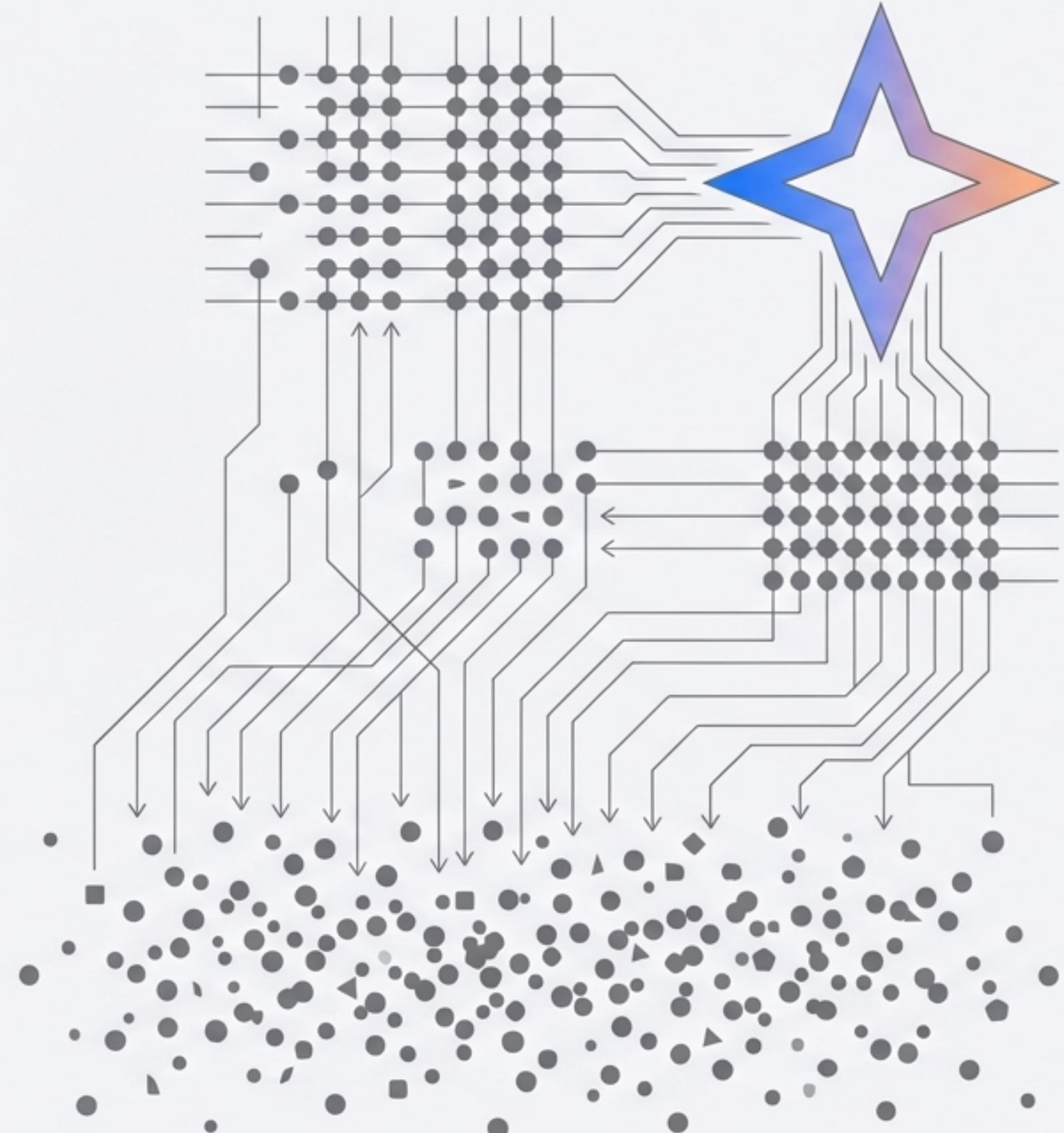


# Fundamentos de la IA Generativa

## Del Mecanismo al Despliegue

- El Motor: Arquitectura Transformer  
Atención, auto-atención y capas secuenciales.
- El Combustible: Inferencia y Probabilidad  
Muestreo, tokens y predicción probabilística.
- El Vehículo: Ecosistema Gemini  
API, Vertex AI y despliegue escalable.



