



Universidad Nacional Autónoma de México
Instituto de Fisiología Celular
Unidad de Bioinformática y Manejo de la Información

“Me lo dijo ChatGPT”

Oportunidades y limitaciones de los *chatbots* en el marco de la investigación científica

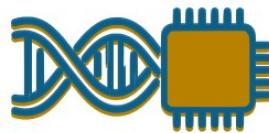


Dr. Carlos Alberto Peralta Alvarez
12 de Junio de 2025

- ¿Qué son los modelos grandes del lenguaje (LLMs)?
- Oportunidades, riesgos y limitaciones de los *chatbots* en la investigación biomédica

¿Qué son los modelos grandes de lenguaje?





¿Qué son los modelos grandes de lenguaje?

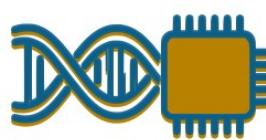
¿Con qué puedo ayudarte?

Pregunta lo que quieras

+ Herramientas

0





¿Qué son los modelos grandes de lenguaje?

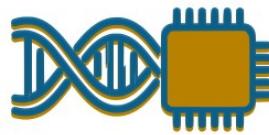


¿Con qué puedo ayudarte?

Pregunta lo que quieras

+ Herramientas

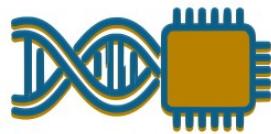




¿Qué son los modelos grandes de lenguaje?

- México es un país _____
- Depende de la _____
- Cachún cachún _____
- Omnis cellula e _____

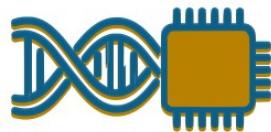




LLMs cómo predictores de texto

- México es un país megadiverso
- Depende de la especie
- Cachún cachún rá rá
- Omnis cellula e cellula





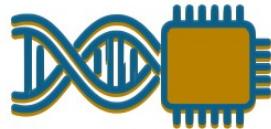
LLMs cómo predictores de texto

- LLMs – modelos grandes de lenguaje
 - Predictores de texto
 - Reconocimiento de patrones

The screenshot shows a Google search interface with a dark theme. At the top, the word "Google" is written in its signature white font. Below the search bar, there is a list of suggested search queries, each preceded by a magnifying glass icon. The queries are:

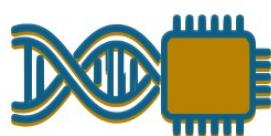
- mejores
- mejores películas netflix
- mejores restaurantes cdmx
- mejores universidades de mexico
- mejores películas
- mejores películas de terror
- mejores afores 2025
- mejores series de netflix
- mejores celulares gama media 2025
- mejores perfumes para hombre
- mejores playas de mexico

At the bottom of the search interface, there are two buttons: "Buscar con Google" and "Voy a tener suerte". A small link "Denunciar predicciones ofensivas" is located at the very bottom right.



LLMs cómo predictores de texto basados en *machine learning*

- LLMs – modelos grandes de lenguaje
 - Predictores de texto
 - Reconocimiento de patrones
- Aprendizaje de máquina
 - Patrones de idiomas
 - Generación de texto
 - Interacción con el usuario



LLMs como predictores de texto basados en *machine learning*

- ## ■ LLMs – modelos grandes de lenguaje

- ## - Predictores de texto

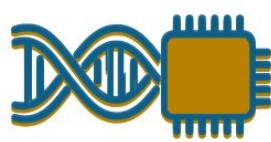
- ## – Reconocimiento de patrones

■ Aprendizaje de máquina

- ## – Patrones de idiomas

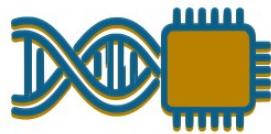
- ## – Generación de texto

- ### – Interacción con el usuario



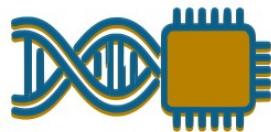
¿Cómo funcionan los LLMs? - ¡Dependen del contexto!





El tamaño del contexto importa

- Gemini 1.5 Pro – 1×10^6 **Tokens**
- GPT-4.1 mini – 1×10^6 **Tokens**
- Claude 3 - 2×10^5 **Tokens**
- Gemini 1.5 – 1.2×10^5 **Tokens**
- Deepseek-r1:1.5b - 1.2×10^5 **Tokens**



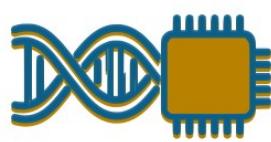
El tamaño del contexto importa...

Abstract

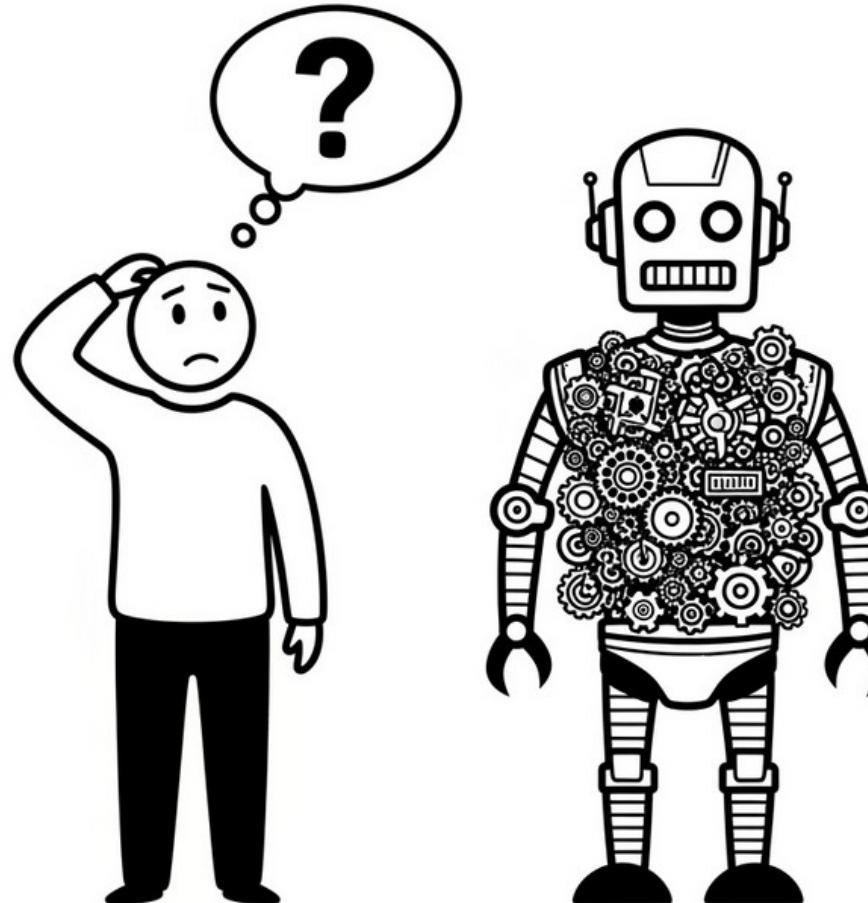
Recent evidence suggests that human gene promoters display gene expression regulatory mechanisms beyond the typical single gene local transcription modulation. In mammalian genomes, genes with an associated bidirectional promoter are abundant; bidirectional promoter architecture serves as a regulatory hub for a gene pair

