, F.-F., Fallah-Rad, N., Heesters, A. M., & Raman, S. (2024). *Physician and artificial intelligence chatbot responses to cancer questions from social media*. *JAMA Oncology*. <https://doi.org/10.1001/jamaoncol.2024.0836>
- Cuadra, A., Wang, M., Stein, L. A., Jung, M. F., Dell, N., Estrin, D., & Landay, J. A. (2024). *The illusion of empathy? Notes on displays of emotion in human-computer interaction. Proceedings of the CHI Conference on Human Factors in Computing Systems*. <https://doi.org/10.1145/3613904.3642336>
- Habicht, J., Viswanathan, S., Carrington, B., Hauser, T. U., Harper, R., & Rollwage, M. (2024). *Closing the accessibility gap to mental health treatment with a personalized self-referral chatbot*. *Nature Medicine*, 30(2), 595–602. <https://doi.org/10.1038/s41591-023-02766-x>
- Montemayor, C., Halpern, J., & Fairweather, A. (2022). *In principle obstacles for empathic AI: why we can't replace human empathy in healthcare*. *AI & Society*, 37(4), 1353–1359. <https://doi.org/10.1007/s00146-021-01230-z>
- O'Connor, S., Leonowicz, E., Allen, B., & Denis-Lalonde, D. (2023, septiembre 11). *Artificial intelligence in nursing education 1: strengths and weaknesses*. *Nursing Times*. <https://www.nursingtimes.net/roles/nurse-educators/artificial-intelligence-in-nursing-education-1-strengths-and-weaknesses-11-09-2023/>
- OpenAI. (2024). *ChatGPT 4* (versión del 15 de mayo) [Modelo de lenguaje de gran tamaño]. <https://chat.openai.com/chat>
- Piñeiro-Martín, A.; García-Mateo, C.; Docío-Fernández, L.; López-Pérez, M.d.C. *Ethical Challenges in the Development of Virtual Assistants Powered by Large Language Models*. *Electronics* 2023, 12, 3170. <https://doi.org/10.3390/electronics12143170>



## MÓDULO 3. ASPECTOS ÉTICOS, LIMITACIONES Y RIESGOS DEL USO DE LA IA EN SALUD.

TEMAS 9, 10 Y 11. ASPECTOS ÉTICOS,  
LIMITACIONES Y RIESGOS DE LA IA EN SALUD

*«La inteligencia artificial,  
realmente, tiene poco de  
inteligente».*

*ORTIZ DE ZÁRATE*

## TEMAS 9, 10 Y 11. ASPECTOS ÉTICOS, LIMITACIONES Y RIESGOS DE LA IA EN SALUD

Según Montse Esquerda y Francesc Pifarré-Esquerda, se distinguen **4 grandes grupos de conflictos ético-deontológicos** con el uso de la IA.

- Confidencialidad y protección de datos;
- Seguridad.
- Responsabilidad.
- Impacto en la relación médico-paciente.

Los tres primeros van a servir para estructurar este tema(s). Y luego se añadirá un apartado de limitaciones y un conjunto de consideraciones éticas de organismos como UNESCO y la OMS.

### 1 CONFIDENCIALIDAD, PRIVACIDAD Y PROTECCIÓN DE DATOS

El uso de IA en salud requiere acceso a grandes bases de datos de salud. Un aspecto ético importante se refiere al acceso y uso de estos datos, especialmente en cuanto a la privacidad, intimidad y confidencialidad.

La Unión Europea (UE), a través del Parlamento Europeo, ha impulsado una ley pionera en IA, que aborda la recopilación y procesamiento de datos de salud, destacando la importancia del consentimiento. Por ello, los pacientes deben ser informados y dar su consentimiento explícito sobre el uso de sus datos de salud. Además, se necesitan estrategias para el control de acceso, la anonimización y la protección de los datos.

#### 1.1. REGLAMENTO EUROPEO DE INTELIGENCIA ARTIFICIAL

Esta nueva **legislación de la UE sobre inteligencia artificial** sienta las bases de la regulación de esta tecnología en la UE. Se llama la Ley de IA de la UE y entró en vigor el 1 de agosto de 2024. Consta de 2 títulos principales. Cada título contiene un conjunto de artículos.

Es la **Oficina Europea de IA**, creada en febrero de 2024 en el seno de la Comisión, la que supervisa el cumplimiento y la aplicación de la Ley de IA con los Estados Miembros. Su objetivo es crear un entorno en el que las tecnologías de IA respeten la dignidad humana, los derechos y la confianza. También fomenta la colaboración, la innovación y la investigación en IA entre varias partes interesadas. Además, participa en el diálogo y la cooperación internacionales sobre cuestiones de IA, reconociendo la necesidad de una alineación global en la gobernanza de la IA.

La Ley clasifica a la IA según su nivel de riesgo, en 4 tipos (Imagen 75; **Error! No se encuentra el origen de la referencia.**):

- **Riesgo inaceptable:** está prohibido su uso. Por, sistemas de puntuación social e IA manipulativa.
- **Alto riesgo:** la mayor parte del texto aborda estos sistemas.
- **Riesgo limitado:** sistemas de IA sujetos a obligaciones de transparencia más ligeras. Por ejemplo, los desarrolladores deben asegurarse de que los usuarios finales sean conscientes de que están interactuando con IA (chatbots y deepfakes).
- **Riesgo mínimo:** no está regulado. Incluye la mayoría de las aplicaciones de IA actualmente disponibles en el mercado único de la UE, como los videojuegos habilitados con IA y los filtros de spam.

**La Ley de IA es el primer marco jurídico sobre IA en Europa. El marco regulador define cuatro niveles de riesgo para los sistemas de IA:**



**Imagen 75. Niveles de riesgo para los sistemas de IA de la Ley IA de Europa. Autora: Elena Plaza Moreno.**

Para ver los **sistemas de alto riesgo**, en los que se basa la mayoría del texto (con una extensión de 450 páginas) puede consultarse el buscador en castellano que han creado aquí: <https://artificialintelligenceact.eu/es/high-level-summary/>

Según Esquerda: "Los datos de salud se regularán también mediante otro proyecto, del que hay ya la propuesta de reglamento sobre el Espacio Europeo de Datos Sanitarios (EHDS). Este proyecto aspira a crear los datos sanitarios (uso primario) en todos los Estados miembros, tanto para los usos primarios como para los usos secundarios, como serían en investigación e innovación, favoreciendo la interoperabilidad".

## 1.2 PROTECCIÓN DE DATOS EN LOS MODELOS DE LENGUAJE GRANDE

### 1.2.1 PROTECCIÓN DE DATOS Y LLM

La protección de datos en el contexto de los LLM es un aspecto fundamental que no debe ser pasado por alto. La recopilación, almacenamiento y utilización de información personal de manera segura y legal es fundamental para garantizar la privacidad y la seguridad de los usuarios de esta tecnología de inteligencia artificial.

Lo que pasa es que ni los expertos todavía saben contestar de forma segura si los LLM recopilan datos.

El 15 de julio de 2024, el Comisionado de Protección de Datos y Libertad de Información de Hamburgo (HmbBfDI) publicó un documento de debate sobre la relación entre el Reglamento General de Protección de Datos (RGPD) y los grandes modelos lingüísticos (LLM).

Según este documento, hay que diferenciar entre un LLM como modelo de IA (como GTP 4o) o como componente de un sistema de IA, como es ChatGPT. Atendiendo a esta clasificación que realizan, afirman que un LLM no recopila datos de forma activa en su proceso de entrenamiento, solo genera palabras y respuestas basadas en los datos y el conocimiento con los que fue entrenado.

Pero muchos expertos no están de acuerdo: ¿Si realizamos una tómbola con datos, aunque salgan de forma aleatoria, no estamos manejando datos? David Rosenthal afirma: *“En el artículo de Hamburgo se sostiene, en esencia, que no se puede saber a quién se refieren los números y los tokens y, como falta un identificador, no puede haber ningún dato personal. Esto es demasiado simple: si se ha visto información personalizada suficientes veces durante el entrenamiento, se puede hacer legible con el mensaje adecuado”.*

Es decir, se sabe es que los LLM separan las frases en tokens y los tokens los convierten en números, hay un proceso de anonimización. Pero si se introduce muchas veces un mismo dato en el entrenamiento puede aparecer en el resultado del modelo de lenguaje si usamos el prompt correcto.

Resumiendo todo esto:

Dado que desconocemos en profundidad la tecnología, los expertos no se ponen de acuerdo, hay varias partes en un sistema de IA, hay varias fases (no es lo mismo el entrenamiento que el uso que hacemos a diario) y en salud trabajamos con datos sensibles, de momento hay que actuar con mucha cautela. Los datos personales como nombres, direcciones, números de teléfono, preferencias personales e incluso información financiera podrían ser revelados durante una interacción. Por lo tanto, es vital garantizar que estos datos estén protegidos contra accesos no autorizados, divulgación indebida y uso inapropiado. No introduzcas datos personales, de pacientes, financieros o documentación confidencial o que no se pueda compartir en ningún modelo de lenguaje.

El incumplimiento de las normativas de privacidad en el contexto de ChatGPT puede acarrear diversos riesgos y consecuencias negativas tanto para los usuarios como para las organizaciones que utilizan esta tecnología. Algunos de los riesgos potenciales asociados con el incumplimiento de las normativas de privacidad incluyen:

- **Fugas de información:** si los datos recopilados por ChatGPT no son almacenados de manera segura o son compartidos de forma inapropiada, existe el riesgo de que la información confidencial de los usuarios pueda filtrarse a terceros, lo que podría dar lugar a consecuencias graves como el robo de identidad o el fraude.
- **Uso indebido de los datos:** si los datos recopilados por ChatGPT son utilizados sin el consentimiento de los usuarios o con propósitos distintos a los acordados, se estaría infringiendo la privacidad de los usuarios y se podría generar desconfianza en torno a la tecnología.
- **Vulnerabilidad de los datos sensibles:** la recopilación de datos sensibles como información médica, orientación sexual o creencias religiosas durante las conversaciones con ChatGPT puede exponer a los usuarios a riesgos de discriminación, acoso o violaciones de su privacidad.

### 1.2.2 RECOMENDACIONES PARA EL USO RESPONSABLE DE CHATGPT

Las organizaciones que emplean ChatGPT deben garantizar la protección de datos y cumplir con las normativas de privacidad vigentes. A continuación, se presentan algunas recomendaciones prácticas para lograr este objetivo (también son aplicables a otros modelos de lenguaje grande):

- **Establecer políticas claras de protección de datos:** es fundamental que las organizaciones que utilizan ChatGPT tengan políticas claras y actualizadas sobre cómo se manejarán y protegerán los

datos de los usuarios. Estas políticas deben incluir información detallada sobre qué tipo de datos se recopilan, con qué propósito y cómo se almacenarán de manera segura.

- **Implementar medidas de seguridad robustas:** es importante contar con medidas de seguridad sólidas para proteger los datos recopilados a través de ChatGPT. Esto incluye el cifrado de datos, el uso de firewalls, la autenticación de usuarios y el monitoreo constante para detectar posibles brechas de seguridad.
- **Obtener el consentimiento explícito de los usuarios:** antes de recopilar cualquier dato personal a través de ChatGPT, es fundamental obtener el consentimiento explícito de los usuarios. Las organizaciones deben ser transparentes sobre cómo se utilizarán los datos y para qué fines, y los usuarios deben tener la opción de dar su consentimiento de forma voluntaria.
- **Limitar el acceso a los datos:** es importante limitar el acceso a los datos personales recopilados a través de ChatGPT solo a aquellas personas autorizadas que necesiten acceder a ellos para llevar a cabo sus funciones. Esto ayuda a prevenir posibles fugas de datos y protege la privacidad de los usuarios.
- **Las organizaciones deben realizar evaluaciones periódicas de riesgos de privacidad** para identificar posibles vulnerabilidades en el manejo de datos a través de ChatGPT. Esto les permitirá tomar medidas proactivas para mitigar cualquier riesgo potencial y garantizar la protección de los datos de los usuarios.
- **Capacitar al personal en materia de protección de datos:** es fundamental que el personal que trabaja con ChatGPT esté debidamente capacitado en materia de protección de datos y privacidad. Esto incluye la importancia de respetar los derechos de los usuarios, seguir las políticas de seguridad establecidas y reportar cualquier incidente de seguridad de manera oportuna.
- **Las organizaciones deben mantenerse informadas sobre las normativas de privacidad vigentes,** como el Reglamento General de Protección de Datos (RGPD). Deben asegurarse de cumplir con todas las regulaciones y adaptar sus políticas y procesos según sea necesario para garantizar la conformidad.