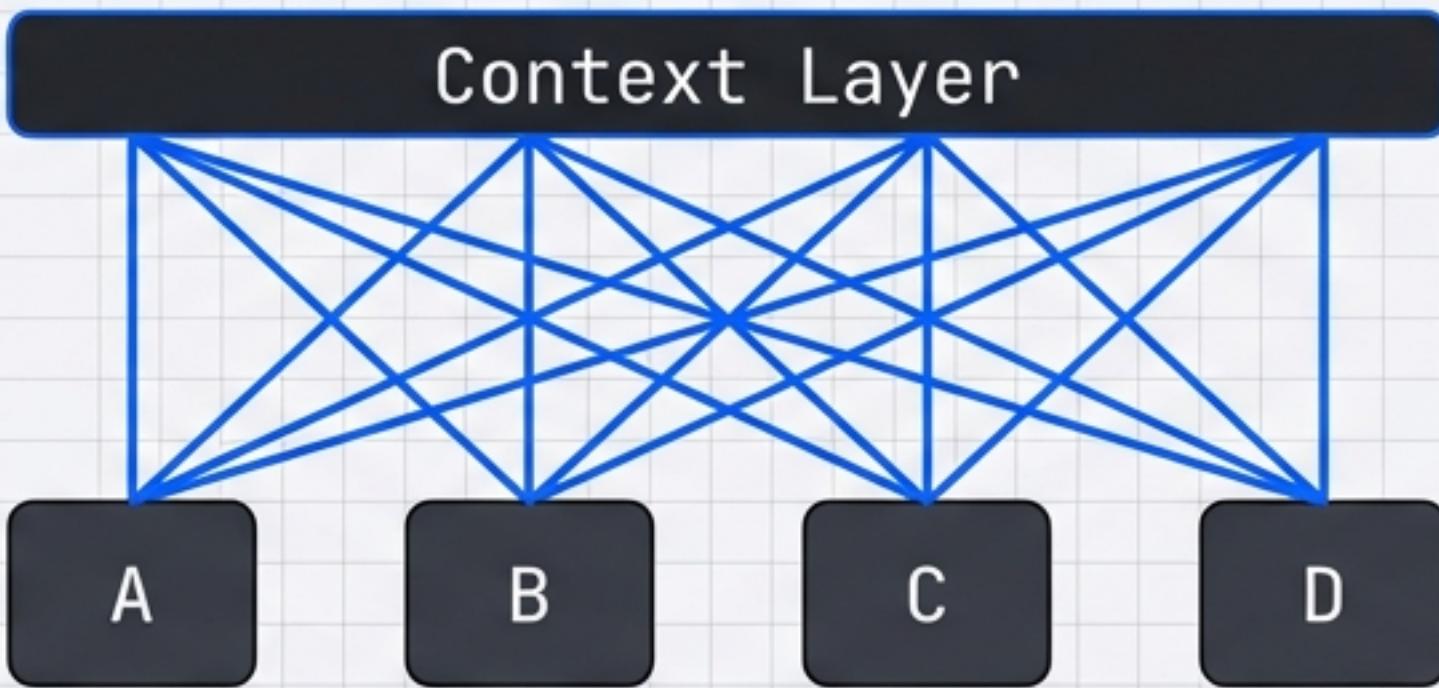
# El Cambio de Paradigma: De la Secuencia al Paralelismo

ANTERIOR: RNN / LSTM (Secuencial)



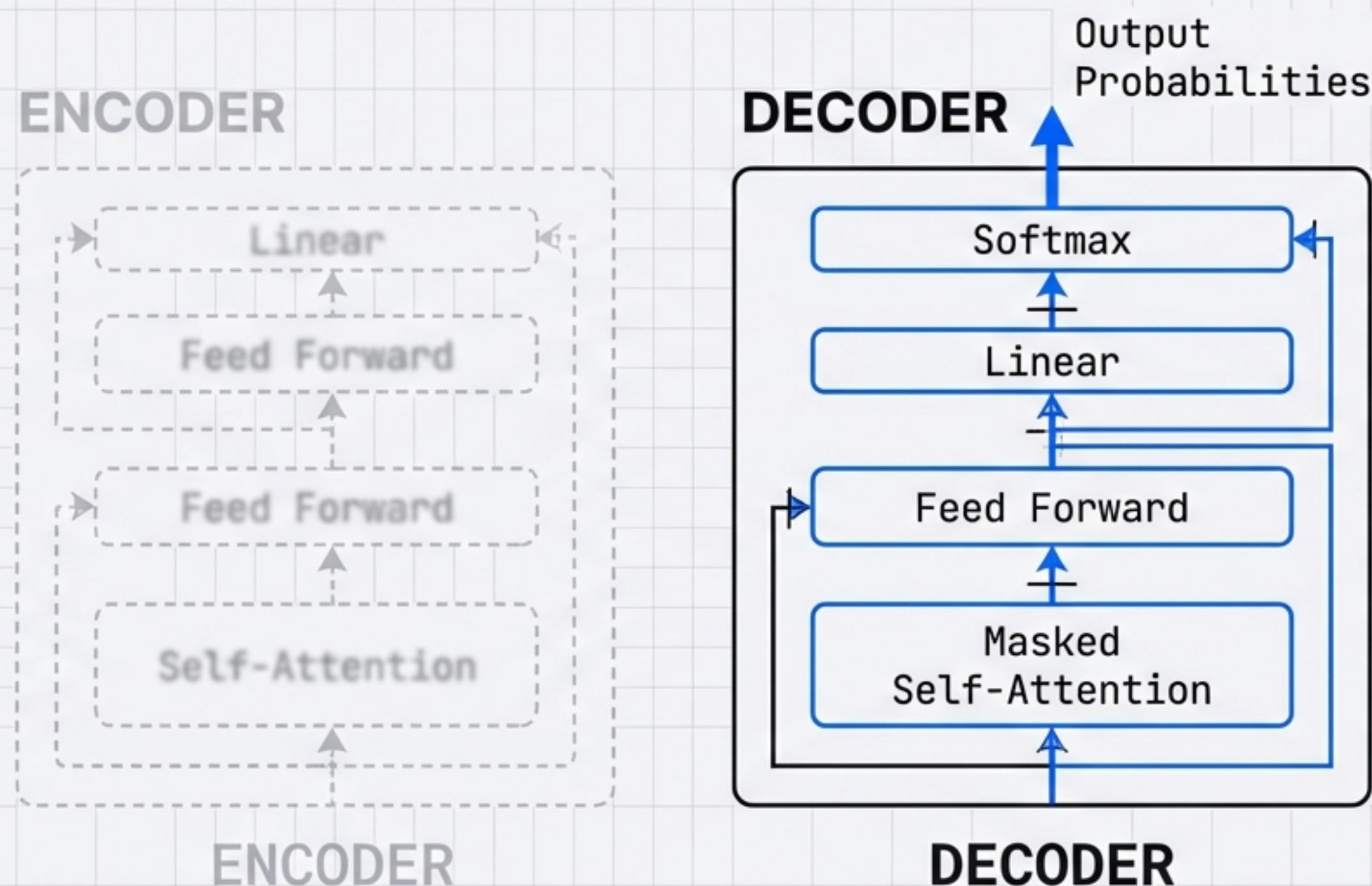
Cuello de botella secuencial. Pérdida de contexto en largas distancias.

