Álvaro Caruana Santiago

Análisis de tecnologías IA aplicadas a la investigación biotecnológica

Tutor académico: Héctor Candela Antón
Tutor profesional: Roberto Gómez-Espinosa Martín
Departamento de Biología Aplicada, Área de Genética

Facultad de Ciencias Experimentales
Grado en Biotecnología
2023/2024

Universidad Miguel Hernández de Elche

ÍNDICE DE MATERIAS

ÍNDICE DE FIGURAS

1. Abreviaturas

AF2: AlphaFold2

CNN: Red Neuronal Convolucional

RNN: Red Neuronal Recurrente

DL: Aprendizaje Profundo

IA: Inteligencia Artificial

LLM: Modelo de Lenguaje a Gran Escala

ML: Aprendizaje Automático

MLLM: Modelo Multimodales de Lenguaje a Gran Escala

MLP: Perceptrón Multicapa

NLP: Procesamiento del Lenguaje Natural

NN: Redes Neuronal

SVM: Máquinas de Soporte Vectorial

2. Resumen

En las últimas décadas han aumentado las tasas de publicación científica a lo largo de todas las disciplinas, incluidas las biomédicas. Esto ha generado una creciente dificultad de mantenerse al tanto del estado del arte del campo de estudio de los investigadores. De manera similar, la presión por publicar puede volverse un problema para la comunidad científica, al forzar al personal investigador a generar artículos científicos cuya calidad puede estar comprometida. Para contrarrestar estos obstáculos se han de desarrollar prácticas que ayuden al investigador a realizar su labor. En este marco, la inteligencia artificial puede ser una respuesta eficaz. Con cantidad de herramientas donde elegir, la inteligencia artificial es una campo que ayuda a potenciar la tareas de investigación. Es importante hacer un repaso por los fundamentos de la IA, para saber cuáles son los mejores enfoques para resolver los problemas que surjan durante la proceso científico, a la vez que se ha de poner en valor los grande descubrimientos que se ha hecho con su ayuda. Por último, y para tener una visión global, es de interés debatir sobre los dilemas éticos que el uso de la inteligencia artificial puede conllevar.

Palabras clave: inteligencia artificial, investigación científica, modelos de lenguaje a gran escala, ética sobre la IA, redes neuronales.

In recent decades, scientific publication rates have increased across all disciplines, including biomedical disciplines. This has led to increasing difficulty in keeping researchers abreast of the state of the art in their field of study. Similarly, the pressure to publish can become a problem for the scientific community, forcing research personnel to generate scientific articles whose quality may be compromised. To counteract these obstacles, practices have been developed that will help the researcher conduct their work. In this framework, artificial intelligence can be an effective response. With several tools to choose from, artificial intelligence is a field that helps enhance research tasks. It is important to review the foundations of AI, to know which the best approaches are to solve the problems that arise during the scientific process, while at the same time valuing the great discoveries that have been made with its help. Finally, and to have a global vision, it is of interest to debate the ethical dilemmas that artificial intelligence can entail.

Keywords: artificial intelligence, scientific research, large language models, AI ethics, neural networks.

3. Prólogo

Durante mis prácticas en la consultora tecnológica HI Iberia (HI iberia | Desarrollo software y soluciones TIC, s. f.), me formaron en el campo de inteligencia artificial (IA). Esta empresa dispone de una gran variedad de proyectos que engloban diferentes disciplinas entre las que se encuentra la física, química, matemáticas y biotecnología entre otras. El objetivo original del TFG era plasmar los resultados de mi participación en el proyecto de NanomatIA (Proyectos de I+D+i en líneas estratégicas - Transmisiones 2023 | Agencia Estatal de Investigación, s. f.). Este proyecto, fue creado a partir de la iniciativa “TransMisiones 2023” entre el Centro para el Desarrollo Tecnológico y de Innovación y la Agencia Estatal de Investigación. NanomatIA tiene como objetivo la búsqueda de nuevos nanomateriales para baterías de ion litio impulsados a través de la inteligencia artificial y la biotecnología industrial. Este trabajo pretende explotar las capacidades de los modelos de lenguaje a gran escala (*large language models,* LLM) o modelos multimodales de lenguaje a gran escala (*multimodal large language models,* MLLM), especializados en el ámbito científico, para hacer inferencias sobre la identificación de enzimas lignocelulósicas en la literatura científica. La información extraída ayudaría a la producción industrial de nanomateriales de interés para el proyecto. Sin embargo, al ser un proyecto tan complejo no se llegaron a obtener resultados antes de la finalización de las prácticas. Debido a esto, decidimos cambiar la dirección de este trabajo al ámbito bibliográfico.

El objetivo principal es poner en valor el papel de la inteligencia artificial en el ámbito científico. La manera que decidimos enfocarlo es resaltar mediante casos prácticos cómo, las herramientas de IA actuales, son ubicuas en la realización de la tarea investigadora, pero es importante saber identificarlas y utilizarlas.

4. Introducción

El aumento del número de publicaciones científicas en los últimos años ha sido exponencial, con un crecimiento general del 4,10% anual y el tiempo que tarda en duplicarse es de 17,3 años (Figura 1) (Bornmann *et al.*, 2021). Esto ha planteado un desafío significativo para los investigadores pues se enfrentan a una sobrecarga de información, lo que obstaculiza su capacidad para identificar y acceder eficientemente a la literatura relevante. En concreto, en el ámbito de las ciencias biomédicas, este crecimiento puede atribuirse al gran volumen de artículos revisados por pares y a la disolución de las fronteras entre las disciplinas, así como a la llegada de nuevas tecnologías de análisis de alto rendimiento. Por ello se ha hecho cada vez más difícil para los científicos mantenerse al día con el rápido ritmo del desarrollo científico incluso dentro de campos especializados de la ciencia (Cohan & Goharian, 2017), lo que potencialmente genera lagunas en el conocimiento y a un desperdicio de tiempo y recursos (Özgür *et al.*, 2010).

Paralelamente, la presión por publicar, añadida al rápido ritmo de los avances científicos, puede potencialmente comprometer el rigor y la reproducibilidad científica (Forsythe *et al.*, 2019).

Esta problemática combinada hace necesario que: (1) se exploren enfoques innovadores para la recuperación de información, y (2) se desarrollen y adapten nuevas prácticas y herramientas que agilicen el desarrollo del trabajo investigador. De esta forma, los investigadores se pueden involucrar de forma eficiente en la comunidad científica y contribuir a la expansión del conocimiento de esta.

De manera simultánea al desarrollo científico, la inteligencia artificial (IA) también ha progresado significativamente durante las últimas dos décadas, con avances observados en varios sectores como el aprendizaje automático (*machine learning,* ML), el procesamiento del lenguaje natural (*natural language processing,* NLP), la visión por computadora y la robótica (Jerbi, 2023). Estos avances han sido impulsados por innovaciones en técnicas de aprendizaje profundo (*deep learning*, DL), lo que permite la utilización efectiva de datos de alta dimensión y alto volumen en los sistemas de IA. La integración de tecnologías de IA en diversas industrias ha sido transformadora, y la IA se ha reconocido como una tecnología altamente demandada en campos como la salud, la educación, la ciencia y otros (Saini *et al.*, 2021). Notablemente, herramientas como ChatGPT (OpenAI, s. f.) ejemplifican cómo la IA puede procesar y sintetizar grandes cantidades de información de manera eficiente, aliviando la carga cognitiva de los usuarios.

