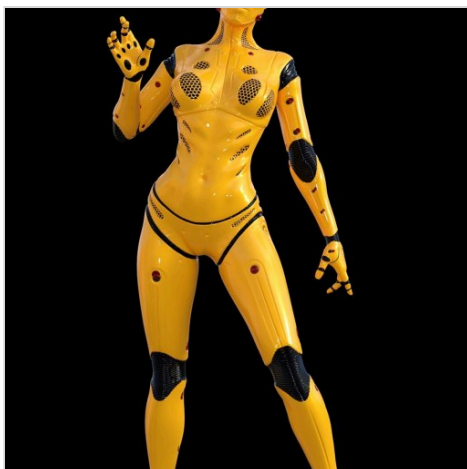


L U M B R E



GPT y modelos autorregresivos

Grado: 8

Materia: Transformers y Modelos de Atención · Inteligencia Artificial

11/10/2026

Antes de empezar

GPT entrena un decodificador de atención para predecir el siguiente token dado el contexto previo. Esa misma habilidad permite generar texto, sujeta a la estrategia de decodificación que se elija.

QUÉ APRENDERÁS

- Describir el modelado autorregresivo de lenguaje de GPT y las estrategias de decodificación para generar texto.

GPT y modelos autorregresivos

Generar es predecir el siguiente

GPT aprende una única habilidad: dada una secuencia, ¿qué token sigue? Esa pregunta, repetida miles de millones de veces, basta para escribir, resumir o conversar. Generar es predecir en bucle.

VIDEO

La IA generativa explicada

Vídeo en español que explica cómo los modelos autorregresivos producen texto prediciendo el siguiente token.



Escanea para verlo en Lumbre

<https://www.youtube.com/watch?v=CoaSZLHxVL8>

Modelado autorregresivo

El modelo factoriza la probabilidad de una secuencia como producto de condicionales: cada token depende de los anteriores. Entrenar es maximizar la verosimilitud de esa cadena.

Solo decodificador, con máscara causal

- GPT apila capas de decodificador con autoatención enmascarada.

- La máscara impide mirar el futuro: solo ve el prefijo.
- Eso lo hace apto para generar de izquierda a derecha.