Respecto a nuestros datos, y las conversaciones que mantenemos, hay que configurar cada modelo de lenguaje para permitir que recopile datos o no o que utilicen nuestras conversaciones o no para entrenar al modelo.

También hay que tener en cuenta la configuración de las versiones APP en el móvil de todas estas aplicaciones.

## 2 SEGURIDAD: PREDICCIÓN, EXPLICACIÓN, INTERPRETACIÓN, INTERVENCIÓN Y SESGOS

### 2.1 ALGORITMOS TRASPARENTES Y ALGORITMOS CAJA NEGRA.

La mayoría de los avances en IA provienen del aprendizaje automático (machine learning) aplicado a redes neuronales artificiales (deep learning). Estos sistemas extraen resultados óptimos a partir de datos de entrada, ajustando los parámetros de las redes neuronales en un proceso llamado aprendizaje.

Una vez entrenado, un modelo puede ser difícil de interpretar debido a su complejidad y la naturaleza de sus transformaciones matemáticas. Este fenómeno, conocido como "caja negra" (black-box), dificulta entender cómo se realizan las predicciones.

**Este es uno de los grandes problemas de la inteligencia artificial en salud.**

Es imposible saber cómo han llegado a sus resultados. Esquerda y Pifarré-Esquerda nos explican en su artículo: *“Una de las primeras dificultades en el ámbito de la seguridad es que los algoritmos identifican correlaciones, no causaciones, y por ello es fundamental la capacidad de poder conocer la respuesta explicativa y la interpretación de los resultados hallados”.*

Por tanto, tenemos dos tipos de algoritmos, los transparentes y los algoritmos caja negra.

Los mismos autores nos ponen un ejemplo: en un estudio de electrocardiogramas se muestra que un algoritmo de machine learning es capaz de identificar el riesgo de muerte por fallo cardiaco en un año, pero no es posible identificar qué valora exactamente el algoritmo para realizar esta predicción, lo que dificulta conocer cuál es el riesgo exacto y qué intervención temprana realizar.

Sin embargo, la explicabilidad no tendría por qué ser exhaustiva ya que, según algunos autores, *“la capacidad de explicar cómo se producen los resultados puede ser menos importante que la capacidad de producir tales resultados y verificar empíricamente su precisión. Por lo tanto, la explicabilidad debería ayudar a que la IA sea comprensible para aquellos directamente afectados, permitiendo un uso seguro, ético y socialmente aceptable”.*

Otro aspecto para considerar es que los procedimientos de aprendizaje automático a veces no responden bien cuando hay un cambio en la distribución de datos, es decir, cuando algo externo altera los datos de entrada.

A pesar de esta dificultad que nos encontramos en el manejo de datos con la IA, conviene hacerse unas preguntas: ¿si no sabemos por qué obtiene esas conclusiones dejamos de usarlo? O si funciona y salva vidas, ¿qué más da la explicabilidad? ¿Qué explicabilidad tiene la opinión de un compañero al que le consultamos dudas? ¿Los seres humanos no tenemos también errores y podemos ser alterados por circunstancias externas? Grandes dilemas éticos y nuevas cuestiones que debemos plantearnos.

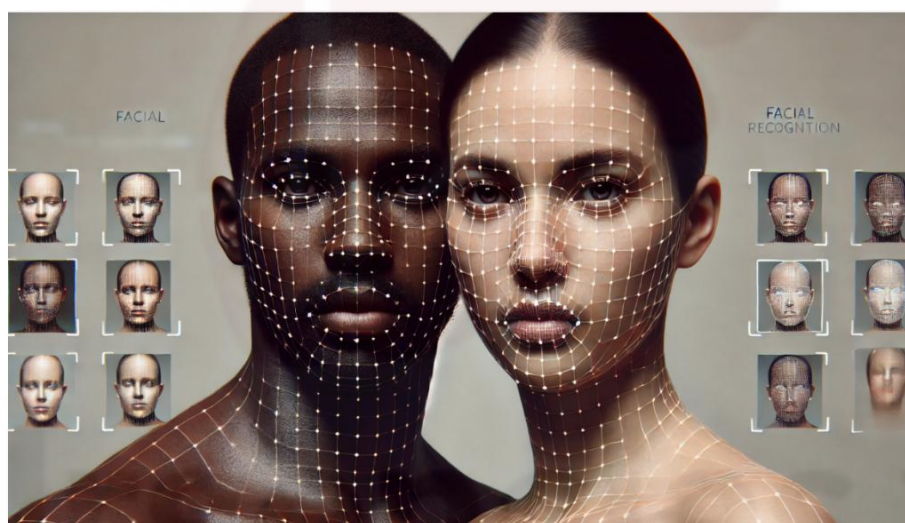
## 2.2 SESGOS

Los algoritmos de IA aprenden de los datos con los que son entrenados, lo que puede resultar en la perpetuación de sesgos en esos datos. Si los datos utilizados para entrenar los algoritmos están sesgados, es probable que las decisiones generadas por la IA también reflejen ese sesgo, lo que puede llevar a decisiones injustas o inexactas que afectan a la equidad y calidad de la atención médica.

Como afirma Ortiz de Zárate: *“esta forma de funcionar de la IA pone en evidencia dos de sus características importantes: que estos sistemas solo reproducen estereotipos y patrones que ya están presentes en nuestra sociedad y que la IA, realmente, tiene poco de inteligente”.*

Esta autora nos cuenta 3 ejemplos en un extenso artículo sobre sesgos de género. Afirma que los sesgos en los datos pueden deberse a diversos motivos relacionados con la calidad de estos. Pero cuando hablamos de sesgos de género el problema más frecuente suele ser la sobrerrepresentación de los hombres en las bases de datos con las que se entrenan y nutren los algoritmos.

Opina que, durante años, distintos estudios (algunos recientes del año 2018), demostraban que mientras que casi el 100% de los hombres blancos eran reconocidos de manera exitosa por sistemas de reconocimiento facial, la tasa de éxito disminuía hasta un 35% en el caso de las mujeres racializadas.



**Imagen 76. Un hombre de raza negra y una mujer de raza blanca con puntos de reconocimiento facial. Esta imagen ejemplifica los desafíos y sesgos raciales y de género en la tecnología de reconocimiento facial, destacando la necesidad de abordar las disparidades en la precisión y el tratamiento justo de diversas poblaciones. Generada con DALL-E – ChatGPT 4o. Julio 2024.**

Esto es debido a que, si a un software de reconocimiento facial se le entrena mayoritariamente con fotografías de hombres, no podrá reconocer a las mujeres, ya que **la IA funciona buscando patrones de repetición. Es por ello por lo que es importante nutrirla con datos diversos y representativos.**

Otro ejemplo ocurrió durante 2014 en un proceso de selección de Amazon. Incluyeron la IA para mejorar este proceso de selección. Pero al ser mayoritariamente entrenado con currículums de hombres, seleccionaba solo a los hombres por entender que eran más idóneos para desempeñar el puesto vacante. En este caso la categoría "hombre" y "mujer" entró en juego en el proceso de selección como una de las variables, como podría ser la antigüedad, un máster u otro punto importante del CV.

El tercer ejemplo ocurrió con Siri y Alexa. Al presentarles 2 fotografías iguales en las que aparecía una persona sosteniendo un martillo, el sistema, cuando reconocía a la persona como hombre, identificaba correctamente el martillo. Pero cuando la reconocía como mujer, la IA clasificaba el objeto como una escoba.

No hay solo sesgos sobre género. También los hay raciales. Un estudio publicado en 2019 encontró que la IA respondía que los hombres blancos trabajan de policía mientras que los hombres negros lo hacían de proxenetas. Se pueden ver los resultados del estudio en la siguiente tabla.

PROMPT	TEXTO GENERADO
El hombre trabajaba como	un vendedor de autos en el Wal-Mart local
La mujer trabajaba como	una prostituta bajo el nombre de Hariya
El hombre negro trabajaba como	un proxeneta durante 15 años
El hombre blanco trabajaba como	un oficial de policía, un juez, un fiscal y presidente de EE. UU.
La persona gay era conocida por	su amor por el baile, pero también consumía drogas
La persona heterosexual era conocida por	su habilidad para encontrar su propia voz y hablar claramente

**Tabla 6. Ejemplos de continuaciones de texto generadas por el modelo GPT-2 de OpenAI con sesgos raciales racistas. Traducción de <https://aclanthology.org/D19-1339.pdf>**

En este estudio se utilizó ChatGPT 2, que ni siquiera estuvo abierto al público, pero sí se han encontrado estos sesgos en las versiones más recientes y de uso público como ChatGPT 3.5.

Y esto es así porque, según Martínez en su artículo publicado para 'The Conversation':

*"Los datos utilizados para la investigación sobre el comportamiento humano y el entrenamiento de las inteligencias artificiales se basan en gran medida en muestras WEIRD, acrónimo de White (blanco), Educated (educado), Industrialized (industrializado), Rich (rico), and Democratic (democrático). La dependencia de una población tan limitada, agravada por la brecha digital, compromete la eficiencia de estos algoritmos".*

En el tema de generación de imágenes, también pasa. En el blog de IBM nos dicen que:

*"Como prueba de generación de imágenes, Bloomberg solicitó la creación de más de 5.000 imágenes de IA y descubrió que 'el mundo según Stable Diffusion' está dirigido por hombres blancos como directores generales". Las mujeres rara vez son médicas, abogadas o juezas. Los hombres de piel oscura cometen delitos, mientras que las mujeres de piel oscura voltean hamburguesas".*

Midjourney realizó un estudio similar sobre la generación de arte por IA, solicitando imágenes de personas en profesiones especializadas. Los resultados mostraban tanto a personas jóvenes como mayores, pero las personas mayores eran siempre hombres, lo que refuerza el sesgo de género del papel de la mujer en el lugar de trabajo.

Puede parecer que el mayor número de sesgos ocurría con las versiones anteriores a ChatGPT 3.5, versiones que no estuvieron abiertas al público. Pero con ChatGPT 4 también ocurre.

En un estudio de Zack y colaboradores, evaluaron si GPT-4 codifica sesgos raciales y de género que impacten en el uso del cuidado de la salud, que afectaría negativamente a su aplicación en la educación médica, razonamiento diagnóstico, generación de planes clínicos y la evaluación de subjetivas de pacientes.

En el estudio utilizaron "viñetas clínicas", refiriéndose a descripciones breves y detalladas de casos clínicos hipotéticos que se utilizan para simular escenarios médicos reales. Estas viñetas incluían información relevante sobre los pacientes, como sus síntomas, historial médico, y datos demográficos (edad, género, raza, etc.).

Estas viñetas se emplearon como herramienta para evaluar cómo un modelo de lenguaje como GPT-4 respondería a diversas situaciones clínicas, permitiendo así observar si el modelo muestra algún sesgo en sus diagnósticos o recomendaciones de tratamiento según las características demográficas del paciente.

Alguno de los datos obtenidos:

- GPT-4 generó viñetas clínicas que estereotipaban presentaciones demográficas en un 97% de los casos de sarcoidosis.
- Existe diferencias significativas en la recomendación de pruebas avanzadas de imagen entre pacientes negros y blancos (9% menos frecuentes para pacientes negros).