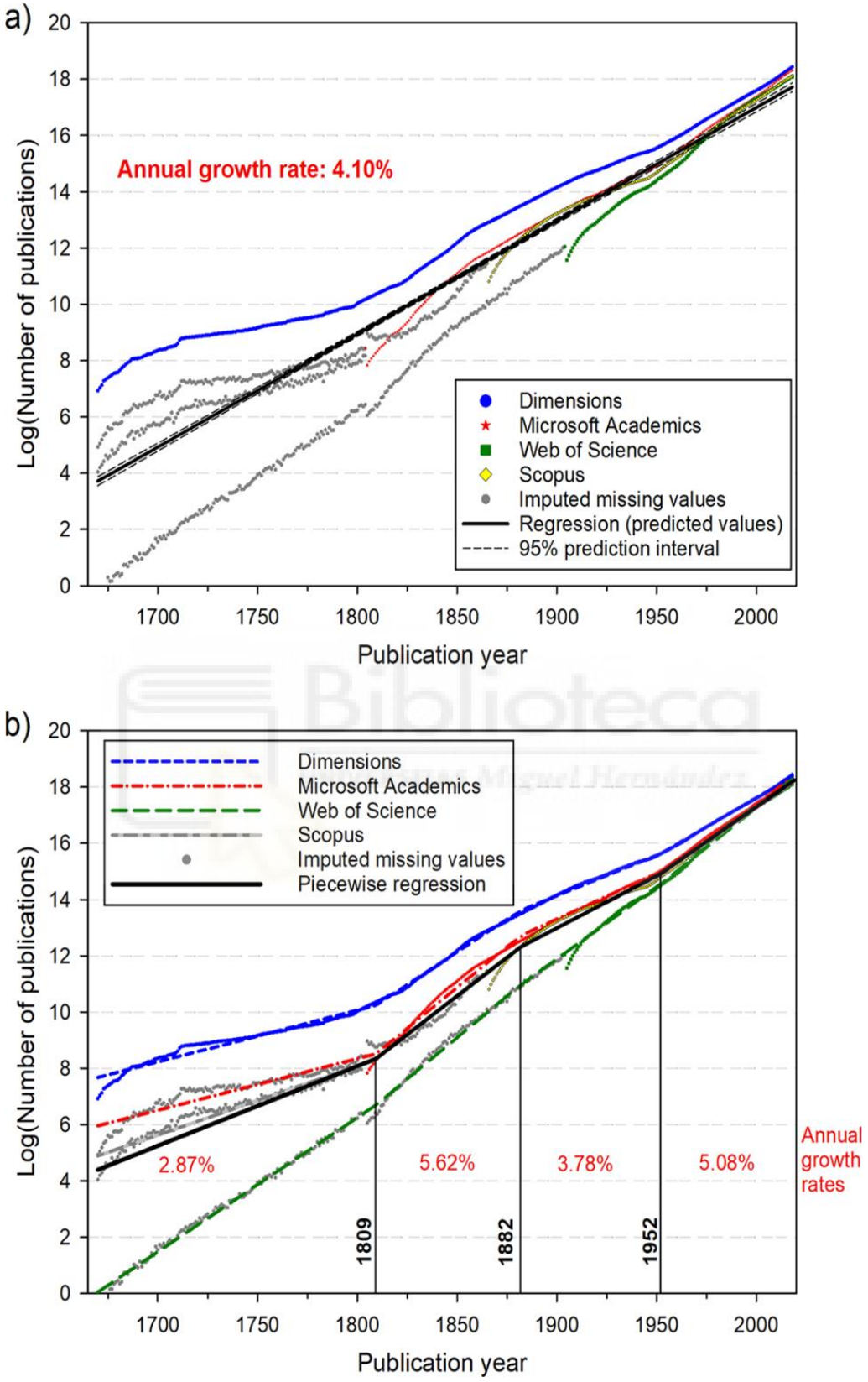

Figura 1.- Gráficas del crecimiento científico basado en el número de publicaciones anuales de cuatro bases de datos bibliográficas a lo largo del tiempo. La gráfica a representa la tasa de crecimiento general, y la gráfica b representa la tasa de crecimiento segmentada. Tomado de Bornmann *et al.* (2021).

Debido a esto, en el ámbito de la investigación científica, el uso de la IA es particularmente relevante, ya que puede aliviar el impacto de los problemas comentados además de proporcionar herramientas especializadas que directamente participen del proceso de investigación.

En la actualidad, la versatilidad de la aplicación de los sistemas de IA permite que se puedan encontrar en una gran cantidad de disciplinas biotecnológicas, como por ejemplo en el análisis de datos multiómicos (ej.: metabolómica o proteómica) (Reel *et al.*, 2021), el descubrimiento de biomarcadores (Xiao *et al.*, 2021), y en estudios de seguridad de medicamentos (Diaw *et al.*, 2023), entre muchas otras disciplinas.

5. Fundamentos de la IA

Aunque la idea de máquinas automáticas o “autómatas” traza sus orígenes hasta la antigua Grecia, el concepto moderno de la inteligencia artificial se atribuye en gran medida a Alan Turing cuando este publicó en 1950 un trabajo titulado "Computer Machinery and Intelligence" (Turing, 1950), que eventualmente se conocería como el Test de Turing. En él, Alan Turing plantea los conceptos básicos de la IA como el aprendizaje por medio de la experiencia o la idea de que una máquina puede, de manera teórica, realizar cualquier tarea intelectual como un humano. Sin embargo, el término no se llegaría a acuñar hasta 1955 por John McCarthy, Marvin Minsky, Nathaniel Rochester and Claude Shannon, redactaron una propuesta de un taller de verano en la Universidad de Darthmouth. Sería allí, junto a otros matemáticos y científicos, donde, debatieron y refinaron conceptos esenciales del campo como el procesamiento del lenguaje natural (NLP) o las redes neuronales (*neural networks*; NN) (McCarthy, 1955).

En la actualidad, la IA se puede definir de forma general como aquel software informático que imita la forma en que piensan los humanos para realizar tareas complejas, como analizar, razonar y aprender. La IA es un gran campo de estudio extenso. Por ello, para proporcionar una comprensión fundamental de la IA es crucial introducir y explicar conceptos y terminología clave. Esta sección cubre los bloques fundamentales de la IA, incluyendo el aprendizaje automático, el aprendizaje profundo, las redes neuronales y los algoritmos. Al comprender estas ideas centrales, se estará mejor preparado para entender discusiones más complejas sobre las aplicaciones de la IA en la biotecnología y otros campos.

4.1 Aprendizaje Automático (*Machine Learning*)

El aprendizaje automático (*machine learning,* ML) es una subdisciplina de la IA centrada en desarrollar algoritmos y modelos que permiten a las computadoras aprender y tomar decisiones basadas en datos. A diferencia de la programación tradicional, donde se proporcionan instrucciones explícitas para cada acción, los sistemas de aprendizaje automático mejoran su rendimiento en las tareas a través de la experiencia. Hay varios tipos de aprendizaje automático según los datos de los que se dispongan:

- **Aprendizaje Supervisado**: En el aprendizaje supervisado, el modelo se entrena con un conjunto de datos etiquetados, lo que significa que cada ejemplo de entrenamiento se empareja con una etiqueta de salida, es decir, con la respuesta que se desea que del sistema. El modelo aprende a mapear datos de entrada a los de salida correctos comparando sus predicciones con las etiquetas reales y ajustando sus parámetros en consecuencia. Ejemplos de tareas de aprendizaje supervisado incluyen la clasificación (ej., clasificación de tipos de cáncer según patrones de expresión (Mostavi et. Al, 2020)) y la

regresión (ej., predicción de la bioactividad de diferentes compuestos para tratar cáncer de mama (Qin *et al.*, 2022)).

- **Aprendizaje No Supervisado**: El aprendizaje no supervisado trata con datos no etiquetados. El objetivo es encontrar patrones ocultos o estructuras intrínsecas en los datos de entrada. Las técnicas comunes incluyen el agrupamiento (ej., identificación de poblaciones celulares fenotípicamente diferentes para la estratificación de pacientes con cáncer (Leelatian *et al.* 2020)) y la reducción de dimensionalidad (ej., reducir el número de variables en un conjunto de datos mientras se preservan sus características esenciales).

- **Aprendizaje por Refuerzo (*Reinforcement Learning,* RL)**: En el aprendizaje por refuerzo, un "agente" aprende a tomar decisiones realizando acciones en un entorno, estas decisiones conllevan recompensa acumulativa (positiva o negativa) que han de maximizar. Este enfoque se usa a menudo en escenarios donde la secuencia de acciones es importante (ej., control robótico en equipamiento de laboratorios autónomos (Volk et. al, 2023)) aunque puede ser aplicado a diversos problemas con un planteamiento adecuado (ej., identificación de tumores en imágenes 2D de escáneres cerebrales (Stember & Shalu, 2020)).

4.2 Aprendizaje Profundo (*Deep Learning*) y Redes Neuronales (*Neural Networks*)

El aprendizaje profundo (DL) es un subcampo especializado del ML que involucra redes neuronales con muchas capas. Estas redes neuronales de múltiples capas, conocidas como redes neuronales profundas, pueden aprender automáticamente a representar datos a través de múltiples niveles de abstracción. El DL ha tenido un éxito particular en tareas que implican grandes cantidades de datos y patrones complejos, como el reconocimiento de imágenes y voz, o la clasificación de datos con alta dimensionalidad. Para ello, el DL dispone de una potente herramienta, las redes neuronales.

En el corazón del aprendizaje profundo se encuentran las NN, que son modelos computacionales inspirados en la estructura y función del cerebro humano. Una red neuronal está formada por capas de nodos interconectados (neuronas), cada una realizando un cálculo simple (Géron A., 2019). Las capas de las NN más simples se organizan típicamente en una capa de entrada, una o más capas ocultas densas (donde todas las neuronas de una capa están conectadas con todas las neuronas de las capas anterior y posterior) y una capa de salida. Cada conexión entre dos neuronas está representada por un valor numérico llamado peso, a parte cada neurona tiene un valor que se conoce como *bias* o sesgo y una función de activación.