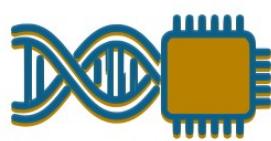
TOKENS	CHARACTERS
12,150	53,394

→ Paper completo

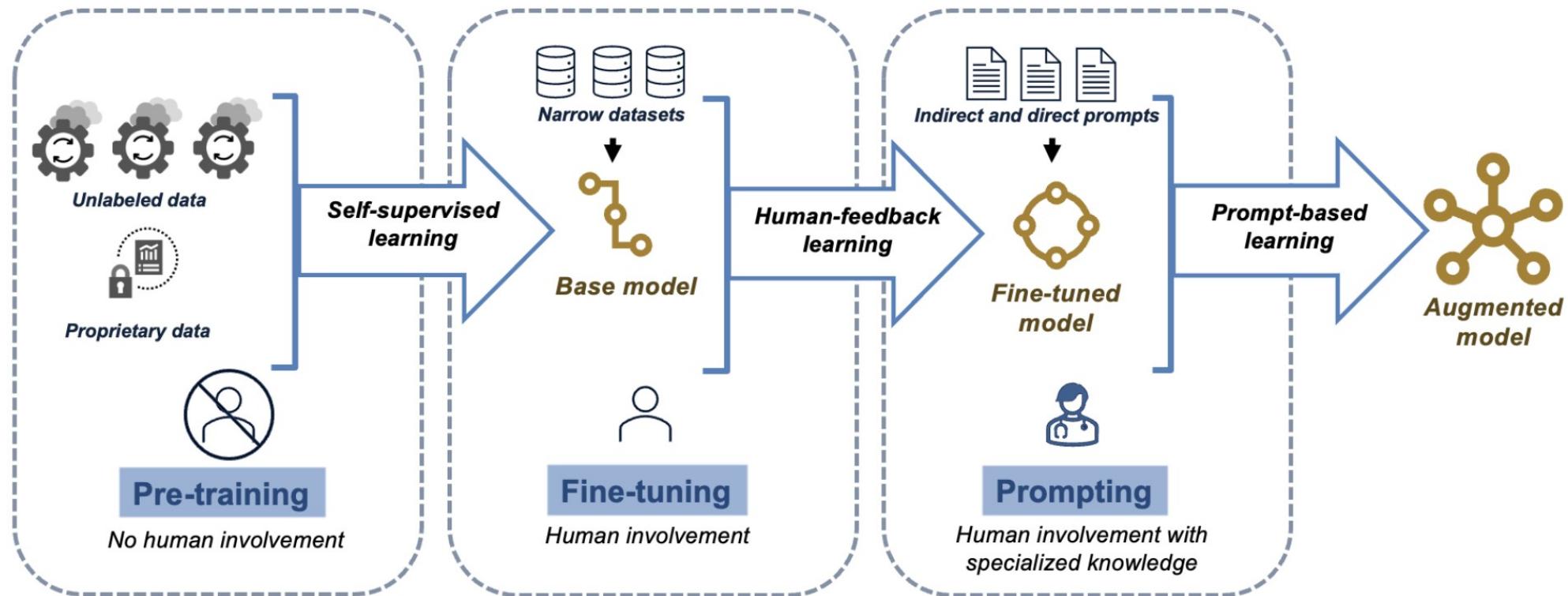


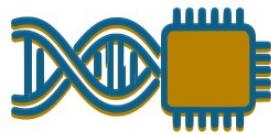
Pero ... ¿ Cómo funcionan?





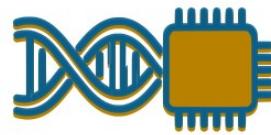
Etapas de la construcción de un LLM



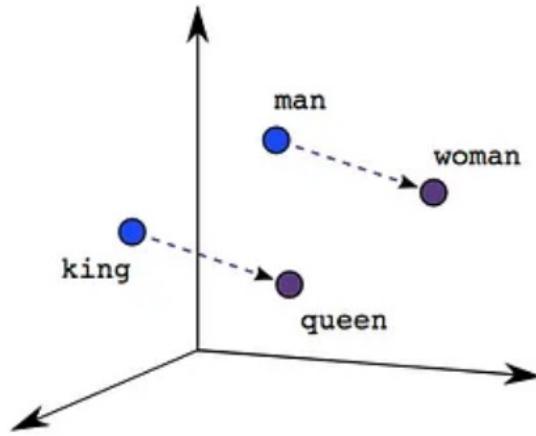


Consecuencias del entrenamiento

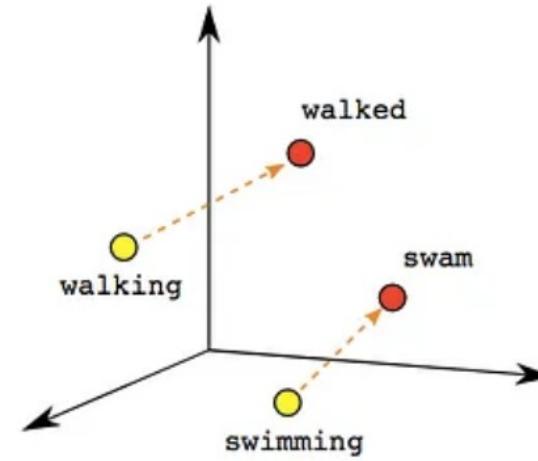
- Identificación de la estructura y sintaxis del lenguaje
- Identificación de patrones de escritura
- Generación de una base de **conocimiento del modelo**
 - Propagación de sesgos y prejuicios
 - Altamente dependiente del tamaño de los datos de entrenamiento
- Identificación de relaciones semánticas



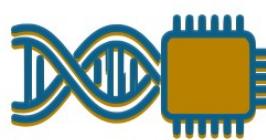
Los embeddings: parte central del funcionamiento de los LLMS



Male-Female

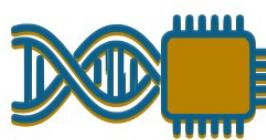


Verb Tense



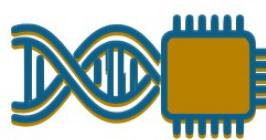
Los embeddings: parte central del funcionamiento de los LLMS





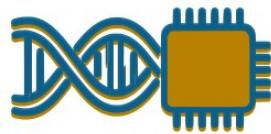
Los embeddings: parte central del funcionamiento de los LLMS