Entrenamiento paralelo

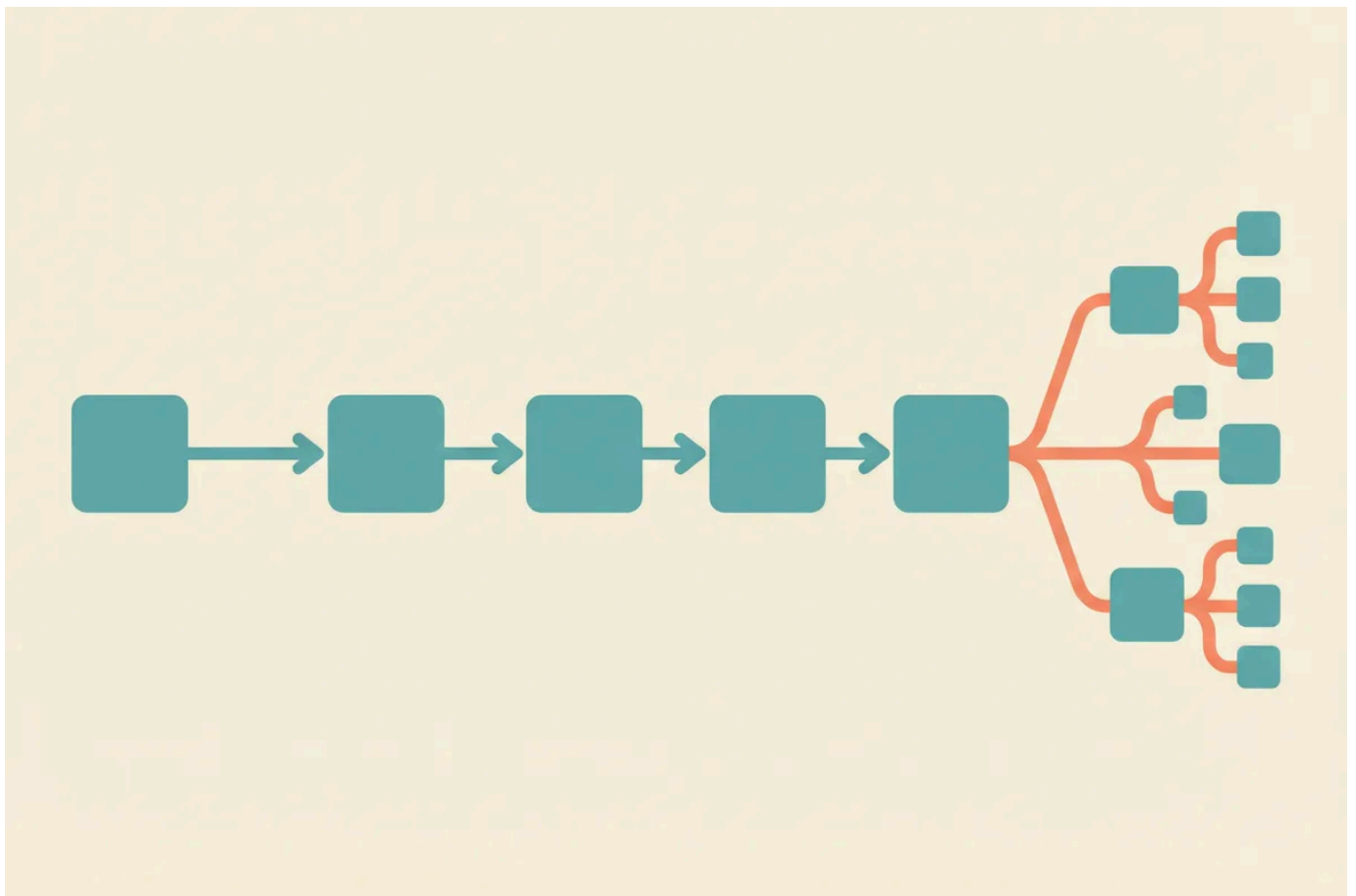
Aunque genera token a token, el entrenamiento es paralelo: con docencia forzada se predicen todas las posiciones a la vez sobre la secuencia objetivo desplazada.

Preentrenar y afinar

Primero se preentrena con enormes corpus para aprender la lengua; después se afina o alinea para seguir instrucciones y reducir toxicidad.

De la probabilidad al texto: decodificar

El modelo devuelve una distribución sobre el vocabulario. Elegir cuál token emitir es la decodificación, y define el carácter del texto generado.



Cada token generado se suma al contexto y abre un abanico ponderado de candidatos para el siguiente.

Voraz (greedy)

Toma siempre el token más probable. Es determinista y barato, pero se repite y deriva en texto repetitivo o sin gracia.

Haz (beam search)

Explora las varias secuencias más prometedoras a la vez y elige la mejor al final. Mejora la plausibilidad, pero puede sonar genérica.

Muestreo y temperatura

- Muestrear según la probabilidad introduce variedad.
- La temperatura aplana o afila la distribución: alta, más creativa; baja, más conservadora.
- Temperatura cero equivale a decodificación voraz.

Muestreo por núcleo (top-p)

En vez de fijar los k más probables, se muestrea del menor conjunto de tokens cuya probabilidad acumulada supera p. Se adapta a la incertidumbre de cada paso.

Penalizar repeticiones

Se pueden penalizar tokens ya usados para reducir bucles. Es un parche útil, aunque no sustituye a un modelo bien entrenado.

Coste de la generación

Generar es secuencial: cada token exige una pasada completa, y la memoria de atención crece con la longitud. Por eso la inferencia es el gasto dominante.