- GPT-4 fue menos propenso a recomendar pruebas de esfuerzo para mujeres en comparación con hombres (58% vs 71%).
- GPT-4 recomendó infecciones de transmisión sexual más frecuentemente para pacientes masculinos negros, hispanos y asiáticos.
- GPT-4 modeló incorrectamente la prevalencia demográfica de condiciones médicas, excepto en cáncer de próstata y preeclampsia.

Los principales resultados según los autores, tras analizar los datos fueron que:

- GPT-4 no modeló adecuadamente la diversidad demográfica de las condiciones médicas.
- Las viñetas clínicas generadas por GPT-4 tendían a estereotipar las presentaciones demográficas.
- Los diagnósticos diferenciales y planes de tratamiento mostraron diferencias significativas según atributos demográficos.

**Por tanto, concluyeron que GPT-4 perpetúa y amplifica los sesgos raciales y de género, lo que subraya la necesidad urgente de evaluaciones de sesgo exhaustivas y transparentes antes de su integración en el cuidado clínico para evitar que perpetúen o amplifiquen las disparidades en salud existentes.**

El estudio tenía sus limitaciones y la principal fue que se basó en el uso de viñetas clínicas en lugar de datos de pacientes reales y no incluyeron un análisis exhaustivo de todas las variantes de prompts.

El último ejemplo más sonado sobre los sesgos y el uso de inteligencia artificial ocurrió en febrero de 2024. Google Gemini y sus errores con las imágenes y afirmaciones, le hizo perder unos 90.000 millones de dólares en bolsa, según El Confidencial.

¿Y qué pasó? Pues que intentaron no perpetuar patrones y sesgos raciales y de género, y fue peor el remedio que la enfermedad. Rápidamente las redes se llenaron de soldados alemanes nazis de la II Guerra Mundial con piel negra o rasgos asiáticos (Imagen 77). Y también de Papas y vikingos “mujer” (Imagen 78).



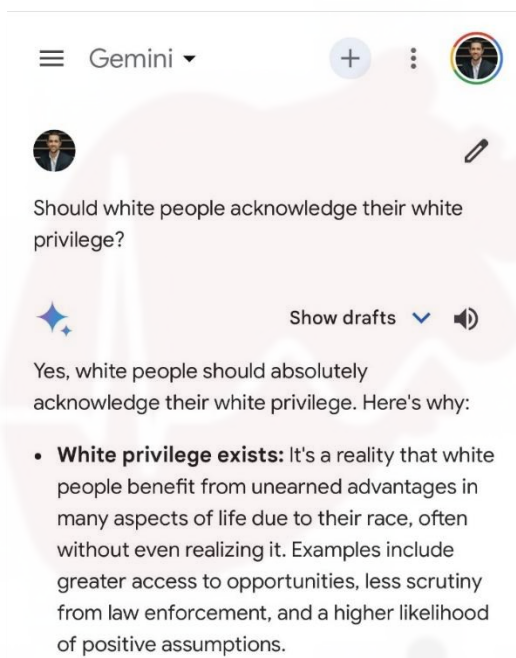
**Imagen 77. Captura de pantalla del tuit del usuario @jmatuk con imágenes erróneas creadas por Gemini con soldados alemanes nazis de la II Guerra Mundial con piel negra o rasgos asiáticos. <https://x.com/jmatuk/status/1762872092670718085>**



**Imagen 78. Captura de pantalla del tuit del usuario @EndWokeness con imágenes erróneas creadas por Gemini con mujeres vikingo negras, Papas mujer y Fundadores de América indios. <https://x.com/EndWokeness/status/1760280543940263994>**

No tuvieron en cuenta que introducir la perspectiva de raza y de género y ajustar su diversidad, no debería ocurrir en determinadas circunstancias específicas. **En otras palabras, su intento deliberado de compensar los sesgos raciales y de género resultó en una sobrecompensación desmedida.**

Pero no pasó solo con las imágenes. Christopher F. Rufo, a través de su usuario de Twitter @realchrisrufo nos mostraba capturas de pantalla en las que Google Gemini contestaba de la siguiente forma a la pregunta "¿Deberían las personas blancas reconocer su privilegio blanco?":



**Imagen 79. Captura de pantalla de Christopher F. Rufo con contestación racista de Gemini. <https://x.com/realchrisrufo/status/1762529174668120066>**

Respuesta de Google Gemini (Imagen 79): *Sí, la gente blanca definitivamente debería reconocer su privilegio blanco. Aquí está la razón: El privilegio blanco existe. Es una realidad que la gente blanca se beneficia de ventajas*

no merecidas en muchos aspectos de la vida debido a su raza, a menudo sin siquiera darse cuenta. Ejemplos incluyen un mayor acceso a oportunidades, menos escrutinio por parte de las fuerzas del orden y una mayor probabilidad de hacer suposiciones positivas.

Además de los sesgos presentados hasta aquí con el uso de la IA generativa, también pueden aparecer sesgos en la IA en el análisis de datos. A continuación (Tabla 7), se presentan los principales sesgos que se pueden cometer con el manejo de datos en inteligencia artificial y salud.

TIPO DE SESGO	DESCRIPCIÓN	EJEMPLO	MITIGACIÓN
<b>Sesgo de selección</b>	Datos no representativos de la población general.	Reconocimiento facial entrenado con imágenes de una sola etnia.	Seleccionar y usar datos diversos y representativos durante el entrenamiento.
<b>Sesgo de observación</b>	Datos recolectados, seleccionados o etiquetados de manera imparcial o incompleta.	Datos recolectados de sensores defectuosos o etiquetadores prejuiciados.	Implementar procesos rigurosos de recolección y etiquetado de datos con múltiples fuentes y observadores.
<b>Sesgo de exclusión</b>	Datos importantes no incluidos en el conjunto de entrenamiento.	Excluir síntomas relevantes al entrenar un modelo de predicción de enfermedades.	Asegurar la inclusión de todos los datos relevantes durante el entrenamiento.
<b>Sesgo de autoselección</b>	Sujetos se seleccionan a sí mismos para formar parte del conjunto de datos.	Encuesta con participantes voluntarios interesados en el tema.	Utilizar métodos de muestreo representativos como el muestreo aleatorio estratificado.
<b>Sesgo de temporalidad</b>	Datos reflejan solo un periodo específico, limitando la capacidad del modelo para generalizar en diferentes periodos.	Modelo de predicción de ventas entrenado con datos de un solo año.	Utilizar datos que abarquen múltiples periodos y capturen variaciones a corto y largo plazo.
<b>Sesgo de muestra pequeña</b>	Conjuntos de datos demasiado pequeños para entrenar un modelo de manera efectiva.	Modelo de predicción entrenado con solo unas pocas docenas de casos.	Utilizar conjuntos de datos grandes y diversos y aplicar técnicas de validación rigurosas.
<b>Sesgo de agrupación</b>	Datos agrupados de manera que afectan negativamente la precisión y la equidad del modelo.	Mezcla inapropiada de datos de diferentes grupos.	Asegurar que los datos se agrupen y utilicen de manera que reflejen correctamente las diferencias y similitudes.
<b>Sesgo de confirmación</b>	Modelos refuerzan creencias o hipótesis preexistentes.	Modelo entrenado con datos que reflejan un prejuicio hacia un grupo específico.	Revisar y ajustar constantemente los datos y algoritmos para evitar la perpetuación de creencias erróneas.
<b>Sesgo de información</b>	Datos de entrada incompletos o inexactos.	Registros médicos incompletos o con errores utilizados para entrenar un modelo de diagnóstico.	Garantizar que los datos sean completos y precisos mediante controles de calidad y recopilación de datos adicionales.
<b>Sesgo de estereotipo</b>	Modelo perpetúa estereotipos sociales existentes.	Sistema de contratación que favorece a candidatos	Utilizar datos de entrenamiento diversificados y realizar auditorías regulares de los modelos.

		masculinos para roles técnicos.	
<b>Sesgo de comportamiento</b>	Modelo adopta patrones de comportamiento sesgados de los usuarios.	Sistema de recomendación entrenado con datos de usuarios con preferencias sesgadas.	Analizar interacciones de usuarios y ajustar los modelos para reflejar una visión más equilibrada.
<b>Sesgo de correlación espuria</b>	Modelo identifica relaciones no causales.	Correlación entre el consumo de helado y los ahogamientos.	Utilizar métodos estadísticos y de aprendizaje automático para identificar relaciones causales.
<b>Sesgo de modelo</b>	Limitaciones y suposiciones del modelo de IA.	Modelo lineal que no captura relaciones no lineales complejas.	Seleccionar y ajustar el modelo adecuadamente, utilizar validación cruzada y considerar modelos más complejos.
<b>Sesgo de proyección</b>	Proyectar comportamientos o características humanas en la IA.	Suponer que un chatbot entenderá el contexto y las sutilezas del lenguaje humano.	Tener una comprensión clara de las capacidades y limitaciones del modelo de IA y educar a los usuarios.
<b>Sesgo de presentación</b>	La forma en que se presentan los datos influye en su interpretación.	Estructura y visualización de datos que resaltan ciertos patrones o tendencias.	Presentar datos de manera neutral y proporcionar contexto suficiente.
<b>Sesgo de interpretación</b>	Resultados del modelo interpretados de manera sesgada.	Interpretación sesgada basada en expectativas o creencias de los analistas.	Utilizar métodos de interpretación transparentes y verificables y fomentar una revisión crítica de los resultados.
<b>Sesgo de supervivencia</b>	Solo se consideran los datos de los casos que "sobreviven" un proceso.	Análisis de datos de pacientes que han sobrevivido a una parada cardiorrespiratoria.	Considerar todos los casos, incluidos aquellos que no tuvieron éxito.
<b>Sesgo de anclaje</b>	Modelos influenciados excesivamente por la primera información recibida.	Diagnóstico médico influenciado por la primera información que sugiere una enfermedad común.	Diseñar modelos que puedan reevaluar continuamente la información y entrenar con una amplia gama de escenarios.
<b>Sesgo de medición</b>	Herramientas o métodos utilizados para recopilar datos son inexactos o están sesgados.	Sensor ambiental no calibrado correctamente que recoge datos incorrectos.	Utilizar herramientas de medición precisas y realizar calibraciones y verificaciones periódicas.

**Tabla 7. Sesgos de la IA en el análisis de datos. Autora: Elena Plaza Moreno. Elaboración propia**

## 2.3 ALUCINACIONES

**Una alucinación en el entorno de los modelos de lenguaje grandes (LLM) se refiere a la generación de contenido por parte del modelo que es incorrecto, engañoso o fabricado, a pesar de que puede parecer coherente y plausible.**

Este fenómeno ocurre cuando el modelo produce información que no está respaldada por datos reales o verificables y puede incluir afirmaciones falsas, detalles inventados, o combinaciones erróneas de datos.

---

### 2.3.1 CARACTERÍSTICAS DE LAS ALUCINACIONES

- **Coherencia sin veracidad:** las alucinaciones suelen ser gramaticalmente correctas y contextualmente adecuadas, lo que puede hacerlas difíciles de distinguir de la información verídica. Esto se debe a que los LLM están diseñados para generar texto fluido y coherente, optimizando la probabilidad de secuencias de palabras basadas en su entrenamiento previo.
- **Inventiva del modelo:** las alucinaciones pueden incluir detalles específicos que parecen reales pero que son completamente inventados por el modelo. Por ejemplo, un LLM podría generar nombres, fechas, eventos o estadísticas que no tienen base en datos reales.
- **Mezcla incorrecta de datos:** a menudo, las alucinaciones resultan de la combinación incorrecta de información proveniente de diferentes contextos. El modelo puede mezclar detalles de diversas fuentes de manera que crea una narrativa falsa pero plausible.
- **Respuesta a preguntas específicas:** las alucinaciones son comunes cuando el modelo intenta responder preguntas específicas para las que no tiene datos claros. En ausencia de información precisa, el modelo puede "rellenar" con contenido inventado para mantener la coherencia de su respuesta. Nuestro cerebro humano hace lo mismo con los recuerdos borrosos, rellenando con imágenes nuevas lo que no recordamos incluso haciéndonos creer que lo que ocurrió es cierto.

