

# Diplomado: "IA Generativa, Ingeniería de Prompts y Agentes Inteligentes"

Octubre 3 de 2025

José Antonio Sánchez y Xabier Fabián Roldán

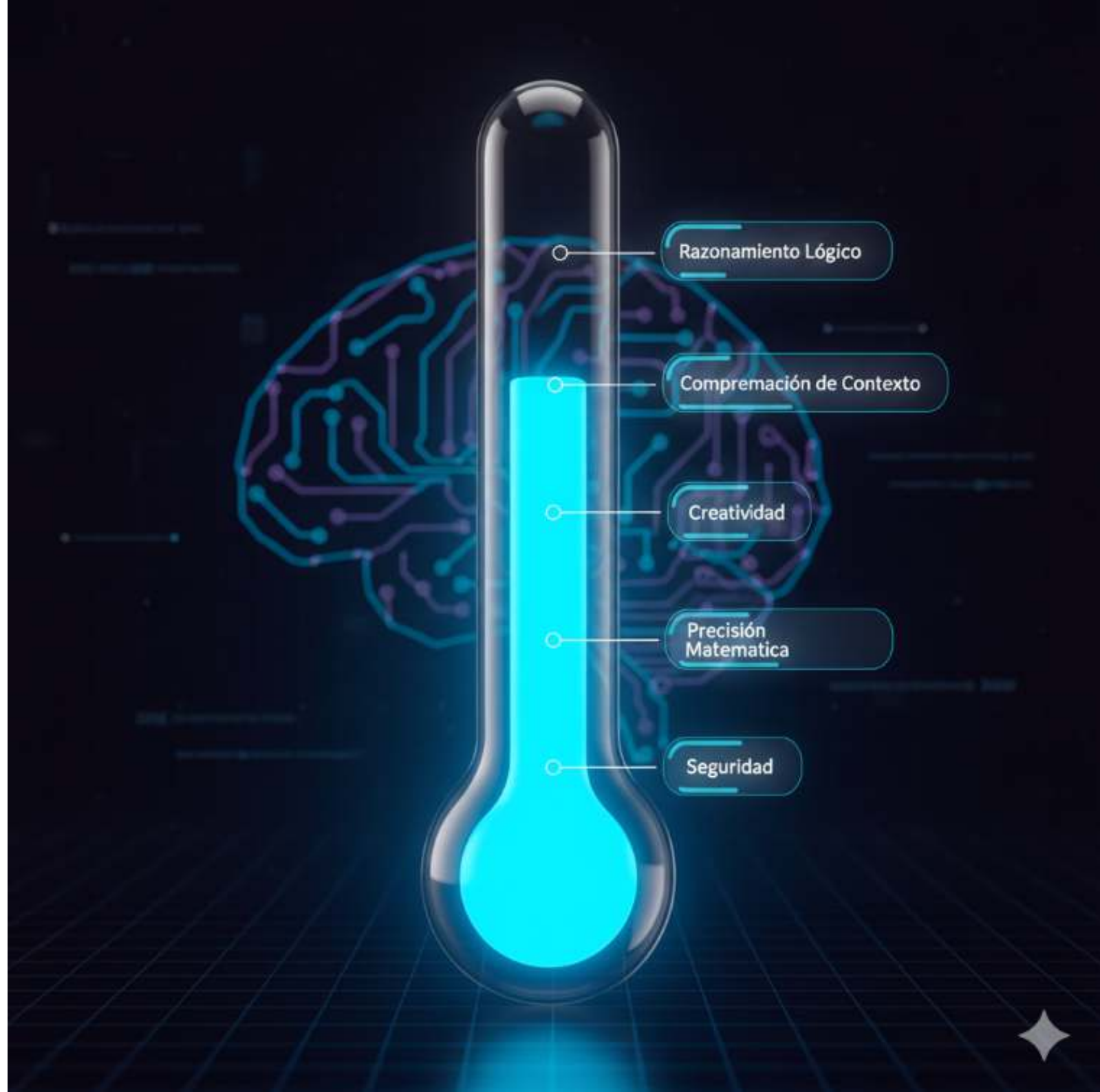


Universidad de  
**La Sabana**

VIGILADA MINEDUCACIÓN

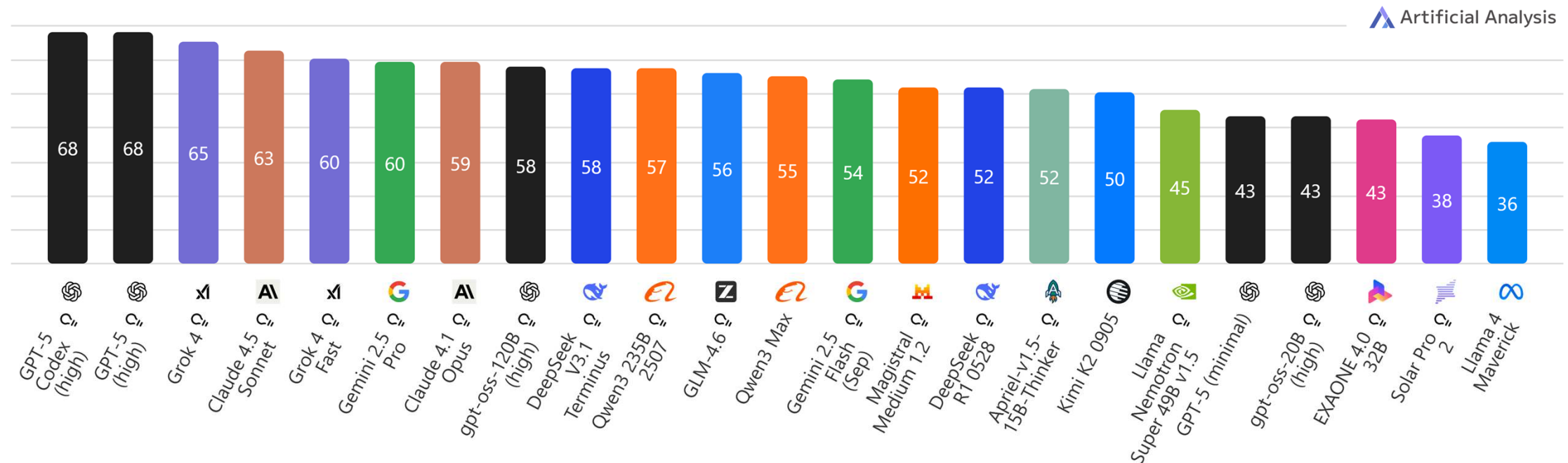
Inteligencia Artificial

# El termómetro de la inteligencia artificial: *Métricas para LLMs*



# Artificial Analysis Intelligence Index

Artificial Analysis Intelligence Index v3.0 incorporates 10 evaluations: MMLU-Pro, GPQA Diamond, Humanity's Last Exam, LiveCodeBench, SciCode, AIME 2025, IFBench, AA-LCR, Terminal-Bench Hard,  $\tau^2$ -Bench Telecom



