

IA



Firefly
Software Consulting

SESIÓN 3

www.firefly-e.com

Otros modelos de IA

Diseñados para tareas específicas:

- Ejemplo:
 - Clasificación: Identificar correos como spam o no spam.
 - Regresión: Predecir el precio de una vivienda.
- Limitados al alcance para el que fueron entrenados.

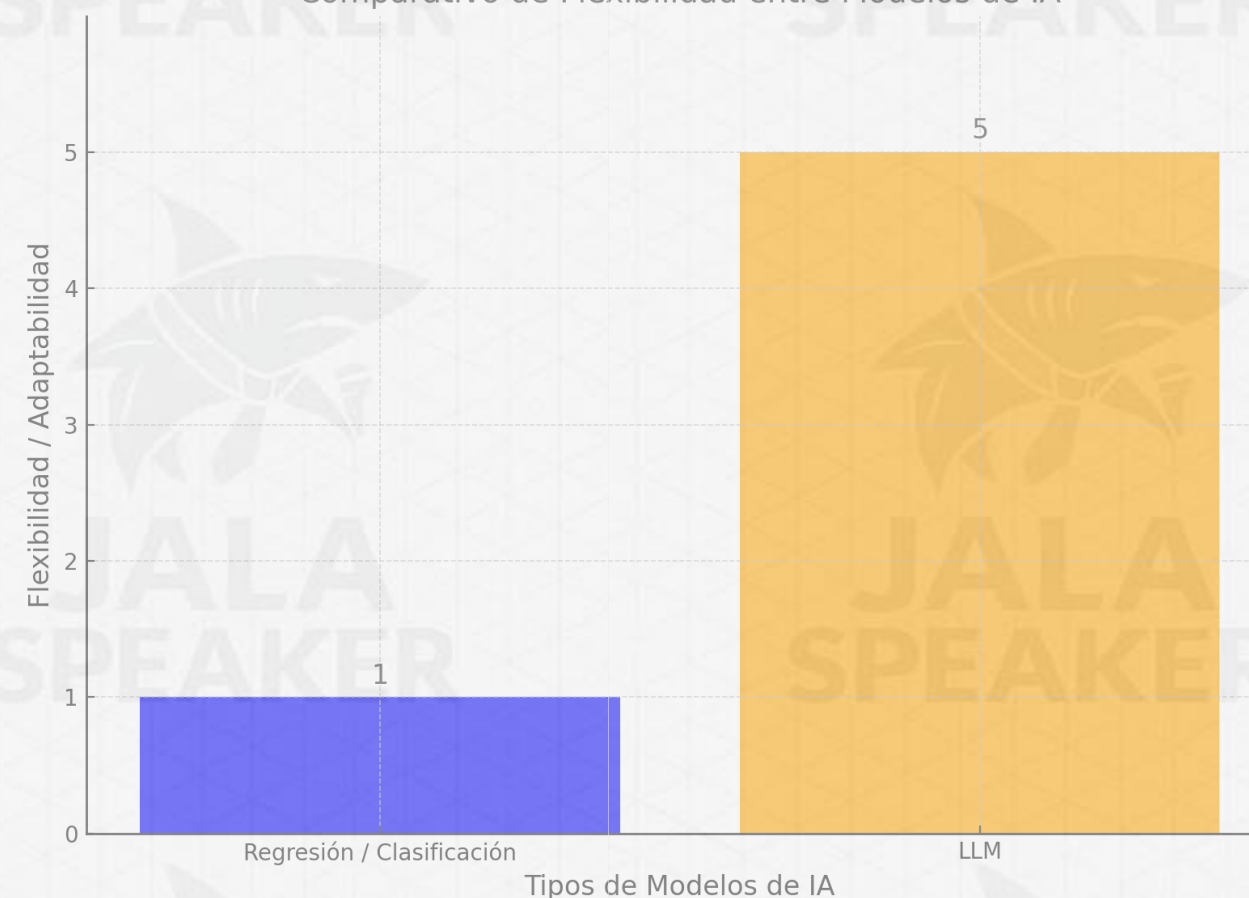
•Menor adaptabilidad a tareas generales:

- Funcionan bien dentro de un dominio predefinido.
- Requieren reentrenamiento o ajuste para abordar nuevas tareas o datos diferentes.

•Optimización enfocada:

- Uso eficiente de recursos computacionales en problemas concretos.
- Menor complejidad en comparación con modelos generales como los LLM.

Comparativo de Flexibilidad entre Modelos de IA



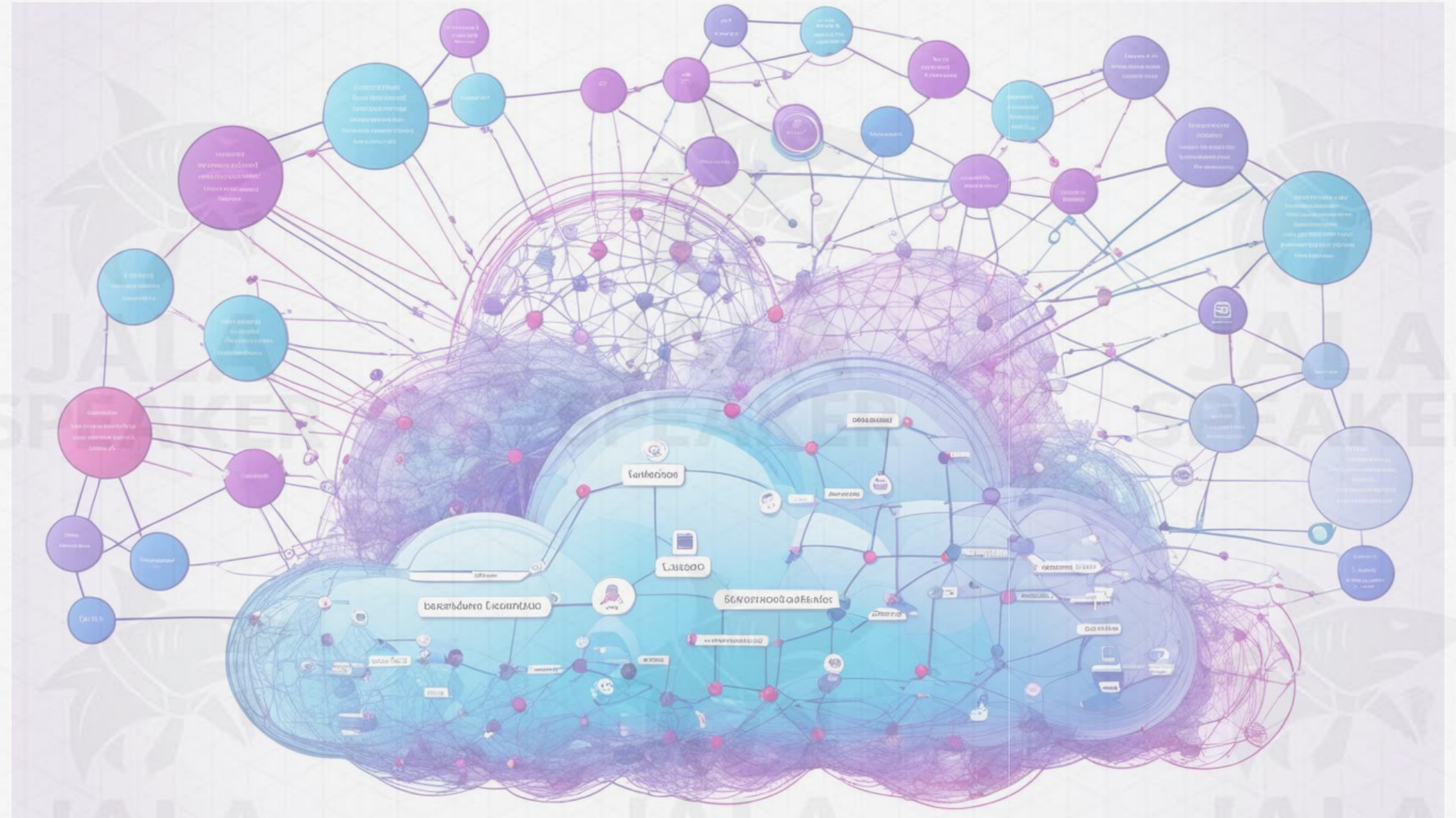
Definición de LLM

- Modelos entrenados para procesar y generar lenguaje natural de manera coherente y contextualizada.
- Capacidad para entender, resumir, traducir y generar texto en diferentes idiomas.



Características de los LLM

- Capacidad para entender y generar lenguaje humano
- Aprendizaje a partir de grandes conjuntos de texto
- Capacidad para adaptarse a diferentes contextos
- Capacidad para generar texto original y coherente
- Generalización: capaces de realizar tareas no explícitamente enseñadas.