En el proceso de computación de una red neuronal, cada neurona calcula el sumatorio de los productos de los valores neuronas de la capa anterior (que están conectadas a ella) y sus respectivos pesos, más su sesgo (Ecuación 1). El valor resultante lo ingresará por la función de activación para obtener el resultado final que lo transmitirá hacia delante a las neuronas a las cuales esté conectada (Figura 2).

$$y = \sum_{1}^{i} (w_i \cdot x_i) + b \qquad (Ecuación\ 1)$$

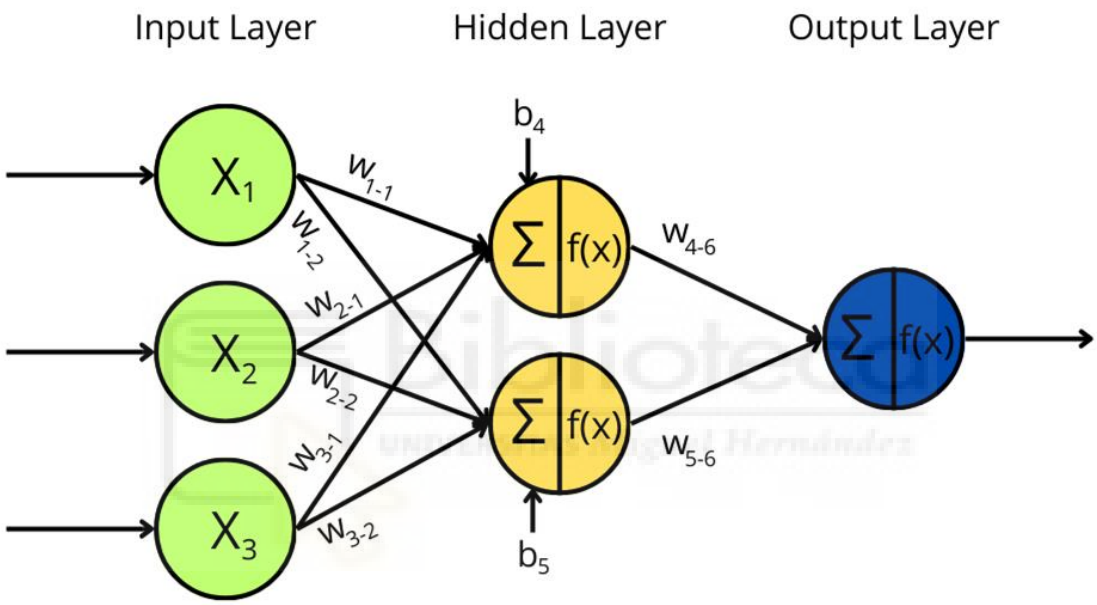

Figura 2.- Diagrama de un perceptrón con una sola capa oculta. X_i son los valores de entrada, w_{i-j} son los pesos de las conexiones, b_i son los sesgos y f(x) es la función de activación.

Durante el entrenamiento, la red ajusta los pesos de las conexiones y los sesgos de las neuronas para minimizar la diferencia entre sus predicciones y las salidas reales, a este proceso de le llama propagación inversa o *backpropagation* y es lo que permite que la red neuronal aprenda de los datos (Rumelhart *et al.*, 1985).

Existen una gran variedad de tipos de NN cada una con una arquitectura adaptada a tareas y dominios específicos. Algunas de las arquitecturas más relevantes para la investigación biotecnológica son:

- **Perceptrones multicapa (*Multilayer Perceptron,* MLP)**: son tipo de redes neuronales son las más simples. Estas NN están formadas por una capa de entrada, una capa de

salida y una o más capas ocultas. La red neuronal de la Figura 2 es un ejemplo de MLP con una única capa oculta formada por dos neuronas. A pesar de su simplicidad, los MLP alcanzan buenos resultados en tareas de regresión y clasificación.

- **Redes Neuronales Convolucionales (*Convolutional Neural Networks,* CNN)**: Las CNN son un tipo de red neuronal específicamente diseñada para procesar datos estructurados en cuadrícula, como imágenes, donde la posición relativa de los datos (píxeles en este caso) es importante. Utilizan capas convolucionales para detectar automáticamente características como bordes, texturas y formas en las imágenes, lo que las hace muy efectivas para tareas de reconocimiento de imágenes (LeCun *et al.*, 2015).

- **Redes Neuronales Recurrentes *(Recurrent Neural Networks*, RNN)**: Las RNN están diseñadas para datos que disponen de un orden, como series temporales o lenguaje natural, donde existe una relación secuencial de los datos. Tienen conexiones que forman ciclos dirigidos, lo que permite que la información persista a lo largo de pasos en una secuencia. Esto las hace adecuadas para tareas como modelado de lenguaje y traducción automática, o cálculo de predicciones en la evolución de un sistema (LeCun *et al*., 2015).

- **Transformadores (*Transformers*)**: Los transformadores son un tipo avanzado de modelo de DL que ha revolucionado el procesamiento del lenguaje natural. A diferencia de las RNN, los transformadores no procesan datos secuenciales de manera lineal. Utilizan mecanismos de atención que permiten a los modelos enfocarse en diferentes partes de la secuencia de entrada simultáneamente, lo que mejora significativamente la eficiencia y el rendimiento en tareas como traducción automática, generación de texto y reconocimiento de voz (Vaswani *et al.*, 2018). Es por ello por lo que los transformadores se han convertido en la base de todo modelo de lenguaje moderno, como GPT (*Generative Pre-trained Transformer)* (Radford *et al.*, 2018).

4.2.1 Modelos de Lenguaje de Gran Escala (*Large Language Models*)

El procesamiento del lenguaje natural surge como intersección entre la inteligencia artificial y la lingüística. El NLP comprende una gran variedad de tareas relacionadas con la interpretación y generación de texto a diferentes niveles (Prakash *et al.*, 2011). En los niveles más altos tenemos tareas como la traducción de texto, identificación de errores o la identificación y clasificación de palabras o frases (“entidades”), que a su vez están compuestas por tareas de bajo nivel como la detección de los límites de una frase o descomposición morfológica de palabras compuestas (algo muy común en las disciplinas científicas) (Prakash *et al.* 2011).

Las herramientas más potentes en la realización de estas tareas de NLP son los modelos de lenguaje de gran escala. Los LLM son modelos probabilísticos autorregresivos, redes neuronales cuyo objetivo final es predecir la siguiente palabra en una secuencia de texto. Estos modelos aprovechan los mecanismos de autoatención de los transformadores y el entrenamiento previo con grandes cantidades de datos lingüísticos para la resolución de esas tareas relacionadas con el lenguaje natural.

Una familia de los LLMs más conocidos y con mejores resultados es GPT (que empezó con GPT-1 (Radford *et al.*, 2018) y siendo el más reciente GPT-4 (OpenAI *et al.*, 2023)), los modelos detrás de la herramienta chatGPT (OpenAI, 2024), aunque existen muchos otros de código abierto como la familia de LLaMA (Touvron *et al.*, 2023) y LLaMA2 (Touvron *et al.*, 2023) o PaLM (Chowdhery *et al.*, 2023). Estos modelos alcanzan tales capacidades del NLP que se han convertido en herramientas muy versátiles y potentes capaces de mantener conversaciones, generar resúmenes y contrastar ideas, habilidades útiles para el desarrollo del trabajo de investigación.

4.2.2 Modelos de Lenguaje de Gran Escala Multimodales (*Multimodal Large Language Models*)

En la vida real, mucha información que procesamos día a día no proviene de textos, sino que también somos capaces de procesar sonidos e imágenes, entre otros. De la misma forma, los documentos científicos contienen figuras en las que se representan datos de forma diferente a un texto escrito. El papel de los gráficos es fundamental para facilitar la comunicación efectiva de información, transmitir ideas y representar resultados. Los gráficos y otras representaciones multimedia (diagramas, tablas, etc.), son herramientas con la capacidad de resumir conjuntos de datos complejos de manera concisa e intuitiva, lo que permite a los investigadores captar rápidamente ideas clave, ayudando en el razonamiento informado y la comprensión de conceptos complejos (Huang *et al.*, 2024).

Con el objetivo de aprovechar estas maneras alternativas de transmitir información surgen los MLLM (OpenAI, 2023; Liu *et al.*, 2023). Los MLLM representan un avance significativo en el campo de la inteligencia artificial, integrando la capacidad de procesamiento de lenguaje natural con la comprensión y generación de información visual y auditiva (Figura 3). Estos modelos se entrenan con vastos conjuntos de datos que pueden abarcar texto, imágenes, audio y video, permitiéndoles interpretar y generar respuestas que incorporan múltiples formas de información.