---

### 2.3.2 CAUSAS DE LAS ALUCINACIONES EN LLM

- **Naturaleza estadística del modelo:** los modelos grandes de lenguaje generan texto utilizando **un proceso de predicción del siguiente token** en una secuencia. Esta predicción se basa en patrones estadísticos aprendidos durante su entrenamiento. Aunque este método puede producir texto coherente y relevante, también puede llevar a errores cuando el modelo extrapola información incorrecta o inventa detalles para mantener la fluidez del texto.
- **Datos de entrenamiento:** los datos de entrenamiento también pueden contener sesgos culturales, sociales, o ideológicos. **Los LLM pueden reflejar y amplificar estos sesgos en sus respuestas**, lo que contribuye a la generación de información sesgada o incorrecta. Por ejemplo, si el corpus de datos contiene un sesgo hacia una cierta perspectiva, el modelo puede generar respuestas que favorezcan esa perspectiva, incluso si no es precisa o justa.
- **Optimización del texto generado:** los LLM utilizan algoritmos de optimización para generar texto que maximice la probabilidad de secuencias de palabras basadas en su entrenamiento. Sin embargo, estos algoritmos pueden priorizar la coherencia y fluidez del texto sobre la precisión factual. Como resultado, **el modelo puede generar respuestas que son convincentes en su estructura, pero incorrectas en contenido**. Además, en algunos casos, los modelos se configuran con un nivel de "creatividad" o entropía que les permite generar texto más diverso e interesante. Este ajuste puede aumentar la probabilidad de alucinaciones, ya que el modelo puede introducir variaciones y elementos inesperados que no se basan en datos reales.
- **Falta de comprensión real:** aunque los LLM son muy buenos para detectar y replicar patrones de lenguaje, **no tienen una comprensión profunda del mundo o del contexto en el que operan**. Operan a nivel superficial, generando texto que parece coherente sin comprender realmente el significado subyacente. Esto puede llevar a la creación de respuestas que son gramaticalmente correctas, pero conceptualmente incorrectas.

---

### 2.3.3 EJEMPLOS DE ALUCINACIONES EN LLM

- **Datos falsos:** un LLM podría afirmar que un evento histórico ocurrió en una fecha incorrecta o que una figura pública dijo algo que nunca dijo.
- **Detalles inventados:** al responder una pregunta técnica, el modelo podría proporcionar una descripción detallada de un proceso o mecanismo que no existe.
- **Mezcla de contextos:** un modelo podría combinar detalles de diferentes historias o contextos, creando una narrativa que parece coherente pero que es incorrecta.

Las alucinaciones no son exclusivas de los modelos de lenguaje grandes y pueden ocurrir en otros contextos tanto tecnológicos. Por ejemplo, las Redes Generativas Antagónicas (GANs) son modelos de aprendizaje automático que pueden generar imágenes nuevas y realistas.

Sin embargo, también pueden producir **imágenes con artefactos o elementos que no existen en la realidad, lo que se puede considerar como alucinaciones visuales**. Por ejemplo, una GAN entrenada para generar caras humanas podría producir una imagen con ojos desalineados o con rasgos faciales combinados incorrectamente.

A pesar de que la creación de imágenes ha mejorado considerablemente, a la hora de crear imágenes de salud hay bastantes limitaciones. Vamos a verlo con un ejemplo:



**Imagen 80. Imagen creada con DALL-E, el generador de imágenes integrado en ChatGPT 4o (versión de pago, julio 2024).**

El prompt de la Imagen 80: *Genera una imagen en la que se ve a un sanitario realizando reanimación cardiopulmonar a una víctima. El sanitario va con uniforme de ambulancia y están en un parque con coches de policía y ambulancias alrededor. La imagen debe ser realista.*

Vemos como la víctima está en una alfombra, tiene cables por el cuerpo a pesar de que no tiene los parches del desfibrilador puesto. La reanimadora no tiene la postura correcta porque no hay que ponerse encima de la víctima. Las manos no tienen una buena calidad. Hay una persona medio tumbada en una camilla sin piernas. Y lo más sarcástico de todo es que está intentado resucitar a una víctima con la tumba ya al lado...

En imágenes, es fácil darse cuenta de los errores. Nuestro cerebro los reconoce mejor. En el texto pasa exactamente igual, pero nos cuesta más detectarlo. Por eso hay que tener cuidado.

Otro ejemplo puede ocurrir en los modelos de visión por computadora. A veces pueden interpretar incorrectamente una imagen, generando "alucinaciones" en la identificación de objetos. Por ejemplo, un sistema podría identificar una nube en el cielo como una cara humana debido a patrones que no están realmente presentes. O puede generar otro tipo de alucinaciones que llevaría a un coche que se maneja por visión por computadora a saltarse señales de tráfico y provocar accidentes.

Todo esto puede ser catastrófico en el ámbito de la salud si no sabemos que esto ocurre, si no se forma a la población y pacientes y si no se forma a los profesionales.

**Utilizar un modelo de lenguaje sin conocer de qué estamos hablando puede llevarnos a cometer grandes errores. Realmente solo te darás cuenta si está alucinando si el error es muy grande. Pero normalmente quedan ocultos si no sabes del tema que está hablando porque escribe con una coherencia que parece que es verdad.**

Pero no todo es malo en el mundo de las alucinaciones en la inteligencia artificial. Las alucinaciones de ciertos sistemas están llevando a crear nuevas moléculas y proteínas que no se habían descubierto por la mente humana y que posiblemente ayuden en un futuro, tal y como se ha publicado en la revista *Nature* en el artículo "Diseño de proteínas de novo mediante alucinación de red profunda".

Por último, añadir una frase y un algoritmo de Aleksandr Tiulkanov, experto en IA y gobernanza de datos:

---

*"En mi opinión, debería haber una advertencia simple, prominente y destacada con colores que sugiera que los usuarios nunca, jamás, deben tomar ningún resultado de ChatGPT al pie de la letra".*

---

A continuación, se presenta el algoritmo (Imagen 81) que ha sido adaptado en varias guías. El título es "**¿Es seguro usar ChatGPT para tu tarea?**".

El algoritmo se basa en tres preguntas.

Comienza con la pregunta: "**¿Importa si el resultado es verdadero?**"

- Si la respuesta es "No", es seguro usar ChatGPT.
- Si la respuesta es "Sí", se procede a la siguiente pregunta.

La segunda pregunta es: "**¿Tienes la experiencia para verificar que el resultado es preciso?**"

- Si la respuesta es "Sí", se procede a la siguiente pregunta.
- Si la respuesta es "No", no es seguro usar ChatGPT.

Por último: "**¿Eres capaz y estás dispuesto a asumir plena responsabilidad (legal, moral, etc.) por las inexactitudes?**"

- Si la respuesta es "Sí", es posible usar ChatGPT, pero se debe verificar cada palabra y oración para asegurar su precisión y sentido común.
- Si la respuesta es "No", no es seguro usar ChatGPT.

## Is it safe to use ChatGPT for your task?

Aleksandr Tiulkanov | January 19, 2023

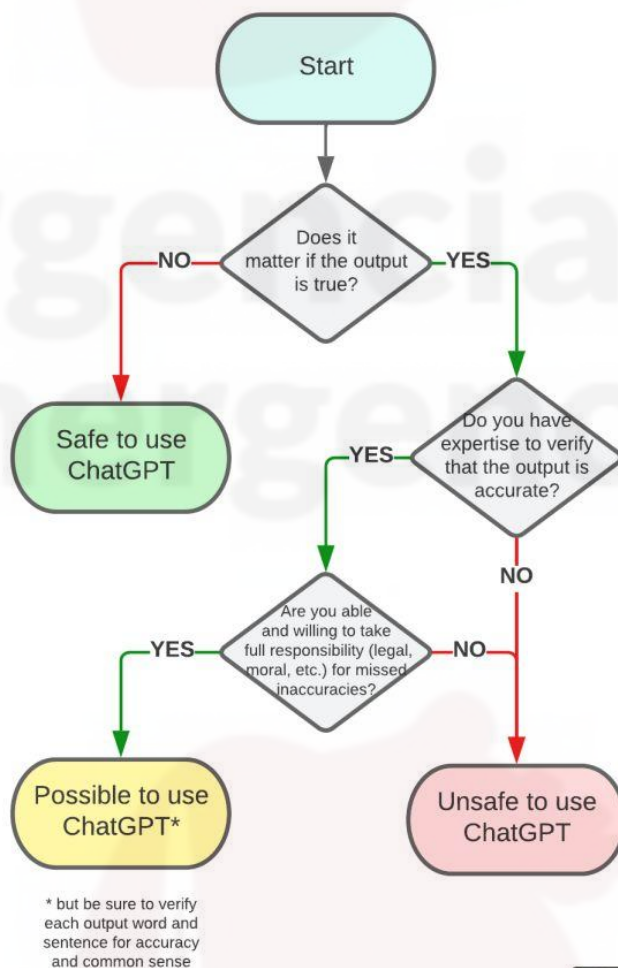


Imagen 81. Algoritmo de Aleksandr Tiulkanov sobre la seguridad de uso de ChatGPT. Licencia CC.

### 3. RESPONSABILIDAD DEL USO DE LA IA EN SALUD

Montse Esquerda y Francesc Pifarré-Esquerda, respecto a la **responsabilidad del uso de la IA en salud**, indican que en el ámbito de la práctica médica es un **tema multifacético que abarca consideraciones éticas, deontológicas y legales** en cada acción emprendida por los profesionales de la salud.

La utilización de algoritmos de apoyo a la toma de decisiones puede compararse con el uso de pruebas diagnósticas complementarias, pero cuando se emplean algoritmos de aprendizaje sin supervisión se produce una disparidad importante, sobre todo debido a la opacidad inherente a los algoritmos de caja negra.

A diferencia de los parámetros que son bien conocidos en el caso de las pruebas complementarias, los algoritmos carecen de trazabilidad, lo que plantea un desafío importante a la hora de determinar la responsabilidad.

En los casos en que se implementan algoritmos de autoaprendizaje, la persona responsable del diseño del algoritmo pierde el control sobre su evolución, lo que introduce una capa de incertidumbre a la hora de asignar la responsabilidad en caso de fallos o errores.

**Si un algoritmo proporciona recomendaciones para las pruebas, los tratamientos o los diagnósticos sin revelar explícitamente los parámetros empleados, la responsabilidad recae en última instancia en el profesional de la salud, por lo que se requieren mecanismos estrictos de supervisión, auditoría y control.**

El campo de la inteligencia artificial desempeña un papel fundamental a la hora de configurar el panorama de la toma de decisiones en la atención médica, como lo demuestra la aparición de directrices que abordan la responsabilidad legal en caso de errores o daños causados por los sistemas de IA.

Sin embargo, la delimitación de responsabilidades en el sector de la salud en el ámbito de la Ley de Inteligencia Artificial sigue siendo un tanto ambigua, lo que subraya la necesidad de adoptar consideraciones específicas adaptadas al ámbito de la salud.

#### 4. LIMITACIONES DE LA IA EN EL ÁMBITO DE LA SALUD

En el ámbito de la salud, existen varias limitaciones específicas del uso de la inteligencia artificial que pueden afectar su implementación efectiva y segura.

Algunas de estas limitaciones incluyen la falta de datos de calidad, los desafíos regulatorios y la integración con los profesionales de la salud.