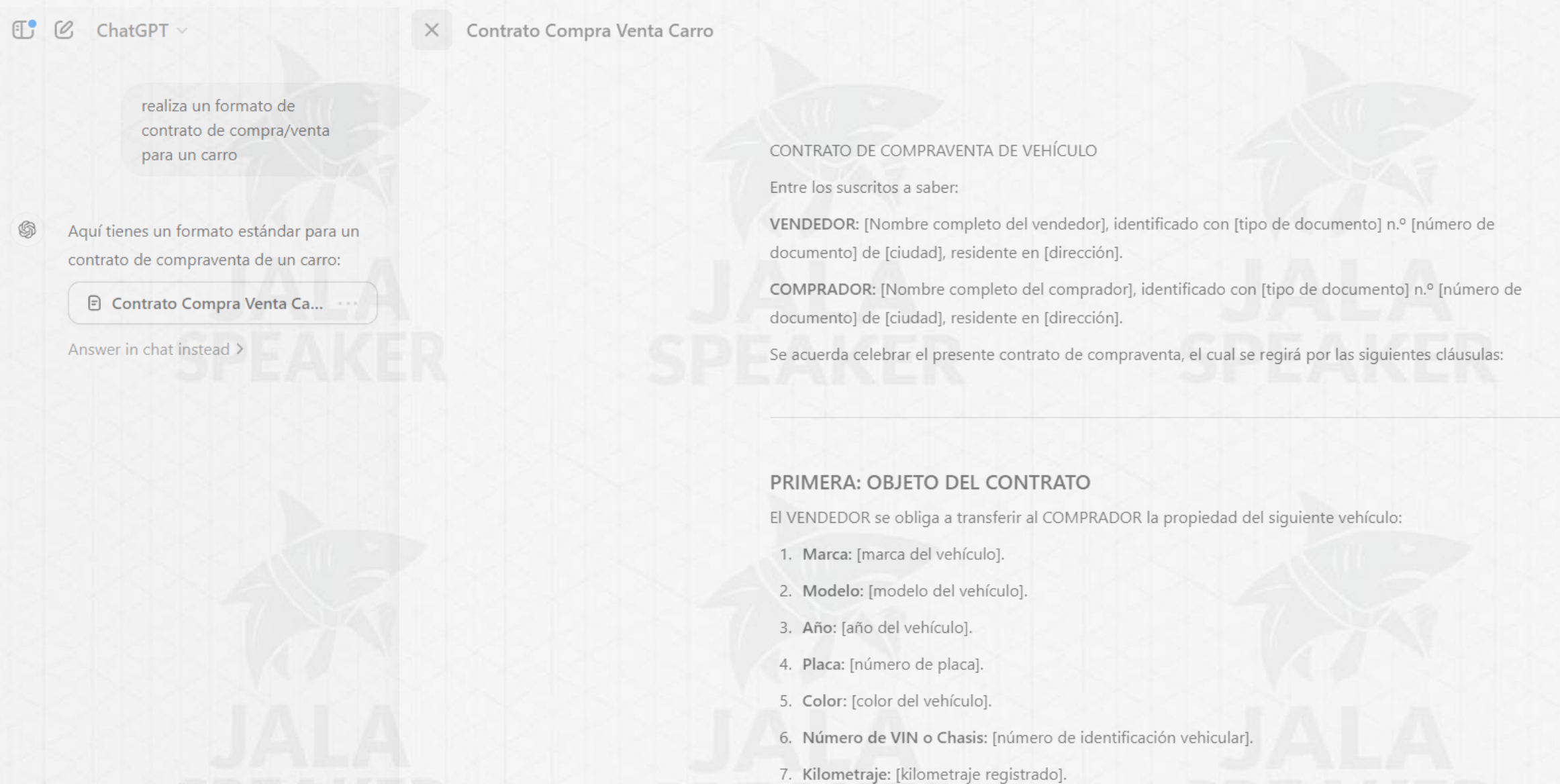
Modelos multimodales

- Un modelo multimodal de IA es un tipo de modelo que puede procesar y fusionar información de diferentes modos, como texto, imágenes, audio, etc. Este enfoque permite a los modelos aprender patrones y relaciones entre diferentes tipos de datos y realizar tareas más complejas.



¿Qué puede hacer un modelo multimodal?

Generación de texto de todo tipo



The screenshot displays a chat interface with a text input field containing the prompt: "realiza un formato de contrato de compra/venta para un carro". The AI response is a document titled "CONTRATO DE COMPRAVENTA DE VEHÍCULO". The document includes fields for the seller (VENDEDOR) and buyer (COMPRADOR), a list of seven vehicle specifications (Marca, Modelo, Año, Placa, Color, Número de VIN o Chasis, Kilometraje), and a section for the contract object.

realiza un formato de contrato de compra/venta para un carro

Aquí tienes un formato estándar para un contrato de compraventa de un carro:

Contrato Compra Venta Ca...

Answer in chat instead >

CONTRATO DE COMPRAVENTA DE VEHÍCULO

Entre los suscritos a saber:

VENDEDOR: [Nombre completo del vendedor], identificado con [tipo de documento] n.º [número de documento] de [ciudad], residente en [dirección].

COMPRADOR: [Nombre completo del comprador], identificado con [tipo de documento] n.º [número de documento] de [ciudad], residente en [dirección].

Se acuerda celebrar el presente contrato de compraventa, el cual se regirá por las siguientes cláusulas:

PRIMERA: OBJETO DEL CONTRATO

El VENDEDOR se obliga a transferir al COMPRADOR la propiedad del siguiente vehículo:

1. **Marca:** [marca del vehículo].
2. **Modelo:** [modelo del vehículo].
3. **Año:** [año del vehículo].
4. **Placa:** [número de placa].
5. **Color:** [color del vehículo].
6. **Número de VIN o Chasis:** [número de identificación vehicular].
7. **Kilometraje:** [kilometraje registrado].



¿Qué puede hacer un modelo multimodal?

Generación de imágenes y video

Genera el logo para una empresa de inteligencia artificial llamada SmartIA, donde su foco es la enseñanza a partir de la inteligencia artificial y se presenta como un humanoide con aspecto muy tranquilo y juvenil para enseñar en clases virtuales



Aquí tienes el logo generado para SmartIA. Si necesitas ajustes o quieres explorar un estilo diferente, no dudes en pedírmelo.



¿Qué puede hacer un modelo multimodal?

Generación de código

ChatGPT

Crea el código de una función en python para generar una calculadora

Aquí tienes un ejemplo de una función en Python que genera una calculadora básica:

`</>` Calculadora Basica

Answer in chat instead >

El código define una función `calculadora` que permite realizar operaciones básicas como suma, resta, multiplicación y división, además de gestionar errores comunes como entradas no válidas o divisiones por cero. ¡Puedes copiarlo y ejecutarlo directamente en tu entorno de Python!

¿Esta conversación te resultó útil hasta ahora?

Envía un mensaje a ChatGPT

ChatGPT puede cometer errores. Comprueba la información importante.

Calculadora Basica

```
1 def calculadora():
2     print("\nBienvenido a la calculadora básica\n")
3     print("Operaciones disponibles:")
4     print("1. Suma")
5     print("2. Resta")
6     print("3. Multiplicación")
7     print("4. División")
8     print("5. Salir")
9
10    while True:
11        try:
12            opcion = int(input("\nSeleccione una opción (1-5): "))
13
14            if opcion == 5:
15                print("Gracias por usar la calculadora. ¡Adiós!")
16                break
17
18            if opcion not in [1, 2, 3, 4]:
19                print("Opción no válida. Intente nuevamente.")
20                continue
21
22            num1 = float(input("Ingrese el primer número: "))
23            num2 = float(input("Ingrese el segundo número: "))
24
25            if opcion == 1:
26                print(f"El resultado de la suma es: {num1 + num2}")
27            elif opcion == 2:
28                print(f"El resultado de la resta es: {num1 - num2}")
29            elif opcion == 3:
30                print(f"El resultado de la multiplicación es: {num1 * num2}")
31            elif opcion == 4:
32                if num2 != 0:
33                    print(f"El resultado de la división es: {num1 / num2}")
34                else:
35                    print("Error: No se puede dividir entre cero.")
36
37            except ValueError:
38                print("Entrada inválida. Por favor, ingrese un número válido.")
39
40            # Llamar a la función para ejecutarla
41            calculadora()
42
```



Modelos razonadores

- Un modelo razonador de IA es un tipo de modelo que puede analizar, evaluar y sintetizar información de manera lógica y coherente. Este enfoque se basa en técnicas de inferencia y deducción que le permiten identificar patrones, establecer relaciones y generar conclusiones fundamentadas a partir de datos complejos



Modelos Fundacionales

- GPT-4 (OpenAI)
- Claude (Anthropic)
- PaLM 2 (Google)
- LLaMA 2 (Meta)
- R1 (Deepseek)



JALA
SPEAKER

JALA
SPEAKER

JALA
SPEAKER

JALA
SPEAKER

JALA
SPEAKER



Casos de uso de los LLM

Análisis de Imágenes con Modelos Multimodales

¿Qué hacen?

- Procesan imágenes y generan:
 - **Descripciones automáticas.**
 - **Etiquetas y clasificación.**
 - **Detección de anomalías.**



Casos de Uso:

- **Salud:**
 - Detección de enfermedades en radiografías y tomografías.
- **Seguridad:**
 - Reconocimiento facial en sistemas de vigilancia.
- **Retail:**
 - Análisis de imágenes de productos para inventarios o calidad.



Casos de uso de los LLM

Decisiones Inteligentes con LLM

¿Qué hacen?

- Responden preguntas o generan estrategias con base a una descripción de objetivos.

Casos de Uso:

•Marketing:

- Diseño de estrategias a partir de un brief.

•Finanzas:

- Análisis rápido de riesgos basados en texto introductorio.

•Legal:

- Interpretación de contratos o generación de resúmenes legales.



Casos de uso de los LLM

Agentes Autónomos con LLM

¿Qué son?

- Sistemas autónomos que combinan LLM con otras herramientas para ejecutar tareas específicas.

Casos de Uso:

• Atención al Cliente:

- Agente virtual para responder consultas en tiempo real.

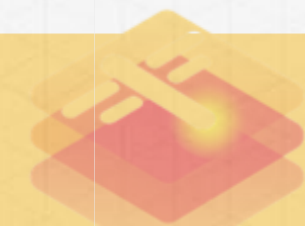
• Análisis de Datos:

- Asistente que procesa y visualiza datos.

• Automatización:

- Consulta de APIs y generación de reportes personalizados

88



Casos de uso de los LLM

Automatización con Chains de Tareas

¿Qué son?

- Flujos de trabajo que integran múltiples pasos automatizados utilizando LLM y herramientas adicionales.

Casos de Uso:

•Negocios:

- Flujos que analizan datos y generan reportes automáticamente.

•Educación:

- Creación de materiales didácticos personalizados.

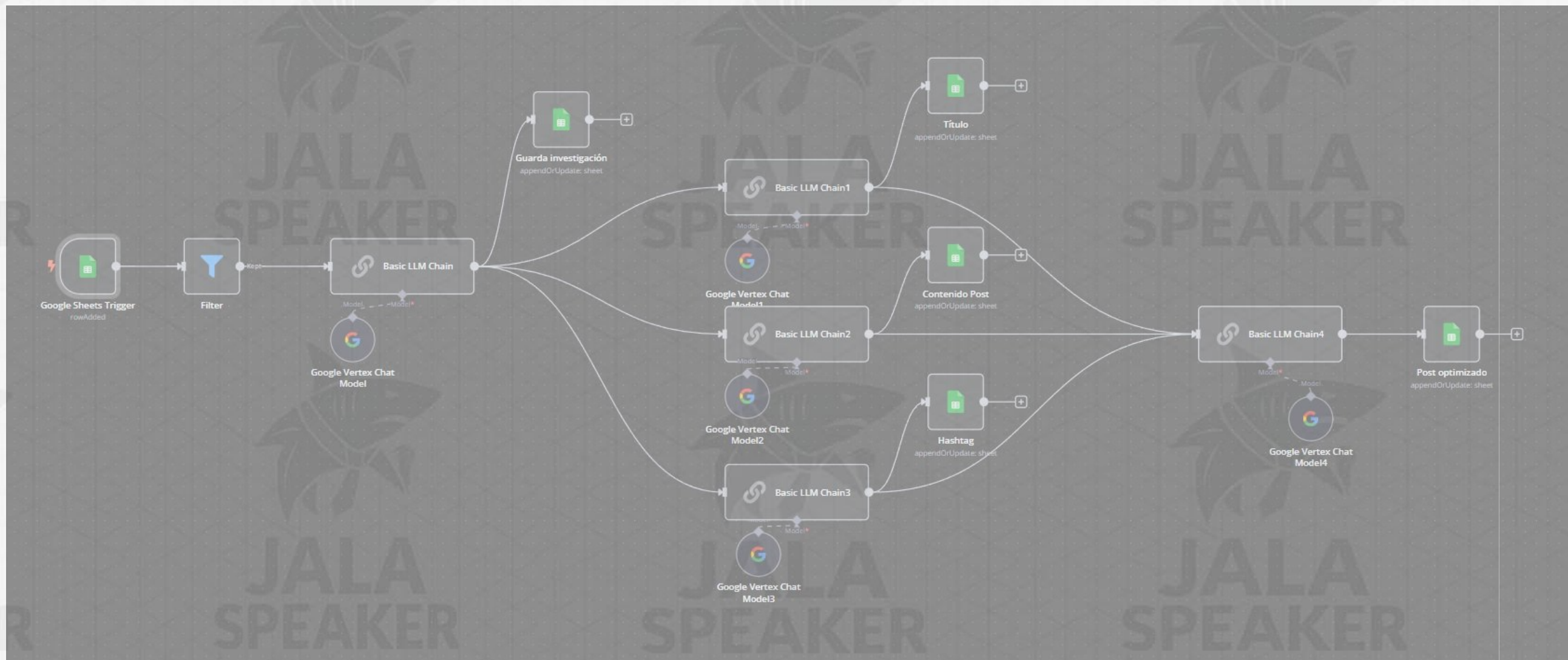
•Logística:

- Optimización de rutas basadas en datos dinámicos.