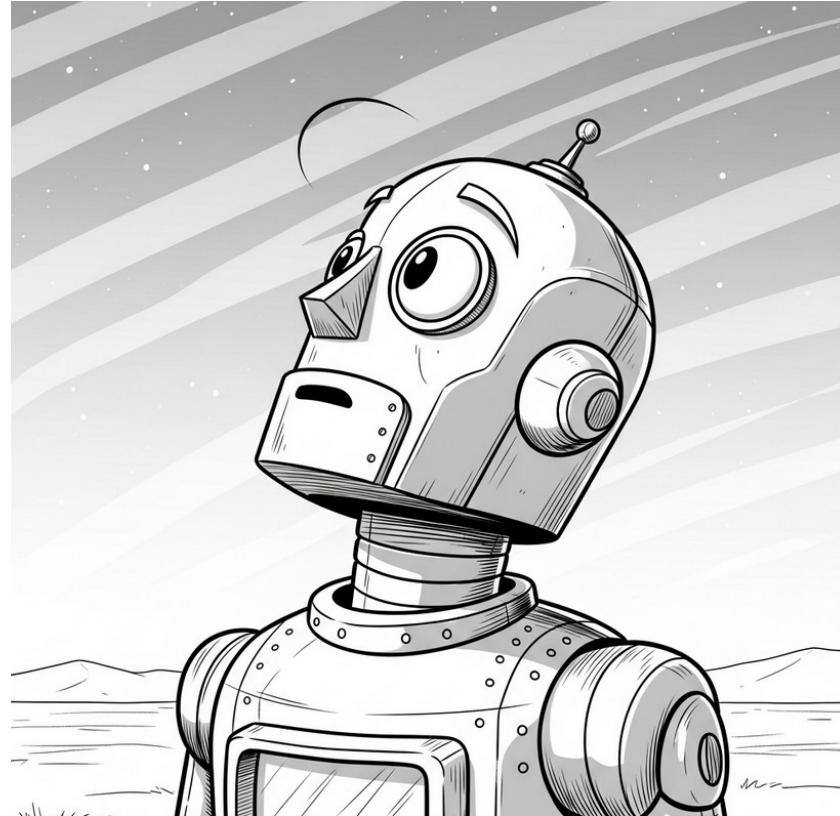
Los embeddings: parte central del funcionamiento de los LLMS





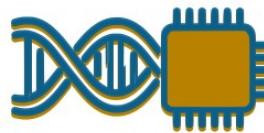
En resumen los LLMs son...

- Algoritmos de predicción y generación de texto
 - Calidad del texto generado depende de:
 - Complejidad del modelo
 - Proceso de entrenamiento
 - **Datos de entrenamiento**
 - **Objetivo del modelo**
- **NO** son inteligencias artificiales que entienden lo que se les pregunta o contestan



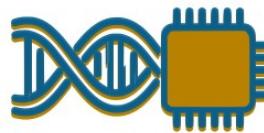
Oportunidades y limitaciones de los LLMs en la investigación





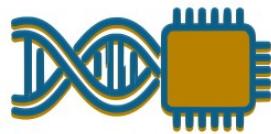
Oportunidades de uso de los chatbots en investigación biológica

- 1) Consulta
- 2) Documentación
- 3) Asistencia de escritura
- 4) Generación de hipótesis



Oportunidades de uso de los chatbots en investigación biológica

- 1) Consulta
- 2) Documentación
- 3) Asistencia de escritura
- 4) Generación de hipótesis



Consulta – Pregunta directa a un *chatbot*

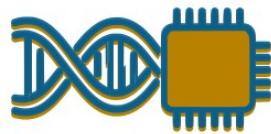
- Oportunidad
 - Obtener información rápidamente sobre un tema en particular

¿Con qué puedo ayudarte?

Pregunta lo que quieras

+ Herramientas





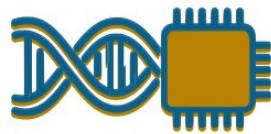
Consulta – Pregunta directa a un *chatbot*

- Oportunidad
 - Obtener información rápidamente sobre un tema en particular

- Limitantes
 - Existen temas “prohibidos”
 - Poca profundidad de la respuesta
 - Respuesta errónea
 - Alucinaciones

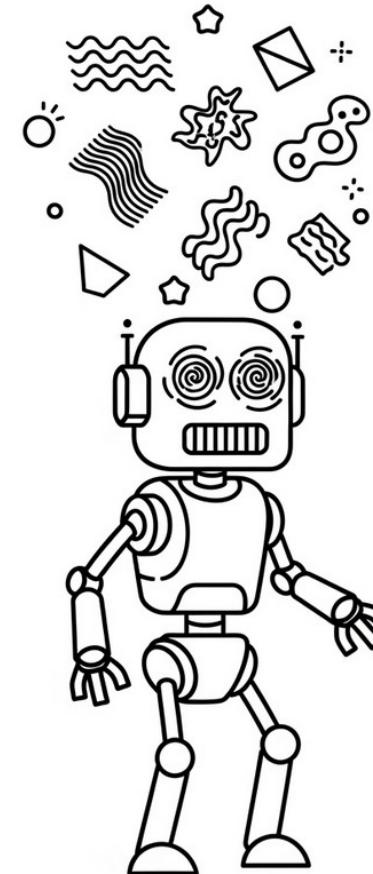
¿Con qué puedo ayudarte?

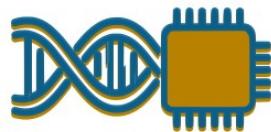




Abordando el problema de las limitaciones

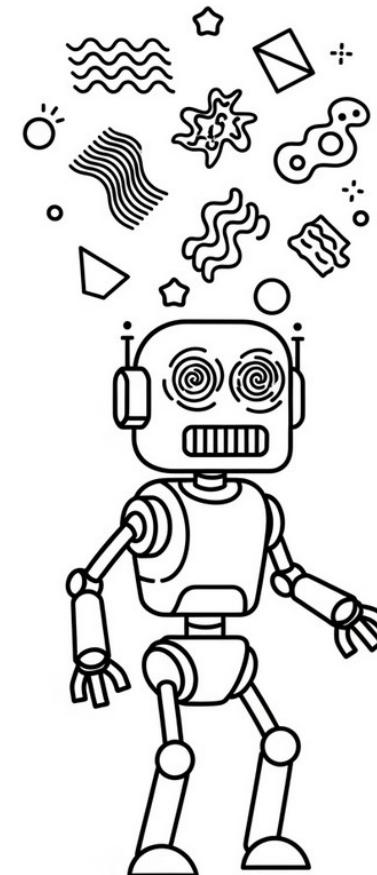
- **Poca profundidad en la respuestas, alucinaciones y errores**
 - Respuestas superficiales
 - Errores conceptuales o de razonamiento
 - Argumentos falsos que parecen **reales**

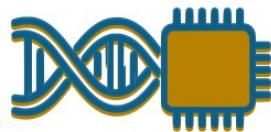




Abordando el problema de las limitaciones

- **Poca profundidad en la respuestas, alucinaciones y errores**
 - Respuestas superficiales
 - Errores conceptuales o de razonamiento
 - Argumentos falsos que parecen **reales**
 - **Citas (referencias) falsas**
- **Causas:**
 - Conjunto de entrenamiento inadecuado
 - ***Prompt*** inadecuado



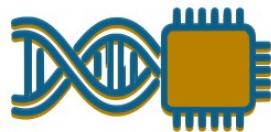


Abordando el problema de las limitaciones - Prompts

- **Malos prompts:**

- H3K27me3 cáncer de próstata
 - ¿Cuál es el papel de H3K27me3 en cáncer de próstata?