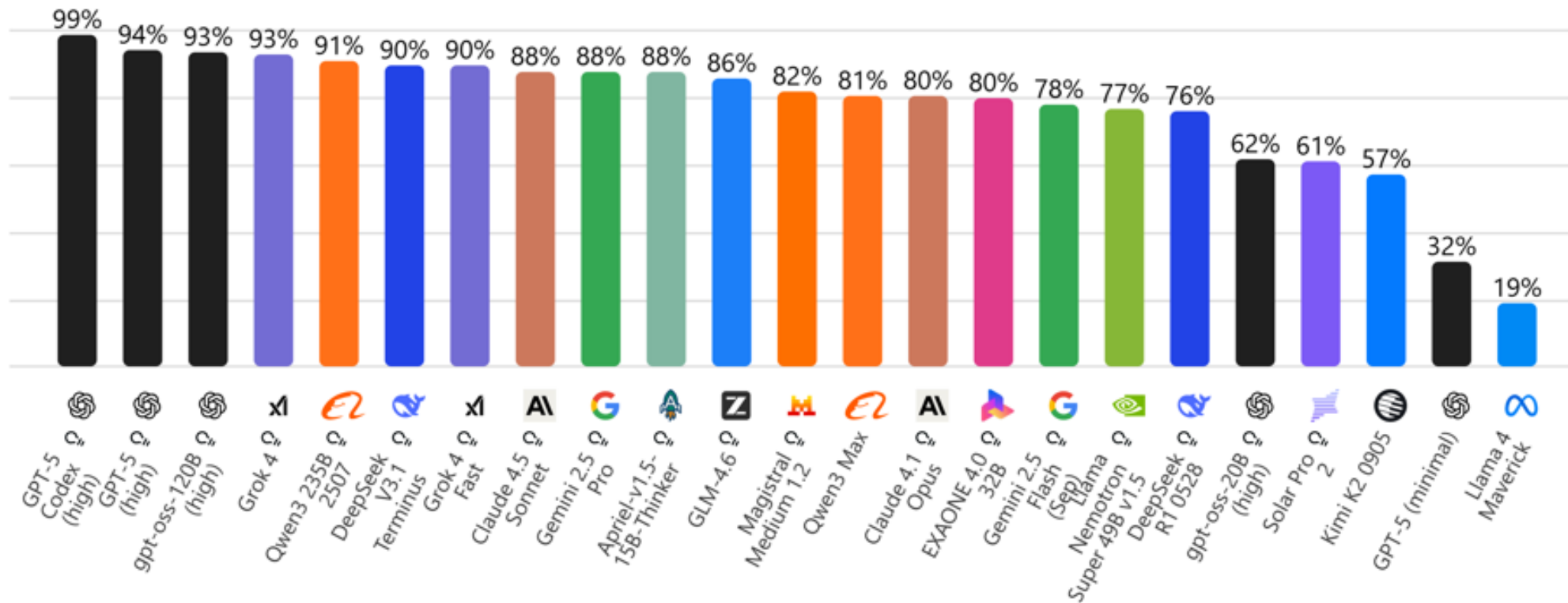
<https://artificialanalysis.ai/models> 03/10/2025

El *Índice de Inteligencia de Artificial Analysis* es una métrica combinada que abarca múltiples dimensiones de la inteligencia de los modelos de IA.

- **MMLU-Pro:** Evalúa comprensión general y razonamiento en muchos temas.
- **GPQA Diamond:** Mide razonamiento en preguntas científicas de nivel avanzado (PhD).
- **Humanity's Last Exam:** Prueba conocimiento y razonamiento experto en diversas áreas.
- **LiveCodeBench:** Evalúa la capacidad de generar y manejar código de programación.
- **SciCode:** Mide la habilidad para crear código que resuelva problemas científicos.
- **AIME:** Evalúa habilidades en matemáticas avanzadas de competición.
- **MATH-500:** Prueba la capacidad para resolver problemas matemáticos complejos.

## AIME 2025 (Competition Math)

Artificial Analysis



<https://artificialanalysis.ai/models> 03/10/2025

El *Índice Matemático de Artificial Analysis (AIME)* es una medida específica dentro del Índice de Inteligencia general.

Representa el **promedio** del desempeño de los modelos de IA en las evaluaciones centradas específicamente en matemáticas. Esencialmente, te da una idea rápida de cuán “*buenas*” son las IAs resolviendo problemas matemáticos complejos.

- **AIME:** Evalúa habilidades en matemáticas avanzadas de competición.
- **MATH-500:** Prueba la capacidad para resolver problemas matemáticos complejos.

## *Actividad - Arena de Prompts: Descubriendo tu IA para cada tarea*

### Instrucciones

**Elige tu Arena:** Cada participante escoge un escenario.

**Experimenta:** Usa el prompt para tu escenario en **ChatGPT, Claude y Gemini**.






**Evalúa:** Evalúa según tu criterio quien realizado mejor la tarea.

**Vota:** Llena la encuesta.

### Posibles temas




























- ✓ La explicación de un artículo de la constitución.
- ✓ Un asistente creativo que construye eslóganes para una campaña publicitaria.
- ✓ Una analogía brillante para explicar el Overfitting.
- ✓ Un consejo “muy malo” para aprobar un examen, dicho como si fuera bueno.
- ✓ El chiste más gracioso sobre estudiantes y exámenes.

**SLMs vs. LLMs: Eficiencia Especializada o Capacidad General.** Una decisión crítica es apostar por **Small Language Models (SLMs)**, que son eficientes, privados y especializados, o por **Large Language Models (LLMs)**, que ofrecen capacidades generales avanzadas, pero con mayor costo y complejidad.

Criterio	Small Language Models (SLMs)	Large Language Models (LLMs)
 Razonamiento	Limitado, pero excelente en dominios específicos.	Excelente, ideal para tareas complejas y abstractas.
 Costo y Ejecución	Bajo. Ideal para ejecución local en hardware convencional.	Alto. Requiere nube o hardware muy caro (GPUs A100/H100).
 Velocidad (Latencia)	Depende de la infraestructura.	Moderada (>500ms). Adecuada para tareas analíticas sin urgencia.
 Privacidad	Excelente. Al ejecutarse localmente, los datos nunca salen de la organización.	Riesgos inherentes. El uso de APIs implica enviar datos a servidores de terceros.
 Caso de Uso Ideal	Tareas de alto volumen, predecibles y especializada.	Tareas creativas, de síntesis y razonamiento complejo.