NUEVO: TRANSFORMER (Paralelo)



Atención Global. Acceso simultáneo a toda la información.

# La Arquitectura: Encoder, Decoder y la Realidad de los LLMs



## Decoder-Only Architecture

La mayoría de los LLMs modernos (Gemini, GPT) descartan el Encoder. Se especializan en predecir el siguiente token basándose puramente en la secuencia generada previamente.

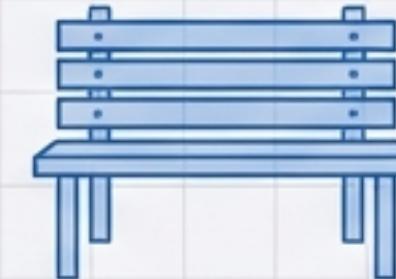
# El Corazón del Motor: Mecanismo de Self-Attention

Resolviendo la ambigüedad mediante contexto.

El gato se sentó en el tapete porque estaba cansado.



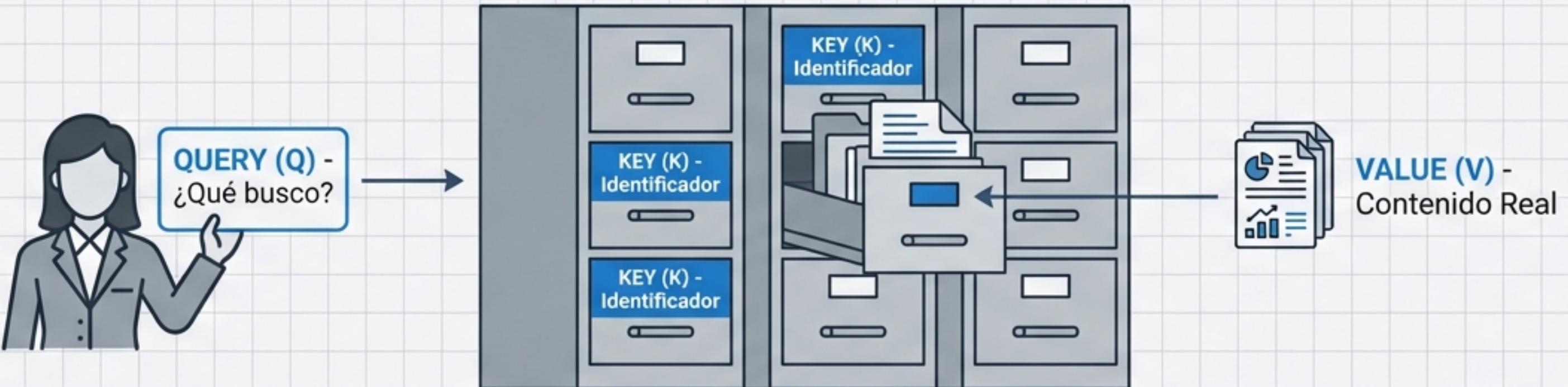
Banco (Institución)



Banco (Mueble)

Sin atención, la palabra 'banco' es solo un vector estático.  
Con atención, su valor cambia dinámicamente según sus vecinos.

# La Tríada Matemática: Query, Key y Value



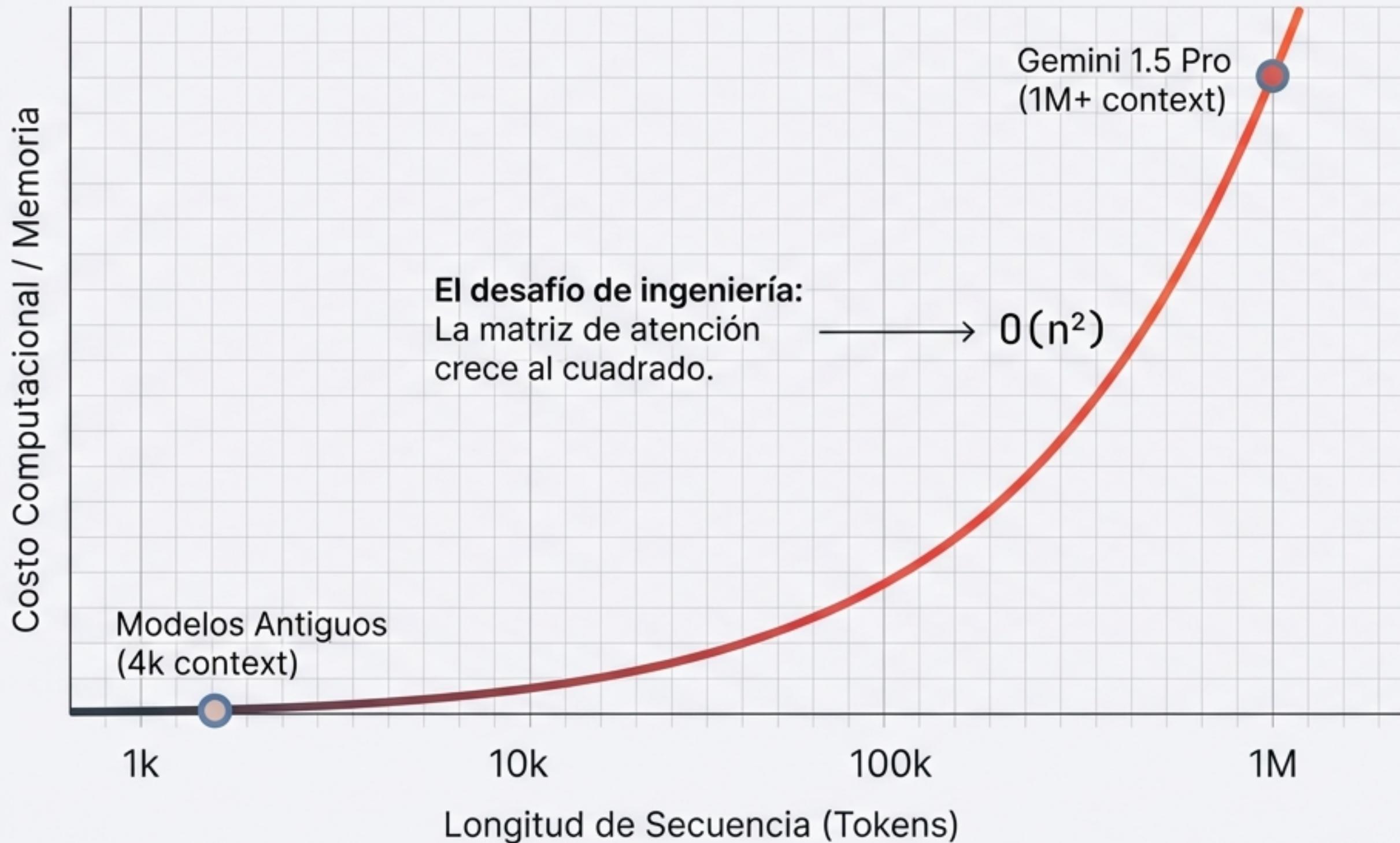
$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q} \cdot \mathbf{K}^T}{\sqrt{d_k}}\right) \cdot \mathbf{V}$$