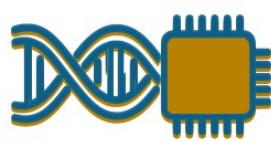




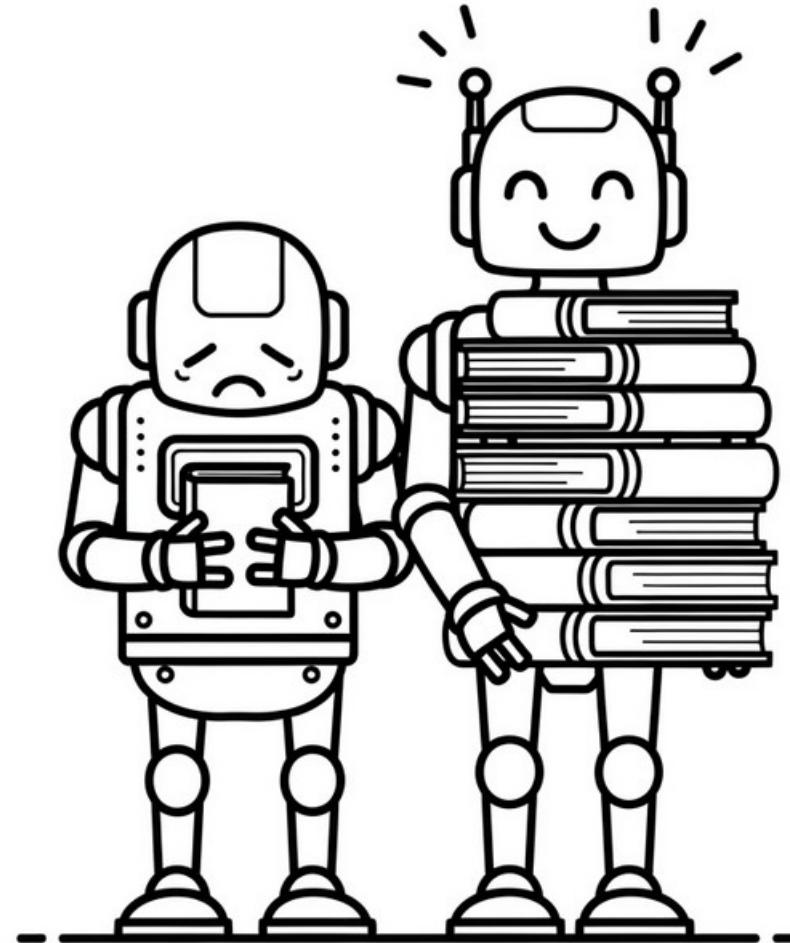
Abordando el problema de las limitaciones - Prompts

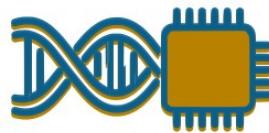
- **Malos prompts:**
 - H3K27me3 cáncer de próstata
 - ¿Cuál es el papel de H3K27me3 en cáncer de próstata?
- **Alternativa:**
 - Describe a nivel de un estudiante de doctorado de biología molecular, **¿Cuál es el efecto a nivel de la organización en tres dimensiones de la cromatina cuándo se ven alterados los niveles de H3K27me3 en biopsias de tejido tumoral de cáncer de próstata respecto al tejido sano.** Organiza la respuesta en una lista numerada.





Los LLMs dependen del contexto





Abordando el problema de las limitaciones - Prompts



Tutoriales UBMI IFC

Recursos para la comunidad



Aprende a crear un prompt

Para obtener los mejores resultados de una consulta a un modelo de lenguaje, es muy importante utilizar el *prompt* adecuado.

Un *prompt* es la instrucción que le damos al modelo de lenguaje para recibir un resultado, este puede ser una pregunta, una instrucción o simples palabras (como si usáramos un motor de búsqueda como Google), sin embargo, la calidad de los resultados, dependerá en gran medida de nuestra instrucción inicial.

En las siguientes líneas el modelo de lenguaje mas avanzado al momento de escribir este tutorial (Gemini 2.5 Pro), te explica lo que constituye "un buen prompt"

El contenido de la respuesta fué verificado.

Al final del documento puedes encontrar el prompt que se usó para generar esta información.



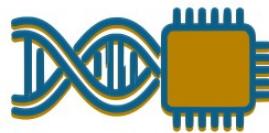
Prompt malo vs. prompt bueno ✨

Mal Ejemplo de Prompt: 🤖

"Háblame de la proteína Tau."

Buen Ejemplo de Prompt: 👍

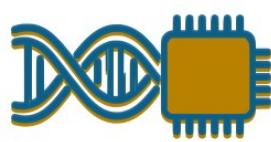




El *prompt* no lo es todo

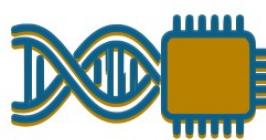
- “Hice un buen *prompt* pero la respuesta sigue siendo mala”