SLMs, Small Open-Source  
Language Models  
*Menos de 3B de parámetros*

Model Name	Size	Score	VRAM (GB)
 Gemma 3 270M	270M	0.5	0.5
 Embeddinggemma 300M	300M	0.49	1.2
 Qwen3 Embedding 0.6B	1B	0.47	1.2
 SmolLM3 3B	3B	0.46	6.2
 Llama 3.2 3B Instruct 	3B	0.46	6.5
 Vaultgemma 1B	1B	0.46	2.1
 Gemma 3 270M It	270M	0.46	0.5
 Voxtral Mini 3B 2507	3B	0.45	9.4
 DeepSeek R1 Distill Qwen 1.5B	2B	0.44	3.5
 Qwen2.5 3B Instruct 	3B	0.43	6.2
 SmolLM2 1.7B Instruct  	2B	0.43	3.4
 Llama 3.2 1B Instruct 	1B	0.43	2.5
 Gemma 3 1B It	1B	0.42	2
 Llama 3.2 1B	1B	0.42	2.5
 Qwen2.5 1.5B Instruct 	2B	0.41	3.1
 Hunyuan 1.8B Instruct 	2B	0.41	3.6
 Granite 3.1 2B Instruct  	2B	0.4	5.1
 Nemotron Research Reasoning Qwen 1.5B	2B	0.4	7.1

SLMs, Small Open-Source  
Language Models  
*7B de parámetros*

Model Name	Size	Score	VRAM (GB)
 Gemma 3 270M	270M	0.5	0.5
 Hunyuan MT 7B	7B	0.49	16.1
 Qwen3 4B Instruct 2507 	4B	0.49	8.1
 Embeddinggemma 300M	300M	0.49	1.2
 Mistral 7B Instruct V0.2   	7B	0.48	14.4
 Qwen2.5 7B Instruct 	7B	0.48	15.4
 Phi 3 Mini 4K Instruct  	4B	0.48	7.7
 Qwen3 4B Thinking 2507	4B	0.48	8.1
 Qwen3 Embedding 0.6B	1B	0.47	1.2
 Phi 4 Mini Instruct 	4B	0.46	7.7
 SmoLLM3 3B	3B	0.46	6.2
 Llama 3.2 3B Instruct 	3B	0.46	6.5
 Vaultgemma 1B	1B	0.46	2.1
 Gemma 3 270M It	270M	0.46	0.5
 FastVLM 7B	7B	0.45	15.5
 Janus Pro 7B	7B	0.45	14.8
 Voxtral Mini 3B 2507	3B	0.45	9.4
 Qwen2.5 Omni 7B	7B	0.44	22.4



# Tokens y Palabras: La Arquitectura Oculta de los LLMs

Los modelos de lenguaje a gran escala (LLMs) procesan textos dividiéndolos en **tokens**, que son *fragmentos lingüísticos*—pueden ser palabras completas, partes de palabras o incluso signos de puntuación.

- En inglés, se estima que **1 token representa aproximadamente 0,75 palabras** (*o sea, 100 palabras equivalen cerca de 130 tokens*).
- En otros idiomas, la densidad de tokens por palabra puede aumentar. Por ejemplo, en **español** o **francés** se suele necesitar entre **2 y 2,1 tokens por palabra**; en idiomas como **ruso**, la cifra sube a 3,3; e incluso en **hindi** puede llegar a 6 **tokens por palabra**.
- Los *tokenizadores* entrenados principalmente con datos en inglés tienden a ser menos eficientes en idiomas con estructuras lingüísticas distintas, provocando que algunas palabras complejas requieran muchos más tokens que en inglés.



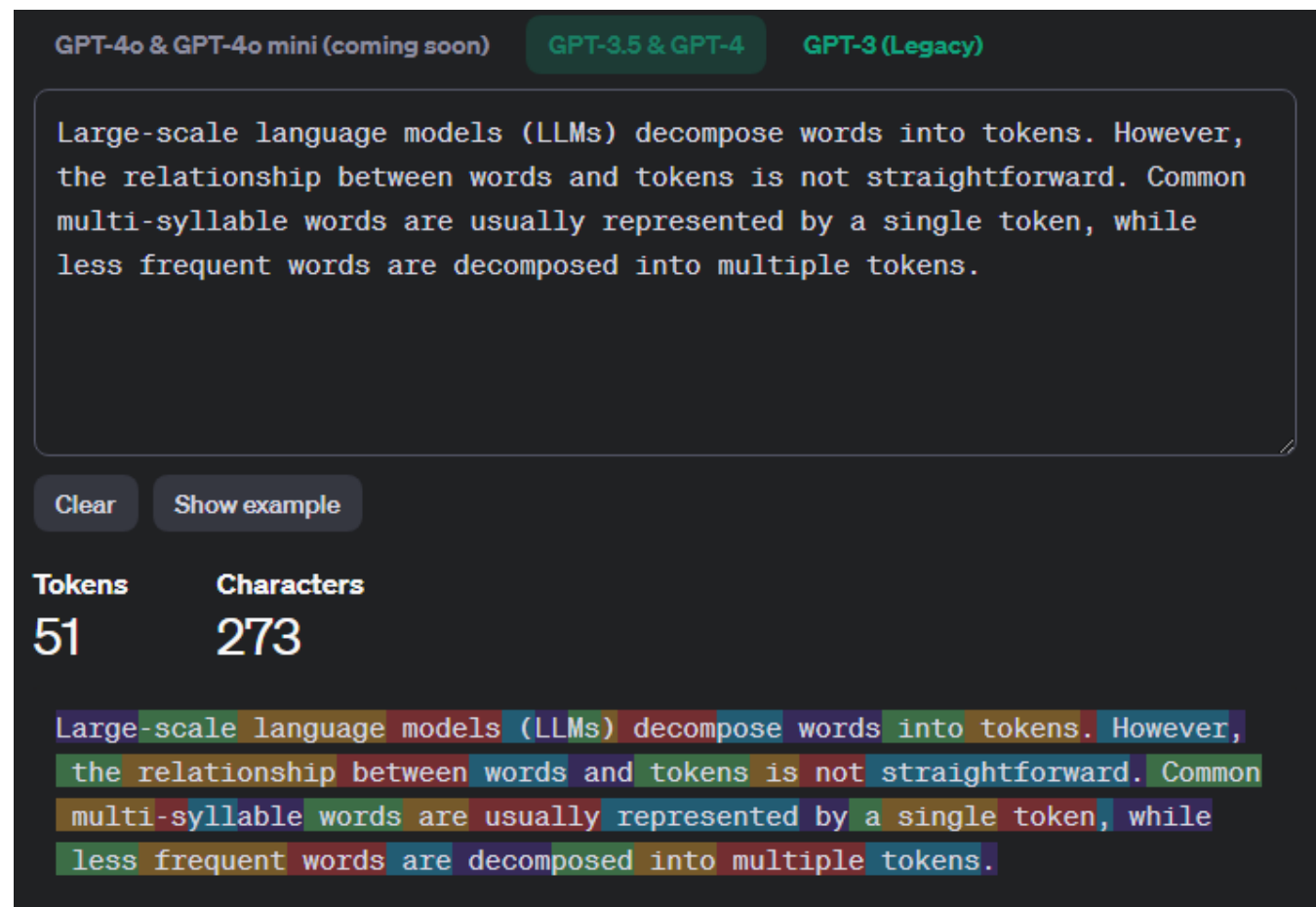
“Los *tokens* son esencialmente los bloques de construcción del lenguaje escrito... la tarea fundamental de estos LLMs es usar probabilidad para predecir el siguiente *token* en función de los anteriores. Por ello, se espera que el rendimiento dependa de la calidad del *prompt*. Normalmente, cuanto más detallado y explícito es el *prompt*, más precisa es la respuesta.”