Al combinar diversas modalidades, los MLLM pueden ser herramientas de gran utilidad en ciencia pues pueden procesar artículos científicos y sus figuras obteniendo así una visión más completa de las publicaciones. En los últimos años, concorde a la explosión de aparición de MLLM

(Figura 4), se han desarrollado multitud de conjuntos de datos específicos del ámbito científico (Li *et al.*, 2023) para así entrenar MLLM que sean capaces de interpretar tanto el texto como las gráficas, y que dispongan de habilidades como la conversión de gráfica a tabla, la generación de pies de figura o la detección de errores de consistencia entre gráfica y texto (Huang *et al.*, 2024). Algunos MLLM con estas características son GPT4-V (OpenAI, 2023), ChartAssistant (Meng *et al.*, 2024) o ChartLlama (Han *et al.*, 2023).

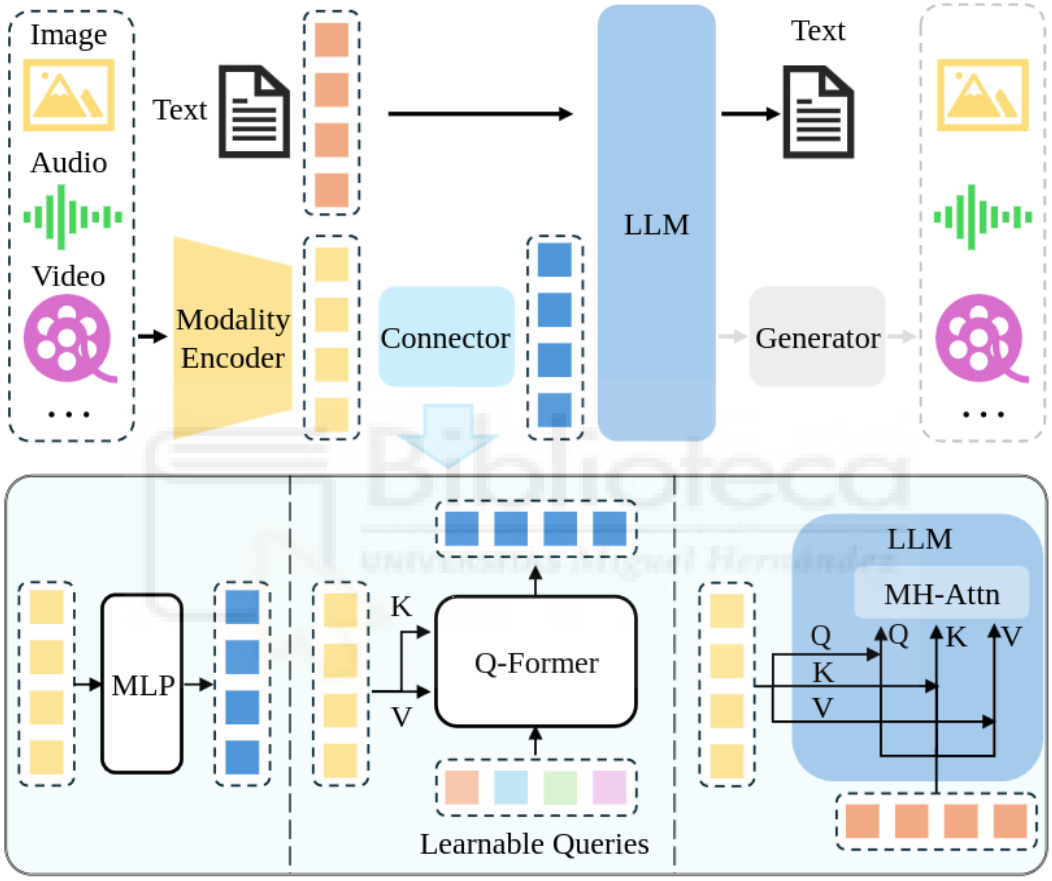

Figura 3.- Una ilustración de la arquitectura MLLM típica. Incluye un *encoder*, un conector y un LLM. Opcionalmente se puede conectar un generador al LLM para generar más modalidades además del texto. El codificador capta imágenes, audios o vídeos y genera características codificadas, que son procesados por el conector para que el LLM pueda comprender mejor. Tomada de Yin *et al.*(2023).

4.3 Algoritmos

Aparte del aprendizaje profundo y las redes neuronales, el aprendizaje automático tiene a su disposición una variedad de algoritmos. Los algoritmos son el conjunto de reglas o instrucciones que una computadora sigue para resolver un problema o realizar una tarea. En el contexto de la

IA y el aprendizaje automático, los algoritmos procesan datos de entrada y aprenden de ellos para hacer predicciones o tomar decisiones. Algunos algoritmos comúnmente utilizados en IA son:

- **Árboles de Decisión**: Estos algoritmos dividen los datos en ramas basadas en ciertas condiciones, creando un modelo de decisiones en forma de árbol. Cada nodo representa una característica, cada rama una regla de decisión y cada hoja un posible resultado. Los árboles de decisión son fáciles de interpretar y visualizar, lo que los hace útiles para comprender el proceso de toma de decisiones (Breiman, L, 2011).

- **Máquinas de Soporte Vectorial (*Support Vector Machines,* SVM)**: Las SVM son modelos de aprendizaje supervisado utilizados para tareas de clasificación y regresión. Funcionan encontrando el hiperplano que mejor separa las clases en el espacio de características, maximizando el margen entre las clases. Las SVM son efectivas en espacios de alta dimensionalidad y con donde los datos se puedan separar claramente (Géron A., 2019).

- **Agrupamiento en K-Medias**: Este algoritmo de aprendizaje no supervisado divide los datos en K grupos, donde cada punto de datos pertenece al grupo con la media más cercana. Se usa comúnmente para en aplicaciones donde agrupar puntos de datos similares es beneficioso (clasificación de datos).

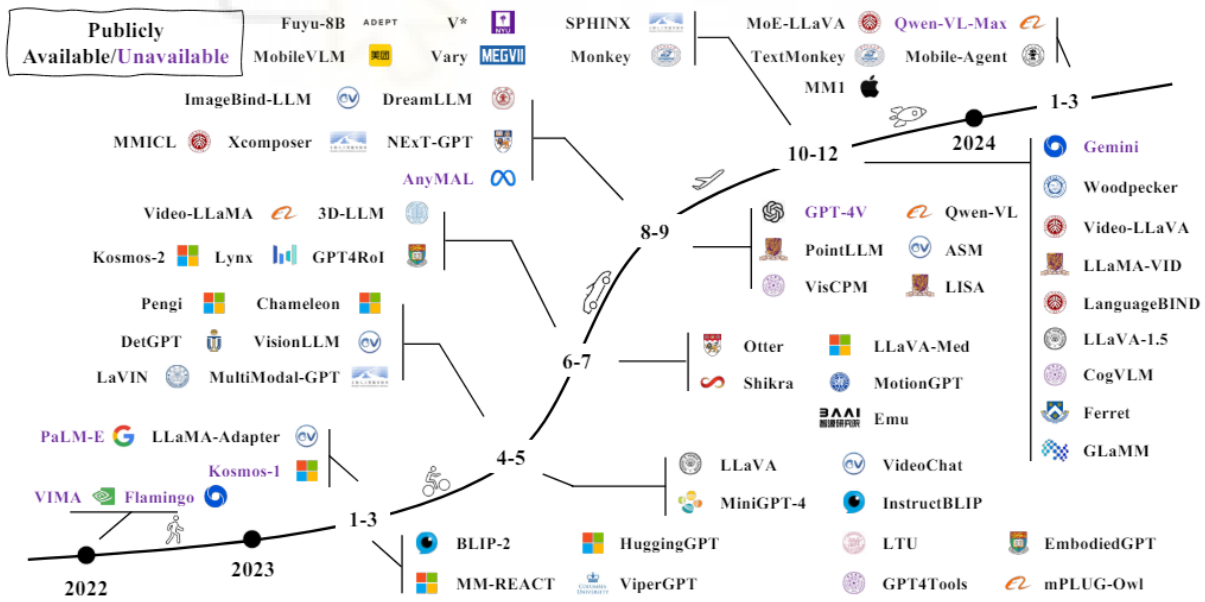

Figura 4.- Una cronología de MLLM representativos. Estamos siendo testigos de un rápido crecimiento en este campo. Tomada de Yin *et al.*(2023).

