LORA Y QLORA

START

HOW TO PLAY

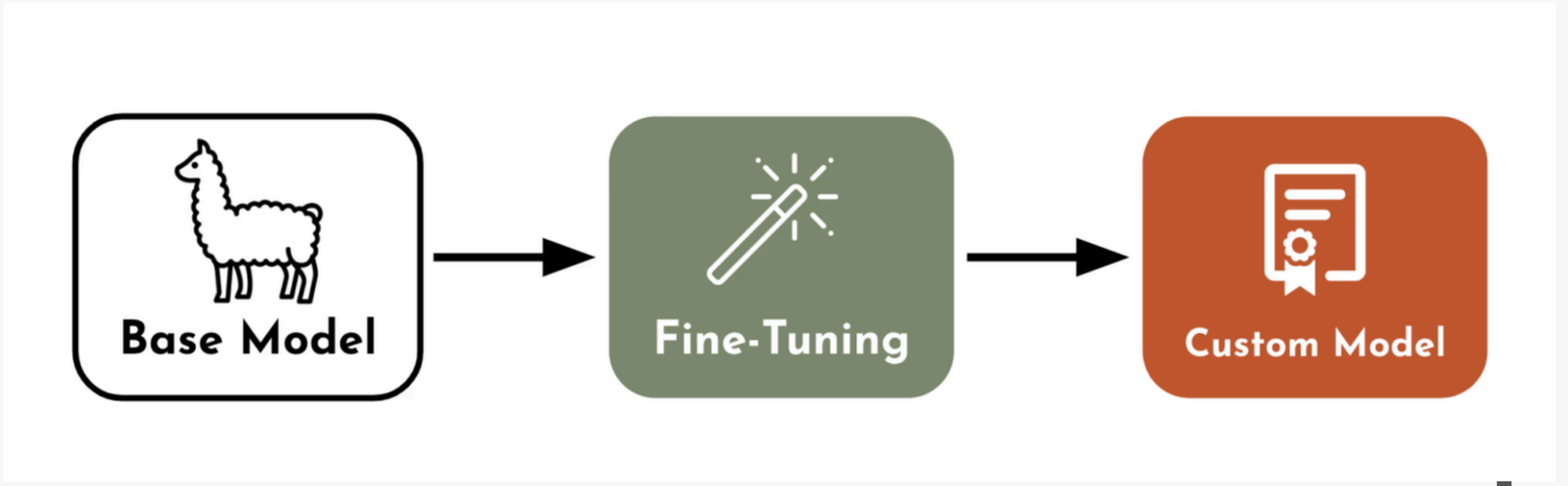
THERE ARE 3 DIFFERENT TYPES OF GAMES!

LOS LLMS HAN REVOLUCIONADO LA IA, PERO SU ENTRENAMIENTO Y AJUSTE FINO (FINE-TUNING) REQUIERE ENORMES RECURSOS.

PARA ADAPTARLOS A TAREAS ESPECIFICAS (TRADUCCIÓN, CHATBOTS, CLASIFICACIÓN, ETC.), NECESITAMOS METODOS MAS LIVIANOS Y EFICIENTES.

AQUÍ SURGEN LORA Y QLORA, QUE PERMITEN REENTRENAR MODELOS GIGANTES USANDO GPUS COMUNES.

FINE-TUNING



REENTRENAR UN MODELO PREENTRENADO CON NUEVOS DATOS ESPECÍFICOS

LORA (LOW-RANK ADAPTATION)