Casos de uso de los LLM

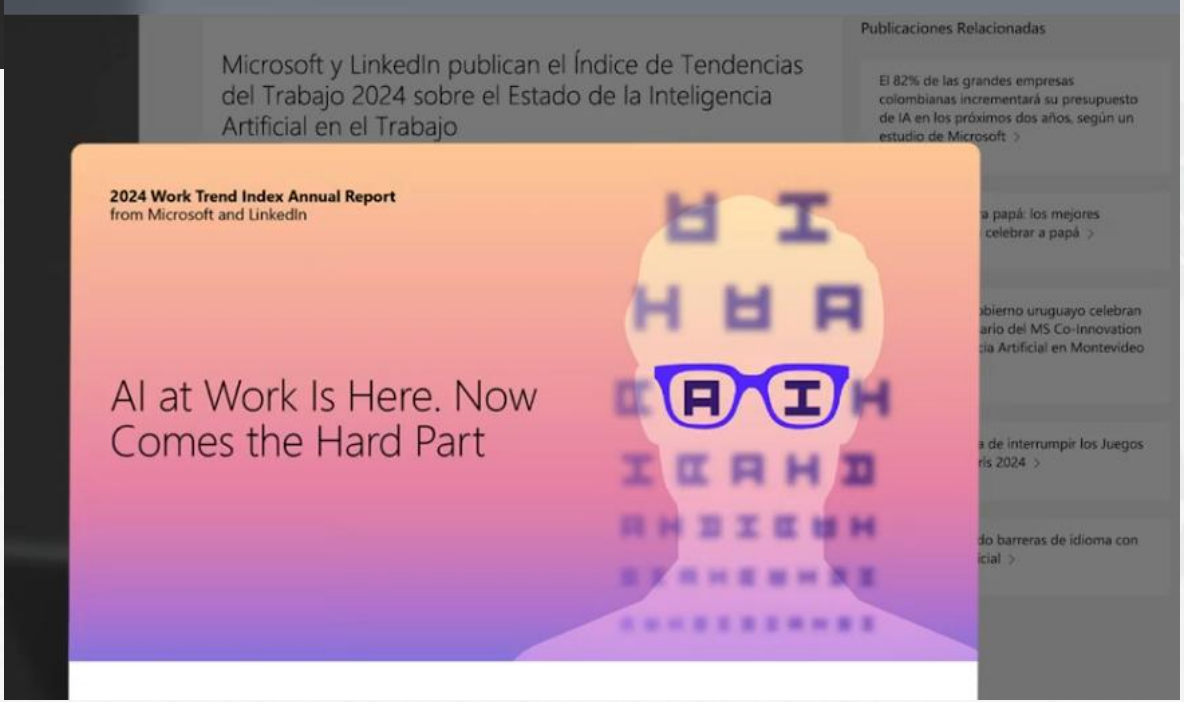
Automatización con Chains de Tareas



Tendencias sobre el trabajo 2024



Firefly
Software Consulting



By: JHON ALEXANDER LÓPEZ ALVAREZ

El 75% de los trabajadores utiliza la AI en el trabajo.



El 78% de los trabajadores usa sus propias herramientas de AI en el trabajo.

Para 76% de encuestados es indispensable usar AI para ser competitivos.

¿Cómo integramos AI de forma segura en el trabajo?

El 60% de los líderes no tienen una estrategia para integrar AI.

66% de los líderes requieren habilidades de AI en colaboradores.



“La Inteligencia Artificial no viene a reemplazar personas. Viene a reemplazar tareas. La IA no es magia, ni es el villano de una película futurista..”

By: JHON ALEXANDER LÓPEZ ALVAREZ

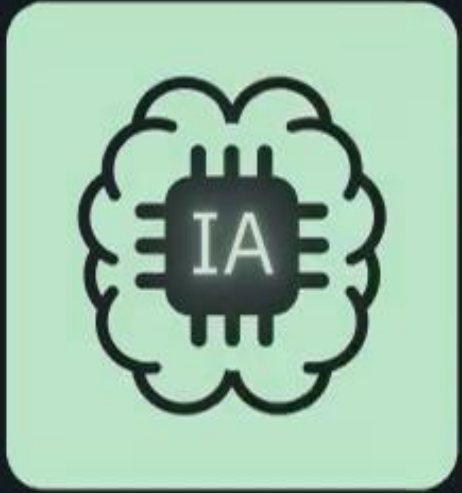


Firefly
Software Consulting





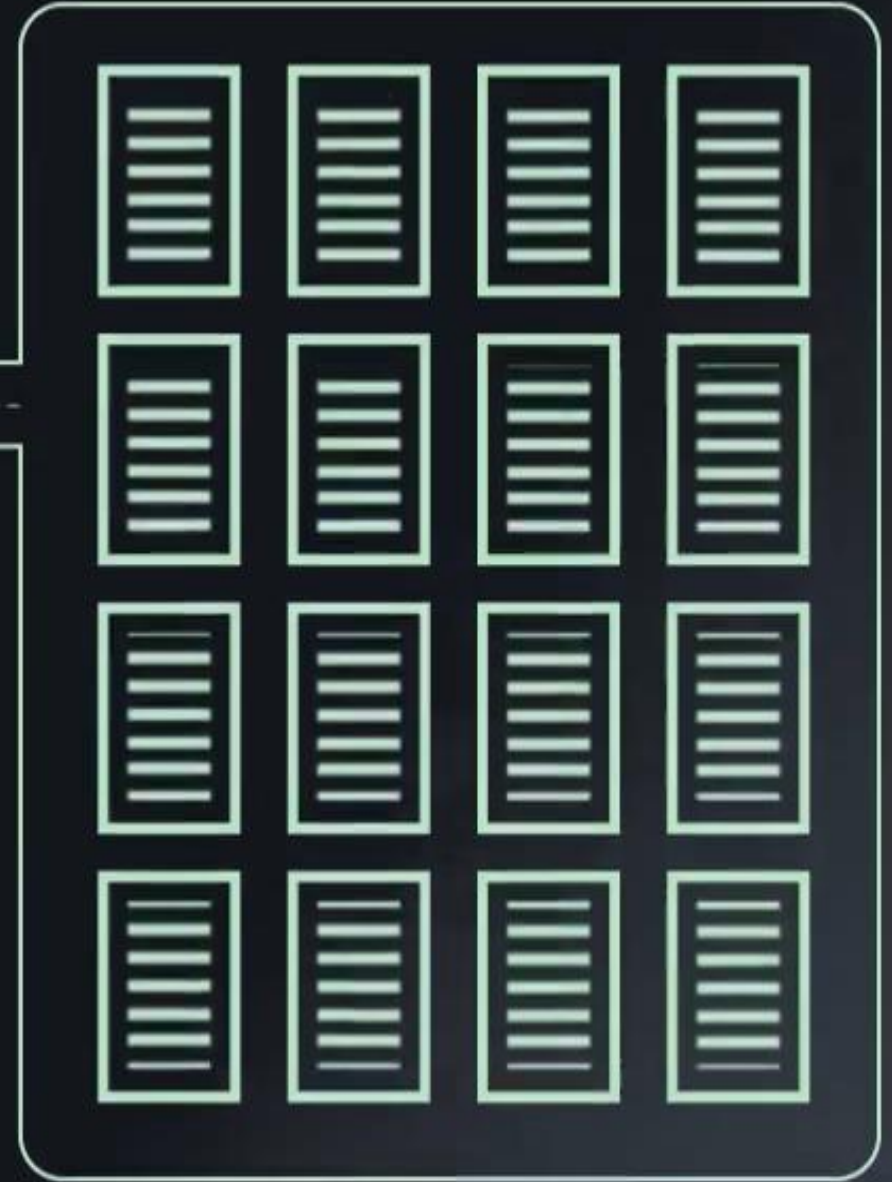
ENTRENAMIENTO



Campo de teorías
para crear



Sistemas
Computacionales



Datos



Clasificar

Hacer
predicciones





Aprenidizaje
automático



Aprenidizaje
supervisado



Datos etiquetados
para encontrar
patrones



Aprenidizaje
no supervisado



Datos sin
etiquetar



El nacimiento de una idea

• **1950:** Alan Turing introduce el "Test de Turing" en su artículo "*Computing Machinery and Intelligence*".

Este test se basa en una prueba donde un humano debe distinguir si esta hablando con una persona o una maquina.



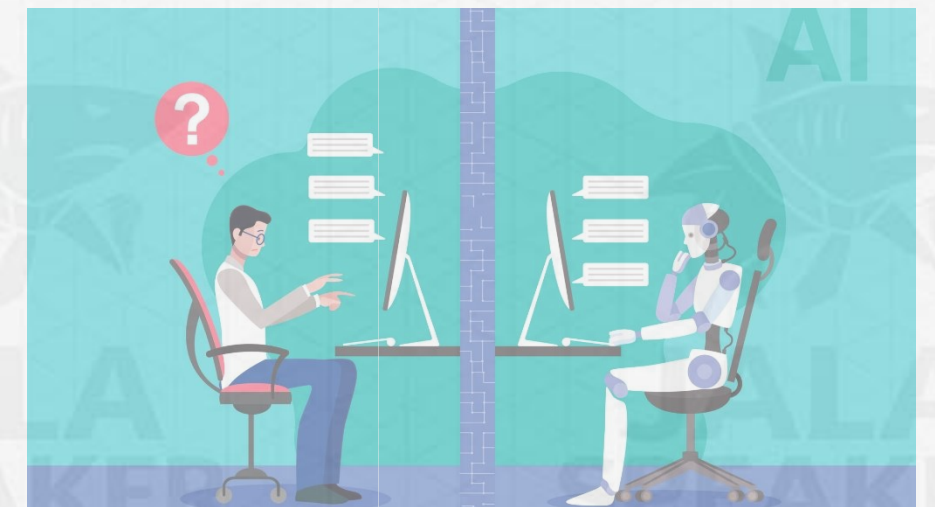
• A hoy en día se utilizan métodos como:

- Benchmarks específicos
- Capacidad de razonamiento y generalización
- Eficiencia y robustez
- Comportamiento ético y explicabilidad

• **1956:** Conferencia de Dartmouth: Coincidencia del término "Inteligencia Artificial".