Similitud (Producto Punto)

Normalización a Probabilidad (0-1)

Extracción de Información Ponderada

# La Realidad Técnica: Complejidad y Costos



## El Cuello de Botella

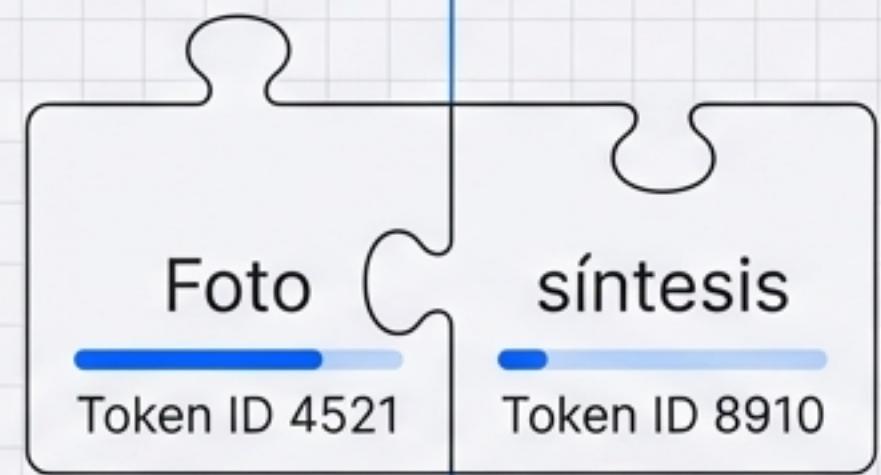
Cada vez que duplicamos la longitud del contexto, el trabajo se cuadriplica.

Técnicas como 'Flash Attention' y 'Sparse Attention' son necesarias para hacer viable la ventana de 1 millón de tokens.

# Inferencia: El Arte de la Predicción Autoregresiva

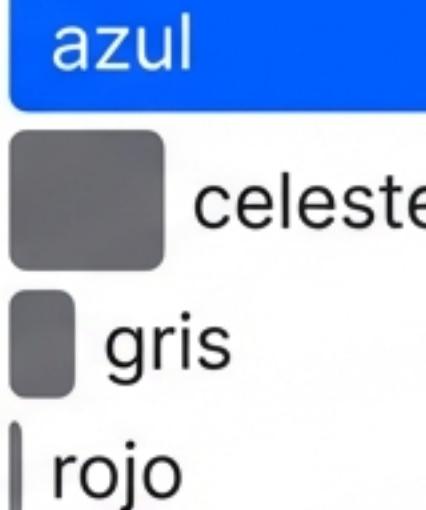
Tokenización (Sub-palabras)