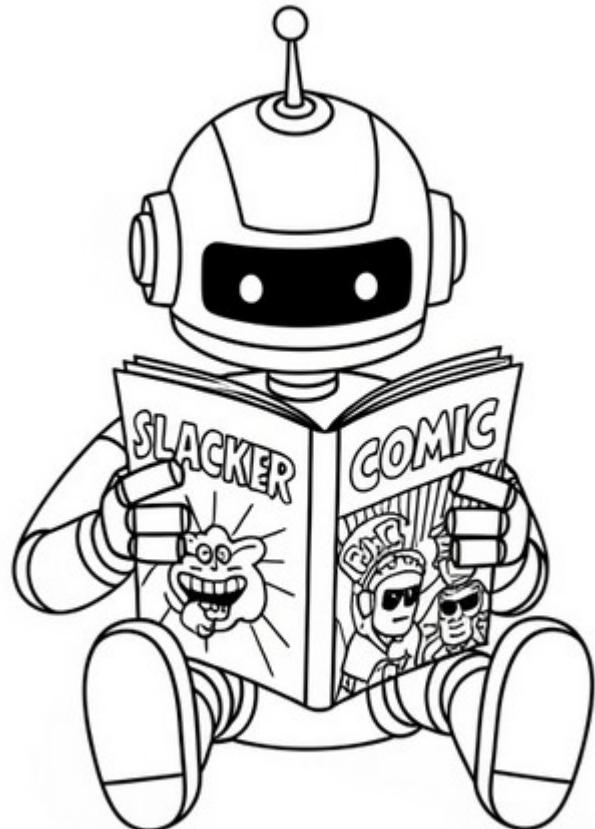


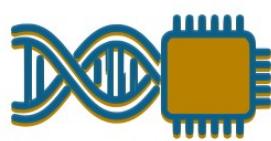
El prompt no lo es todo



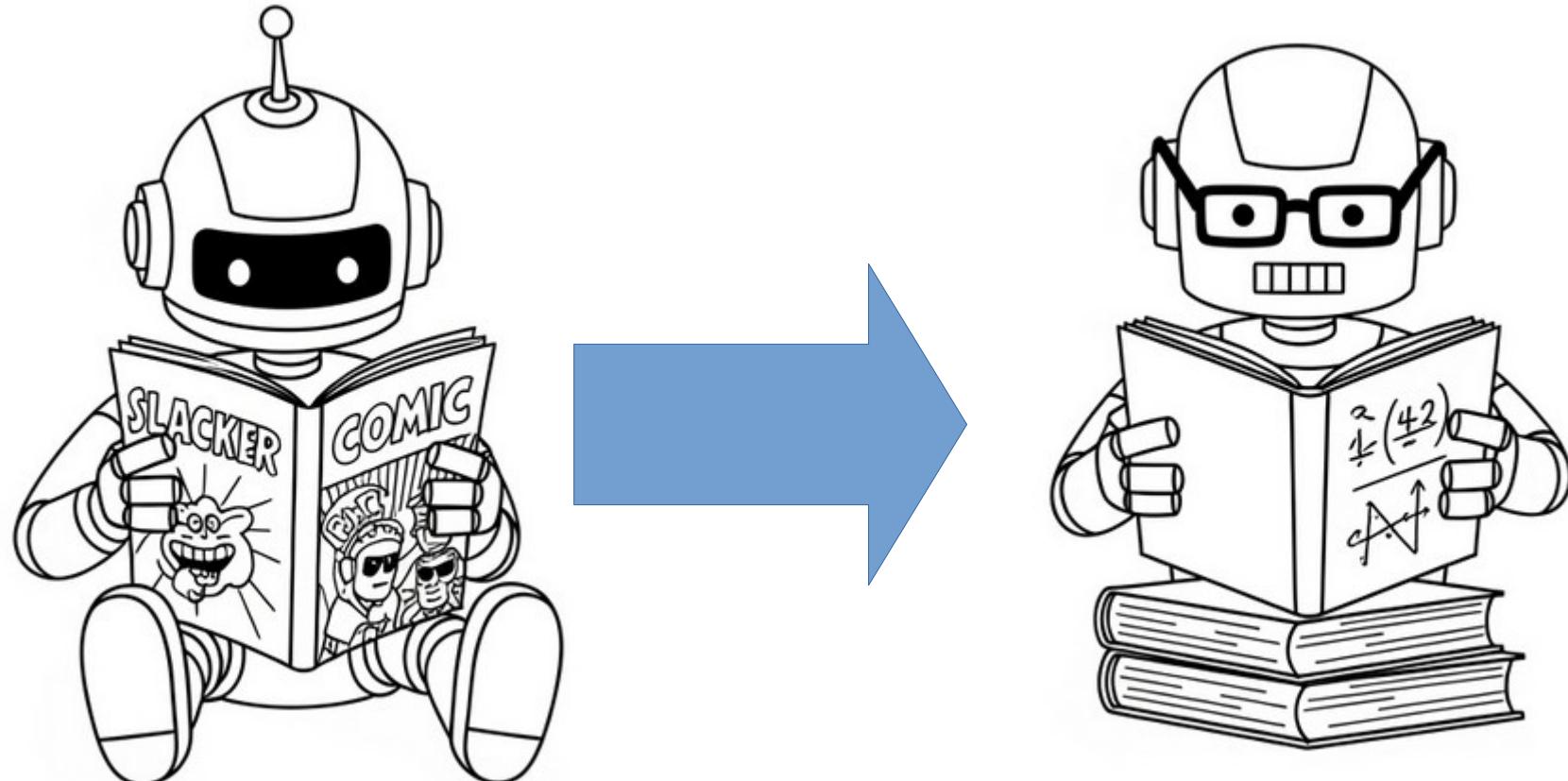


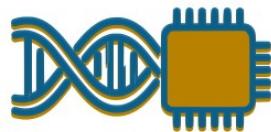
El problema viene “de origen”





El problema viene “de origen”





Añadiendo información externa – “Búsqueda web”

perplexity



Gemini

2.5 Pro (versión preliminar) ▾

ChatGPT ▾

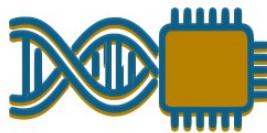
¿Con qué puedo ayudarte?

Pregunta lo que quieras

+ Herramientas

0 ↑

- Crea una imagen
- Busca en la web
- Escribe o programa
- Ejecuta la investigación a fondo Queda/n 5
- Piensa más tiempo



Búsqueda web

- **Prompt – en ChatGPT con búsqueda activada**

- Describe a nivel de un estudiante de doctorado de biología molecular, ¿Cuál es el efecto a nivel de la organización en tres dimensiones de la cromatina cuando se ven alterados los niveles de H3K27me3 en biopsias de tejido tumoral de cáncer de próstata respecto al tejido sano. Organiza la respuesta en una lista numerada.

nih.gov

Global levels of H3K27me3 track with differentiation in vivo and are deregulated by MYC in prostate cancer

ABSTRACT Cancer cells and stem cells share a number of biological characteristics including abundant amounts of...

nih.gov

Histone modifications and chromatin organization in prostate cancer - PMC

Increasing evidence suggests that alterations of histone modifications play important roles during prostate...

nih.gov

Chromatin and Epigenetic Dysregulation of Prostate Cancer Development, Progression, and Therapeutic...

For instance, a chromatin region marked by methylation in lysine 4 of the histone 3 tail (typically marked as H3K4me1..

nih.gov

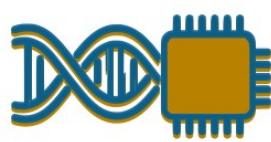
Global analysis of H3K27me3 as an epigenetic marker in prostate cancer progression

ABSTRACT Background: H3K27me3 histone marks shape the inhibition of gene transcription. In prostate cancer, the...

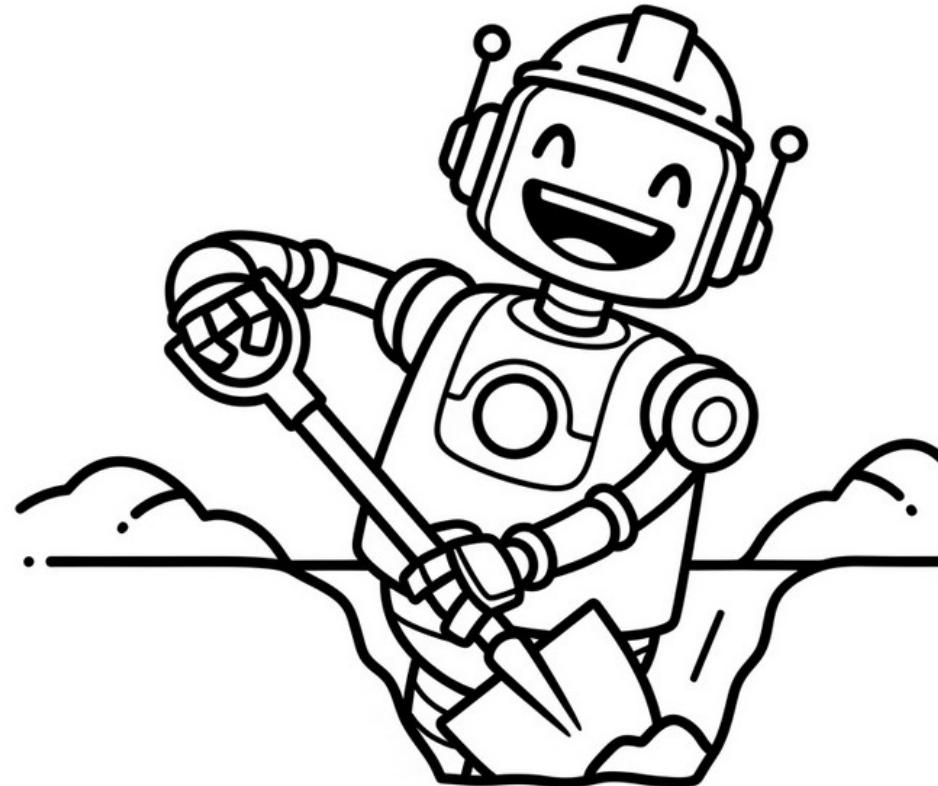
nih.gov

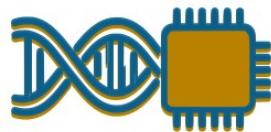
The association between Histone 3 Lysine 27 Trimethylation (H3K27me3) and prostate cancer:...