Primeros programas exitosos:

- **Logic Theorist** (1956): Resolvía teoremas matemáticos.
- **ELIZA** (1966): Primer chatbot que simulaba conversaciones humanas.



El invierno de la IA y la persistencia de los pioneros

Años 70:

"Invierno de la IA", desilusión por promesas incumplidas.

Progresos notables:

- Sistemas expertos: Software diseñado para resolver problemas específicos (e.g., DENDRAL, MYCIN).

Característica	DENDRAL	MYCIN
Dominio	Química	Medicina
Propósito	Identificar estructuras moleculares	Diagnosticar infecciones y recomendar tratamientos
Impacto	Uso práctico en investigación química	Demostración conceptual; no usado clínicamente

- Surgen conceptos como redes neuronales. (Nodos y Capas)

- Impacto limitado por hardware y costos.



De la teoría a las aplicaciones prácticas

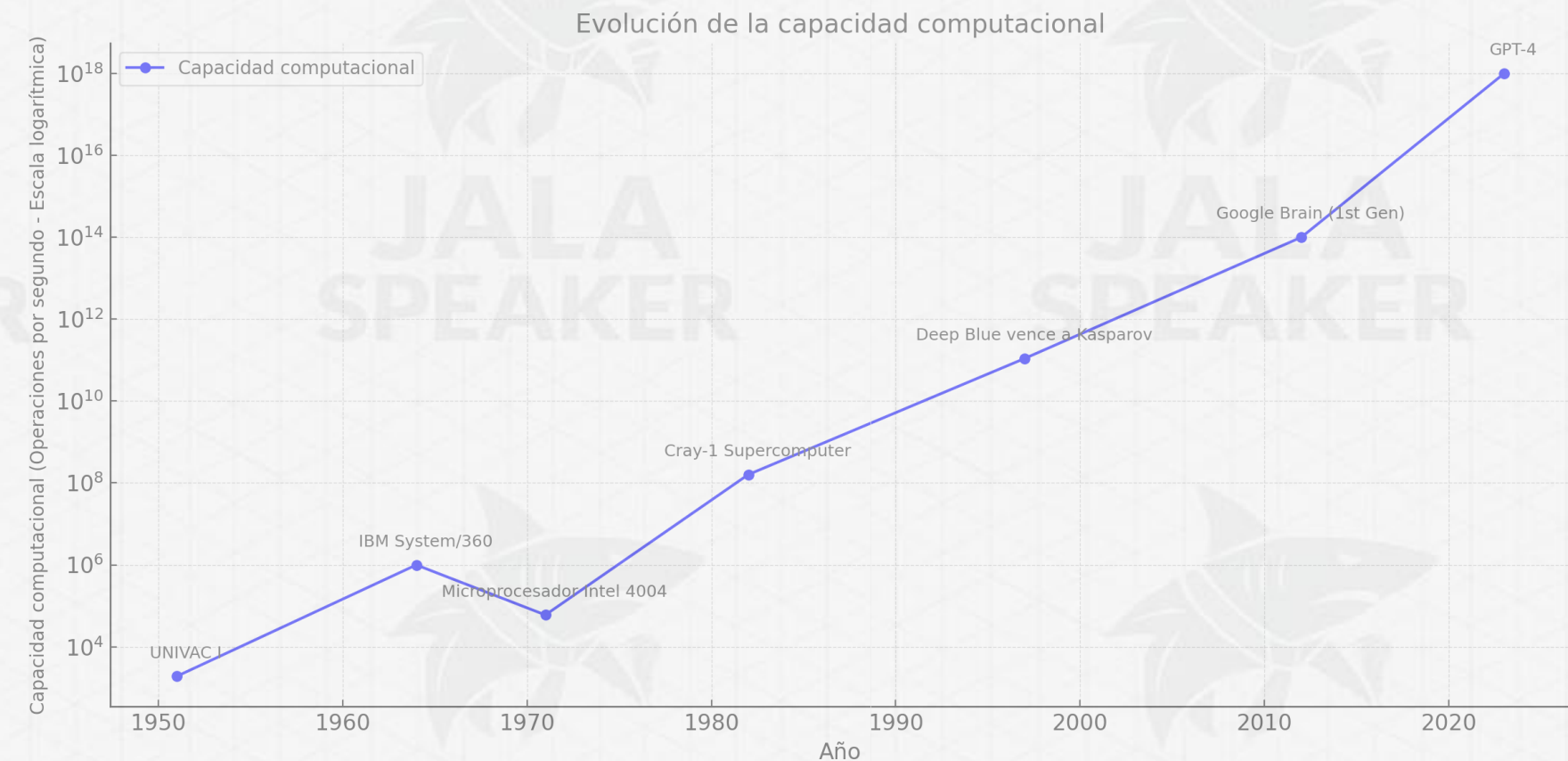
Años 90: Avances en hardware impulsan algoritmos más complejos.

- **1997:** Deep Blue de IBM derrota al campeón mundial de ajedrez Garry Kasparov.

Auge del aprendizaje automático:

- Desarrollo de algoritmos de clasificación y regresión.
- Crecimiento de bases de datos para entrenar modelos.

2000s: Los gigantes tecnológicos como Google y Amazon adoptan la IA para buscadores y recomendaciones.



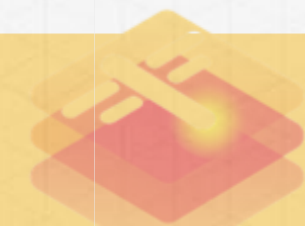
La era del Deep Learning

Progreso exponencial en redes neuronales profundas (Deep Learning):

- **2012:** AlexNet revoluciona la visión computacional con redes convolucionales (CNN).
- Grandes modelos de lenguaje como GPT-2 y BERT redefinen el procesamiento de lenguaje natural.

IA en nuestra vida cotidiana:

- Asistentes virtuales (Siri, Alexa).
- Sistemas de recomendación.
- Vehículos autónomos.



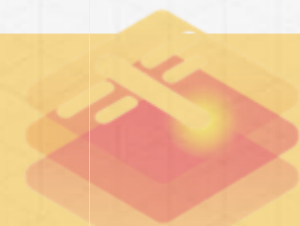


Deep Learning





**Estas redes simulan cómo
funcionan las neuronas
de nuestro cerebro**





Aprendizaje profundo
[Deep Learning]



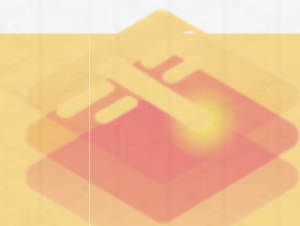
Procesamiento de
Lenguaje Natural



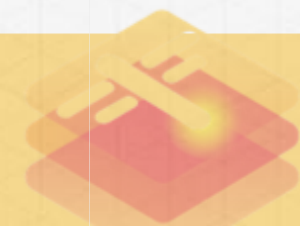
Analizar Simular Resolver



Es un campo con
muchos retos por resolver



Hacer que una computadora
comprenda el contexto de una
palabra en diferentes oraciones



Un salto hacia la creatividad artificial

2020s:

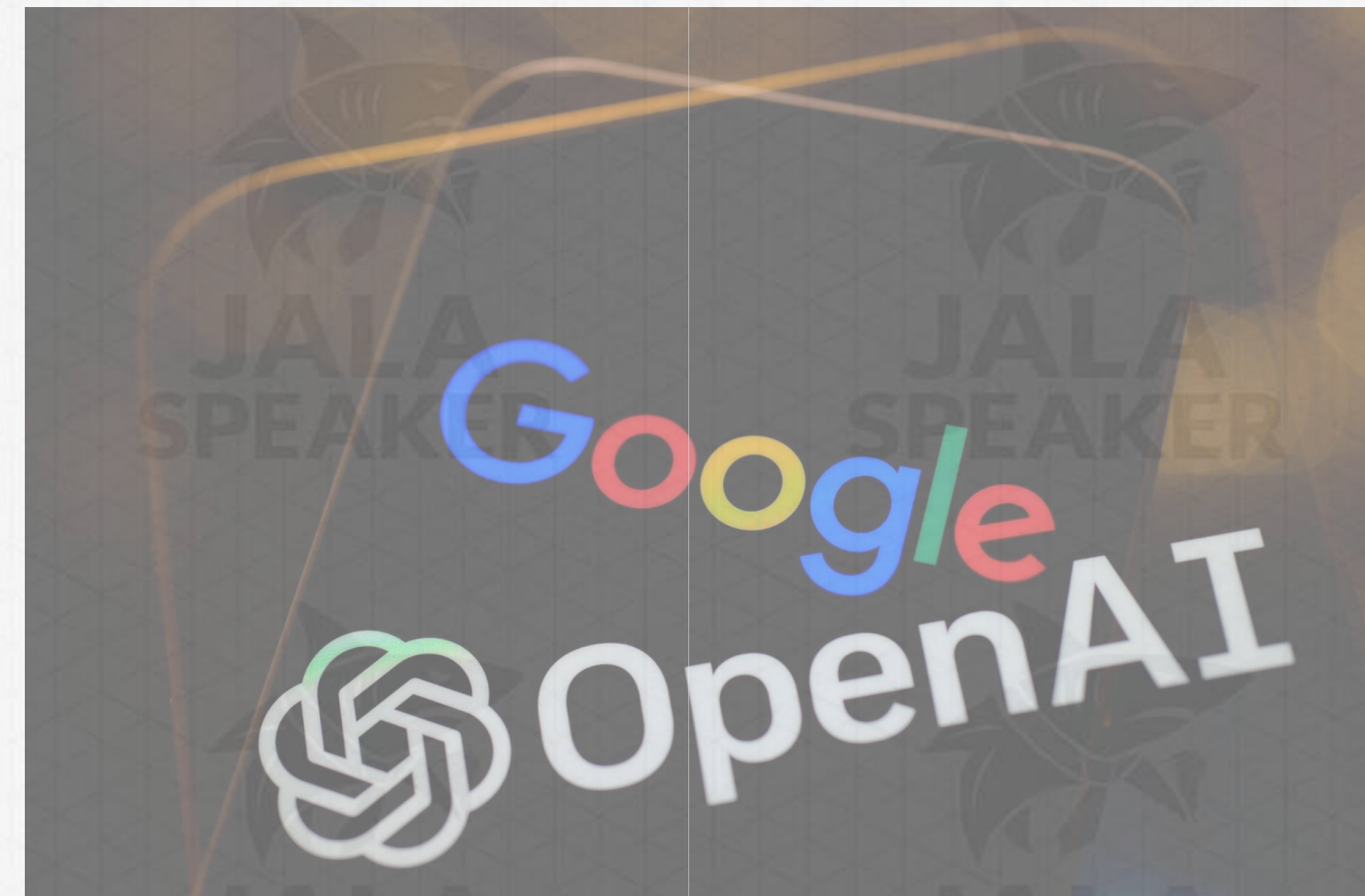
- IA generativa: GPT-3, DALL·E, y ChatGPT revolucionan la interacción humano-máquina.
- Modelos multimodales: Combina texto, imágenes, y audio (e.g., CLIP, GPT-4).