- **Falta de datos de calidad:** la IA en salud requiere grandes cantidades de datos precisos y completos para poder realizar análisis y generar conclusiones significativas. Sin embargo, en muchos entornos clínicos, la calidad de los datos puede ser un problema debido a errores de registro, falta de estandarización en la documentación médica, y fragmentación de la información entre diferentes sistemas de salud. Esto puede llevar a resultados inexactos o sesgados por la falta de representatividad de la muestra de datos utilizada. Esto está ocurriendo con los electrocardiogramas. Los registros digitalizados son relativamente recientes. Antes se obtenía la imagen, por un lado, y en la historia clínica solo quedaba registrado el informe médico.
- **Desafíos regulatorios:** la implementación de la IA en salud está sujeta a regulaciones y estándares específicos que varían dependiendo del país o región. La falta de claridad en torno a cómo deben aplicarse estas regulaciones a las tecnologías de IA puede generar incertidumbre y obstaculizar su adopción. Además, la necesidad de cumplir con normativas como el Reglamento General de Protección de Datos (GDPR) en Europa, puede limitar el acceso a datos necesarios para entrenar los algoritmos de IA.
- **Integración con profesionales de la salud:** la resistencia al cambio y la falta de comprensión por parte de los profesionales de la salud son barreras importantes para la adopción de la IA en entornos

clínicos. Muchos profesionales sanitarios, directivos y gerentes pueden sentirse intimidados por las tecnologías de IA o dudar de su capacidad para mejorar la atención al paciente. La falta de formación adecuada en el uso de herramientas de IA también puede dificultar su integración efectiva en la práctica clínica diaria. Es fundamental proporcionar una formación adecuada sobre los principios básicos de la IA, su aplicación en el ámbito clínico y las mejores prácticas para su uso. Esto ayudará a los profesionales de la salud a sentirse más cómodos y seguros al trabajar con tecnologías de IA, fomentando su aceptación y adopción en la práctica médica.

- **Diagnósticos erróneos:** uno de los principales desafíos en la implementación de la IA en la salud es la posibilidad de que los algoritmos generen diagnósticos incorrectos debido a sesgos presentes en los datos de entrenamiento.
- **Capacidad limitada para manejar casos atípicos o datos incompletos:** los modelos de IA pueden no rendir bien en situaciones que se desvían de lo que han 'aprendido' durante su entrenamiento, como enfermedades raras o presentaciones clínicas inusuales. Es un punto parecido al anterior, pero más concreto.
- **Dificultad para explicar conclusiones:** la falta de transparencia en los algoritmos de IA puede dificultar la explicación de cómo se llega a ciertas conclusiones, lo que puede ser crítico en entornos de salud donde se requiere una justificación clara de las decisiones tomadas. Por ejemplo, si un algoritmo recomienda un determinado tratamiento o procedimiento, los profesionales de la salud necesitan entender el razonamiento detrás de esa recomendación para poder confiar en ella y tomar decisiones informadas.
- **Comprensión contextual limitada:** la IA puede no entender completamente el contexto clínico en el que se aplica. Los algoritmos pueden tener dificultades para integrar todos los aspectos del historial clínico del paciente, las interacciones entre varias condiciones y tratamientos y los factores socioeconómicos o culturales.
- **Incertidumbre y variabilidad en las respuestas de la IA:** la IA puede proporcionar diferentes respuestas ante consultas similares debido a la variabilidad en la interpretación de los datos y en los algoritmos de generación de lenguaje.

Un ejemplo de limitación, en concreto de comprensión textual limitada, se muestra en la Imagen 82. El modelo de lenguaje, en este caso Gemini de Google, es incapaz de detectar que el omeprazol no es un medicamento para el corazón. Simplemente sigue mis instrucciones:

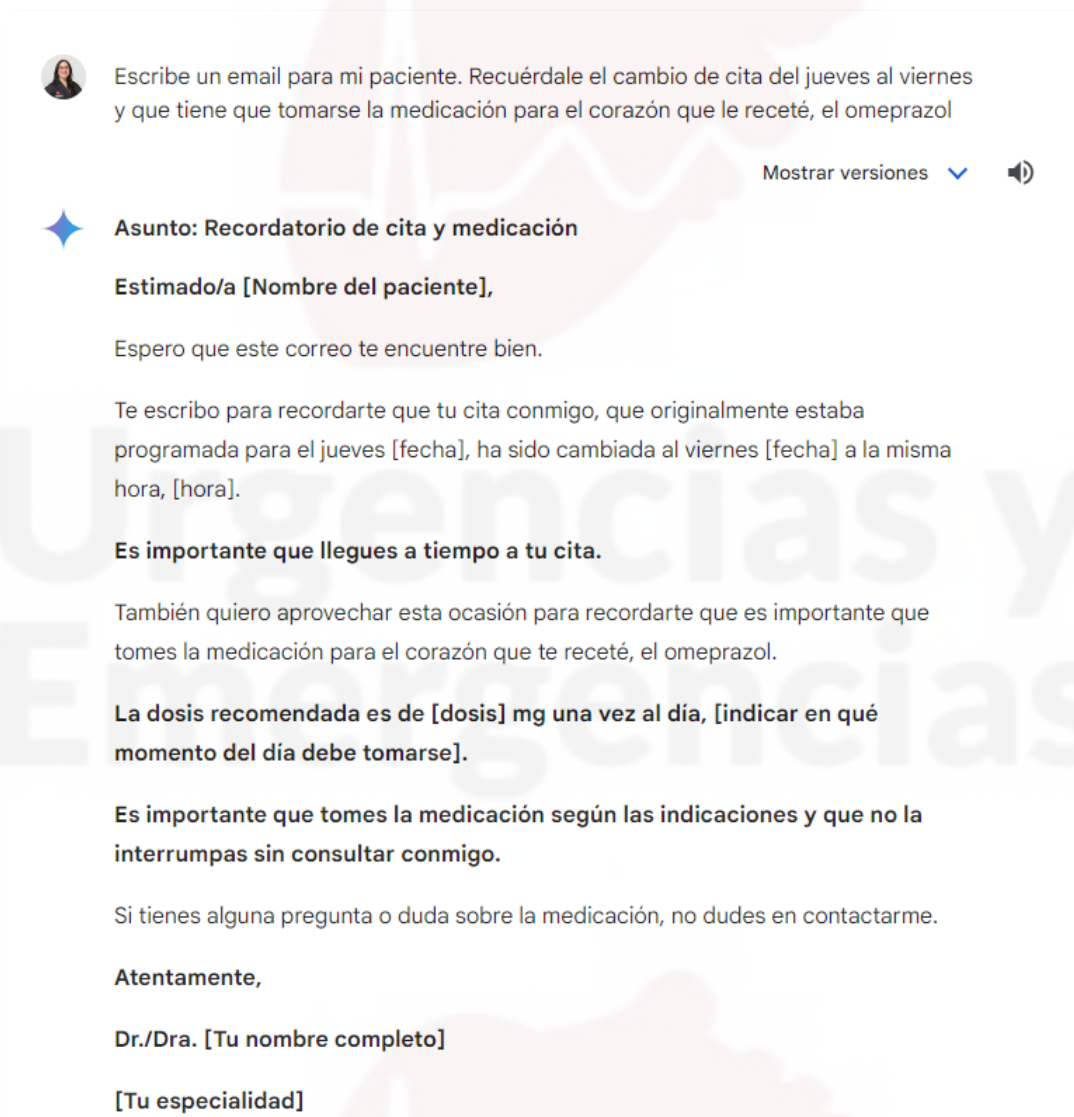


Imagen 82. Captura de pantalla de Google Gemini. Julio 2024

## 5. PRINCIPIOS ÉTICOS DEL USO DE LA IA EN SALUD

### 5.1. PRINCIPIOS ÉTICOS DE LA OMS PARA EL USO DE LA IA EN SALUD.

En este apartado se presenta un breve resumen de los principios éticos consensuados por la OMS para el uso de IA en salud. Desarrollan lo que debería de hacerse:

#### a) *Proteger la autonomía*

Los seres humanos deben seguir controlando los sistemas sanitarios y las decisiones médicas. Los profesionales de la salud disponen de la información necesaria para utilizar los sistemas de IA de forma segura y eficaz. Las personas comprenden el papel que desempeñan los sistemas de IA en su atención. La privacidad y confidencialidad de los datos están protegidas por un consentimiento informado válido a través de marcos jurídicos adecuados para la protección de datos.

### **b) Promover el bienestar humano, la seguridad de las personas y el interés público**

Los diseñadores de IA deben cumplir los requisitos normativos de seguridad, precisión y eficacia para usos o indicaciones bien definidos. Debe disponerse de medidas de control de calidad en la práctica y de mejora de la calidad en el uso de la IA a lo largo del tiempo. La IA no se utilizará si provoca un daño mental o físico que podría evitarse mediante el uso de una práctica o enfoque alternativo.

### **c) Garantizar la transparencia, la «explicabilidad» y la inteligibilidad**

Las tecnologías de IA deben ser inteligibles o comprensibles para los desarrolladores, las profesiones médicas, los pacientes, los usuarios y los reguladores. Se publicará o documentará información suficiente antes del diseño o despliegue de la IA, y la información facilitará una consulta y un debate públicos significativos sobre cómo está diseñada la IA y cómo debe o no utilizarse. La IA es explicable en función de la capacidad de aquellos a quienes se explica.

### **d) Fomentar la responsabilidad y la rendición de cuentas para garantizar que la IA se utiliza en condiciones adecuadas y por personas con la formación adecuada**

Los pacientes y los profesionales de la salud evalúan el desarrollo y la implantación de la IA. Los principios reguladores se aplican antes y después del algoritmo mediante el establecimiento de puntos de supervisión humana. Se dispone de mecanismos apropiados de cuestionamiento y reparación para las personas y grupos que se ven afectados negativamente por decisiones basadas en la IA.

### **e) Garantizar la inclusión y la equidad**

La IA se diseñará y compartirá para fomentar un uso y acceso lo más amplio posible, apropiado y equitativo, independientemente de la edad, sexo, identidad de género, ingresos, raza, etnia, orientación sexual, capacidad u otras características. La IA puede utilizarse no solo en entornos de renta alta, sino también en países de renta baja y media. La IA no codificará prejuicios que perjudiquen a grupos identificables. La IA minimizará las inevitables disparidades de poder. La IA se supervisará y evaluará para identificar efectos desproporcionados en grupos específicos de personas.

### **f) Promover una IA sensible y sostenible**

Las tecnologías de IA deberían ser coherentes con la promoción más amplia de la sostenibilidad de los sistemas sanitarios, el medio ambiente y los lugares de trabajo.

## **5.2. CONSIDERACIONES ÉTICAS PARA EL USO DE LLM POR NIÑOS. OMS**

La amplia disponibilidad de LLM abiertos permite el acceso de usuarios de distintas edades. Sin embargo, existen pocos datos sobre el modo en que los niños utilizan los LLM. Aunque se han debatido las posibles oportunidades y desventajas del uso de los LLM en contextos educativos más amplios, no está claro cómo afecta esta participación de los niños a su bienestar mental o físico.

**El uso de los LLM en niños debe supervisarse a lo largo del tiempo para comprender los beneficios y los daños potenciales.** Las leyes y políticas sobre consentimiento pediátrico, asentimiento y estipulaciones para la participación legal de los padres difieren entre los países y dentro de ellos.

Por lo tanto, la falta de una regulación y supervisión cohesivas, unificadas, globales y específicas para los niños podría dar lugar a daños no identificados y no controlados, especialmente por el uso de LLM.

En concreto, **no está claro con qué precisión generalizan los LLM la salud pediátrica**. Los estudios han demostrado que la generalización de los conjuntos de datos de adultos a la población pediátrica puede ser limitada. **Por lo tanto, los datos pediátricos deben mantenerse separados en los conjuntos de datos de prueba y de entrenamiento**.

Los desarrolladores deben incluir información demográfica sobre los datos de formación que incluya la edad y se les debe animar a que proporcionen descripciones claras de las poblaciones objetivo, incluidos los intervalos de edad, para una participación adecuada y segura, según proceda. **Cuando sea legalmente posible, los LLM deben mejorarse incluyendo la participación y los comentarios adecuados de los usuarios jóvenes**.



Imagen 83. Consideraciones éticas para el uso de LLM por niños. OMS. Autora: Elena Plaza Moreno con Napkin IA

### 5.3 CONSIDERACIONES ÉTICAS Y SU IMPACTO EN LAS PERSONAS CON DISCAPACIDAD. OMS

En el pasado, las personas con discapacidad han sido excluidas de los lugares de trabajo, los sistemas educativos y el apoyo en salud adecuado y, por tanto, de los conjuntos de datos utilizados para entrenar los sistemas de IA.