DISCUSSION The progression of prostate cancer, like other cancers, is facilitated by the epigenetic silencing of tumor-..



Búsqueda “profunda”





Búsqueda “profunda”

Gemini

2.5 Pro (versión preliminar) ▾

Describe a nivel de un estudiante de doctorado de biología molecular, ¿Cuál es el efecto a nivel de la organización en tres dimensiones de la cromatina cuando se ven alterados los niveles de H3K27me3 en biopsias de tejido tumoral de cáncer de próstata respecto al tejido sano.

ChatGPT ▾

¿Con qué puedo ayudarte?



Video

Deep Research

Canvas



Describe a nivel de un estudiante de doctorado de biología molecular, ¿Cuál es el efecto a nivel de la organización en tres dimensiones de la cromatina cuando se ven alterados los niveles de H3K27me3 en biopsias de tejido tumoral de cáncer de próstata respecto al tejido sano.



Herramientas



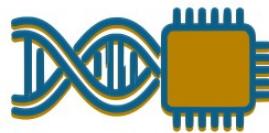
>Create an image

Search the web

Write or program

Execute research in the background
Queda/n 5

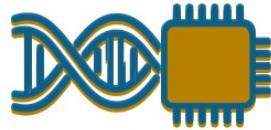
Think more time



Búsqueda profunda

- ***Deep Research – Gemini 2.5 Pro***
 - Reporte de 20 cuartillas
 - 46 referencias
 - La mayoría eran de artículos científicos

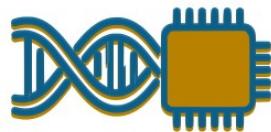




Ya tengo información de mejores fuentes . . . ¿y ahora qué?

- **VERIFICAR , VERIFICAR, VERIFICAR**
 - Todos los modelos son susceptibles a alucinaciones, incluso los mas avanzados.
 - Una respuesta coherente no implica una respuesta verdadera o precisa





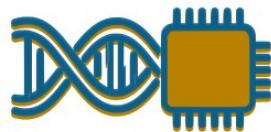
ChatGPT ayuda, pero el experto resuelve

```
#!/bin/bash

while true; do
    read -p "Press Enter to continue: " userInput

    # Check if the user input is NOT empty
    if [[ -n "$userInput" ]]; then
        echo "Skipping this iteration..."
        continue # Skip to the next iteration of the loop
    fi

    # --- Your code for the current iteration goes here ---
    echo "Running the main logic for this iteration..."
    # Example: perform some action
    sleep 1
    echo "Iteration complete."
    # -----
done
```



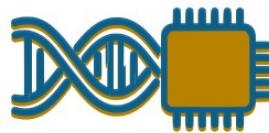
ChatGPT ayuda, pero el experto resuelve

```
#!/bin/bash

while true; do
    read -p "Press Enter to continue: " userInput < /dev/tty
    # Check if the user input is NOT empty
    if [[ -n "$userInput" ]]; then
        echo "Skipping this iteration..."
        continue # Skip to the next iteration of the loop
    fi

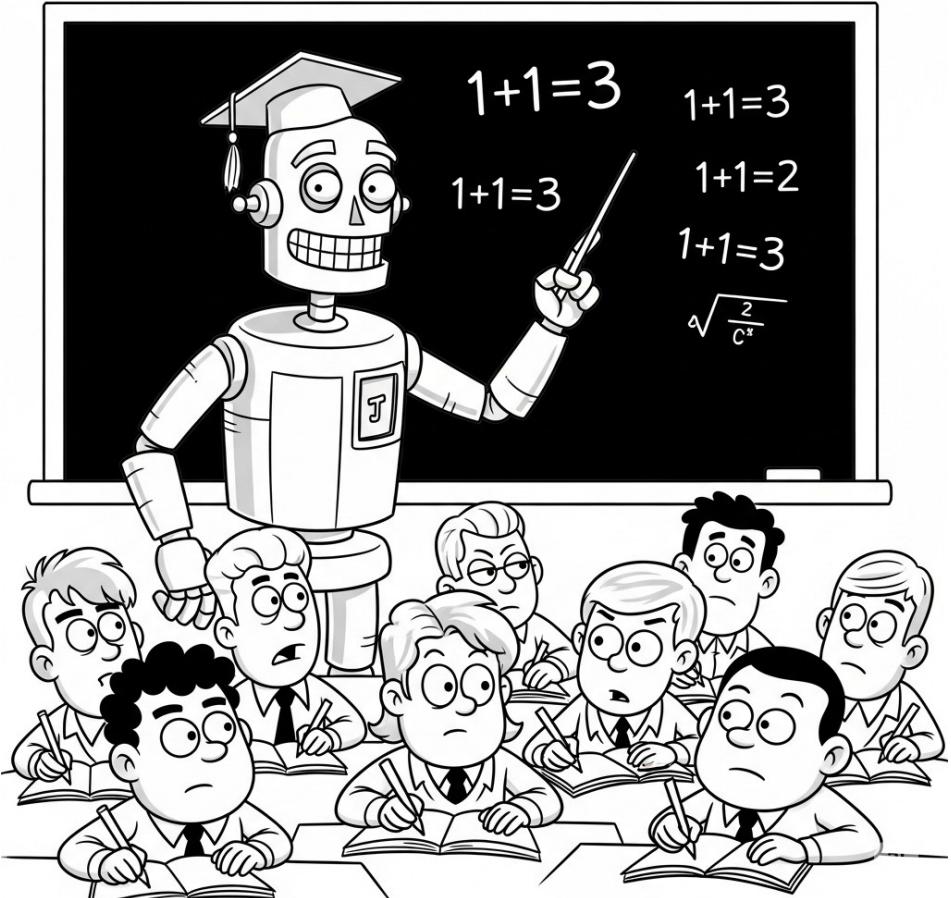
    # --- Your code for the current iteration goes here ---
    echo "Running the main logic for this iteration..."
    # Example: perform some action
    sleep 1
    echo "Iteration complete."
    # -----
done
```

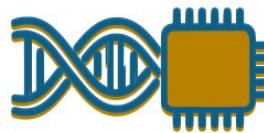




¿ Podemos aprender de los LLMs si no podemos confiar en ellos?

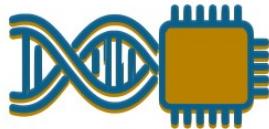
- **Los LLMs NO sustituyen al aprendizaje tradicional**
 - Demasiado riesgo de información imprecisa, falsa o superficial
 - “Me lo dijo ChatGPT”
- **Los LLMs benefician más al experto que al aprendíz**
 - Procurar preguntar cosas que podamos verificar





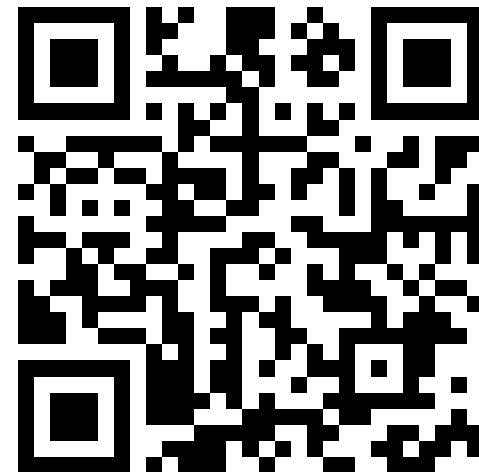
Oportunidades de uso de los chatbots en investigación biológica

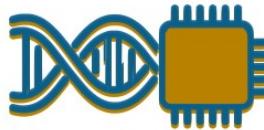
- 1) Consulta
- 2) Documentación
- 3) Asistencia de escritura
- 4) Generación de hipótesis



Los LLMs son una excelente forma de obtener artículos interesantes

- Herramientas enfocadas específicamente a obtener artículos a partir de **consultas en lenguaje natural (*Prompts*)**
 - **AI2 QA Scholar**
 - Elicit
 - Scispace
 - Concensus
 - Undermind





Los LLMs son una excelente forma de obtener artículos interesantes

❖ Ai2 Scholar QA

Beside nucleotide sequence what others levels of information does the dna molecule encode, such as thermodynamically , non-sequential motifs, k-mer abundance , local structures and others that i do not know

Are these answers helpful?