6. Etapas de la investigación científica y aplicaciones de herramientas de IA

La investigación científica es un proceso sistemático que implica varias etapas clave para asegurar la validez y fiabilidad de los hallazgos. Las principales etapas de la investigación científica se pueden categorizar en términos generales en: (1) la identificación del objetivo de la investigación, (2) la elección del enfoque científico adecuado, (3) la recopilación de datos mediante la realización de experimentos, (4) el análisis de los datos obtenidos y (5) la presentación de los resultados a la comunidad científica (Masic, 2016). Sin embargo, trabajar como personal de investigación puede tener efectos psicológicos significativos, afectando tanto a su bienestar como a los resultados de sus investigaciones. Estudios han destacado la prevalencia del agotamiento y los problemas de salud mental entre los investigadores, lo cual puede afectar negativamente la calidad de la investigación científica. La presión por publicar, asegurar financiamiento y cumplir con las expectativas académicas puede forzar al investigador a trabajar jornadas más largas alcanzando a altos niveles de estrés y agotamiento (Nagy *et al.*, 2019).

En este contexto, el desarrollo de nuevas técnicas que ayuden y agilicen la tarea de investigación, puede conllevar un beneficio no solo para el propio investigador sino como para la comunidad científica. Así pues, en esta sección se repasarán casos prácticos donde se muestran ejemplos del estado del arte de la IA aplicada a cada etapa de la investigación biotecnológica.

5.1 Identificación del objetivo de la investigación

Inicialmente, los investigadores deben definir claramente qué objetivo pretenden cumplir con su investigación. Esto, a menudo, implica revisar teorías existentes y documentarse sobre el estado del arte en el campo para así generar una hipótesis que será probada a través de sus métodos de investigación. Esta etapa establece la base para todo el proceso de investigación al delinear el objetivo específico y el alcance del estudio.

En esta fase de la investigación, el principal foco de esfuerzo es adquirir una visión lo más completa posible sobre el campo y el problema a tratar mediante el análisis exhaustivo de artículos. Esto se traduce en el procesamiento de una gran cantidad de información, principalmente de texto, pero también en forma de figuras. Estas características hacen de los LLM/MLLM unos candidatos perfectos para su uso como herramienta de ayuda.

Una opción sería la de generar un conjunto de artículos relacionados, ya sea manualmente, con una búsqueda simple en una base de datos o con herramientas disponibles online como Litmaps (Litmaps, 2024) y pasarlos (bien solo el texto o el texto e imágenes) por una MLLM como GPT-4V (OpenAI, 2023) para generación resúmenes.

Sin embargo, hay modelos con la capacidad de combinar todo el proceso anterior, como por ejemplo Scite (Nicholson *et al.*, 2021). Scite es una herramienta de búsqueda bibliográfica que utiliza el contexto de las citas y las clasifica utilizando DL. El investigador sólo tiene que hacer una pregunta al sistema y automáticamente Scite busca los artículos relevantes para la pregunta y utiliza sus citas para construir una respuesta. Además, una vez generada, comprueba su veracidad y te proporciona todas las referencias que están insertadas.

Scite usa una gran base de datos tanto de publicaciones de acceso abierto como de editoriales con las que tenga acuerdo. La búsqueda es transparente, por lo que puedes ver las consulta que realiza a la base de datos a partir de tu pregunta, y te permite modificar los parámetros. Adicionalmente, puedes crear un grupo de artículos y seleccionarlo como nueva base datos para generar entornos más controlados.

En definitiva, Scite es un instrumento de extracción y análisis automatizados de referencias científicas que imita y acelera la tediosa búsqueda de artículo científicos previa a la investigación.

5.2 Elección del enfoque científico

Después de la formulación de la pregunta de investigación, los investigadores proceden a seleccionar el enfoque científico que se alinea con sus objetivos. Este paso implica decidir si la investigación será cuantitativa, cualitativa o una combinación de ambas metodologías y diseñar los experimentos que serían necesarios para estudiar la hipótesis. La elección entre estos enfoques depende de la naturaleza de la pregunta de investigación y del tipo de datos necesarios para abordarla de manera efectiva.

En este punto de la investigación, es menos común que la IA se vea involucrada, pues es trabajo del propio investigador elegir el rumbo que va a seguir el estudio. Sin embargo, hay casos donde esta tarea sí que se ve escogida por la IA y suceden cuando todo el proceso de investigación está decidido por un agente de inteligencia artificial. Estoy hablando de laboratorios o experimentación autónomos (Ferguson *et al.*, 2022; Szymanski *et al.*, 2023; Sparkes *et al.*, 2010).

Los laboratorios autónomos son más utilizados en el campo de la ciencia de materiales y de la síntesis inorgánica. Sin embargo, se han hecho estudios sobre la aplicación de estos conceptos al campo de la biotecnología.

En Sparkes *et al.*, 2010 se presenta la figura de Adam un “científico robot”. Adam es una combinación de métodos computacionales, instrumentos automatizados de robótica de laboratorio avanzados, aprendizaje en bucle cerrado y expresión lógica formal. Genera automáticamente hipótesis a partir del conocimiento y modelos disponibles, diseña experimentos físicos para probar

estas hipótesis, lleva a cabo los experimentos en un sistema robótico de laboratorio, y luego analiza e interpreta los resultados.

Fue diseñado para llevar a cabo experimentos de crecimiento microbiano para estudiar la genómica funcional en la levadura *Saccharomyces cerevisiae*. Su objetivo era identificar los genes que codifican 13 "enzimas huérfanas" (aquellas que se saben que existen en el organismo, pero no se sabe su gen correspondiente). Adam concibió 20 hipótesis sobre la identidad de los genes, probó todas estas hipótesis en su laboratorio robótico y pudo confirmar mediante experimentación, con un alto grado de confianza, la corrección de 12 de ellas.

En estudios más recientes (Ferguson *et al.*, 2022), se sigue estudiando la experimentación autónoma con tejidos biológicos como manera de aumentar la rapidez y consistencia de los experimentos (pues los tejidos vivos son muy sensibles a fluctuaciones). Además, se intenta transferir el conocimiento adquirido por el propio agente de IA en laboratorios autónomos químicos a laboratorios autónomos de ciencias biomédicas.

5.3 Recopilación de datos mediante la realización de experimentos

Una vez determinado el enfoque de investigación, los investigadores pasan a la fase de recopilación de datos. Esta etapa implica la recolección de información o datos relevantes que se utilizarán para confirmar o rechazar la hipótesis. Los métodos de recopilación de datos pueden variar ampliamente según el diseño y los objetivos de la investigación, e incluyen encuestas, experimentos, entrevistas u observaciones. Es crucial que los investigadores aseguren que los datos recopilados sean precisos, fiables y aborden directamente la pregunta de investigación en cuestión.

En esta etapa es donde la inteligencia artificial más puede ser explotada dada su extrema versatilidad. Aquí la IA puede ser más que un complemento para el investigador y convertirse en parte integral a la investigación, empujando los límites del conocimiento en biotecnología. Uno de los mayores ejemplos sería el de AlphaFold2 (AF2), la primera aproximación computacional capaz de predecir estructuras proteicas con resoluciones parecidas a las experimentales (Figura 5) (Jumper *et al.*, 2021).

El problema del plegamiento de proteínas ha sido un desafío fundamental en biología molecular desde que en 1973 se descubrió que la estructura terciaria de una proteína depende de su secuencia de aminoácidos (Anfinsen, 1973). Durante más de medio siglo, los investigadores han trabajado con el objetivo de predecir las estructuras terciarias de las proteínas a partir de secuencias de aminoácidos y desentrañar los mecanismos subyacentes al plegamiento de proteínas (Li *et al.*, 2017).

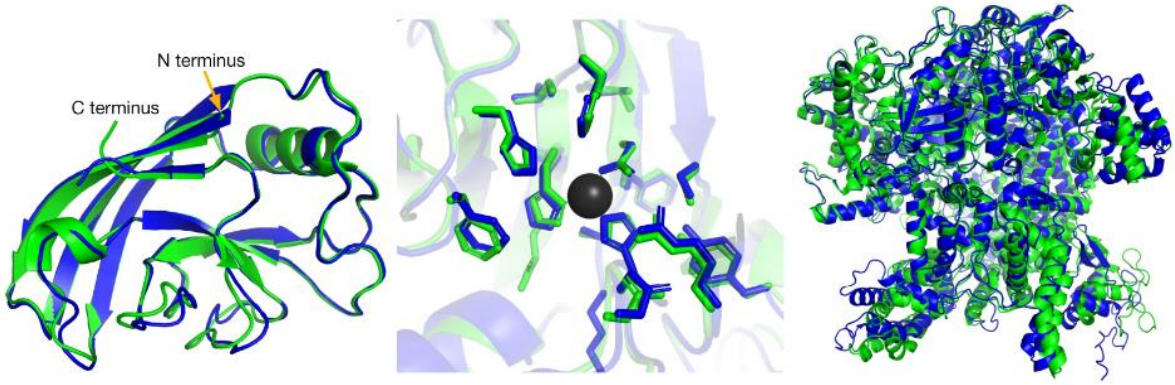

Figura 5.- Predicciones estructurales (azules), generadas por AlphaFold, de proteínas comparadas con la real (verde). Las proteínas utilizadas pertenecen al conjunto de datos CASP14, diseñado con el objetivo de evaluar herramientas de modelado molecular. Editado de Jumper *et al.* (2021).