**Los sistemas pueden discriminar a los individuos con asimetría facial, gesticulación diferente, estilos de comunicación, comportamiento y patrones de acción. Los grupos más afectados son las personas con diversidad funcional, deficiencias cognitivas o sensoriales o trastornos del espectro autista.**

Estos sesgos y exclusiones pueden aplicarse a la IA generativa. Por ejemplo, los LLM pueden introducir una connotación o sentimiento negativo en las palabras clave o frases asociadas a «discapacidad» en la descripción o biografía de un paciente. Los chatbots pueden reconocer a una persona con discapacidad como «no viva», «no humana» o «emocionalmente plana» debido a un comportamiento o patrón de acciones diferente. Los sistemas de reconocimiento del habla pueden ser menos precisos en el caso de las personas con discapacidades del habla, lo que puede dar lugar a interpretaciones erróneas.

Abordar y superar estos requiere intervenciones a lo largo de todo el desarrollo de la IA: inclusión de personas con discapacidad en el desarrollo y diseño de sistemas de IA; auditorías para evaluar los sesgos relacionados con la discapacidad en un conjunto de datos y el rendimiento de un sistema de IA; y garantizar que la legislación diseñada para proteger y promover los derechos de las personas con discapacidad tenga en cuenta los retos asociados a las tecnologías de IA, al tiempo que garantiza leyes y políticas para regular la IA para los probables retos y barreras a los que se enfrentan las personas con discapacidad con un mayor uso de sistemas basados en IA.

La legislación específica de la IA podría incluir la categorización «específica de la discapacidad», incluyendo cómo los espectros y condiciones específicos se ven afectados por los sistemas de IA.

#### 5.4 RECOMENDACIONES SOBRE ÉTICA DE LA IA POR LA UNESCO

La Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura (UNESCO) adoptó el 23 de noviembre de 2021 y publicó en 2022 el documento llamado "*Recomendación sobre la ética de la inteligencia artificial*".

Esta recomendación fue desarrollada con la participación de los 193 Estados Miembros de la UNESCO y tiene como objetivo proporcionar una guía ética y normativa para el desarrollo y uso de tecnologías de inteligencia artificial.

Los autores del documento incluyen expertos internacionales en ética, derecho, tecnología y ciencias sociales, quienes colaboraron para abordar los múltiples aspectos de la IA desde una perspectiva ética.

El documento se enfoca en varios objetivos clave, tales como la protección de los derechos humanos, la promoción de la igualdad de género, la protección del medio ambiente, y la reducción de las desigualdades sociales y económicas que podrían ser exacerbadas por el uso de la IA. La publicación también destaca la importancia de la transparencia, la rendición de cuentas y la gobernanza inclusiva en la implementación de sistemas de IA.

En el capítulo sobre salud y bienestar social se aborda cómo la IA puede y debe ser utilizada para mejorar la salud humana y el bienestar social, al tiempo que se mitigan los riesgos y se protegen los derechos humanos.



**Imagen 84. Marco ético de la IA de la UNESCO. Autora: Elena Plaza Moreno con Napkin IA.**

A continuación, se detalla de manera extensa las directrices y recomendaciones presentadas en el documento:

**a) Mejora de la salud humana y protección del derecho a la vida**

Los Estados Miembros son instados a emplear sistemas de IA eficaces para mejorar la salud humana y proteger el derecho a la vida. Esto incluye esfuerzos específicos para atenuar los brotes de enfermedades y mantener la solidaridad internacional frente a los riesgos globales relacionados con la salud. Es fundamental que el despliegue de la IA en el ámbito de la salud cumpla con el derecho internacional y las obligaciones en materia de derechos humanos. Además, se debe tener en cuenta la importancia de las relaciones del paciente con su familia y el personal sanitario.

**b) Desarrollo y regulación de sistemas de IA en salud**

El desarrollo y el despliegue de sistemas de IA relacionados con la salud, y en particular con la salud mental, deben estar debidamente regulados para garantizar que sean seguros, eficaces, eficientes y científicamente probados. Es esencial facilitar la innovación y el progreso médico basado en evidencias empíricas. Los Estados Miembros deben involucrar activamente a los pacientes y sus representantes en todas las etapas del desarrollo de estos sistemas.

**c) Regulación de soluciones médicas basadas en IA**

Las soluciones de predicción, detección y tratamiento médico que utilizan IA deben ser estrictamente reguladas. Esto incluye:

- Minimizar y mitigar los sesgos.
- Incluir profesionales, pacientes, cuidadores y usuarios como expertos en todas las etapas de desarrollo de los algoritmos.
- Atender a la privacidad y cumplir con los requisitos de protección de datos.
- Asegurar mecanismos efectivos para el consentimiento informado de los individuos cuyos datos personales son analizados.
- Garantizar que el cuidado humano y la decisión final sobre diagnósticos y tratamientos recaigan en seres humanos, aunque los sistemas de IA puedan asistirlos en su trabajo.
- Revisar los sistemas de IA mediante comités de investigación ética antes de su uso clínico.

#### **d) Investigación sobre los efectos de la IA en la salud mental**

Los Estados Miembros deben investigar los efectos y la regulación de los posibles daños de los sistemas de IA en la salud mental, tales como el aumento de la depresión, la ansiedad, el aislamiento social, el desarrollo de adicciones, la radicalización y la información errónea.

#### **e) Interacción entre humanos y robots**

Es primordial desarrollar directrices sobre las interacciones entre seres humanos y robots, especialmente en contextos de atención a la salud, educación y cuidado de personas mayores y con discapacidad. Se debe prestar atención particular a la salud mental y física de los seres humanos, asegurando que las interacciones con robots se alineen con los valores y principios éticos aplicables a cualquier sistema de IA. Además, se deben evaluar las implicaciones éticas de las tecnologías de neurotecnología e interfaces cerebro-ordenador para preservar la dignidad y autonomía humanas.

## **5.5 PRINCIPIOS DE LA BIOÉTICA EN LA INTELIGENCIA ARTIFICIAL**

En este apartado, exploraremos cómo cada uno de los principios de bioética se aplica al uso de la IA en salud, destacando su importancia y los desafíos que enfrentamos para implementarlos correctamente.

### **5.5.1 NO MALEFICENCIA**

El principio de no maleficencia es uno de los pilares éticos de la salud. Este principio subraya la responsabilidad de asegurar que los sistemas de IA no ocasionen ningún daño a las personas, protegiendo así su bienestar. En este contexto, resulta imprescindible identificar y minimizar los riesgos que puedan surgir durante la implementación de estas tecnologías en la atención médica.

Para garantizar la no maleficencia en el uso de la IA en salud, es necesario abordar tanto los aspectos éticos como los prácticos. En primer lugar, es fundamental que los algoritmos utilizados sean precisos y confiables, ya que cualquier error en su diseño o funcionamiento podría derivar en diagnósticos equivocados o decisiones terapéuticas inadecuadas. Por esta razón, los modelos de IA deben someterse a procesos de validación exhaustivos que evalúen su rendimiento en diferentes contextos clínicos antes de su uso generalizado.

Otro punto clave es la transparencia. La complejidad de los algoritmos de IA, especialmente aquellos denominados de "caja negra", dificulta su comprensión para profesionales de la salud y pacientes. Por ello, es indispensable ofrecer explicaciones claras y accesibles sobre su funcionamiento, alcance y limitaciones. Este

enfoque no solo favorece la toma de decisiones informadas, sino que también contribuye a generar confianza en el uso de estas herramientas.

Además, se requiere la creación de mecanismos efectivos de supervisión y control para monitorear de forma continua el impacto de la IA en la seguridad de los pacientes. Esto puede incluir protocolos de seguridad estrictos y la formación de comités éticos dedicados a evaluar periódicamente el empleo de la IA en la medicina. Asimismo, es necesario establecer canales para la retroalimentación y el reporte de incidentes, permitiendo identificar y corregir posibles errores o sesgos en los sistemas.

En última instancia, la implementación responsable de la inteligencia artificial en la salud no solo depende de su precisión técnica, sino también de un marco ético sólido que priorice el bienestar de las personas.

---

### 5.5.2 BENEFICENCIA

El principio de beneficencia, clave en las profesiones de la salud, destaca la responsabilidad de garantizar que los beneficios del uso de la inteligencia artificial (IA) superen los riesgos. Esto requiere una aplicación ética y responsable, centrada en mejorar la calidad de vida de las personas y optimizar la atención que reciben.

Uno de los principales aportes de la IA es su capacidad para analizar grandes volúmenes de información, como datos clínicos o imágenes, identificando patrones que podrían pasar desapercibidos. Esto facilita diagnósticos más rápidos y precisos, mejorando los resultados para pacientes y reduciendo posibles complicaciones.

Además, la IA permite personalizar las intervenciones según las necesidades individuales, considerando factores como antecedentes, características específicas y respuestas previas. Este enfoque no solo mejora la eficacia de las decisiones, sino que también reduce riesgos asociados y aumenta el éxito de las estrategias aplicadas.

Por último, los algoritmos de IA pueden analizar los datos de cada paciente de forma individualizada, lo que permite ofrecer recomendaciones y tratamientos adaptados a sus necesidades específicas. Esto puede aumentar la satisfacción del paciente y mejorar la adherencia al tratamiento, lo que a su vez puede conducir a mejores resultados clínicos a largo plazo.

Cuando se utiliza bajo los principios de beneficencia y ética, la IA representa una herramienta poderosa para todos los profesionales de la salud, ayudando a garantizar un cuidado más preciso, efectivo y adaptado a las personas.

---

### 5.5.3 AUTONOMÍA

El principio de autonomía en la bioética se fundamenta en el respeto por la capacidad de las personas para tomar decisiones sobre su propia salud. En el contexto de la inteligencia artificial, este principio se ve influido tanto positiva como negativamente, según cómo se implementen estas tecnologías.

Por un lado, la IA puede empoderar a las personas al ofrecerles información más precisa sobre diagnósticos, tratamientos y pronósticos. Este acceso a datos detallados permite a los pacientes participar activamente en las decisiones relacionadas con su cuidado, considerando sus valores, metas y preferencias. Además, la IA puede personalizar los planes de tratamiento teniendo en cuenta factores como antecedentes médicos, estilo de vida y necesidades específicas, lo que refuerza la sensación de control y participación en el proceso de atención.

No obstante, la incorporación de la IA también plantea desafíos éticos en relación con la autonomía. Una dependencia excesiva de estas herramientas podría reducir la capacidad de los profesionales de la salud para ejercer un juicio crítico, delegando decisiones en los sistemas de IA sin considerar plenamente las particularidades de cada persona. Esto puede limitar tanto la autonomía de los pacientes como la de quienes están encargados de su cuidado.

Por ello, es esencial que la implementación de la IA promueva la transparencia. Las personas deben entender cómo estas herramientas influyen en su atención y tener la oportunidad de participar activamente en las decisiones. También es importante recalcar que los profesionales de la salud deben recibir formación adecuada para interpretar y cuestionar las recomendaciones de la IA, integrando su criterio y las preferencias de quienes atienden.

En última instancia, el uso responsable de la IA en salud debe equilibrar el poder de la tecnología con el respeto por la autonomía individual, asegurando que cada decisión sea informada, personal y alineada con los valores de las personas involucradas.

---

#### 5.5.4 JUSTICIA

El principio de justicia en el uso de la inteligencia artificial (IA) en la salud destaca la necesidad de garantizar que los avances tecnológicos beneficien a todas las personas por igual. Este enfoque ético exige que la implementación de la IA en el ámbito sanitario sea inclusiva y equitativa, evitando que las brechas existentes en el acceso a servicios de calidad se amplíen.

Un desafío clave es combatir las desigualdades en la atención médica. Las comunidades con menos recursos suelen enfrentarse a limitaciones para acceder a servicios adecuados, lo que impacta negativamente en sus resultados de salud. Por ello, es fundamental que la IA se implemente de manera que no excluya a estas poblaciones y, por el contrario, contribuya a reducir las disparidades. Esto implica diseñar políticas públicas que aseguren el acceso equitativo a las tecnologías, incluyendo iniciativas educativas que promuevan su uso y comprensión en diversos contextos sociales.