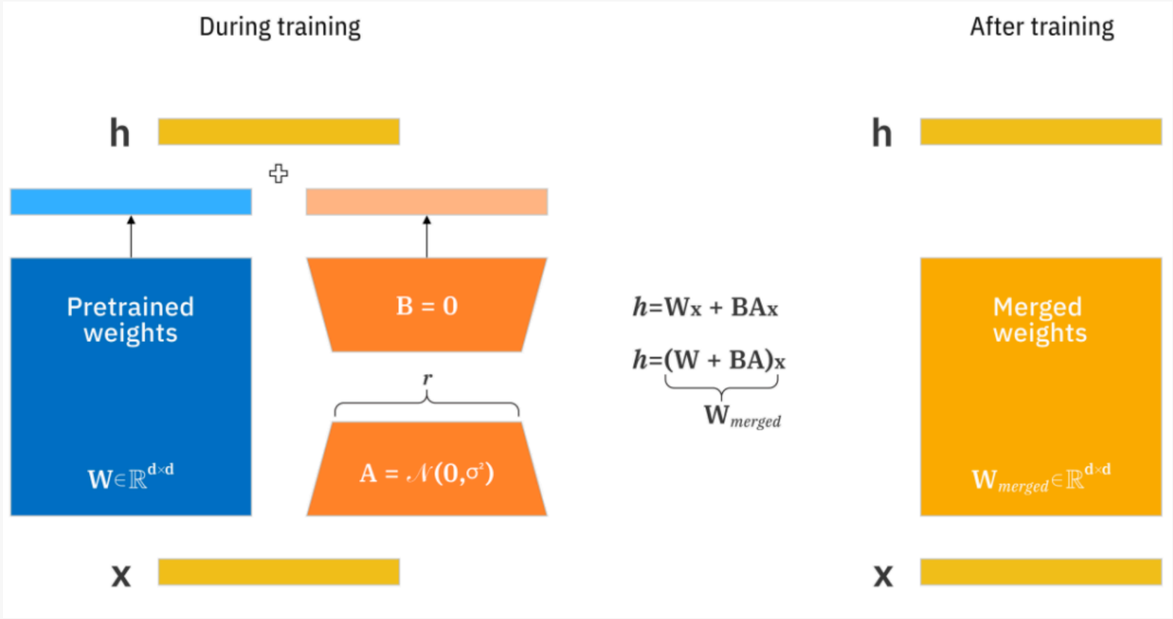
ANSWER THE QUESTION

EN LUGAR DE MODIFICAR TODOS LOS PESOS, SOLO SE ENTRENAN DOS PEQUEÑAS MATRICES A Y B, ADAPTADORES DE BAJO RANGO

A. MATRICES A Y B

LOS PARÁMETROS WPERMANECEN CONGELADOS, MIENTRAS QUE SOLO SE OPTIMIZAN AY B.

LOS PESOS ORIGINALES SE CONGELAN, Y SE AJUSTA UN SUBCONJUNTO, ¸REDUCIENDO DRÁSTICAMENTE LA MEMORIA Y TIEMPO.