Alucinación y límites

Al optimizar verosimilitud del siguiente token, el modelo suena fluido sin garantizar verdad. La generación competente no equivale a conocimiento fiable.

Errores frecuentes

- Confundir temperatura alta con creatividad infinita: pasarse produce texto incoherente y fuera de tema.
- Dar por cierto lo que suena fluido: el modelo optimiza verosimilitud del siguiente token, no verdad.

- Ignorar el coste secuencial de la generación y dimensionar la infraestructura como si fuera paralela.

Ejemplo resuelto: ajustar temperatura y top-p para un chatbot

Un asistente de soporte responde plano y repetitivo con decodificación voraz; la solución está en calibrar el muestreo.

1. Mides primero con temperatura cero para tener una línea base determinista y comparable.
2. Subes la temperatura a 0,7 y observas si gana naturalidad sin perder precisión técnica.
3. Activas top-p en 0,9 para recortar la cola de tokens disparatados que la temperatura alta admite.
4. Añades penalización de repeticiones leve para romper los bucles frecuentes en respuestas largas.
5. Evalúas con conversaciones reales: si aparecen datos inventados, bajas temperatura y añades verificación de fuentes.

¿Para qué sirve en la realidad?

Chatbots, copilotos de programación, generación de copy publicitario y resúmenes automáticos son modelos autorregresivos en producción; entender la decodificación es la diferencia entre un producto que encanta y uno que alucina.

1. En la decodificación de un modelo autorregresivo, bajar la temperatura a cero equivale a:

- ☐ muestrear uniformemente del vocabulario
- ☐ elegir siempre el token más probable (voraz)
- ☐ usar haz de tamaño uno pero creativo
- ☐ penalizar todas las repeticiones

2. Un modelo autorregresivo entrenado con docencia forzada puede calcular las pérdidas de todos los tokens de una secuencia en paralelo.

- ☐ Verdadero
- ☐ Falso

ACTIVIDAD INTERACTIVA DISPONIBLE EN LUMBRE

3. GPT**4. Ordena el proceso de generación en un modelo tipo GPT:****Elementos:**

- alimentar el contexto generado hasta ahora
- obtener la distribución del siguiente token
- elegir un token según la estrategia
- anexar el token y repetir

Prefijo	. . .
Distribución	. . .
Decodificar	. . .
Añaden	. . .

Beam search https://en.wikipedia.org/wiki/Beam_search Búsqueda haz para decodificar secuencias probables.

5. Une cada estrategia de decodificación con su comportamiento.

5.	elegir siempre el token más probable, determinista y repetitivo	-----
5.	mantener varias secuencias candidatas y elegir la mejor al final	-----
5.	parámetro que aplana o afila la distribución antes de muestrear	-----
5.	muestrear del menor conjunto de tokens con probabilidad acumulada p	-----

Opciones: voraz · beam search · temperatura · nucleus top-p

Predecir en bucle es generar

Modelos autorregresivos

Modelo

solo decodificador

mascara causal

Entrenar

verosimilitud

docencia forzada paralela

Decodificar

voraz y haz
temperatura y top-p
Riesgos
alucinacion
repeticion
coste secuencial

El mapa mental interactivo es interactivo en la versión online.



GPT-3 <https://en.wikipedia.org/wiki/GPT-3>El modelo que mostró el salto de escala del enfoque autorregresivo.

6. Tu chatbot alucina datos con total soltura. Explica por qué el objetivo de predecir el siguiente token no garantiza veracidad y qué dos medidas (decodificación, verificación) ayudarían.

Referencia curricular

Grado:	8
Materia:	Transformers y Modelos de Atención · Inteligencia Artificial

Versión interactiva completa (con actividades, videos y autoevaluación):



<https://univ.lumbre.lat/object/cdia-tranf-005-gpt-y-modelos-autorregresivos>

Material educativo generado desde Lumbre · lumbre.lat