Fotosíntesis



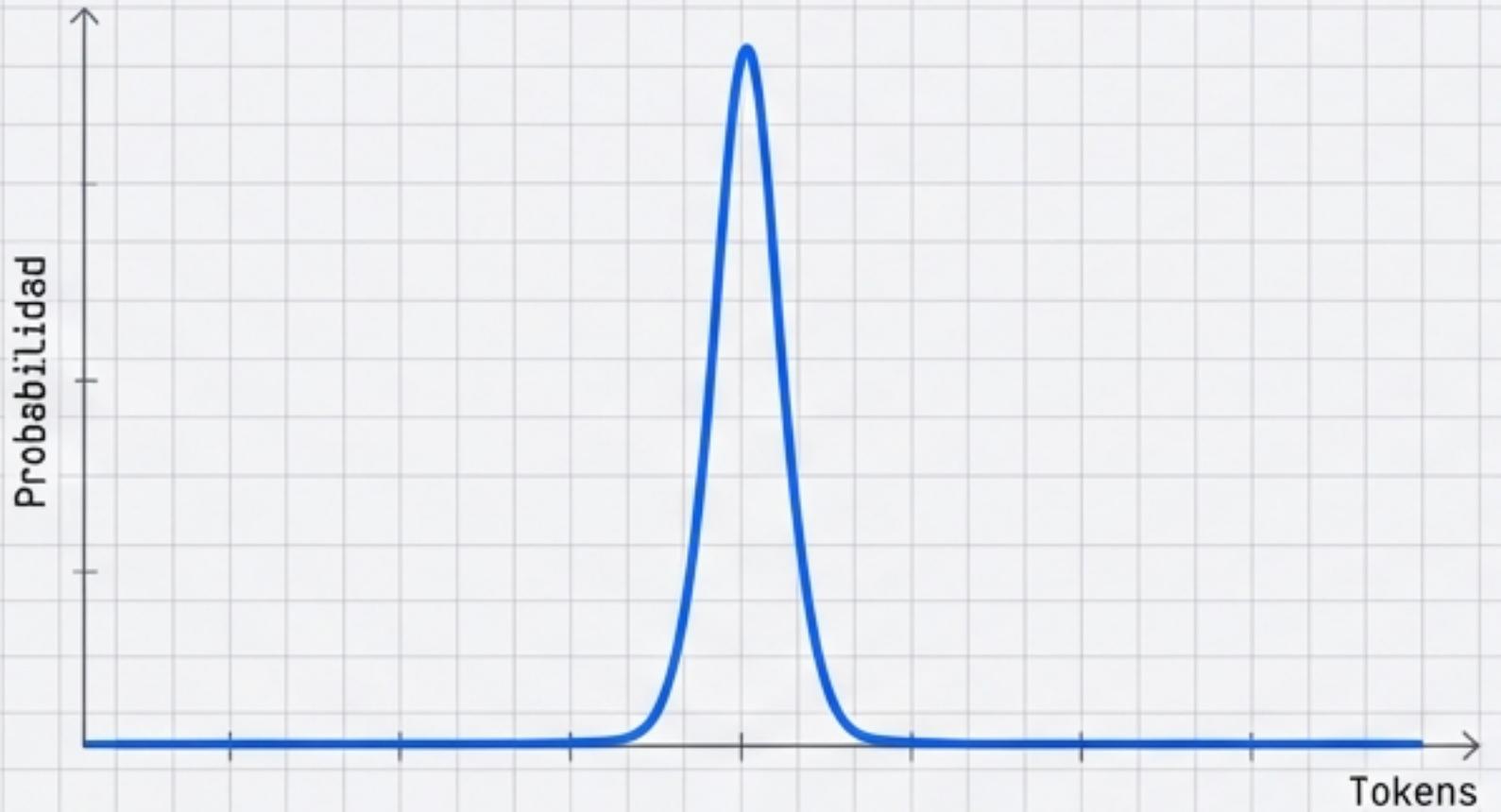
El cielo de verano es de color

Debug View



# Controlando la Entropía: Estrategias de Muestreo

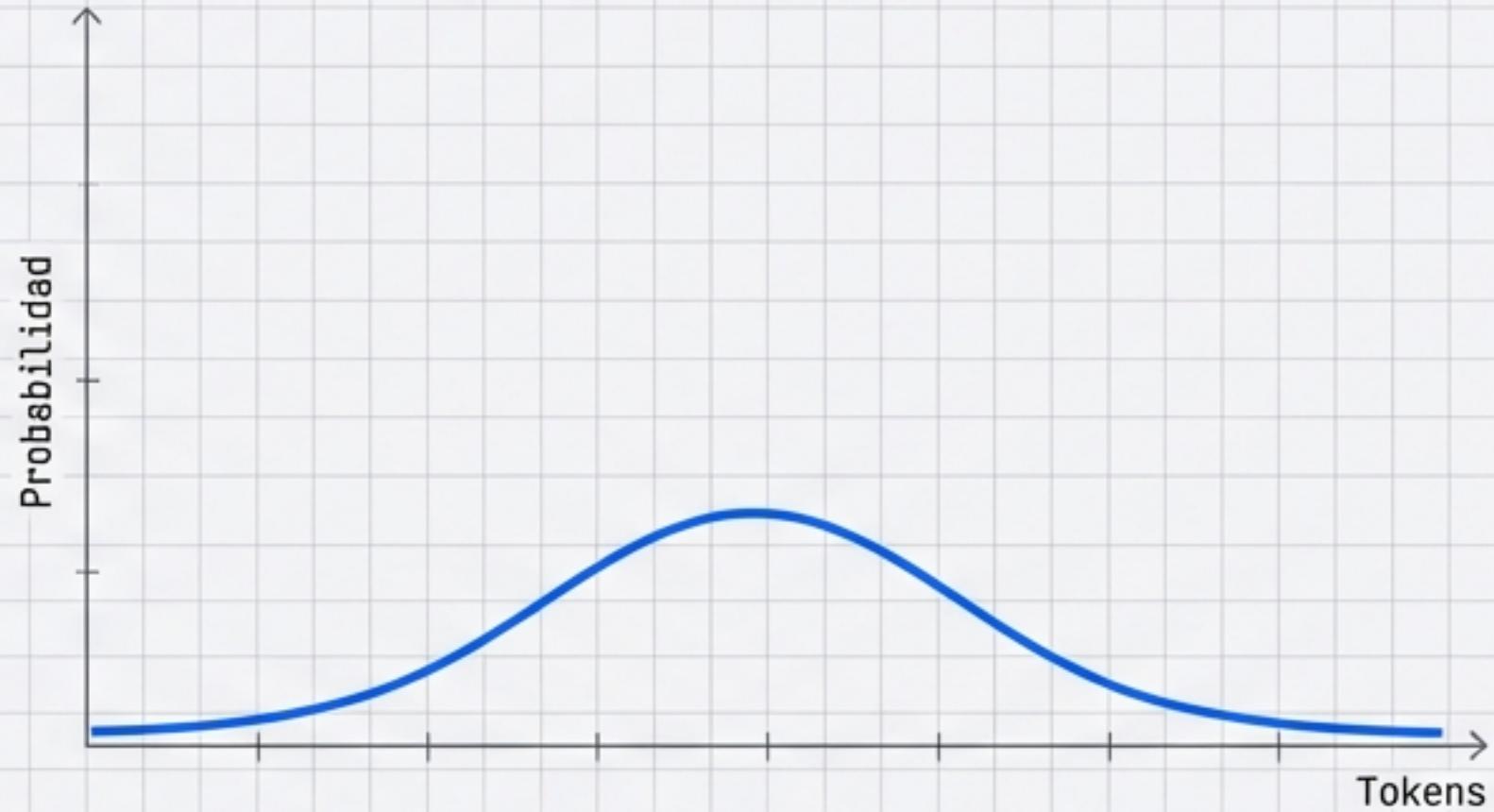
Temperatura Baja ( $< 0.5$ )



**Determinista / Preciso**

Ideal para: Código, Matemáticas, Datos Fácticos.