-8	-2	-6	6
-4	-1	-3	3
28	7	21	-21
24	6	18	-18

=

-2
-1
7
6

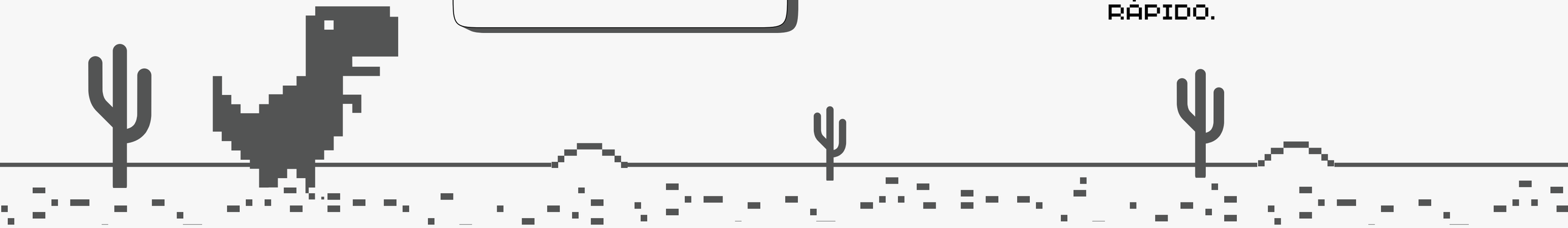
X

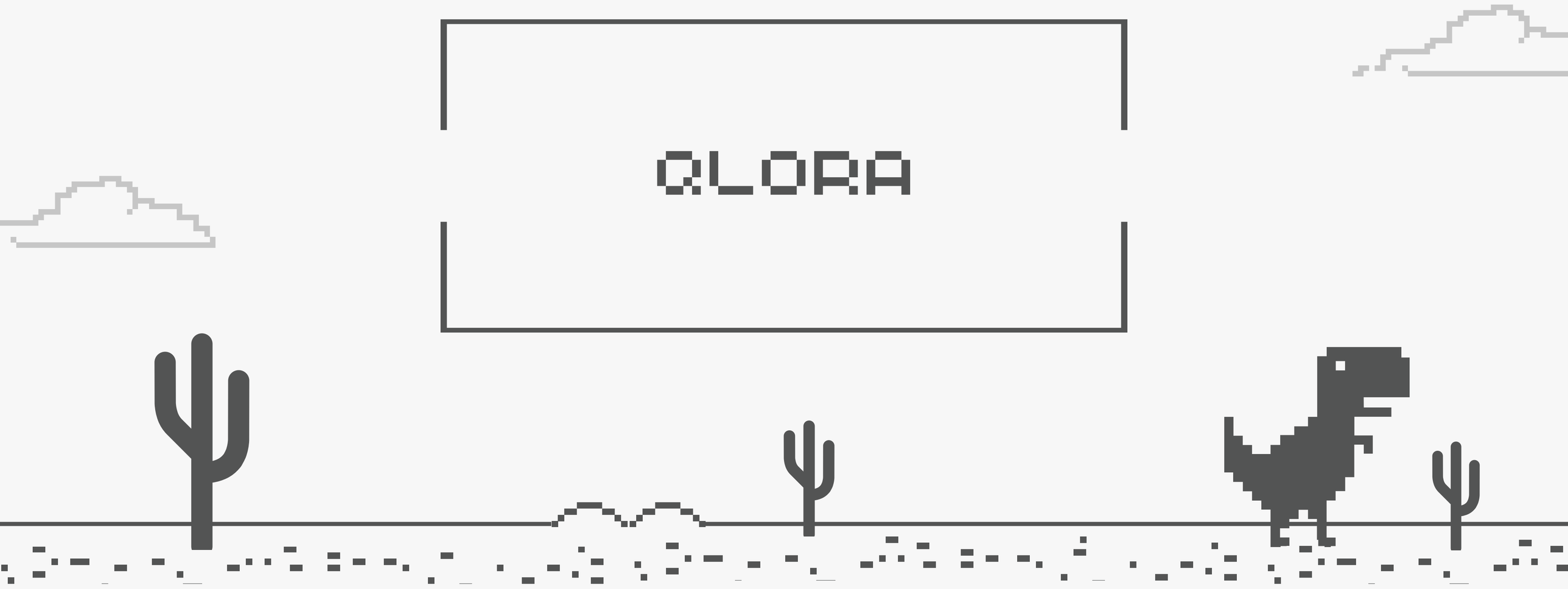
4	1	3	-3

SI UN MODELO TIENE MILLONES DE PARÁMETROS, LORA ENTRENA SOLO UN 1–2% DE ELLOS.

BENEFICIO CLAVE

BENEFICIO CLAVE: MISMO RENDIMIENTO, PERO CON MUCHA MENOS VRAM Y ENTRENAMIENTO MÁS RÁPIDO.