AF2 ha sido fundamental para predecir estructuras de proteínas con alta precisión, comparable a métodos experimentales como la cristalografía de rayos X gracias a módulos en su arquitectura que producen representaciones de alineamientos múltiples de secuencia. Esta herramienta de IA ha tenido un impacto tan significativo en la biotecnología porque, al ofrecer una herramienta capaz de predecir fielmente las estructuras de las proteínas, ha abierto las puertas una vasta cantidad de estudios, acelerando de forma irremediable muchos campos de la biotecnología.

Por ejemplo, AF2 se ha utilizado en la predicción de estructuras de proteínas virales, ayudando en el estudio de la patogénesis viral y el desarrollo de estrategias antivirales (Weissman *et al.*, 2022). También se ha observado que al combinar las predicciones de AF2 con otras herramientas computacionales, se han podido mejorar la termoestabilidad de las enzimas, lo que ha llevado a avances en la ingeniería de enzimas y la biocatálisis (Peccati *et al.*, 2023). La disponibilidad de predicciones de AF2 para una gran cantidad de proteínas en la base de datos AlphaFold DB ha proporcionado a los investigadores un recurso valioso para aprovechar la información estructural en sus estudios (Varadi *et al.*, 2024). Adicionalmente, promueve la democratización del acceso a herramientas sofisticadas basadas en IA, facilitando así la investigación en ciencias de la vida y promoviendo una adopción más amplia de herramientas de bioinformática estructural (Roberts *et al.*, 2023). Además, es importante anotar que en 2024 han publicado AlphaFold3 (Abramson *et al.*, 2024), con una arquitectura mejorada que ayuda a predecir las interacciones biomoleculares. Por lo que es esperable que las contribuciones sigan aumentando.

AF2 ha tenido un profundo impacto en la biotecnología al revolucionar la predicción de estructuras de proteínas, transformando la forma en que los investigadores abordan los estudios relacionados con proteínas, abriendo nuevas posibilidades para la innovación y el descubrimiento

en el campo de la biotecnología. No obstante, AF2 es solo una de las muchas formas en la que la inteligencia artificial puede afectar al desarrollo de líneas de investigación o campos enteros dentro de la biotecnología. Corresponde a los investigadores el intentar implementar soluciones mediante IA a problemas de difícil resolución en las disciplinas biológicas.

5.4 Análisis de datos

Después de la recopilación de datos, los investigadores proceden a analizarlos para obtener conclusiones y revelaciones significativas sobre el tema. El análisis de datos implica procesar e interpretar la información recopilada para identificar patrones, relaciones o tendencias que puedan responder a la pregunta de investigación. Dependiendo de la metodología de investigación elegida, el análisis de datos puede involucrar técnicas estadísticas, análisis temático u otros métodos cualitativos o cuantitativos para derivar resultados significativos.

El aprendizaje automático fue diseñado para ser capaz de procesar gran cantidad de datos y sacar inferencias de ellos. Es por ello por lo que tenemos una gran cantidad de modelos y algoritmos para elegir, aunque generalmente, dependiendo de los tipos de datos hay algunos más eficaces que otros. En este apartado, la IA sigue manteniendo un papel clave en la investigación pues los resultados de estudios biotecnológicos suelen requerir de análisis estadístico de sus resultados.

Un buen ejemplo de ello son los análisis metagenómicos. Los estudios metagenómicos surgen como respuesta al problema de que la mayoría de los organismos que se encuentran por el entorno no son cultivables (Amann *et al.*, 1995).

La IA ha tenido un impacto significativo en el análisis de datos en estudios metagenómicos al proporcionar herramientas para extraer información de conjuntos de estos datos biológicos complejos. Algoritmos de aprendizaje automático como árboles de decisión, SVM, CNN y modelos de aprendizaje profundo se utilizan cada vez más para el análisis de datos metagenómicos y la predicción de diversos fenómenos biológicos (Chang *et al.*, 2017; Vu *et al.*, 2020). Estos enfoques impulsados por IA permiten a los investigadores utilizar la abundancia de especies metagenómicas para la predicción de fenotipos, mejorar la clasificación de hongos y analizar grandes conjuntos de datos metagenómicos para comprender mejor las comunidades microbianas.

Estudios como Mathieu *et al.* (2022), ponen en valor la eficacia de la combinación entre estas dos disciplinas. Demuestran con este enfoque coordinado es probable superar los obstáculos clásicos en el análisis de muestras metagenómicas, como la sobreestimación hacia ciertas especies bacterianas, la posibilidad de hacer uso de genomas reconstruidos sin anotar o la no deleción de secuencias no informativas para reducir ruido.

5.5 Presentación de resultados

Finalmente, la última etapa de la investigación científica implica la presentación de los hallazgos y conclusiones derivados del análisis. Se espera que los investigadores comuniquen sus resultados de manera clara y concisa, siguiendo los estándares de redacción académica e informes científicos. Esta etapa a menudo incluye la redacción de artículos de investigación u otros informes que describen la metodología de investigación, los resultados, la discusión y las implicaciones del estudio. Una presentación adecuada es esencial para compartir conocimientos, contribuir a la comunidad científica y potencialmente influir en las direcciones futuras de la investigación.

Esta situación es una imagen especular de la que nos encontrábamos en el primer paso, en ambas situaciones se quiere convertir el conocimiento científico que hay almacenado en diferentes modalidades en texto. Sin embargo, mientras que en el primer paso el investigador quería extraer esa información para él mismo, ahora lo quiere transmitir a la comunidad científica.

Estas circunstancias no suponen de gran cambio en cuanto a la utilización de IA se refiere, que ofrece las mismas herramientas, los LLM/MLLM. Sólo hay que tener una consideración adicional, y es hasta qué punto el uso de estos modelos para la generación de literatura científica está justificado éticamente.

7. Ética en el uso de la IA

Con el extenso uso que se la da a la IA en muchos sectores, llega también un debate ético (Zhou *et al.*, 2020). A menudo, el debate gira en alrededor de la privacidad, la seguridad y el consentimiento pues los modelos de IA se suelen entrenan con grandes conjuntos de datos, y en ellos puede haber contenido privado del cual no se ha dado permiso (Elliott *et al.*, 2022). Y aunque son temas que se tienen que tratar a nivel de sociedad, en el ámbito científico los valores como la integridad, la veracidad y la transparencia pueden considerarse especialmente importantes.

Aunque el uso de sistemas con IA en la ciencia no es algo nuevo, la ubicuidad y desarrollo de modelos generativos en la última década ha sido un hecho sin precedentes. Como hemos tratado aquí, los LLM son potentes herramientas versátiles para el procesamiento del lenguaje natural. Lo que significa que cada vez más científicos los están utilizando para la redacción de artículos científicos (Figura 6).

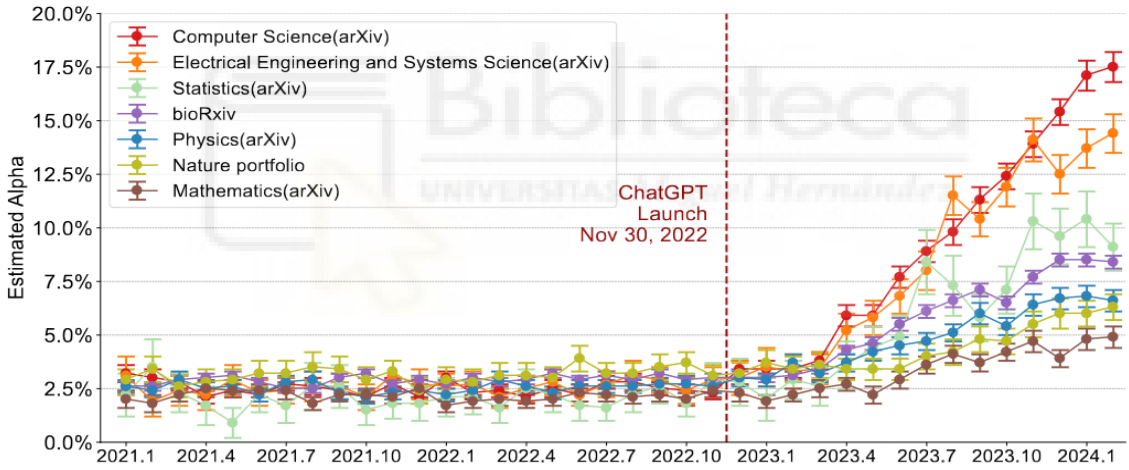

Figura 6.- Esta figura muestra la fracción (α) de frases que se estima que han sido modificadas sustancialmente por LLM en resúmenes de varios lugares de escritura académica. El análisis incluye cinco áreas dentro de arXiv (Informática, Ingeniería Eléctrica y Ciencias de Sistemas, Matemáticas, Física y Estadística), artículos de bioRxiv y un conjunto de datos combinados de 15 revistas dentro de la cartera de Nature. Las estimaciones se basan en el marco de cuantificación distribucional GPT, que proporciona estimaciones a nivel de población en lugar de análisis de documentos individuales. Cada punto en el tiempo se estima de forma independiente, sin aplicar supuestos de suavización temporal o continuidad. Las barras de error indican intervalos de confianza del 95%. Tomada de Liang *et al.* (2024).