Download Report • Disclaimer

Background: DNA as a multi-level information carrier

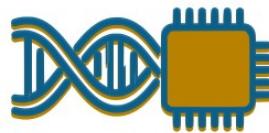
DNA encodes information beyond its primary nucleotide sequence, containing multiple layers of functional information embedded in its physical, structural, and thermodynamic properties. These additional information levels influence crucial biological processes through three-dimensional arrangements and physicochemical features that coexist with the primary genetic code. (10 sources)

Thermodynamic properties and energy landscape

DNA encodes rich thermodynamic information that governs its stability, flexibility, and conformational dynamics:

- **Melting and denaturation properties:** DNA sequences have specific melting temperatures and denaturation profiles that determine where and when DNA strands separate during processes like replication and transcription. These properties can be quantified through melting temperature measurements and vary across the genome, creating a "melting map" with important biological correlations ([Blake et al., 1998](#)) ([Liu et al., 2007](#)).
- **Base stacking energy:** The stacking interactions between adjacent base pairs contribute significantly to DNA stability and can be derived from quantum mechanical calculations of crystal structures. These energies vary depending on the specific dinucleotide sequence ([Gan et al., 2012](#))
- **Duplex stability parameters:** DNA sequences encode information about the thermodynamic stability of the double helix through nearest-neighbor interactions. These parameters allow prediction of the thermal behavior of any DNA duplex from its sequence, with applications ranging from probe design to understanding mutation effects ([Breslauer et al., 1986](#)) ([Sugimoto et al., 1996](#)).
- **Deformation energy and flexibility:** Different DNA sequences exhibit varying resistance to deformation along different axes. The energy

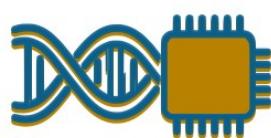




¿Qué hacer con una montaña de papers interesantes?

- ¡Leerlos!
- Que los lea un LLM
 - Uno por uno
 - De forma masiva





No le preguntas a ChatGPT, pregúntale al *paper*

Subir archivos

Añadir desde Drive

Importar código

+

Video Deep Research Canvas

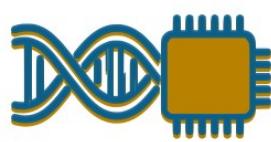
Large_lang...tentials_a

PDF PDF

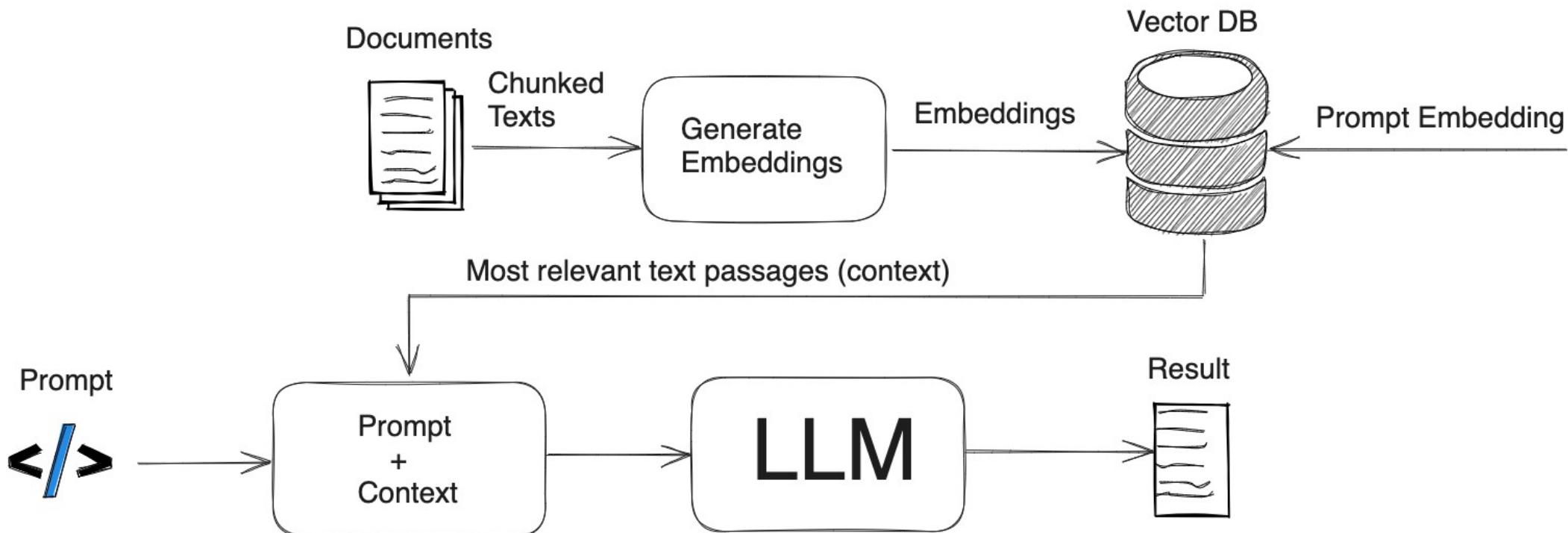
Enumera los hallazgos mas relevantes del artículo en una lista de viñetas, expícalos para alguien que no está familiarizado con machine learning

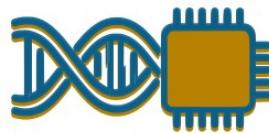
+

Video Deep Research Canvas



Retrieval Augmented Generation (RAG)





RAG en línea (fácil)

- Google notebooklm
 - Máximo 300 fuentes

Bidirectional Promoters, RAG demo

Analíticas Nuevo! Compartir p

Fuentes

+ Añadir Descubrir

Seleccionar todas las fuentes

10075708.txt
10075929.txt
10220171.txt
10337626.txt

Chat

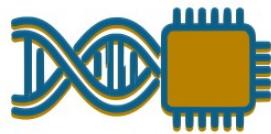
Actualizar

Qué características regulatorias de la expresión génica tienen los promotores bidireccionales?