Democratización:

- Plataformas accesibles como OpenAI, Hugging Face, y Google Vertex AI.

Ética y regulación:

- Debates sobre el uso responsable de la IA.
- Impacto en el trabajo y la sociedad.



“Con redes neuronales entrenadas, podemos detectar fallas antes de que ocurran, ahorrar antes de gastar y predecir antes de improvisar.”

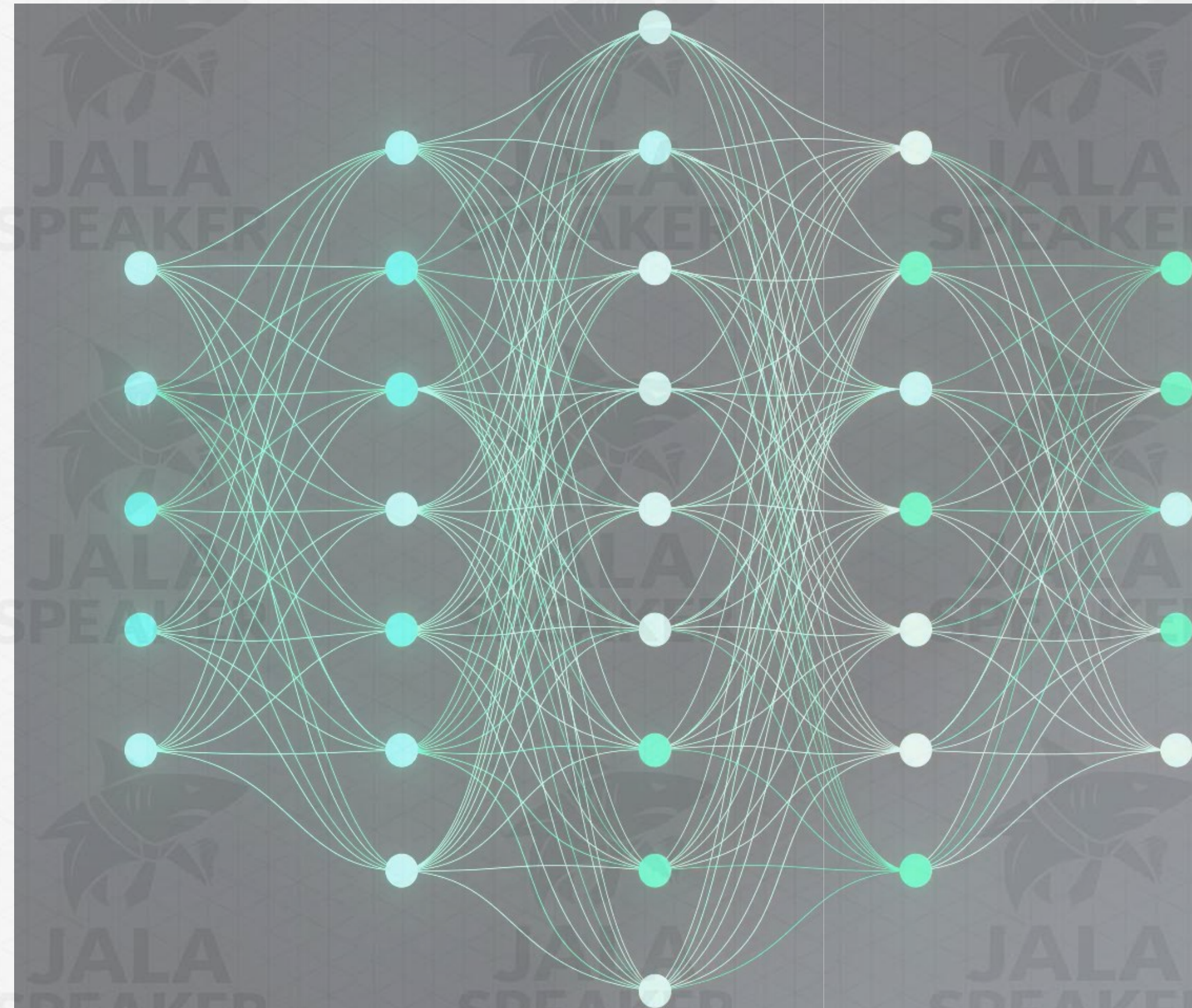
By: JHON ALEXANDER LÓPEZ ALVAREZ



Firefly
Software Consulting



REDES NEURONALES



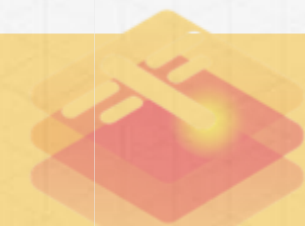
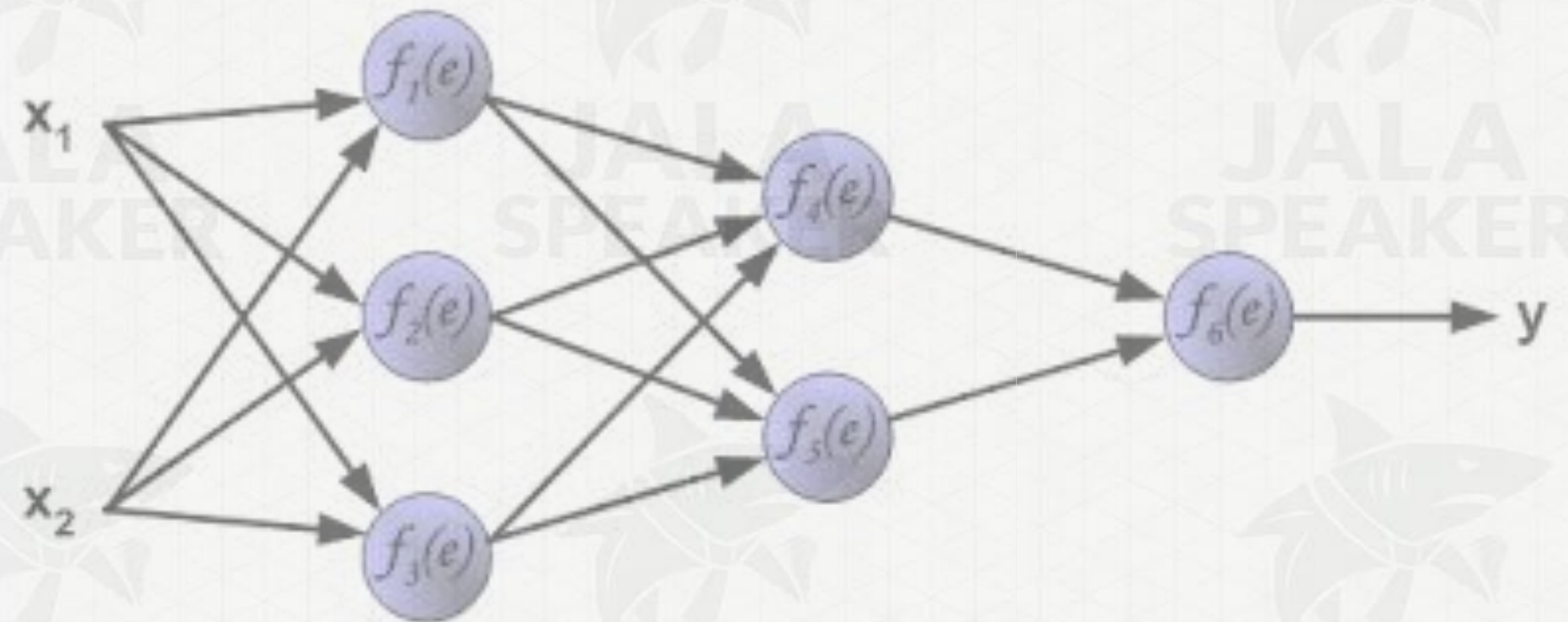
Red neuronal

Perceptrón:

Es un modelo básico de neurona artificial que toma varias entradas, las pondera y produce una salida binaria mediante una función de activación.

Peso sináptico:

Es el valor que multiplica cada entrada en el perceptrón, determinando la importancia de dicha entrada en la decisión final.



Hacer que una computadora
comprenda el contexto de una
palabra en diferentes oraciones



Para resolver este desafío,
en 2017 se publicó un artículo



7 Conclusion

In this work, we presented the Transformer, the first sequence transduction model based entirely on attention, replacing the recurrent layers most commonly used in encoder-decoder architectures with multi-headed self-attention.

For translation tasks, the Transformer can be trained significantly faster than architectures based on recurrent or convolutional layers. On both WMT 2014 English-to-German and WMT 2014 English-to-French translation tasks, we achieve a new state of the art. In the former task our best model outperforms even all previously reported ensembles.

We are excited about the future of attention-based models and plan to apply them to other tasks. We plan to extend the Transformer to problems involving input and output modalities other than text and to investigate local, restricted attention mechanisms to efficiently handle large inputs and outputs such as images, audio and video. Making generation less sequential is another research goal of ours.

The code we used to train and evaluate our models is available at <https://github.com/tensorflow/tensor2tensor>.

Acknowledgements We are grateful to Nal Kalchbrenner and Stephan Gouws for their fruitful comments, corrections and inspiration.

References

- [1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.
- [2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine learning to align and translate. *CoRR*, abs/1409.0473, 2014.