# Tokens y Palabras: La Arquitectura Oculta de los LLMs

Si deseas comprobar cuántos tokens utiliza tu texto, puedes usar herramientas como el tokenizador web de OpenAI.

<https://platform.openai.com/tokenizer>

Nota: Ten en cuenta que el número exacto de tokens depende del modelo y del método de tokenización empleado, por lo que los resultados pueden variar entre diferentes modelos.



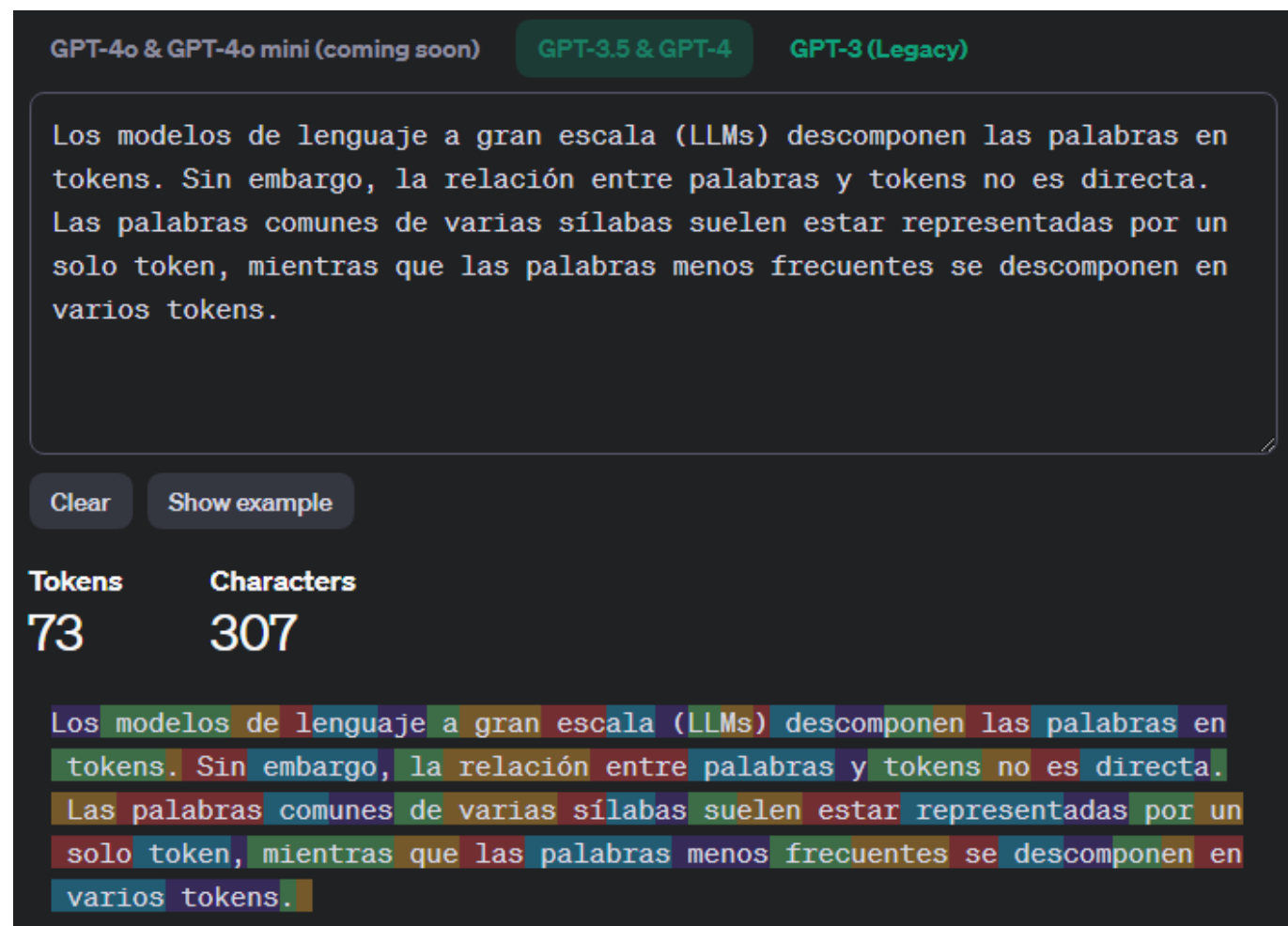
**Ingles:** 39 palabras,  $51/48=1.31$

# Tokens y Palabras: La Arquitectura Oculta de los LLMs

Si deseas comprobar cuántos tokens utiliza tu texto, puedes usar herramientas como el tokenizador web de OpenAI.

<https://platform.openai.com/tokenizer>

Nota: Ten en cuenta que el número exacto de tokens depende del modelo y del método de tokenización empleado, por lo que los resultados pueden variar entre diferentes modelos.



The screenshot shows the OpenAI Tokenizer interface. At the top, there are tabs for 'GPT-4o & GPT-4o mini (coming soon)', 'GPT-3.5 & GPT-4' (selected), and 'GPT-3 (Legacy)'. The main text area contains the Spanish text: 'Los modelos de lenguaje a gran escala (LLMs) descomponen las palabras en tokens. Sin embargo, la relación entre palabras y tokens no es directa. Las palabras comunes de varias sílabas suelen estar representadas por un solo token, mientras que las palabras menos frecuentes se descomponen en varios tokens.' Below the text area are 'Clear' and 'Show example' buttons. The results section shows 'Tokens' as 73 and 'Characters' as 307. At the bottom, the text is displayed with individual tokens highlighted in different colors.