Los promotores bidireccionales son regiones genéticas con **características reguladoras distintivas** que les permiten controlar la expresión de dos genes adyacentes y divergentes 1

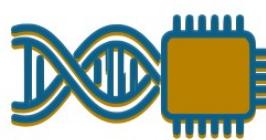
A continuación, se detallan sus principales características reguladoras:

- Organización Genómica Específica:
 - Los promotores bidireccionales se encuentran típicamente entre dos genes que están **dispuestos en una configuración "cabeza con cabeza"** (head-to-head), lo que significa que sus regiones codificantes se transcriben en direcciones opuestas desde el promotor intergénico común 1
 - Los sitios de inicio de la transcripción de estos genes divergentes suelen estar **muy próximos**, a menudo separados



RAG local (difícil)

- Múltiples estrategias desde 100% código hasta 30% código
- Curva de aprendizaje
- Control total
- El límite es tu imaginación ... y presupuesto



Ejemplo de una máquina de RAG local

Hem(A.I.)tólogo

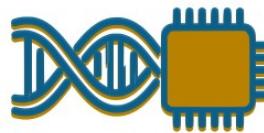
¿Cómo puedo ayudarte hoy?

+ | Búsqueda Web

Microphone icon | User profile icon

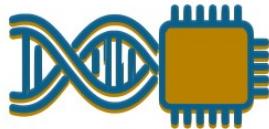


Buscar Colección	
rich_immunology.txt	7.0 MB
turgeon_hematology.txt	3.0 MB
davidson_medicine_24.txt	7.3 MB
williams_hemato_eng.txt	16.6 MB
gartner_histology_english.txt	1.1 MB
rodak_5th.txt	4.3 MB
harrison_21_clean.txt	27.5 MB
Practical Transfusion...	12.3 MB
junqueira_histology_eng.txt	1.7 MB
ellis_anatomy_eng.txt	880.5 KB
carlson_embriology_eng.txt	1.6 MB
guyton_eng.txt	4.7 MB
mollisons-blood-...	17.6 MB
harmening_hemato_hemost...	3.3 MB
The Developing Human...	20.5 MB
Human Embryology and...	22.3 MB



Oportunidades de uso de los chatbots en investigación biológica

- 1) Consulta
- 2) Documentación
- 3) Asistencia de escritura
- 4) Generación de hipótesis



Hacer mas entendibles los textos que escribimos . . . ¡a veces!

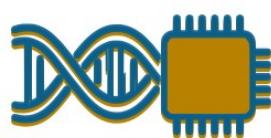
- Copy/Paste en chatbots públicos
- \$ervicios dedicado\$ o adicionale\$
 - Grammarly
 - Gemini → Google docs

La vía de señalización del AR es fundamental en el cáncer de próstata, y no solo se ve afectada pasivamente por los cambios en H3K27me3, sino que es modulada activa y bidireccionalmente por EZH2. Esto ocurre tanto a través de la represión canónica de sus reguladores negativos como mediante la coactivación no canónica del propio AR y sus dianas, a menudo involucrando configuraciones específicas de la cromatina 3D.

- **Regulación Directa de la Expresión de AR:** EZH2 puede transcripción del gen AR a través de mecanismos independientes de la actividad metiltransferasa, uniéndose al promotor de AR (es decir, por H3K27ac, no por H3K27me3).³² Esto crea un bucle de retroalimentación que potencia la señalización de AR.
- **Modulación de la Expresión de Genes Diana de AR:** EZH2 puede potenciar la expresión de genes inducidos por AR, como FKBP5.32 Las interacciones de cromatina asociadas a AR pueden regular sus genes diana.³⁵ EZH2, tanto a través de la depuración de H3K27me3 (afectando la accesibilidad de la cromatina para la transcripción) como

Redacta esta sección de forma estructurada en forma de viñetas

- ↻+ Reformular ←
- ✗ Acortar
- ☰ Desarrollar
- 📁 Más formal



nature machine intelligence

[Explore content](#) ▾ [About the journal](#) ▾ [Publish with us](#) ▾

[nature](#) > [nature machine intelligence](#) > [comment](#) > [article](#)

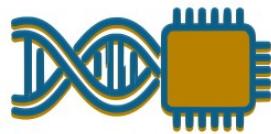
Comment | Published: 13 November 2024

Guidelines for ethical use and acknowledgement of large language models in academic writing

[Sebastian Porsdam Mann](#)✉, [Anuraag A. Vazirani](#)✉, [Mateo Aboy](#), [Brian D. Earp](#), [Timo Minssen](#), [I. Glenn Cohen](#) & [Julian Savulescu](#)

[Nature Machine Intelligence](#) 6, 1272–1274 (2024) | [Cite this article](#)

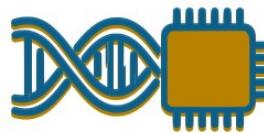
3854 Accesses | 10 Citations | 67 Altmetric | [Metrics](#)



Políticas de uso justo

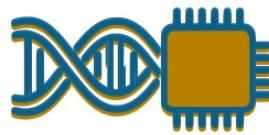
- Papers deben ser escritos por humanos
 - LLMs pueden usarse para, **mejorar claridad, traducción y gramática.**
 - **Encontrar literatura**
 - **Escribir documentación y código (Humano responsable)**
- Revisores deben verificar los papers ellos mismos
 - LLMs pueden usarse para encontrar **inconsistencias lógicas**





Oportunidades de uso de los chatbots en investigación biológica

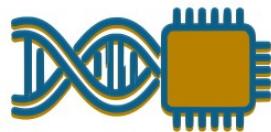
- 1) Consulta
- 2) Documentación
- 3) Asistencia de escritura
- 4) Generación de hipótesis



Generación de Hipótesis

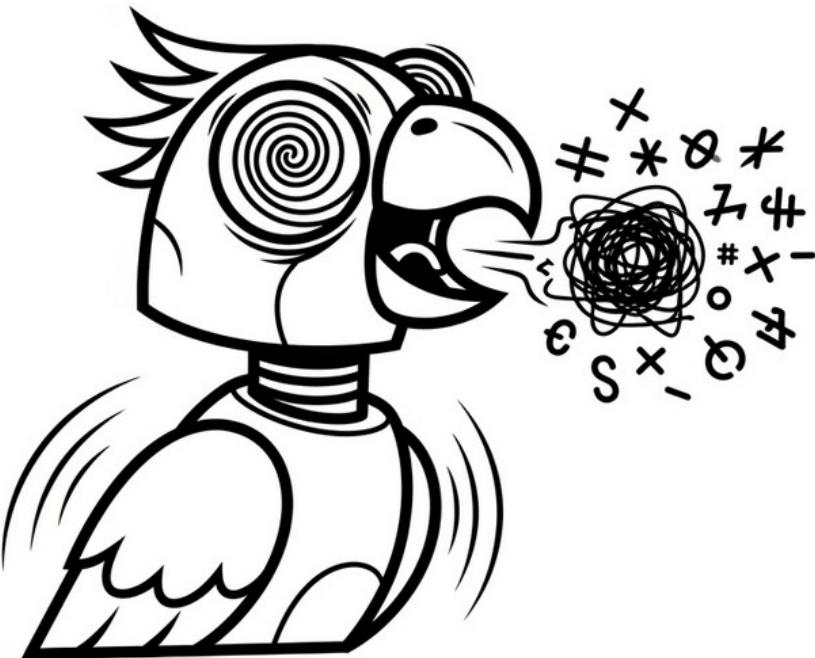
- Todavía no :D





Generación de Hipótesis

- Todavía no :D



Machine Learning Research

Overview Research Highlights Publications Events Work with us

Speech and Natural Language
Processing

Paper | June 2025

The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity

Parshin Shojaee*†, Iman Mirzadeh*, Keivan Alizadeh, Maxwell Horton, Samy Bengio,
Mehrdad Farajtabar



¡Gracias!