QLORA

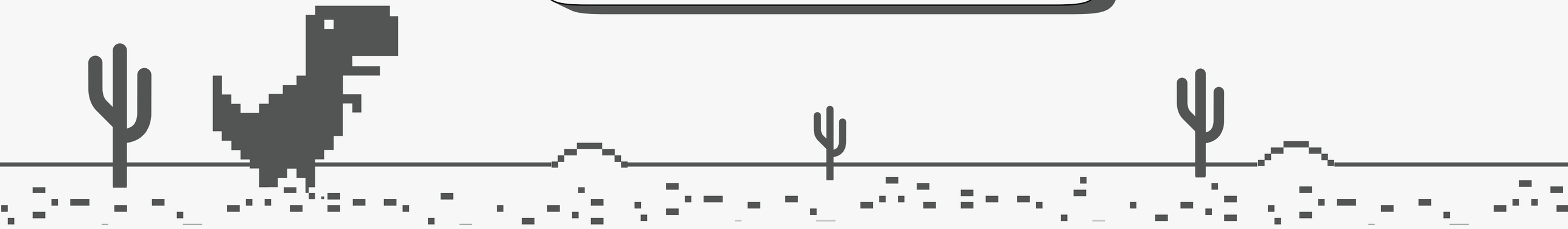
MEJORA AÚN MÁS A LORA

QLORA (QUANTIZED LOW-RANK ADAPTATION)

CUANTIZA EL MODELO BASE A 4 BITS (NF4) PARA REDUCIR MEMORIA X4.

SOLO ENTRENA LOS ADAPTADORES,LORA EN PRECISIÓN MÁS ALTA (FLOAT16 O BFLOAT16).

PUEDE AJUSTAR MODELOS DE HASTA 65 MIL MILLONES DE PARÁMETROS CON UNA SOLA GPU DE CONSUMO (6–24GB).



CAPA LINEAL TÍPICA $y = Wx$

EN LORA, LA MATRIZ DE PESOS SE PARAMETRIZA COMO: $W' = W + \Delta W$ DONDE $\Delta W = AB$

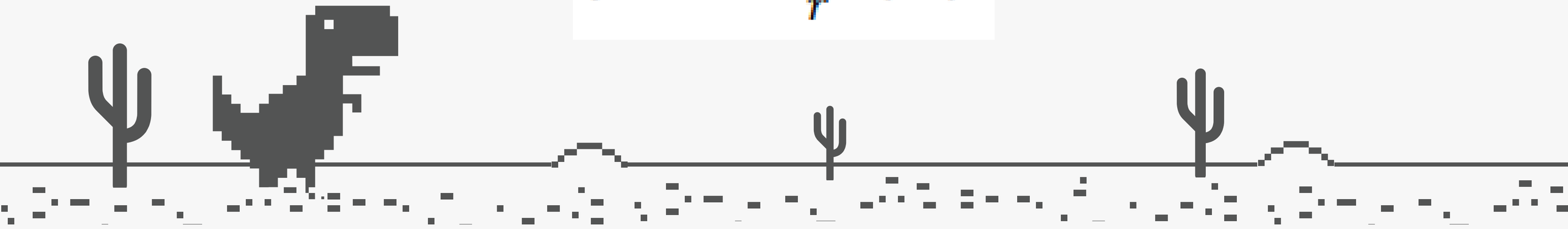
A ES MATRIZ DE PROYECCIÓN DESCENDENTE, DONDE REDUCE LA DIMENSIONALIDAD DE LA ENTRADA x A UN ESPACIO DE DIMENSIÓN r.

B ES MATRIZ DE PROYECCIÓN ASCENDENTE, DONDE RECONVIERTE ESA REPRESENTACIÓN COMPRIMIDA AL TAMAÑO DE SALIDA ORIGINAL.

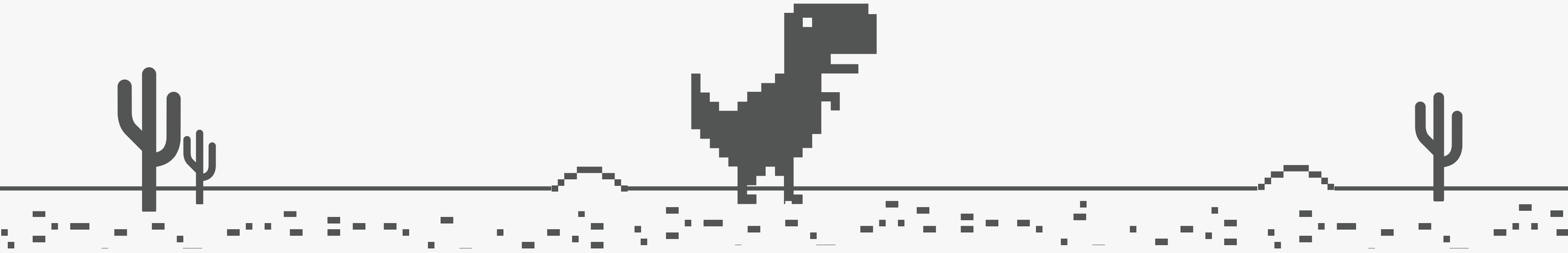
$A \in \mathbb{R}r\text{X}d\text{IN}$ r EL RANGO BAJO ($r \ll$ MIN (dIN, dOUT)).