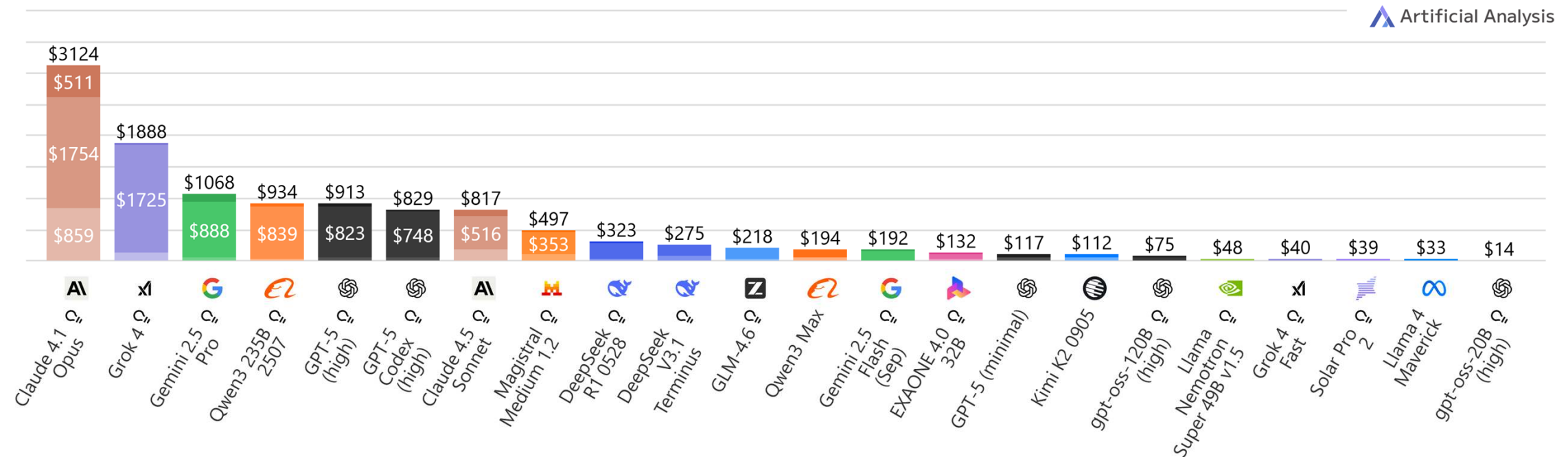
Tokens	Characters
73	307

**Español:** 48 palabras,  $73/48=1.52$

# Cost to Run Artificial Analysis Intelligence Index

Cost (USD) to run all evaluations in the Artificial Analysis Intelligence Index

Input Cost Output Cost Reasoning Cost



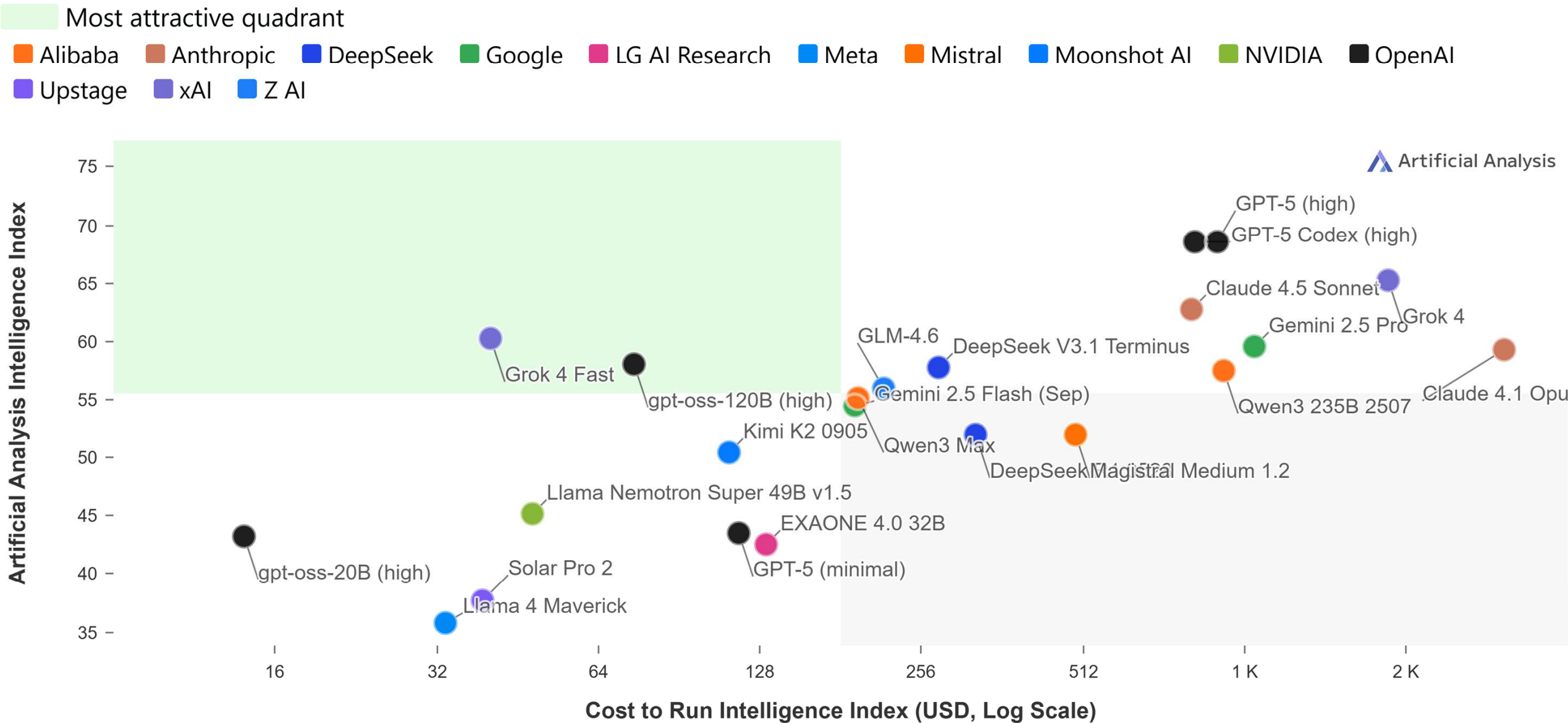
<https://artificialanalysis.ai/models> 03/10/2025

Índice de costo de ejecución de inteligencia de análisis artificial

- El costo de ejecutar las evaluaciones en el Índice de inteligencia de análisis artificial, calculado utilizando los precios de los **tokens** de entrada y salida del modelo y la cantidad de tokens utilizados en las evaluaciones.