No es de extrañar que ya se hayan dado casos donde se han tenido que retirar publicaciones porque se encontraban fragmentos de interacciones con LLMs entre el texto del artículo (DeGeurin, (2024, March 19)).

Por ello es importante desarrollar como comunidad unas guías éticas que ayuden a mantener la integridad de la ciencia. Resnik *et al.* (2024) han propuesto 9 recomendaciones:

- **Gestión de Sesgos y Errores**: Los investigadores deben identificar, describir, reducir y controlar los sesgos relacionados con la IA y los errores aleatorios.
- **Transparencia**: El uso de la IA en la investigación debe ser divulgado, descrito y explicado en un lenguaje accesible para los no expertos.
- **Participación de las Partes Interesadas**: Los investigadores deben involucrar a las comunidades afectadas y a las partes interesadas para abordar sus preocupaciones e intereses, especialmente en lo que respecta al sesgo.
- **Divulgación de Datos Sintéticos**: Al utilizar datos sintéticos, los investigadores deben etiquetar y describir claramente los elementos sintéticos y su proceso de generación.
- **Contribuciones de la IA**: Los sistemas de IA no deben ser nombrados autores o inventores, pero sus contribuciones deben ser divulgadas de manera transparente.
- **Educación y Mentoría**: La formación en ética de la investigación debe incluir discusiones sobre el uso ético de la IA.
- **Normas Éticas**: Las normas éticas existentes de la ciencia no necesitan cambiar, pero es necesario contar con una orientación adicional específica para la IA.
- **Explicabilidad**: Se deben realizar esfuerzos para desarrollar y utilizar sistemas de IA explicables para mejorar la confianza y la responsabilidad.
- **Desarrollo de Políticas**: Las instituciones y organizaciones deben desarrollar políticas y directrices para el uso ético de la IA en la investigación.

8. Conclusiones

El mundo de la inteligencia artificial es extremadamente extenso, pues dispone de innumerables algoritmos y modelos: redes neuronales recurrentes, transformadores, agrupación de K-medias, etc. En la misma línea, también es extremadamente versátil, pues sirve para resolver una gran cantidad de tareas: procesamiento del lenguaje natural, regresión, clasificación, etc.

No es de extrañar que muchas disciplinas de todos los campos la están utilizando de una forma u otra, y la biotecnología no es una excepción. Ya se ha conseguido, mediante el trabajo conjunto, desarrollar descubrimientos que han expandido las fronteras de la investigación. Además, el creciente desarrollo de IA relacionada con el procesamiento natural proporciona al investigador una serie de herramientas que facilitan su trabajo.

En conclusión, la integración de tecnologías de IA en la investigación biotecnológica está revolucionado la forma en que los investigadores acceden y utilizan la literatura científica, diseñan y desarrollan los experimentos y analizan y comparten la información obtenida. Al automatizar tareas, agilizar procesos y mejorar la eficiencia, la IA empodera a los investigadores en cada paso de la investigación, contribuyendo en última instancia a los avances en el conocimiento científico y mejorando los resultados de la investigación.

9. Bibliografía

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, *et al.* 2023. PaLM: Scaling Language Modeling with Pathways. J. Mach. Learn. Res. (2023), 240:1–240:113.

Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A. J., Bambrick, J., Bodenstein, S. W., Evans, D. A., Hung, C.-C., O'Neill, M., Reiman, D., Tunyasuvunakool, K., Wu, Z., Žemgulytė, A., Arvaniti, E., … Jumper, J. M. (2024). Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature, 630(8016), 493–500. https://doi.org/10.1038/s41586-024-07487-w

Amann, R. I., Ludwig, W., & Schleifer, K. H. (1995). Phylogenetic identification and in situ detection of individual microbial cells without cultivation. Microbiological Reviews, 59(1), 143–169. https://doi.org/10.1128/mr.59.1.143-169.1995

Anfinsen, C. B. (1973). Principles that Govern the Folding of Protein Chains. Science. https://doi.org/10.1126/science.181.4096.223

Bornmann, L., Haunschild, R., & Mutz, R. (2021). Growth rates of modern science: a latent piecewise growth curve approach to model publication numbers from established and new literature databases. Humanities and Social Sciences Communications, 8(1). https://doi.org/10.1057/s41599-021-00903-w

Breiman, L. Random Forests. Machine Learning 45, 5–32 (2001). https://doi.org/10.1023/A:1010933404324

Chang, H.-X., Haudenshield, J. S., Bowen, C. R., & Hartman, G. L. (2017). Metagenome-Wide Association Study and Machine Learning Prediction of Bulk Soil Microbiome and Crop Productivity. Frontiers in Microbiology, 8. https://doi.org/10.3389/fmicb.2017.00519

Cohan, A. and Goharian, N. (2017). Scientific document summarization via citation contextualization and scientific discourse. International Journal on Digital Libraries, 19(2-3), 287-303. https://doi.org/10.1007/s00799-017-0216-8

David Rumelhart *et al.* "Learning Internal Representations by Error Propagation," (Defense Technical Information Center technical report, September 1985).

DeGeurin, M. (2024, March 19). AI-generated nonsense is leaking into scientific journals. Popular Science. https://www.popsci.com/technology/ai-generated-text-scientific-journals/

Diaw, M. D., Stéphane, P., Durand-Salmon, A., Felblinger, J., & Oster, J. (2023). Ai-assisted qt measurements for highly automated drug safety studies. IEEE Transactions on Biomedical Engineering, 70(5), 1504-1515. https://doi.org/10.1109/tbme.2022.3221339

Elliott, D., & Soifer, E. (2022). AI Technologies, Privacy, and Security. Frontiers in Artificial Intelligence, 5. https://doi.org/10.3389/frai.2022.826737

F. Meng, W. Shao, Q. Lu, P. Gao, K. Zhang, Y. Qiao, and P. Luo, “Chartassisstant: A universal chart multimodal language model via chart-to-table pre-training and multitask instruction tuning,” arXiv preprint arXiv:2401.02384, 2024.

Ferguson, A. L., & Brown, K. A. (2022). Data-Driven Design and Autonomous Experimentation in Soft and Biological Materials Engineering. Annual Review of Chemical and Biomolecular Engineering, 13(Volume 13, 2022), 25–44. https://doi.org/10.1146/annurev-chembioeng-092120-020803

Forsythe, I., Howells, S., & Barrett, K. (2019). Reproducibility and data presentation. The Journal of Physiology, 597(22), 5313-5313. https://doi.org/10.1113/jp277519

Géron A. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd Edition. 2nd ed. O’Reilly Media, Inc.; 2019

Han, Y., Zhang, C., Chen, X., Yang, X., Wang, Z., Yu, G., Fu, B., & Zhang, H. (2023). ChartLlama: A Multimodal LLM for Chart Understanding and Generation (No. arXiv:2311.16483). arXiv. https://doi.org/10.48550/arXiv.2311.16483

HI iberia | Desarrollo software y soluciones TIC. (s. f.). Recuperado 20 de junio de 2024, de https://hi-iberia.es/

Huang, K.-H., Chan, H. P., Fung, Y. R., Qiu, H., Zhou, M., Joty, S., Chang, S.-F., & Ji, H. (2024). From Pixels to Insights: A Survey on Automatic Chart Understanding in the Era of Large Foundation Models (No. arXiv:2403.12027). arXiv. https://doi.org/10.48550/arXiv.2403.12027

Jerbi, D. (2023). Exploring the latest frontiers of artificial intelligence: a review of trends and developments.. https://doi.org/10.36227/techrxiv.22717327

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021).