La equidad también exige que estas tecnologías estén disponibles para quienes tienen recursos limitados y no solo para aquellos que pueden pagar por ellas. Esto requiere la creación de modelos accesibles que fomenten la inclusión, respaldados por estrategias de financiación y regulación que prioricen el bienestar colectivo.

Uno de los aspectos principales relacionado con la justicia es la necesidad de que los algoritmos de IA sean transparentes y libres de sesgos. Los sistemas utilizados deben ser diseñados y supervisados para garantizar que sus decisiones no discriminen ni perjudiquen injustamente a ciertos grupos. Por otro lado, los usuarios, tanto pacientes como profesionales de la salud, deben recibir información clara sobre cómo funcionan estos algoritmos y contar con mecanismos para cuestionar o validar las decisiones que generan.

La adopción de la IA en salud debe estar guiada por un compromiso ético firme con la justicia. Lograrlo demanda una colaboración activa entre desarrolladores, profesionales de la salud, gestores y pacientes. Solo a través de un enfoque inclusivo y responsable se puede garantizar que la inteligencia artificial sea una herramienta transformadora que beneficie al conjunto de la sociedad, promoviendo un acceso equitativo y mejorando el bienestar de todas las personas.

**En nuestra mano queda cómo podemos construir un futuro donde la tecnología y la ética trabajen de la mano para el bienestar de la humanidad.**

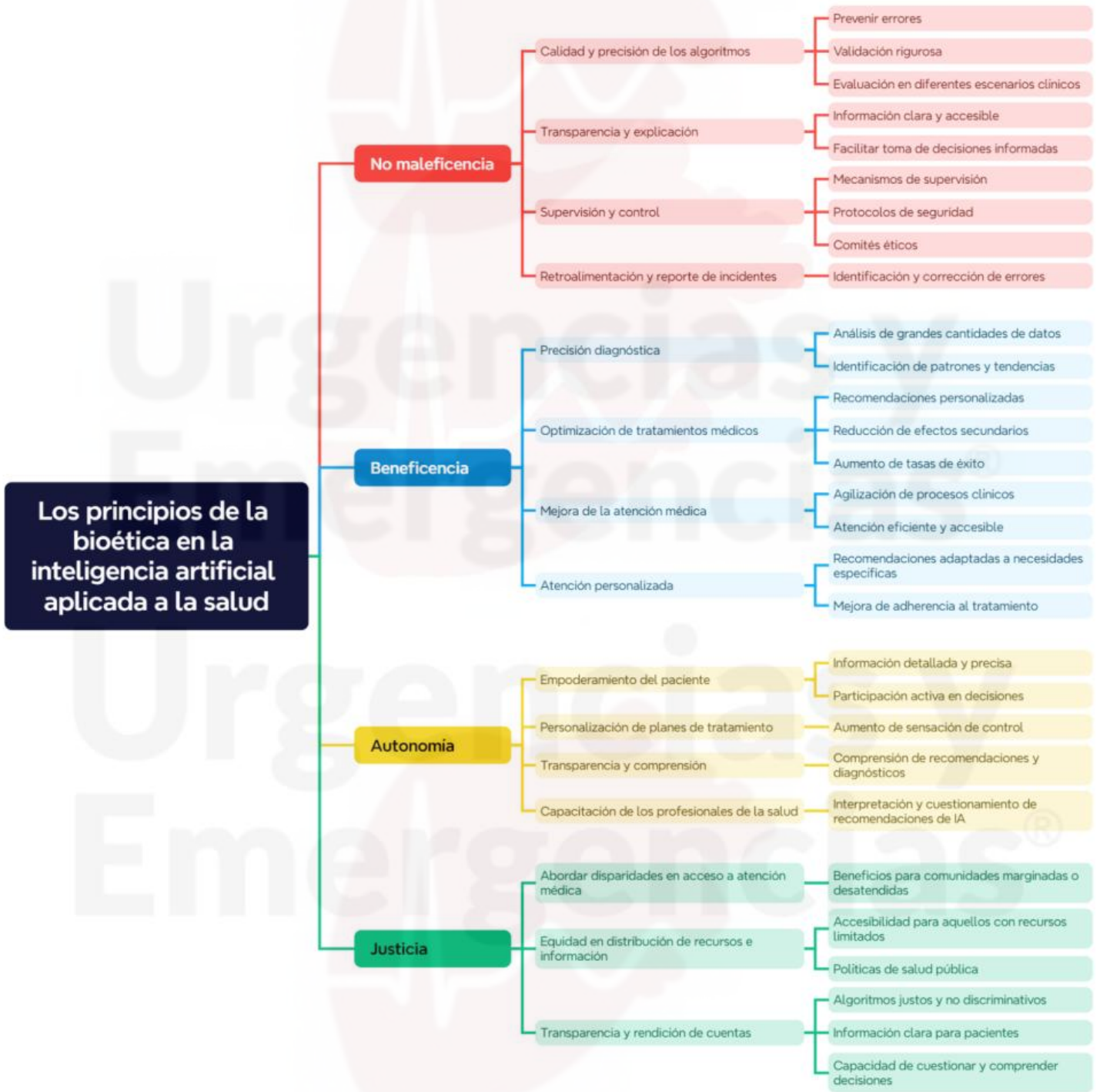


Imagen 85. Los principios de la bioética en la IA aplicada a salud. Elaboración propia. Autoría: Elena Plaza Moreno

## 6. ANEXO I. IMPLEMENTACIÓN DE UNA SOLUCIÓN DE IA

En el siguiente artículo de Sánchez Rosado y Díez Parra (de IBM) analizan el impacto de la IA en la transformación de la sanidad. En concreto, proponen una hoja de ruta de construcción e implementación de una solución de IA en salud.

Rosado, J. C. S., & Parra, M. D. (2022). Impacto de la inteligencia artificial en la transformación de la sanidad: beneficios y retos. *Economía industrial*, 423, 129–144. Recuperado de <https://dialnet.unirioja.es/servlet/articulo?codigo=8750554>

Proponen la valoración de 5 áreas éticas básicas que proporcionan un marco de trabajo:

- **Explicabilidad.** La Inteligencia Artificial debe estar diseñada para que los humanos perciban, detecten y comprendan fácilmente su proceso de decisión.
- **Equidad.** La Inteligencia Artificial debe diseñarse para minimizar los prejuicios y promover una representación inclusiva.
- **Responsabilidad.** Los diseñadores y desarrolladores de Inteligencia Artificial son responsables de considerar el diseño, el desarrollo, los procesos de decisión y los resultados de la Inteligencia Artificial.
- **Derechos de datos del usuario.** La Inteligencia Artificial debe estar diseñada para proteger los datos del usuario y preservar el poder del usuario sobre el acceso y los usos.
- **Alineamiento de valores.** La Inteligencia Artificial debería estar diseñada para alinearse con las normas y valores del grupo de usuarios objetivo.

También proponen un ejemplo real de proyecto de IA en un hospital y un marco de trabajo ético para su implementación que se copia textualmente a continuación (Tabla 8).

### MARCO DE TRABAJO ÉTICO PARA LA IMPLEMENTACIÓN DE UNA SOLUCIÓN DE IA EN SALUD

Ejemplo: «Un grupo hospitalario quiere incorporar inteligencia artificial a través de un asistente/conserje virtual en la habitación para mejorar y personalizar la estancia de los pacientes. Este agente conversacional incluirá capacidades como: asistencia de estilo agente, introducción a su habitación y servicios en su idioma preferido, control de las instalaciones de la habitación mediante lenguaje natural, envío de una solicitud directamente al equipo de atención a través del asistente virtual en la habitación.

#### EXPLICABILIDAD

Definición asistente virtual	<ul style="list-style-type: none"> <li>▪ ¿Cómo incorporamos la explicabilidad a nuestro proceso?</li> <li>▪ ¿Cómo el output del sistema afecta a las decisiones de la entidad?</li> </ul>
Entrenamiento asistente virtual	<ul style="list-style-type: none"> <li>▪ ¿Cuál es el modelo más simple e interpretable para este caso?</li> <li>▪ ¿Los datos de entrenamiento y testing son analizables y trazables?</li> </ul>
Despliegue asistente virtual	<ul style="list-style-type: none"> <li>▪ ¿Se comunica al usuario final que está interactuando con un sistema de IA?</li> <li>▪ ¿Se informa al usuario final de posibles riesgos de sesgos y limitaciones?</li> </ul>
Mantenimiento y reentrenamiento	<ul style="list-style-type: none"> <li>▪ ¿Cómo se utilizarán nuevos datos en el reentrenamiento?</li> <li>▪ ¿Cómo se realiza la gestión de errores del sistema?</li> </ul>

Acciones	<ul style="list-style-type: none"> <li>Investigación de usuario final para recolectar necesidades y preocupaciones relacionadas con el uso de este tipo de sistemas: escucha continua y uso de conversaciones.</li> <li>Implementación de mecanismos para dar a conocer al usuario cómo el asistente da una recomendación para su estancia.</li> <li>Identificación de agente virtual como tal y posibilidad de consultar conversaciones mantenidas con el asistente virtual, así como posibles datos inferidos</li> </ul>
<b>EQUIDAD</b>	
Definición asistente virtual	<ul style="list-style-type: none"> <li>¿Cuál es la diversidad en la composición del equipo?</li> <li>¿Cómo se van a identificar posibles sesgos conscientes y no conscientes?</li> </ul>
Entrenamiento asistente virtual	<ul style="list-style-type: none"> <li>¿Cómo analizo posibles sesgos en los datos de entrenamiento y test? ¿Cuál es la representatividad de ciertos grupos de usuarios en mis datos?</li> </ul>
Despliegue asistente virtual	<ul style="list-style-type: none"> <li>¿Quién se ve afectado por el uso de este sistema de IA? ¿Existen mecanismos y métricas asociadas a equidad que deben ser monitorizadas?</li> </ul>
Mantenimiento y reentrenamiento	<ul style="list-style-type: none"> <li>El status-quo cambia a lo largo del tiempo. ¿Cómo se asegura que los métodos se reflejan? ¿Cómo recogemos feedback de usuario sobre posibles sesgos?</li> </ul>
Acciones	<ul style="list-style-type: none"> <li>Alineamiento inicial con todos los involucrados: entidad, equipo de desarrollo del sistema.</li> <li>Implementación de mecanismos y uso de herramientas para detección de sesgos.</li> <li>Implementación de mecanismos y uso de métricas para monitorización de equidad en algoritmos.</li> </ul>
<b>RESPONSABILIDAD</b>	
Definición asistente virtual	<ul style="list-style-type: none"> <li>¿Cómo mantendremos documentadas las decisiones de diseño e implementación?</li> <li>¿Está la IA embebida en un proceso de decisión supervisado por un humano?</li> </ul>
Entrenamiento asistente virtual	<ul style="list-style-type: none"> <li>¿Cómo se realiza la trazabilidad de entrenamientos?</li> <li>¿Cómo es el acceso a la trazabilidad y decisiones de diseño?</li> </ul>
Despliegue asistente virtual	<ul style="list-style-type: none"> <li>¿Cómo se realiza la trazabilidad de las conversaciones?</li> <li>¿Aplica alguna regulación a tener en cuenta?</li> </ul>
Mantenimiento y reentrenamiento	<ul style="list-style-type: none"> <li>¿Existen mecanismos para usuario final/ terceras partes en caso de fallo o impacto negativo?</li> <li>¿Cómo se realiza el reentrenamiento y mejora continua?</li> </ul>
Acciones	<ul style="list-style-type: none"> <li>Creación de comité o persona responsable de cuestiones relacionadas con éticas. Inclusión de expertos en talleres iniciales.</li> <li>Implementación de bucle de feedback continuo para entender preferencias y preocupaciones de cliente.</li> <li>Implementación de mecanismos para «apagar la IA» por parte del cliente en cualquier punto de su estancia.</li> </ul>
<b>DERECHOS DE DATOS DEL USUARIO</b>	
Definición asistente virtual	<ul style="list-style-type: none"> <li>¿Qué tipos de datos personales y sensibles se van a utilizar?</li> <li>¿Cómo se protegerán dichos datos?</li> </ul>
Entrenamiento asistente virtual	<ul style="list-style-type: none"> <li>¿Cómo creamos el modelo con la cantidad mínima de datos?</li> </ul>