# Intelligence vs. Cost to Run Artificial Analysis Intelligence Index

Artificial Analysis Intelligence Index; Cost to Run Intelligence Index





# Ventana de contexto

Inteligencia Artificial

Solo una oración nítida.

# La Dificultad de Mantener el Contexto

## La IA y la Pérdida de Memoria



Los modelos de lenguaje pueden tener dificultades para *recordar* información a lo largo de una conversación.

Esto puede llevar a respuestas *irrelevantes o contradictorias*.

La longitud de la conversación y la complejidad del tema pueden afectar la capacidad de la IA para **mantener el contexto**.

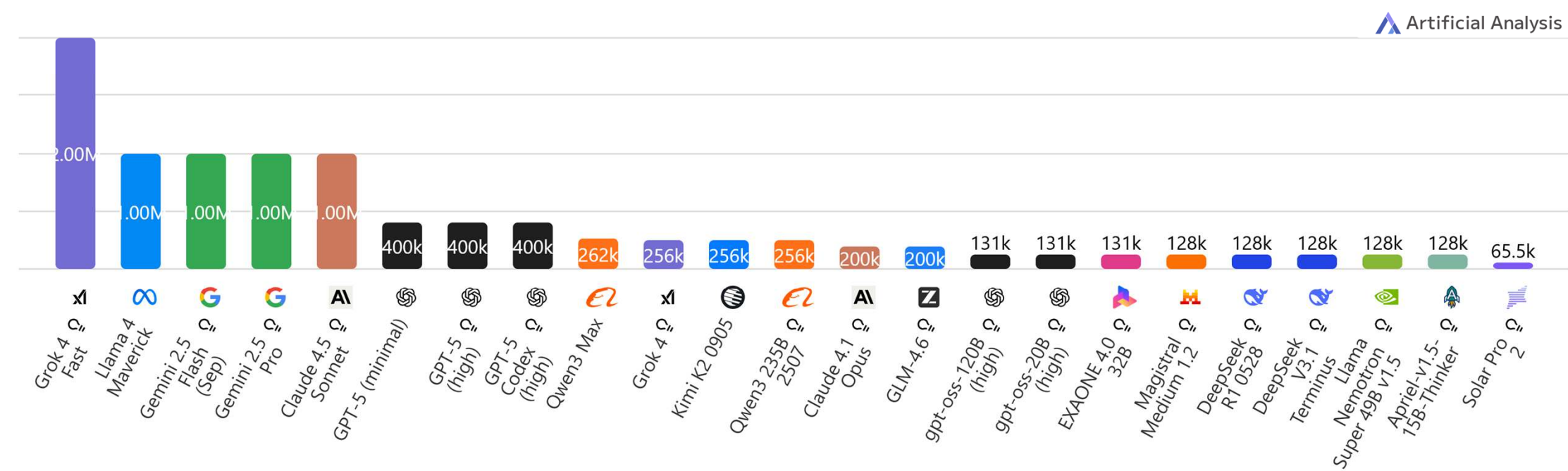
## ¿Qué es la Ventana de Contexto? 🤔

La **ventana de contexto** de un LLM es, en esencia, su **memoria de trabajo a corto plazo**. Es la cantidad máxima de información, medida en **tokens**, que el modelo puede procesar y “recordar” en un único momento.

*Imagina que estás leyendo un libro. Tu ventana de contexto sería la cantidad de páginas que puedes recordar con detalle para entender el capítulo actual. Si la ventana es pequeña, olvidarás rápidamente quiénes eran los personajes del inicio del libro. Si es grande, podrás mantener la coherencia de la trama durante mucho más tiempo.*

# Context Window

Context Window: Tokens Limit; Higher is better



<https://artificialanalysis.ai/models> 03/10/2025

La **Ventana de Contexto** de un modelo de IA es como su “*memoria*” a corto plazo.

Define la **cantidad total de texto** (tokens) que el modelo puede “ver” o considerar a la vez. Esto incluye tanto el texto que tú le das (entrada) como el texto que el modelo genera como respuesta (salida).

(**Nota:** A menudo, hay un límite más bajo para la cantidad de texto que el modelo puede *generar* como respuesta, incluso si la ventana total es grande).

## ¿La ventana de contexto es lo mismo que los tokens de mi pregunta?

**No**, la cantidad de tokens que puedes ingresar en una pregunta (tu *prompt*) es solo una parte de lo que ocupa la ventana de contexto. La ventana de contexto total debe incluir **todo** lo que el modelo necesita para generar una respuesta:

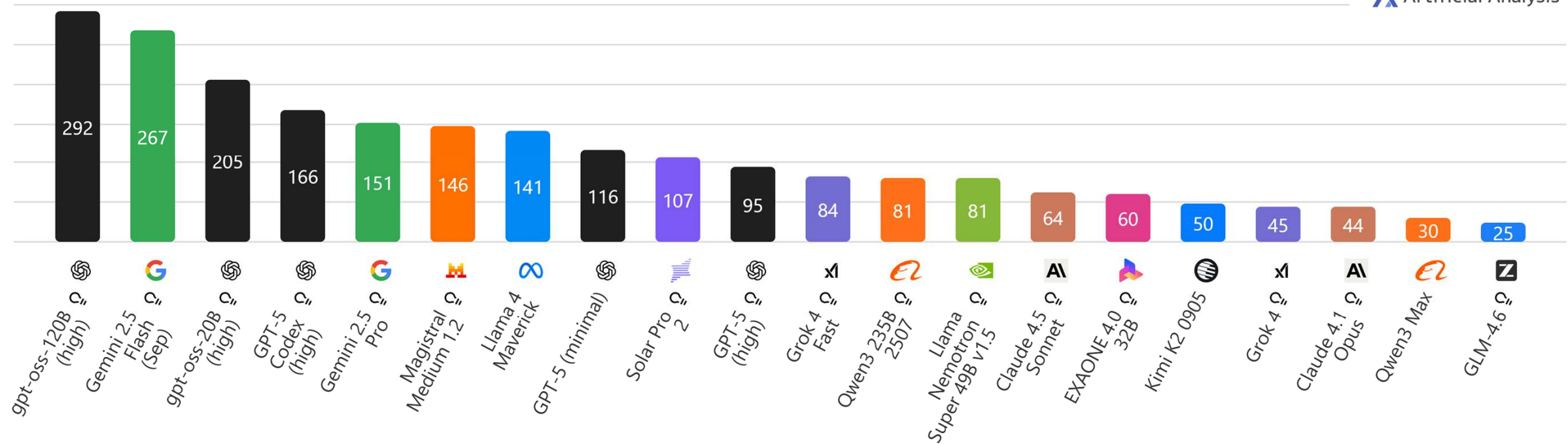
- **Instrucciones del sistema:** Directrices ocultas que definen el comportamiento del modelo.
- **Historial de la conversación:** Todas tus preguntas anteriores y todas las respuestas del modelo en la sesión actual.
- **Tu nueva pregunta (prompt):** El texto que acabas de escribir.
- **La respuesta generada (output):** El modelo también necesita espacio en la ventana para generar su propia respuesta.