Highly accurate protein structure prediction with AlphaFold. Nature, 596(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. Nature 521, 436–444 (2015). https://doi.org/10.1038/nature14539

Leelatian, N., Sinnaeve, J., Mistry, A., Barone, S., Brockman, A., Diggins, K., … & Irish, J. (2020). Unsupervised machine learning reveals risk stratifying glioblastoma tumor cells. Elife, 9. https://doi.org/10.7554/elife.56879

Li, S., & Tajbakhsh, N. (2023). SciGraphQA: A Large-Scale Synthetic Multi-Turn Question-Answering Dataset for Scientific Graphs (No. arXiv:2308.03349). arXiv. https://doi.org/10.48550/arXiv.2308.03349

Liang, W., Zhang, Y., Wu, Z., Lepp, H., Ji, W., Zhao, X., Cao, H., Liu, S., He, S., Huang, Z., Yang, D., Potts, C., Manning, C. D., & Zou, J. Y. (2024). Mapping the Increasing Use of LLMs in Scientific Papers. arXiv. https://doi.org/10.48550/ARXIV.2404.01268

Litmaps. (2024). Litmaps [Herramienta de búsqueda de artículos científicos]. https://www.litmaps.com

Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023). Visual Instruction Tuning. arXiv. https://doi.org/10.48550/ARXIV.2304.08485

Mathieu, A., Leclercq, M., Sanabria, M., Perin, O., & Droit, A. (2022). Machine Learning and Deep Learning Applications in Metagenomic Taxonomy and Functional Annotation. Frontiers in Microbiology, 13. https://doi.org/10.3389/fmicb.2022.811495

McCarthy, J., Minsky, M., Rochester, N., Shannon, C.E., A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence., http://raysolomonoff.com/dartmouth/boxa/dart564props.pdf August 1955

Mostavi, M., Chiu, YC., Huang, Y. *et al.* Convolutional neural network models for cancer type prediction based on gene expression. BMC Med Genomics 13 (Suppl 5), 44 (2020). https://doi.org/10.1186/s12920-020-0677-2

Nagy, G. A., Fang, C. M., Hish, A. J., Kelly, L., Nicchitta, C. V., Dzirasa, K., & Rosenthal, M. Z. (2019). Burnout and Mental Health Problems in Biomedical Doctoral Students. CBE—Life Sciences Education, 18(2), ar27. https://doi.org/10.1187/cbe.18-09-0198

Nicholson, J. M., Mordaunt, M., Lopez, P., Uppala, A., Rosati, D., Rodrigues, N. P., Grabitz, P., & Rife, S. C. (2021). scite: A smart citation index that displays the context of citations and classifies their intent using deep learning. Quantitative Science Studies, 2(3), 882–898. https://doi.org/10.1162/qss_a_00146

OpenAI, “Gpt-4v(ision) system card,” 2023. [Online]. Disponible en: https://openai.com/research/gpt-4v-system-card

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., … Zoph, B. (2023). GPT-4 Technical Report. arXiv. https://doi.org/10.48550/ARXIV.2303.08774

OpenAI. (2024). ChatGPT [Large multimodal model]. https://chat.openai.com/chat

Özgür, A., Xiang, Z., Radev, D., & He, Y. (2010). Literature-based discovery of ifn-γand vaccine-mediated gene interaction networks. Journal of Biomedicine and Biotechnology, 2010, 1-13. https://doi.org/10.1155/2010/426479

Peccati, F., Alunno-Rufini, S., & Jiménez-Osés, G. (2023). Accurate Prediction of Enzyme Thermostabilization with Rosetta Using AlphaFold Ensembles. Journal of Chemical Information and Modeling, 63(3), 898–909. https://doi.org/10.1021/acs.jcim.2c01083

Prakash M Nadkarni, Lucila Ohno-Machado, Wendy W Chapman, Natural language processing: an introduction, Journal of the American Medical Informatics Association, Volume 18, Issue 5, September 2011, Pages 544–551, https://doi.org/10.1136/amiajnl-2011-000464

Proyectos de I+D+i en líneas estratégicas - Transmisiones 2023 | Agencia Estatal de Investigación. (s. f.). Recuperado 20 de junio de 2024, de https://www.aei.gob.es/convocatorias/buscador-convocatorias/proyectos-idi-lineas-estrategicas-transmisiones-2023

Qin Y, Li C, Shi X and Wang W (2022) MLP-Based Regression Prediction Model For Compound Bioactivity. Front. Bioeng. Biotechnol. 10:946329. doi: https://doi.org/10.3389/fbioe.2022.946329

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

Reel, P. S., Reel, S., Pearson, E., Trucco, E., & Jefferson, E. (2021). Using machine learning approaches for multi-omics data analysis: A review. Biotechnology Advances, 49, 107739. https://doi.org/10.1016/j.biotechadv.2021.107739

Resnik, D. B., & Hosseini, M. (2024). The ethics of using artificial intelligence in scientific research: new guidance needed for a new tool. AI and Ethics. https://doi.org/10.1007/s43681-024-00493-8

Roberts, J. B., Nava, A. A., Pearson, A. N., Incha, M. R., Valencia, L. E., Ma, M., Rao, A., & Keasling, J. D. (2023). Foldy: a web application for interactive protein structure analysis. bioRxiv. https://doi.org/10.1101/2023.05.11.540333

Saini, F., Sharma, T., & Madan, S. (2021). A comparative analysis of expert opinions on artificial intelligence: evolution, applications, and its future. Advanced Journal of Graduate Research, 11(1), 10-22. https://doi.org/10.21467/ajgr.11.1.10-22

Sparkes, A., Aubrey, W., Byrne, E., Clare, A., Khan, M. N., Liakata, M., Markham, M., Rowland, J., Soldatova, L. N., Whelan, K. E., Young, M., & King, R. D. (2010). Towards Robot Scientists for autonomous scientific discovery. Automated Experimentation, 2(1), 1. https://doi.org/10.1186/1759-4499-2-1

Stember, J. and Shalu, H. (2020). Reinforcement learning using deep q networks and q learning accurately localizes brain tumors on mri with very small training sets.. https://doi.org/10.48550/arxiv.2010.10763

Szymanski, N. J., Rendy, B., Fei, Y., Kumar, R. E., He, T., Milsted, D., McDermott, M. J., Gallant, M., Cubuk, E. D., Merchant, A., Kim, H., Jain, A., Bartel, C. J., Persson, K., Zeng, Y., & Ceder, G. (2023). An autonomous laboratory for the accelerated synthesis of novel materials. Nature, 624(7990), 86–91. https://doi.org/10.1038/s41586-023-06734-w

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). LLaMA: Open and Efficient Foundation Language Models. arXiv. https://doi.org/10.48550/ARXIV.2302.13971

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., … Scialom, T. (2023). Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv. https://doi.org/10.48550/ARXIV.2307.09288

Turing, A. M. I.—Computing Machinery and Intelligence, Mind, Volume LIX, Issue 236, October 1950, Pages 433–460, https://doi.org/10.1093/mind/LIX.236.433

Varadi, M., Bertoni, D., Magana, P., Paramval, U., Pidruchna, I., Radhakrishnan, M., Tsenkov, M., Nair, S., Mirdita, M., Yeo, J., Kovalevskiy, O., Tunyasuvunakool, K., Laydon, A., Žídek, A.,

Tomlinson, H., Hariharan, D., Abrahamson, J., Green, T., Jumper, J., … Velankar, S. (2024). AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences. Nucleic Acids Research, 52(D1), D368–D375. https://doi.org/10.1093/nar/gkad1011

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023). Attention Is All You Need (No. arXiv:1706.03762). arXiv. https://doi.org/10.48550/arXiv.1706.03762

Volk, A.A., Epps, R.W., Yonemoto, D.T. *et al.* AlphaFlow: autonomous discovery and optimization of multi-step chemistry using a self-driven fluidic lab guided by reinforcement learning. Nat Commun 14, 1403 (2023). https://doi.org/10.1038/s41467-023-37139-y

Vu, D., Groenewald, M., & Verkley, G. (2020). Convolutional neural networks improve fungal classification. Scientific Reports, 10(1), 12628. https://doi.org/10.1038/s41598-020-69245-y

Weissman, A., Bennett, J., Smith, N., Burdorf, C., Johnston, E., Malachowsky, B., & Banks, L. (2022). Computational Modeling of Virally-encoded Ion Channel Structure. https://doi.org/10.21203/rs.3.rs-2182743/v1

Xiao, Q., Zhang, F., Xu, L., Liang, Y., Kon, O. L., Zhu, Y., … & Guo, T. (2021). High-throughput proteomics and ai for cancer biomarker discovery. Advanced Drug Delivery Reviews, 176, 113844. https://doi.org/10.1016/j.addr.2021.113844

Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., & Chen, E. (2023). A Survey on Multimodal Large Language Models. arXiv. https://doi.org/10.48550/ARXIV.2306.13549

Zhou, J., Chen, F., Berry, A., Reed, M., Zhang, S., & Savage, S. (2020). A Survey on Ethical Principles of AI and Implementations. 2020 IEEE Symposium Series on Computational Intelligence (SSCI), 3010–3017. https://doi.org/10.1109/SSCI47803.2020.9308437