Temperatura Alta ( $> 0.8$ )



**Creativo / Diverso**

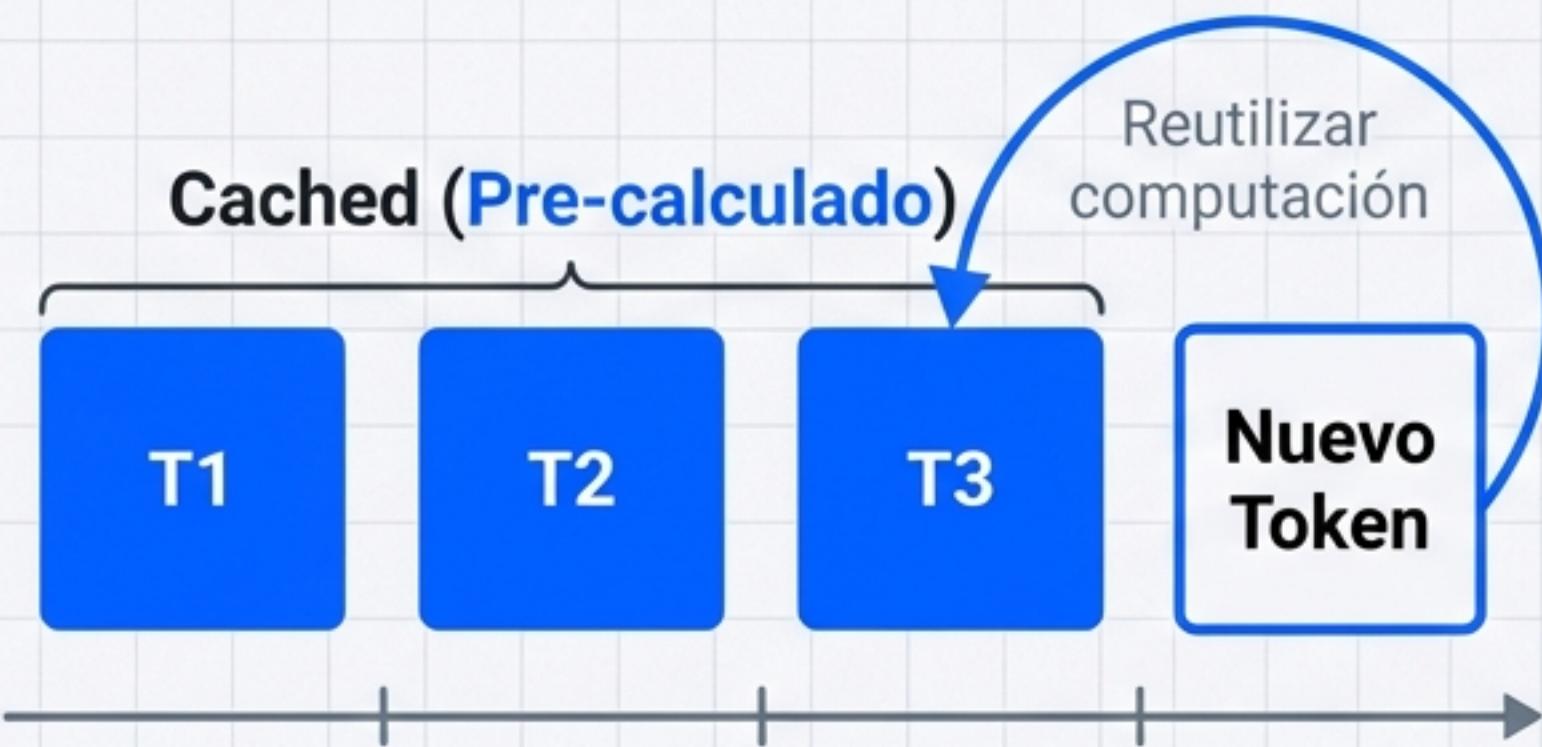
Ideal para: Lluvia de ideas, Escritura creativa, Poesía.

## Otras Palancas:

- **Top-k**: Limita la elección a las 'k' opciones más altas. (JetBrains Mono, Roboto)
- **Nucleus (Top-p)**: Elige del conjunto acumulado de probabilidad 'p'. (JetBrains Mono, Roboto)