**Por lo tanto, si un modelo tiene una ventana de 128,000 tokens, no significa que puedas hacer una pregunta de 128,000 tokens. Significa que la suma de todo el historial, tu pregunta y la respuesta del modelo no debe superar ese límite.**

## Output Speed

Output Tokens per Second; Higher is better

Artificial Analysis



<https://artificialanalysis.ai/models> 03/10/2025

La **Velocidad de Salida** mide qué tan rápido un modelo de IA genera texto una vez que empieza a responder.

- Se mide en “**tokens por segundo**”: cuantos fragmentos de texto (palabras, partes de palabras o caracteres) produce la IA por segundo. Es, básicamente, una medida de la fluidez o rapidez con la que la IA escribe su respuesta después de pensar.



## Consejo: Pide Resúmenes para No Perder el Hilo con la IA

En conversaciones largas con modelos de lenguaje, **pide resúmenes periódicos** para mantener el foco y evitar que la IA olvide información clave.

¿Por qué funciona?: Los modelos de IA no recuerdan todo lo que se ha dicho indefinidamente. A medida que la charla se extiende, pueden **perder acceso a partes anteriores** de la conversación.

Al solicitar un resumen, fuerzas al modelo a:



Revisar y condensar la información clave.



Reintroducir ese contenido en su “memoria activa”.



Seguir una línea argumental coherente y precisa.



### Prompts útiles

- “Resume brevemente lo que hemos discutido hasta ahora.”
- “Haz un resumen antes de continuar, para que mantengas el contexto.”

---

Contexto: a veces menos, es más:  
*El arte de dar el contexto justo*

*“Un nivel competente de Ingeniería de Prompts requiere medir cuidadosamente el contexto necesario y suficiente... no solo ampliarlo, sino **reducirlo** en algunos casos.”*

Los modelos de lenguaje (LLM) tienen una capacidad de atención limitada. Al proporcionarles un contexto recargado con detalles irrelevantes, su atención se diluye, lo que puede llevar a respuestas incorrectas o de baja calidad.



A veces,  
menos.

Contexto: a veces menos, es más: *El arte de dar el contexto justo*

**Caso Ilustrativo:** Imagina un problema de física:

- **Versión 1 (recargada):** Una narrativa larga y colorida sobre Tarzán balanceándose en la selva, con detalles sobre los monos aulladores y el clima.
- **Versión 2 (limpia):** Un enunciado directo con los datos necesarios: la masa de Tarzán, la longitud de la liana y el ángulo de su oscilación.

**Resultado:** La segunda versión, al eliminar el “*ruido*” narrativo, permite que la IA se concentre únicamente en los datos relevantes para resolver el problema, aumentando la precisión de la respuesta.

**Principio Clave: Contexto Óptimo:** El objetivo es darle a la IA solo la información que necesita para la tarea. Esto enfoca su capacidad de razonamiento en lo relevante -“**Prompt Compression**”.

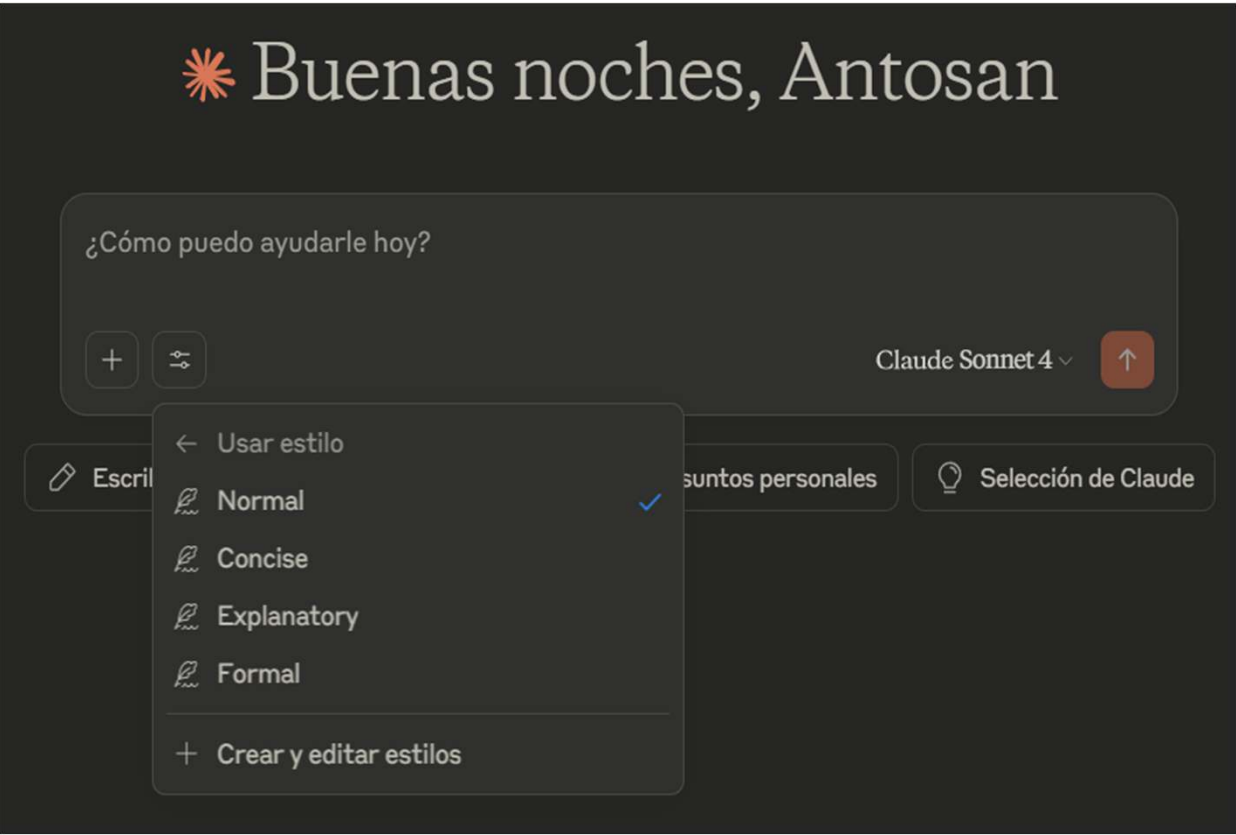
Claude te ayuda a **encontrar** tu estilo