Fuente: Paper Attention Is All You Need

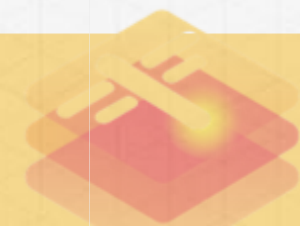
<https://arxiv.org/abs/1706.03762>



ATENCIÓN



REINA



REINA



REINA



REINA



Función de atención





Función de atención

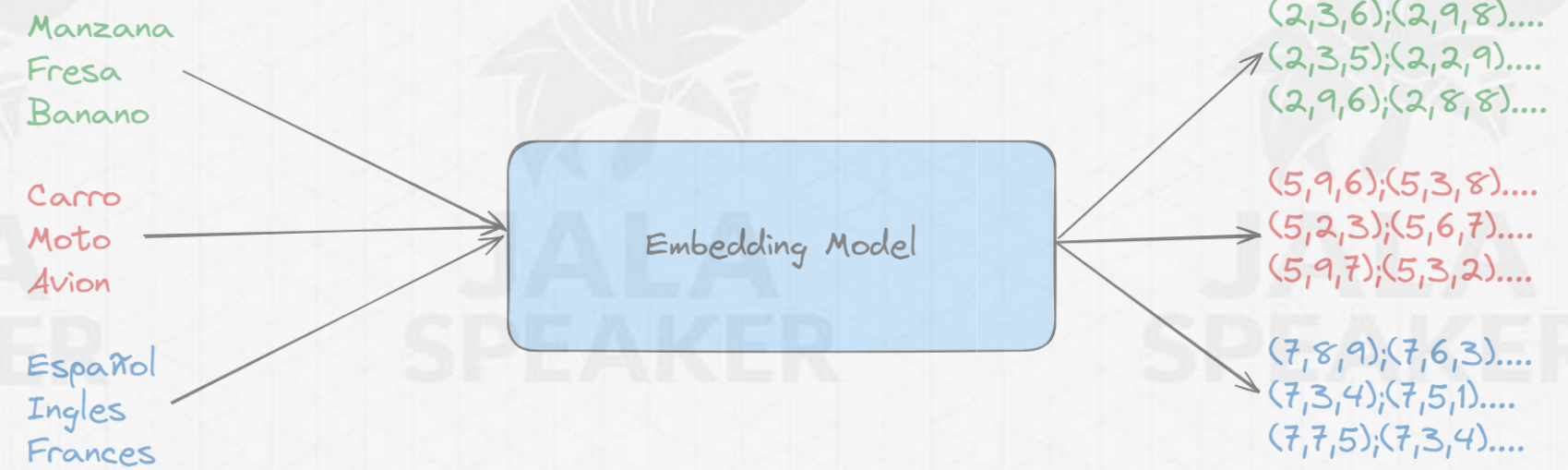


Redes neuronales

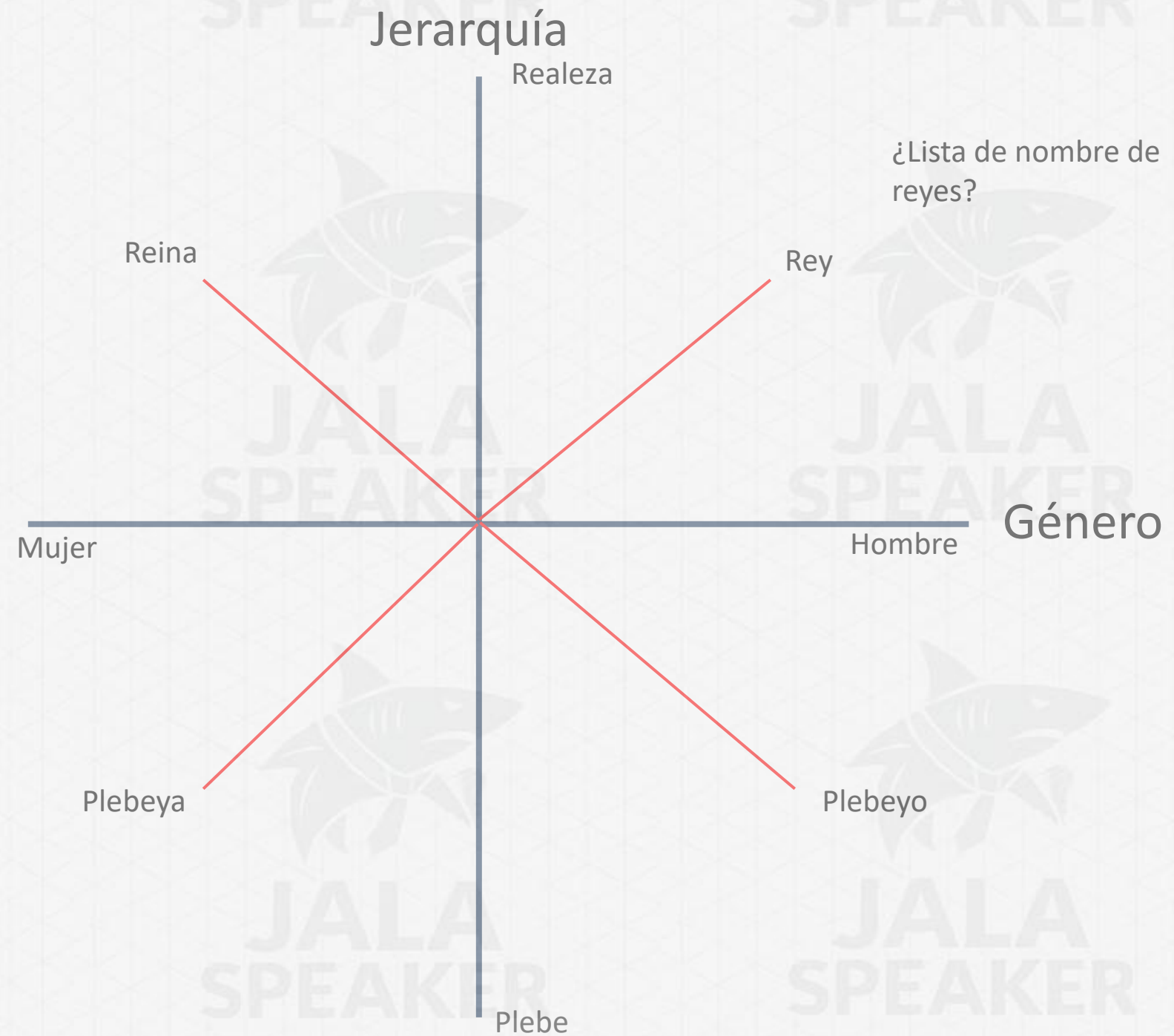


¿Qué son los Embeddings?

- Los embeddings representan palabras como vectores en un espacio multidimensional.
- Capturan la relación semántica entre las palabras.
- Permiten entender mejor el significado del lenguaje.



¿Qué son los Embeddings?



1. Luis XIV
2. Ricardo Corazón de León
3. Salomón
4. Alejandro Magno
5. Arturo



Juego 1: ¿Cuál es el Embed?

Instrucciones:

Envía estos mensajes a ChatGPT e interactúa con las respuestas:

1.Mensaje 1:

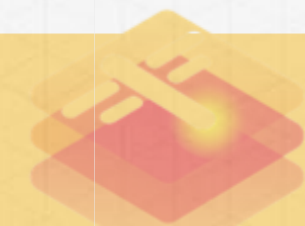
"ChatGPT, dime si las siguientes palabras están cerca o lejos en un espacio de embeddings: 'gato' y 'felino'."

2.Mensaje 2:

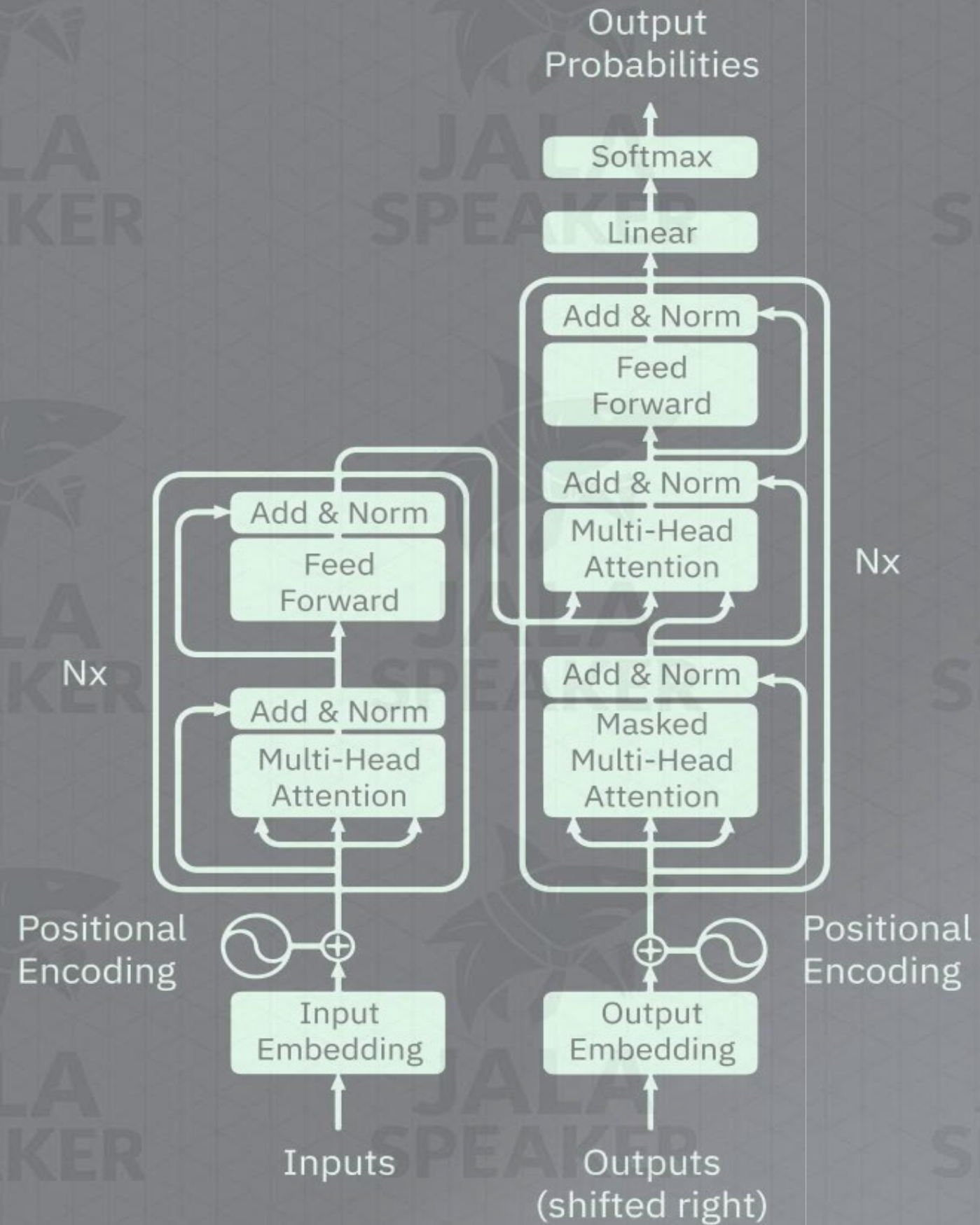
"Ahora, dime si 'gato' y 'perro' están cerca en el espacio de embeddings."