# Ingeniería de Inferencia: Velocidad y Coherencia

## The KV-Cache Concept

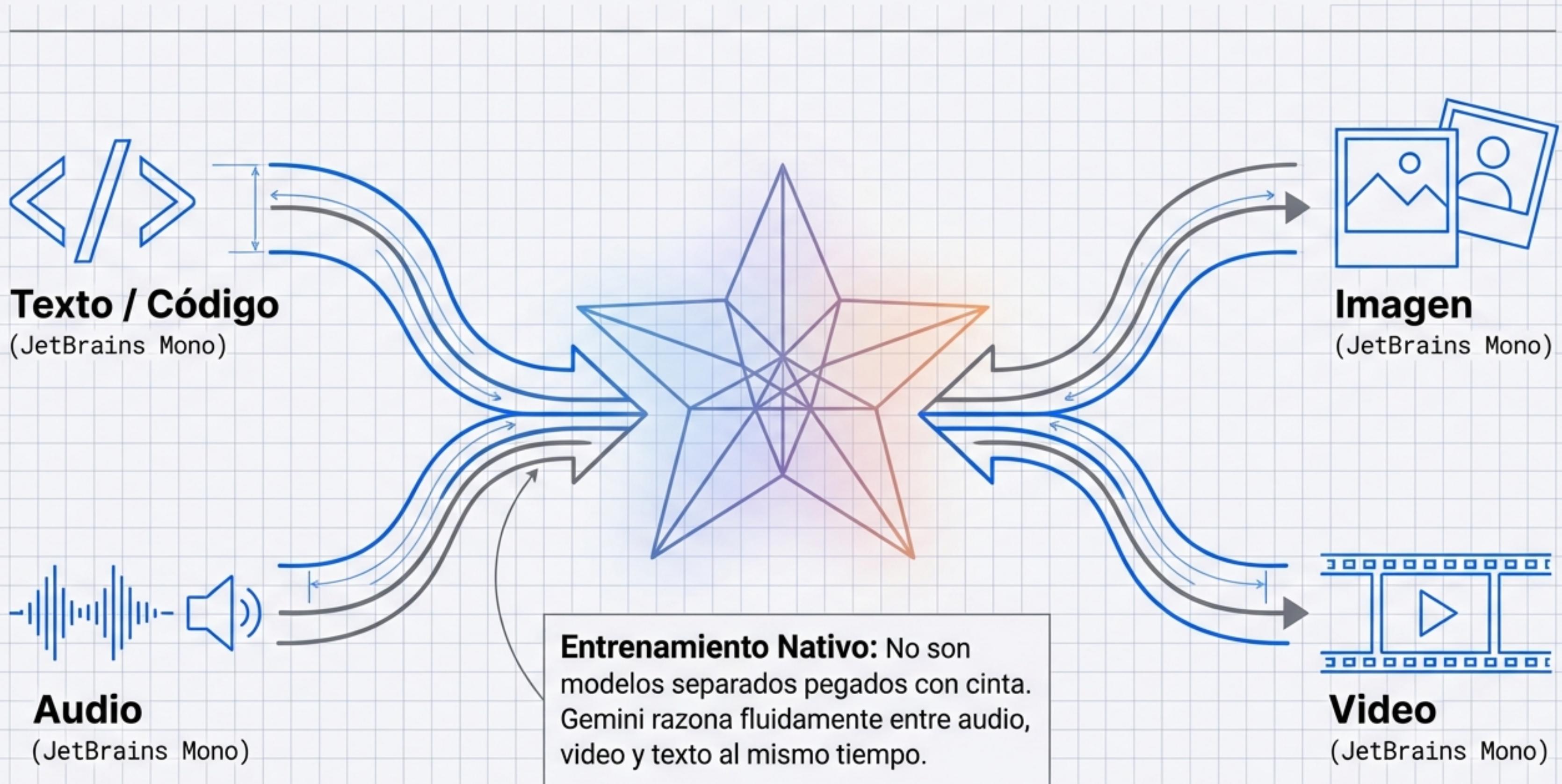


Reutilizamos la computación pasada.  
Costo  $O(1)$  por token generado.

## Mejores Prácticas

- ✓ Usar **KV-Cache** para reducir latencia.
- ✓ Aplicar **Máscaras Causales** (impedir ver el futuro).
- ✗ Asumir que el modelo tiene memoria entre sesiones (Es **Stateless**).

# El Ecosistema Gemini: Nativo Multimodal



# La Familia de Modelos: Eligiendo la Herramienta Correcta

Modelo / Icono	Rol	Especificaciones Técnicas	Mejores Usos
 <b>Gemini 1.5 Pro</b>	El Experto Senior	<ul style="list-style-type: none"><li>• 2M Contexto</li><li>• Razonamiento Complejo</li></ul>	Análisis profundo, contratos legales, coding complejo.
 <b>Gemini 1.5 Flash</b>	Alta Velocidad / Eficiencia	<ul style="list-style-type: none"><li>• 1M Contexto</li><li>• Baja Latencia</li></ul>	Alto volumen, extracción de datos, chatbots rápidos.
 <b>Gemini Nano</b>	Local / On-Device	<ul style="list-style-type: none"><li>• Privacidad Total</li><li>• Sin Internet</li></ul>	Tareas en móvil, sugerencias de teclado, resumen offline.