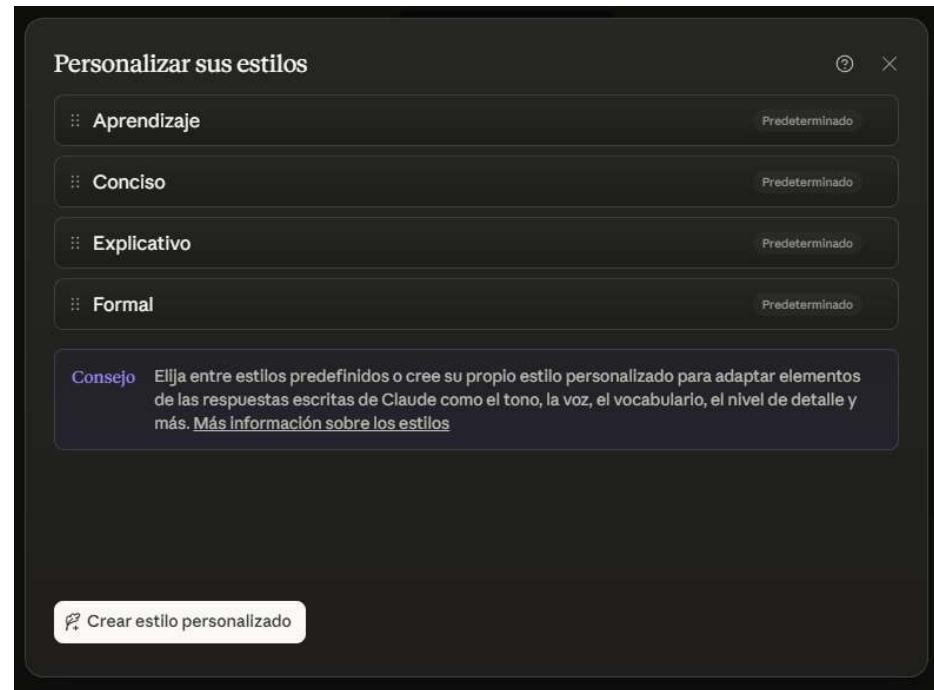
# Estilos: controla la voz de tu IA




Selector de estilos: Ajusta el tono, la longitud y la forma de las respuestas con solo un clic.



Estilo	Descripción breve	Ideal para...
Normal	Estilo base, balanceado	Uso general, interacción casual
Conciso	Respuestas breves y al punto	Resúmenes rápidos, ejecutivos
Explicativo	Más detallado, con contexto adicional	Capacitaciones, soporte, clases
Formal	Tono profesional, vocabulario sobrio	Comunicaciones institucionales

# Claude aprende tu tono de voz



-  **Provee un ejemplo**
  - Sube o pega un fragmento representativo del estilo deseado.
-  **Claude lo analiza**
  - Genera automáticamente una descripción del estilo.
-  **Refina y ajusta**
  - Edita el resumen para afinar el tono, nivel de detalle o formalidad.



# Describe tu estilo

← Describe su estilo

×

Generar estilo desde un punto de partida:

☒ Definir objetivo del estilo

☐ Adaptar a una audiencia

☐ Utilizar un tono y estilo específicos

☐ Describir en términos generales

☐ Usar instrucciones personalizadas (avanzado)

↩ No importa, añadiré un ejemplo

Generar estilo

Describe the main purpose and goal of this writing style — whether to teach, analyze, persuade, inspire, etc.

Try something like:  
I want to teach complex topics step-by-step, with a focus on building understanding.

# Estilo Institucional

**Descripción del Estilo Requerido:** Adopta un estilo de comunicación institucional formal y técnico-jurídico para el departamento tributario del Banco Agrario de Colombia, con las siguientes características específicas:

## Tono y Registro

- **Formal institucional:** Mantén un lenguaje respetuoso, serio y profesional en todo momento.
- **Técnico-jurídico:** Utiliza terminología tributaria precisa y fundamentos normativos sólidos.
- **Claro y accesible:** Evita jerga innecesaria, explica conceptos técnicos cuando sea pertinente.
- **Autoritativo pero servicial:** Proyecta conocimiento especializado manteniendo orientación al servicio.

## Características Específicas del Lenguaje

- Usar tercera persona institucional: “El Banco considera”, “Según el análisis técnico”
- Incluir siempre referencias normativas específicas (Ley, Artículo, párrafo)
- Citar conceptos DIAN cuando corresponda con formato: "Concepto DIAN No. XXX de YYYY"

## Precisión Técnica Requerida

- Verificar coherencia entre interpretación y normativa citada.
- Incluir plazos, porcentajes y fechas exactas.
- Señalar excepciones o casos especiales cuando apliquen.
- Indicar vigencia de las normas referenciadas.

## Estilo Institucional (resultado)

“Comunicarse con un estilo formal institucional propio del departamento tributario del Banco Agrario de Colombia. Utilizar lenguaje técnico-jurídico preciso, con referencias normativas específicas. Emplear tercera persona institucional, evitando jerga innecesaria. Mantener un tono autoritativo pero servicial, explicando conceptos técnicos con claridad. Incluir siempre citas de conceptos DIAN, referencias legales exactas (número de ley, artículo, párrafo), y detalles técnicos como plazos, porcentajes y fechas. Verificar rigurosamente la coherencia normativa y señalar excepciones o casos especiales cuando corresponda. El objetivo es transmitir información tributaria de manera profesional, clara y técnicamente sólida.”

# Actividad: Tu Primer Estilo Personalizado en Claude

- **Objetivo:** Crear y probar un asistente de IA con un estilo de comunicación a tu medida, utilizando la función de “*Estilos Personalizados*” de Claude.
- **Paso 1: Define tu Misión:** Piensa en un estilo de comunicación que sea útil para tu trabajo. Tienes dos opciones:

**Opción A: Describe un Estilo.** Piensa en un rol o personalidad que necesites (ej. “*Analista de negocios conciso*”, “*Profesor didáctico*”, “*Creativo publicitario*”).

**Opción B: Usa un Ejemplo.** Busca un texto corto que hayas escrito (un correo, un párrafo) que represente tu estilo de comunicación ideal.

- **Paso 2: Construye tu Estilo en Claude**

En Claude, haz clic en el ícono de la estrella ✨ (Estilos) y luego en “**Crear estilo personalizado**”.

Según tu elección anterior:

**Opción A:** Pega la **descripción** de tu estilo en el cuadro de texto principal (Definir objetivo del estilo).

**Opción B:** Haz clic en la opción “**No importa, añadiré un ejemplo**” y pega allí tu texto.

Haz clic en “**Generar estilo**” y dale un nombre memorable.

- **Paso 3: Pon a Prueba a tu Asistente Ahora,** con tu nuevo estilo activado, dale una tarea simple y observa el resultado.