$B \in \mathbb{R}d\text{OUT}\text{X}r$,

$$y = Wx + \frac{\alpha}{r} B(Ax).$$

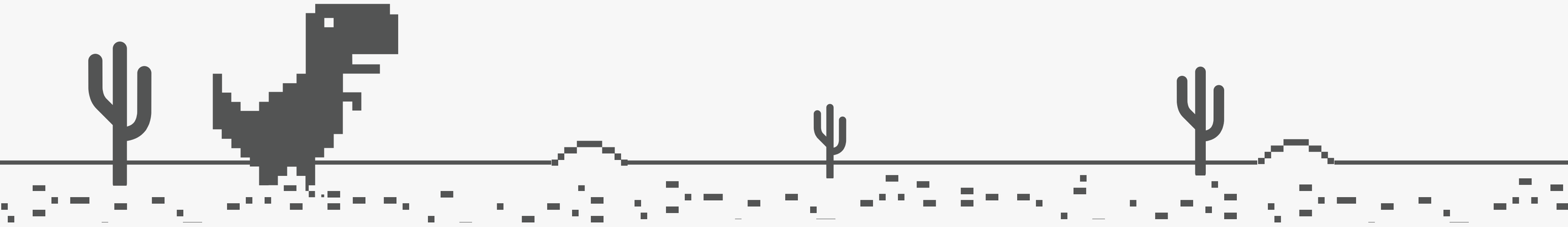


EJERCICIOS



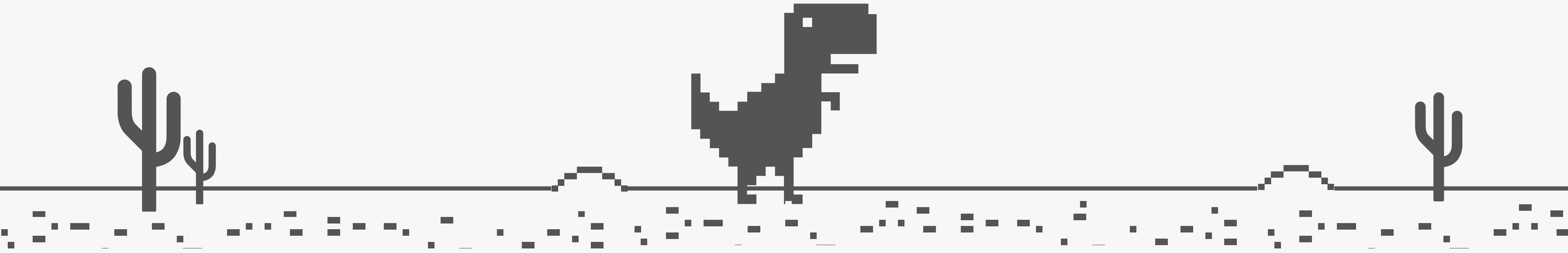
PRUEBAS DE RENDIMIENTO

4050 6GB de VRAM			
Tecnica	Epocas	Tiempo de entrenamiento	Uso de Memoria
LoRA	3	1hr	5.8 Gb
QLoRA	3	30 minutos	0.8 Gb



[SEAN GENTILES XD]

¿PREGUNTAS?



GAME OVER
GRACIAS