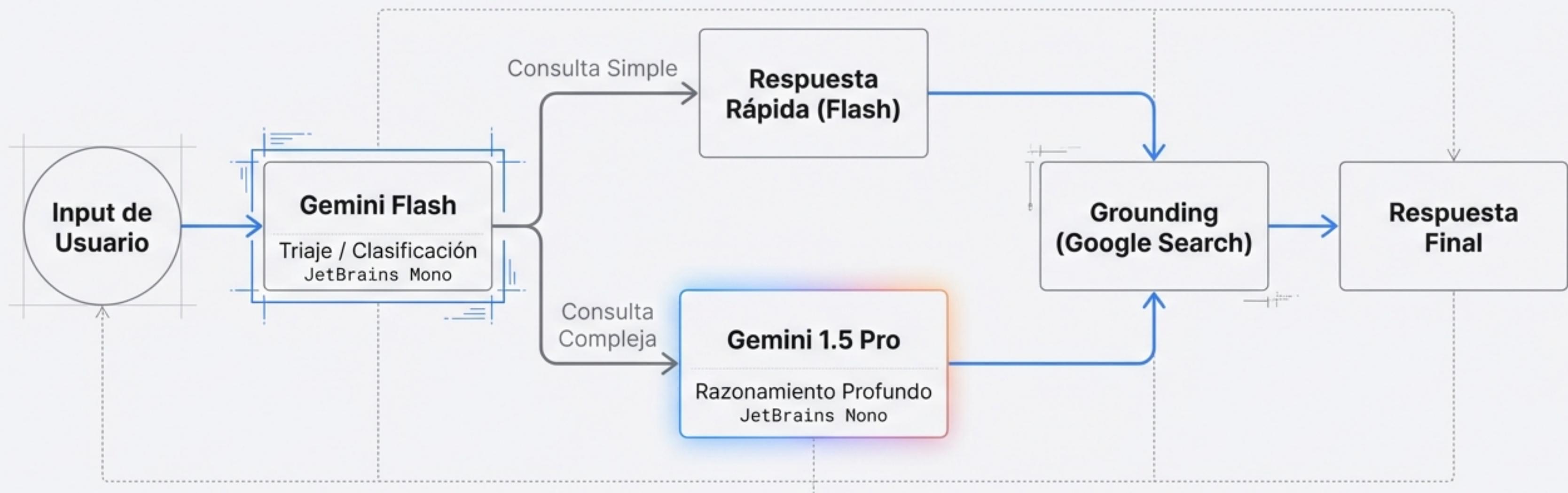
# Superpoderes: Ventana de Contexto Infinita

**1 a 2 Millones de Tokens. ¿Qué significa esto?**



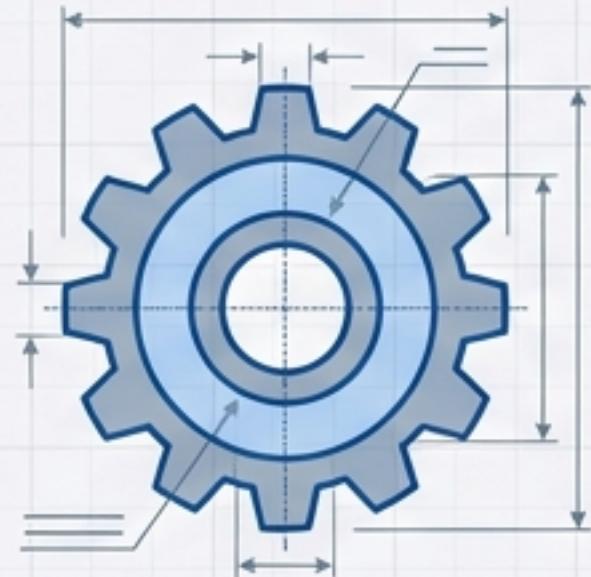
Cambio de Paradigma RAG: Ya no es siempre necesario fragmentar documentos (chunking). Puedes cargar libros enteros o repositorios de código en la memoria activa del modelo ('In-Context Learning').

# Arquitecturas de Producción: Estrategia por Capas (Tiering)



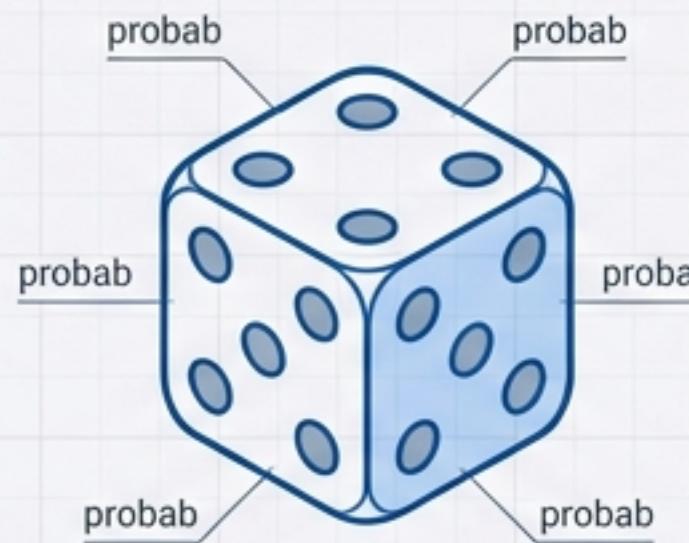
Optimización de Costos: Usar modelos ligeros para decidir y modelos potentes solo para resolver problemas difíciles.

# Resumen y Siguientes Pasos



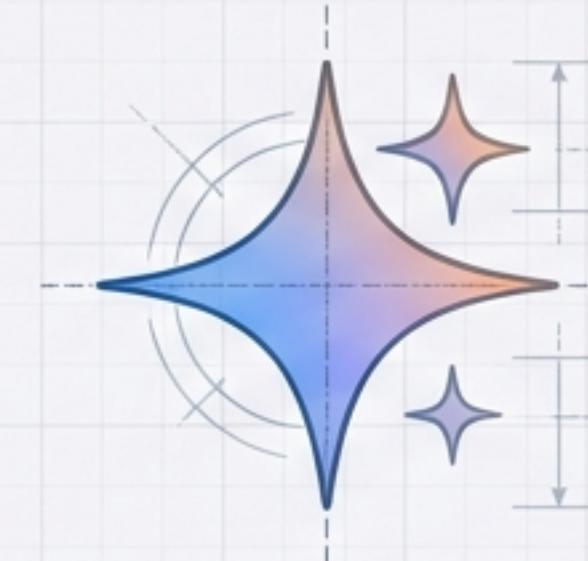
## Arquitectura

El Transformer usa atención paralela. El costo es  $O(n^2)$ ; gestiona tu contexto.



## Inferencia

Los LLMs son probabilísticos. Usa la temperatura para controlar precisión vs. creatividad.



## Gemini

Nativo multimodal. Elige entre Flash (velocidad) y Pro (inteligencia) según el caso.

**Siguiente Módulo (0.2.1): Configuración de Google Cloud y API Keys.**