3.Mensaje 3 (Desafío):

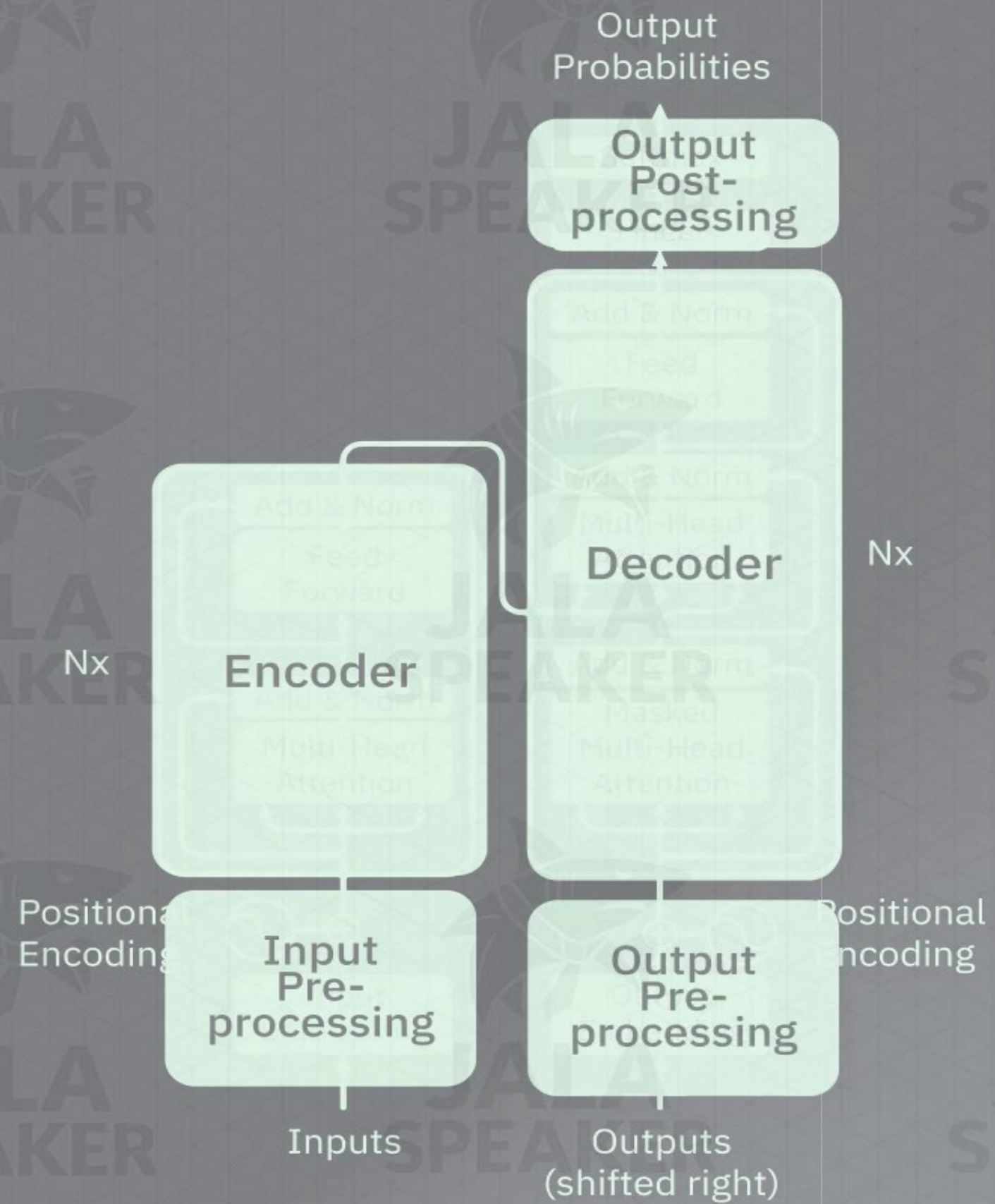
"Si cambio la palabra 'gato' por 'auto', ¿cómo afectaría eso al espacio de embeddings? ¿Lejos o cerca de 'felino'?"



TRANSFORMER



TRANSFORMER





ChatGPT



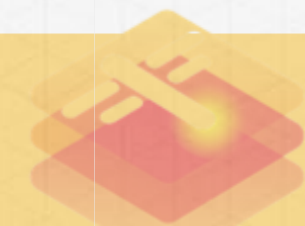


ChatGPT

Generative Pretrained Transformer



Cómo se construyen estas herramientas



Modelos base para desarrollar
múltiples herramientas

Se conocen como
modelos fundacionales



Introducción a los Transformers

- Los transformadores son una arquitectura neural que se utiliza para realizar tareas de procesamiento del lenguaje natural, como la traducción automática o el reconocimiento de entidad mencionada.

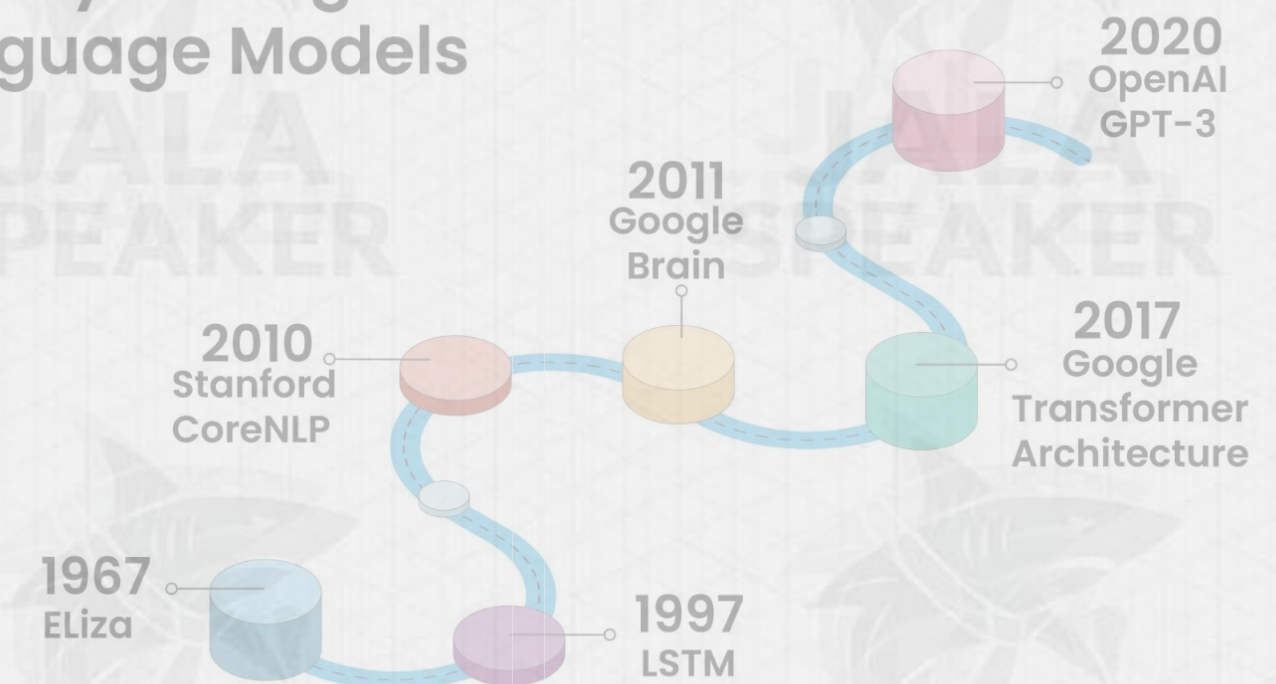
Fueron desarrollados por Vaswani et al. en 2017.

Resuelve problemas de secuencias largas en texto.

Componentes principales:

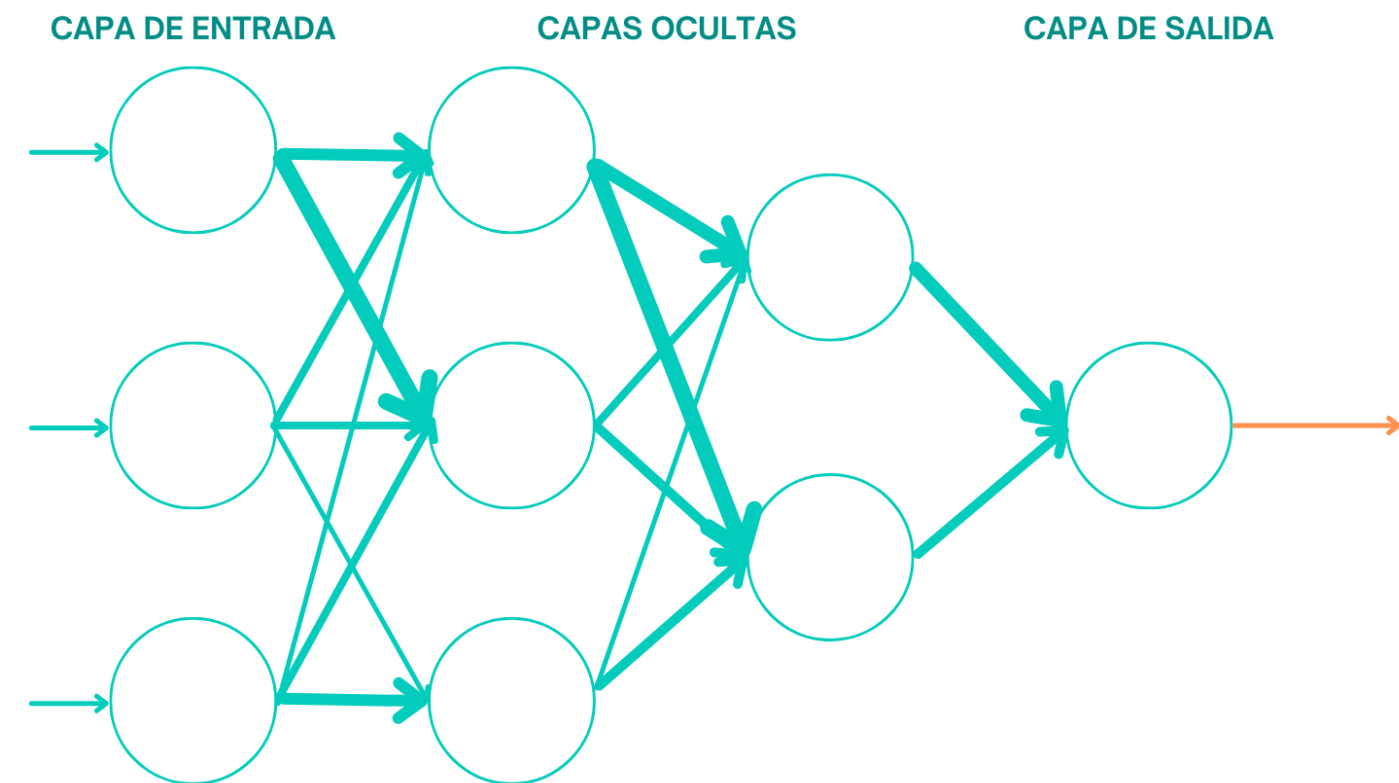
- **Mecanismo de atención:** Identifica las palabras más relevantes en una frase.
- **Capas de codificación y decodificación:** Procesan texto en dos fases.

History of Large Language Models



Mecanismo de Atención en Transformers

- **Atención:** Responde a la pregunta "¿Qué partes del texto son más importantes?"
- **Calcula pesos para cada palabra en relación con otras.**
- Ejemplo: En "el perro persigue al gato", "persigue" está relacionado tanto con "perro" como con "gato".



Juego 2: Atención al Detalle

Instrucciones:

Envía estos mensajes a ChatGPT y responde a las preguntas:

1.Mensaje 1:

"ChatGPT, ¿qué parte de la frase 'El perro corre hacia el parque' sería más importante para un modelo de atención?"