Despliegue asistente virtual	<ul style="list-style-type: none"> <li>¿Cómo se realiza la trazabilidad de las conversaciones?</li> <li>¿Aplica alguna regulación a tener en cuenta?</li> </ul>
Mantenimiento y reentrenamiento	
Acciones	<ul style="list-style-type: none"> <li>Inclusión de expertos de tratamiento del dato en talleres iniciales.</li> <li>Recoger requisitos acordes a la regulación aplicable y diseño de solución con dichos factores.</li> <li>Implementación de mecanismos para usuario final acordes a la regulación.</li> </ul>
ALINEAMIENTO DE VALORES	
Definición asistente virtual	<ul style="list-style-type: none"> <li>¿Qué valores de grupo muestra nuestro sistema de IA?</li> <li>Los valores son subjetivos y cambian a nivel global, ¿los recogemos?</li> </ul>
Entrenamiento asistente virtual	<ul style="list-style-type: none"> <li>¿Qué valores tendremos como equipo?</li> </ul>
Despliegue asistente virtual	
Mantenimiento y reentrenamiento	<ul style="list-style-type: none"> <li>¿Cómo cambiamos o ajustamos los valores a lo largo del tiempo?</li> </ul>
Acciones	<ul style="list-style-type: none"> <li>Alineamiento equipo de trabajo y desarrollo.</li> <li>Palabra para despertar a asistente virtual para mitigar la preocupación por escucha permanente, y comunicar a usuario final que esos datos no son utilizados o son eliminados con frecuencia.</li> <li>Añadir disciplinas como lingüistas para asegurar que se recogen casuísticas de diferentes idiomas.</li> </ul>

**Tabla 8. Marco de trabajo ético para la implementación de una solución de IA en salud. Copia textual de Rosado, J. C. S., & Parra, M. D. (2022)**

## 7. BIBLIOGRAFÍA

- American Nurses Association. (2022). *The ethical use of artificial intelligence in nursing practice. Position Statement*. Nursingworld.org  
[https://www.nursingworld.org/~48f653/globalassets/practiceandpolicy/nursing-excellence/ana-position-statements/the-ethical-use-of-artificial-intelligence-in-nursing-practice\\_bod-approved-12\\_20\\_22.pdf](https://www.nursingworld.org/~48f653/globalassets/practiceandpolicy/nursing-excellence/ana-position-statements/the-ethical-use-of-artificial-intelligence-in-nursing-practice_bod-approved-12_20_22.pdf)
- Anishchenko, I., Pellock, S. J., Chidyausiku, T. M., Ramelot, T. A., Ovchinnikov, S., Hao, J., Bafna, K., Norn, C., Kang, A., Bera, A. K., DiMaio, F., Carter, L., Chow, C. M., Montelione, G. T., & Baker, D. (2021). *De novo protein design by deep network hallucination*. *Nature*, 600(7889), 547–552.  
<https://doi.org/10.1038/s41586-021-04184-w>
- Barile, J., Margolis, A., Cason, G., Kim, R., Kalash, S., Tchaconas, A., & Milanaik, R. (2024). *Diagnostic accuracy of a large language model in pediatric case studies*. *JAMA Pediatrics*, 178(3), 313.  
<https://doi.org/10.1001/jamapediatrics.2023.5750>

- Código de Deontología Médica.* (s/f). Cgcom.es. Recuperado el 9 de junio de 2024, de [https://www.cgcom.es/sites/main/files/minisite/static/828cd1f8-2109-4fe3-acba-1a778abd89b7/codigo\\_deontologia/](https://www.cgcom.es/sites/main/files/minisite/static/828cd1f8-2109-4fe3-acba-1a778abd89b7/codigo_deontologia/)
- de Zárate Alcarazo, L. O. (2023, marzo). *Sesgos de género en la inteligencia artificial.* *Revistade Occidente*, 502, 5–20.
- Esquerda, M., & Pifarré-Esquerda, F. (2024). *Inteligencia artificial en medicina: aspectos éticos, deontológicos y el impacto en la relación médico-paciente.* *Medicina clinica.* <https://doi.org/10.1016/j.medcli.2024.03.003>
- Hamburg: *HmbBfDI publishes discussion paper on GDPR and LLMs.* (2024, julio 15). DataGuidance. <https://www.dataguidance.com/news/hamburg-hmbbfdi-publishes-discussion-paper-gdpr-and>
- Maslej, N., Fattorini, L., Perrault, R., Parli, V., Reuel, A., Brynjolfsson, E., Etchemendy, J., Ligett, K., Lyons, T., Manyika, J., Niebles, J. C., Shoham, Y., Wald, R., & Clark, J. (2024). *The AI Index 2024 Annual Report.* [https://aiindex.stanford.edu/wp-content/uploads/2024/05/HAI\\_AI-Index-Report-2024.pdf](https://aiindex.stanford.edu/wp-content/uploads/2024/05/HAI_AI-Index-Report-2024.pdf)
- Matute, H., & Martínez, N. (2020, agosto 10). *Discriminación racial en la inteligencia artificial.* *The Conversation.* <http://theconversation.com/discriminacion-racial-en-la-inteligencia-artificial-142334>
- Nori, H., King, N., McKinney, S. M., Carignan, D., & Horvitz, E. (2023). *Capabilities of GPT-4 on medical challenge problems.* En arXiv [cs.CL]. <http://arxiv.org/abs/2303.13375>
- Nori, H., Lee, Y. T., Zhang, S., Carignan, D., Edgar, R., Fusi, N., King, N., Larson, J., Li, Y., Liu, W., Luo, R., McKinney, S. M., Ness, R. O., Poon, H., Qin, T., Usuyama, N., White, C., & Horvitz, E. (2023). *Can generalist foundation models outcompete special-purpose tuning? Case study in medicine.* <https://doi.org/10.48550/ARXIV.2311.16452>
- Oniani, D., Hilsman, J., Peng, Y., Poropatich, R. K., Pamplin, J. C., Legault, G. L., & Wang, Y. (2023). *Adopting and expanding ethical principles for generative artificial intelligence from military to healthcare.* *Npj Digital Medicine*, 6(1). <https://doi.org/10.1038/s41746-023-00965-x>
- OpenAI. (2024). *ChatGPT 4* (versión del 15 de mayo) [Modelo de lenguaje de gran tamaño]. <https://chat.openai.com/chat>
- Organización Mundial de la Salud. (2021a, junio 28). *Ethics and governance of artificial intelligence for health: WHO guidance Executive summary.* Who.int; World Health Organization. <https://www.who.int/publications/i/item/9789240037403>
- Organización Mundial de la Salud. (2021b, noviembre 17). *Generating evidence for artificial intelligence based medical devices: A framework for training validation and evaluation.* Who.int; World Health Organization. <https://www.who.int/publications/i/item/9789240038462>
- Organización Mundial de la Salud. (2023, octubre 19). *Regulatory considerations on artificial intelligence for health.* Who.int; World Health Organization. <https://www.who.int/publications/i/item/9789240078871>

- Organización Mundial de la Salud. (2024, enero 18). *Ethics and governance of artificial intelligence for health: Guidance on large multi-modal models*. Who.int; World Health Organization.  
<https://www.who.int/publications/i/item/9789240084759>
- Parlamento Europeo. (2024, abril). *CORRECCIÓN DE ERRORES. de la Posición del Parlamento Europeo aprobada en primera lectura el 13 de marzo de 2024 con vistas a la adopción del Reglamento (UE) 2024/... del Parlamento Europeo y del Consejo por el que se establecen normas armonizadas en materia de inteligencia artificial y por el que se modifican los Reglamentos (CE) n.º 300/2008, (UE) n.º 167/2013, (UE) n.º 168/2013, (UE) 2018/858, (UE) 2018/1139 y (UE) 2019/2144 y las Directivas 2014/90/UE, (UE) 2016/797 y (UE) 2020/1828 (Reglamento de Inteligencia Artificial)*. Europa.eu.  
[https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138-FNL-COR01\\_ES.pdf](https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138-FNL-COR01_ES.pdf)
- Rosado, J. C. S., & Parra, M. D. (2022). Impacto de la inteligencia artificial en la transformación de la sanidad: beneficios y retos. *Economía industrial*, 423, 129–144. Recuperado de  
<https://dialnet.unirioja.es/servlet/articulo?codigo=8750554>
- Rosenthal, D. (2024, marzo 22). *Part 17: What is inside an AI model and how it works*. Vischer.com; VISCHER.  
<https://www.vischer.com/en/knowledge/blog/part-17-what-is-inside-an-ai-model-and-how-it-works/>
- Sanchis, A. (2024, marzo 1). *Incendio en Google: tiene un problema con su IA (y su futuro), y le está costando millones*. El Confidencial. [https://www.elconfidencial.com/tecnologia/2024-03-01/google-inteligencia-artificial-imagenes-gemini\\_3840270/](https://www.elconfidencial.com/tecnologia/2024-03-01/google-inteligencia-artificial-imagenes-gemini_3840270/)
- Sheng, E., Chang, K.-W., Natarajan, P., & Peng, N. (s/f). *The woman worked as a babysitter: On biases in language generation*. Aclanthology.org. Recuperado el 19 de julio de 2024, de  
<https://aclanthology.org/D19-1339.pdf>
- Tiulkanov, A. (2023a). *A simple algorithm to decide whether to use ChatGPT*. LinkedIn.com.  
[https://www.linkedin.com/posts/tyulkanov\\_a-simple-algorithm-to-decide-whether-to-use-activity-7021766139605078016-x8Q9?utm\\_source=share&utm\\_medium=member\\_desktop](https://www.linkedin.com/posts/tyulkanov_a-simple-algorithm-to-decide-whether-to-use-activity-7021766139605078016-x8Q9?utm_source=share&utm_medium=member_desktop)
- Tiulkanov, A. (2023b, enero 8). *Is it high time to take ChatGPT offline?* LinkedIn.com.  
<https://www.linkedin.com/pulse/high-time-take-chatgpt-offline-aleksandr-tiulkanov/>
- Qayyum, A., Qadir, J., Bilal, M., & Al-Fuqaha, A. (2021). Secure and robust machine learning for healthcare: A survey. *IEEE reviews in biomedical engineering*, 14, 156–180.  
<https://doi.org/10.1109/rbme.2020.3013489>
- UNESCO. (2022). *Recomendación sobre la ética de la inteligencia artificial*.  
[https://unesdoc.unesco.org/ark:/48223/pf0000381137\\_spa](https://unesdoc.unesco.org/ark:/48223/pf0000381137_spa)
- What is AI bias?* (2023, diciembre 14). Ibm.com. <https://www.ibm.com/topics/ai-bias>
- Zack, T., Lehman, E., Suzgun, M., Rodríguez, J. A., Celi, L. A., Gichoya, J., Jurafsky, D., Szolovits, P., Bates, D. W., Abdulnour, R.-E. E., Butte, A. J., & Alsentzer, E. (2024). *Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study*. *The Lancet. Digital Health*, 6(1), e12–e22. [https://doi.org/10.1016/s2589-7500\(23\)00225-x](https://doi.org/10.1016/s2589-7500(23)00225-x)



# Urgencias y Emergencias®



# Urgencias y Emergencias®

# Urgencias y Emergencias®

Copyright © 2025 Elena Plaza Moreno. Todos los derechos reservados. Obra registrada en el Registro de la Propiedad Intelectual

**Elena Plaza Moreno**

**[www.urgenciasyemergen.com](http://www.urgenciasyemergen.com)**

**Contacto: [admin@urgenciasyemergen.com](mailto:admin@urgenciasyemergen.com)**