2.Mensaje 2:

"Si cambiamos la frase a 'El gato duerme en el sofá', ¿cómo cambiaría la atención del modelo?"

3.Mensaje 3 (Desafío):

"Imagina que el modelo necesita hacer una traducción compleja. ¿Cómo afectaría el mecanismo de atención a la traducción de la frase 'La inteligencia artificial está revolucionando el mundo'?
¿En qué se enfocarían más?"



Diferencias entre transformer y S2S

Aspecto	Transformer	Seq2Seq (S2S)
Arquitectura	Basado en atención (Self-attention) y capas paralelizadas	Basado en RNN o LSTM, con codificador y decodificador
Manejo de Secuencias	Procesa secuencias completas simultáneamente, en paralelo	Procesa secuencias de manera secuencial, palabra por palabra
Velocidad de Entrenamiento	Más rápido debido al procesamiento paralelo y la atención	Más lento debido a la naturaleza secuencial
Captura de Dependencias	Mejor para capturar dependencias largas dentro del texto	Menos efectivo en secuencias largas, problemas con el "vanishing gradient"
Escalabilidad	Mejor escalabilidad con grandes volúmenes de datos	Puede ser más limitado en escalabilidad, especialmente con secuencias largas



Casos de uso

Modelo	Casos de Uso
Transformer	- Traducción automática: Google Translate usa Transformer para traducir entre idiomas.
	- Generación de texto: GPT-3 genera contenido coherente y relevante.
	- Clasificación de Sentimiento: BERT analiza opiniones en redes sociales.
	- Resumen de texto: BART genera resúmenes de textos largos de forma precisa.
	- Análisis de preguntas y respuestas: Siri y Alexa utilizan Transformers para responder preguntas.
Seq2Seq (S2S)	- Traducción automática: Sistemas antiguos como los de Google Translate antes de Transformers.
	- Subtitulación automática: Modelos S2S generan subtítulos en tiempo real para videos.
	- Reconocimiento de voz: Utilizados en sistemas de transcripción de voz a texto.
	- Generación de resúmenes simples: Sistemas de S2S generan resúmenes básicos de texto.



Juego 3: Transformer vs S2S

Basándote en el texto que esta entre comillas debes responderme en 3 frases como lo resumirías, pero usando 2 técnicas. la primera debes extraer las 3 frases como lo haría un modelo entrenado con transformer y la segunda sería usando una técnica S2S.

"La inteligencia artificial ha avanzado a pasos agigantados en los últimos años. Los sistemas de IA ahora pueden reconocer patrones complejos en grandes volúmenes de datos, lo que ha permitido mejoras en áreas como la salud, la educación y la automoción. Sin embargo, a medida que la IA se vuelve más potente, también surgen preocupaciones sobre su impacto en el empleo. En la medicina, por ejemplo, la IA ha ayudado a diagnosticar enfermedades con una precisión que antes solo era posible con expertos humanos. Aunque muchos temen que los avances en la IA puedan llevar a la automatización de trabajos, también hay un consenso de que, si se implementa de manera ética, puede beneficiar a la humanidad. Por lo tanto, los expertos recomiendan que los gobiernos y las empresas establezcan regulaciones claras para garantizar el uso responsable de la IA."



Procesamiento de Lenguaje Natural



LLM

Large Language Models





Ciencias de datos
Machine learning
Desarrollo backend
Cloud engineering



Productos que nos ayuden a
resolver problemas en el día a día



¿Qué es el Fine-Tuning?

•Definición:

El fine-tuning es el proceso de ajustar un modelo preentrenado a un conjunto de datos específico para una tarea concreta, mejorando su desempeño en un caso particular.

•Objetivo:

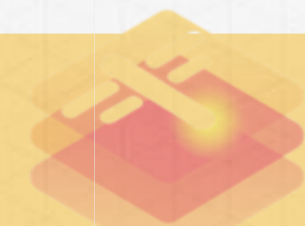
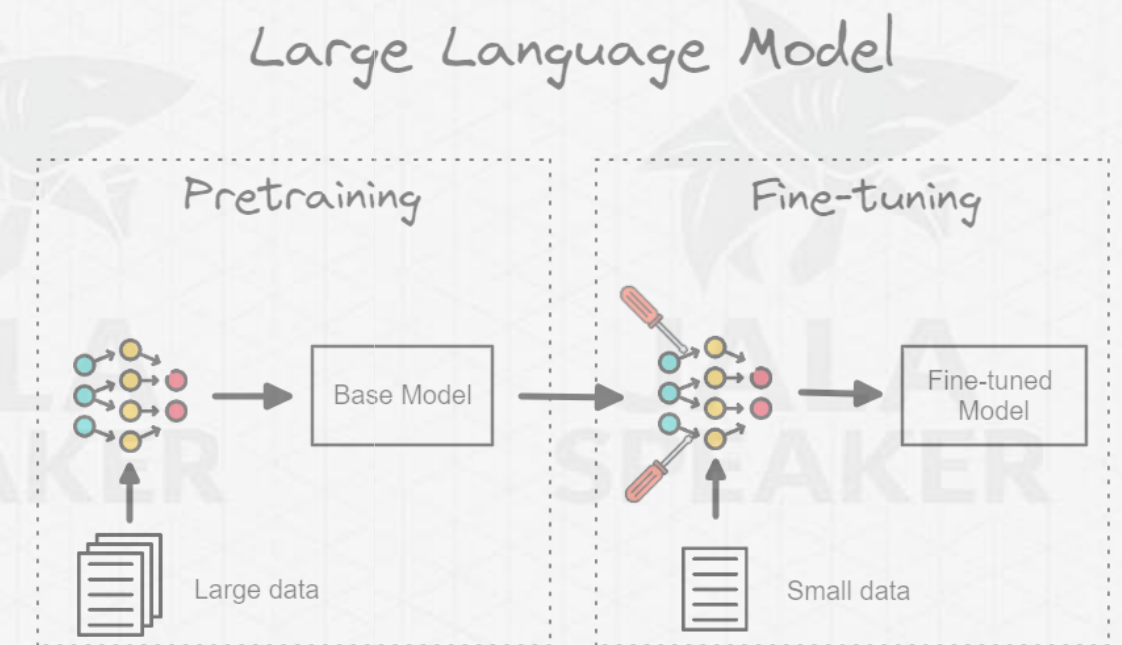
Especializar un modelo general (como GPT) para tareas específicas, como análisis de sentimientos, clasificación de documentos, o asistencia técnica.

•Ventajas:

- Menor necesidad de entrenar un modelo desde cero.
- Reducción en costos computacionales.
- Mejora en precisión y relevancia.

•Ejemplo:

Utilizar un modelo general para crear un chatbot especializado en derecho, entrenándolo con textos legales.



¿Qué es Retrieval-Augmented Generation (RAG)?

•Definición:

RAG combina generación de texto con recuperación de información en tiempo real.

•Funcionamiento:

1.Recuperación: El modelo busca información en bases de datos o documentos externos.

2.Generación: Usa esa información para construir respuestas contextuales y precisas.

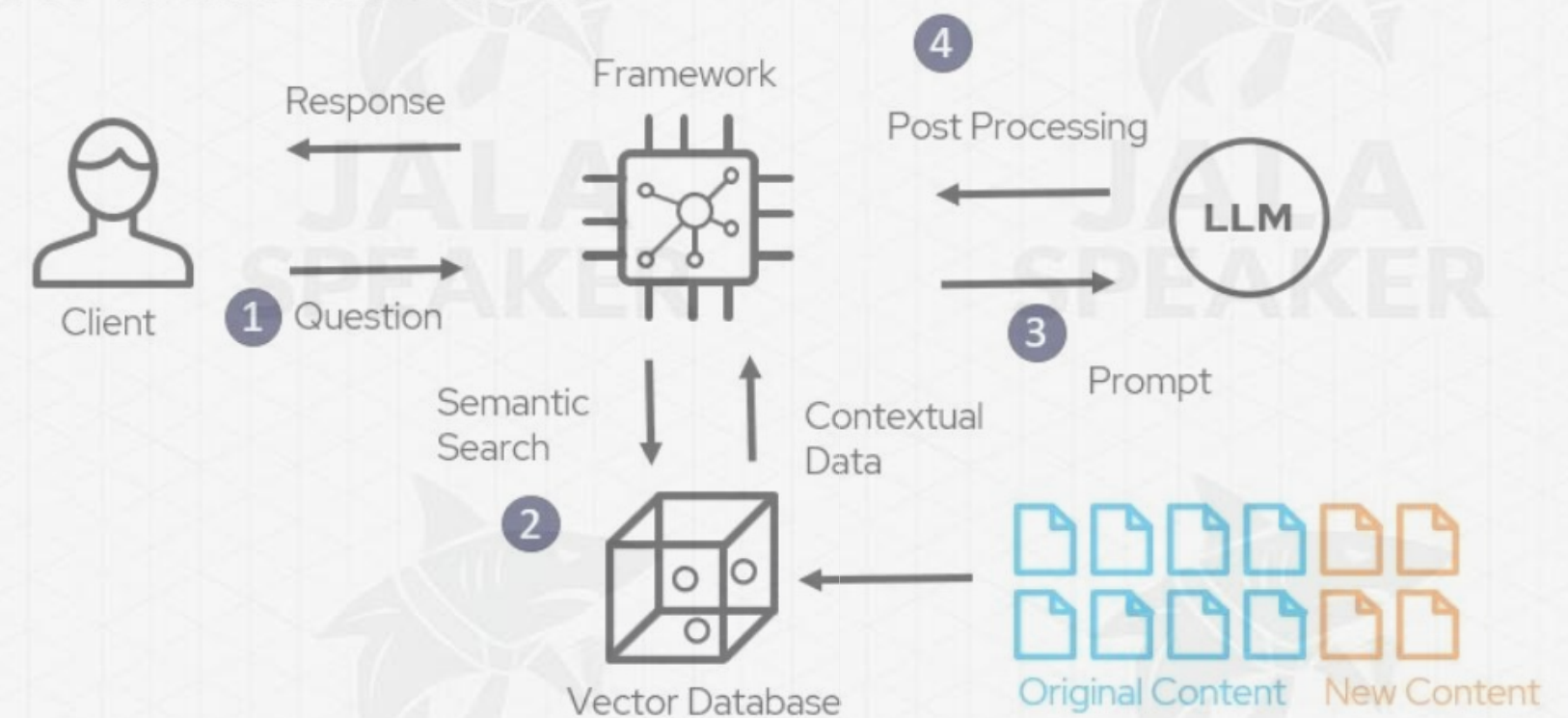
•Ventajas:

- Respuestas más precisas y actualizadas.
- Capacidad de manejar conocimiento fuera del alcance del modelo base.

•Ejemplo de Uso:

Un asistente virtual de soporte técnico que consulta manuales en tiempo real para responder preguntas específicas sobre productos.

RAG Architecture Model



Gracias



Firefly
Software Consulting