

REIVINDICACIONES

1. Un sistema de procesamiento, en donde comprende:

5 al menos una memoria que tiene instrucciones ejecutables almacenadas en esta; y
uno o más procesadores configurados para ejecutar las instrucciones ejecutables para hacer que el sistema de procesamiento:

10 reciba una pluralidad de conjuntos de tokens generados basados en un indicador de entrada y un primer modelo de inteligencia artificial generativa, cada conjunto de tokens en la pluralidad de
15 conjuntos de tokens comprende una secuencia de tokens correspondiente a una respuesta candidata al indicador de entrada;

seleccione, usando un segundo modelo de
20 inteligencia artificial generativa y ajuste recursivo de una distribución objetivo asociada con la pluralidad recibida de conjuntos de tokens, un conjunto de tokens de la pluralidad de
25 conjuntos de tokens; y

emita el conjunto de tokens seleccionado como una respuesta al indicador de entrada.

5 **2.** El sistema de procesamiento de la reivindicación 1, en donde la pluralidad de conjuntos de tokens se organiza en una estructura de datos de árbol.

3. El sistema de procesamiento de la reivindicación 2, en donde un nodo raíz de la estructura de datos de árbol corresponde al indicador de entrada y en
10 donde cada ruta a través de la estructura de datos de árbol corresponde a una secuencia diferente de tokens correspondiente a la respuesta candidata al indicador de entrada.

4. El sistema de procesamiento de la
15 reivindicación 2, en donde:

 la estructura de datos de árbol incluye una pluralidad de niveles, cada nivel correspondiente a un token en la secuencia de tokens, y
20 un número de tokens en un nivel particular en la estructura de datos de árbol está basado en un factor de ramificación asociado con un nivel inmediatamente anterior al nivel particular en la estructura de datos de árbol.

25 **5.** El sistema de procesamiento de la reivindicación 2, en donde una profundidad de la estructura

de datos de árbol corresponde a un parámetro que define un número máximo de tokens generados por un único pase a través del primer modelo de inteligencia artificial generativa.

5 **6.** El sistema de procesamiento de la reivindicación 1, en donde un tamaño de cada conjunto de tokens está basado en una métrica de complejidad computacional asociada con generar un conjunto de tokens objetivo por el segundo modelo de inteligencia artificial generativa.

10 **7.** El sistema de procesamiento de la reivindicación 1, en donde el ajuste recursivo de la distribución objetivo comprende:

 determinar si aceptar o rechazar un primer token en un conjunto de tokens de la pluralidad de conjuntos de tokens; y
15 ajustar una distribución de probabilidad usada para verificar un segundo token en el conjunto de tokens posterior al primer token basado en la determinación de si aceptar o
20 rechazar el primer token.

8. El sistema de procesamiento de la reivindicación 7, en donde para ajustar la distribución de probabilidad, el uno o más procesadores se configuran para hacer que el sistema de procesamiento reste un valor de
25 probabilidad asociado con el primer token de la

distribución de probabilidad basado en determinar rechazar el primer token.

9. El sistema de procesamiento de la reivindicación 1, en donde para seleccionar el conjunto de tokens de la pluralidad de conjuntos de tokens, el uno o más procesadores se configuran para hacer que el sistema de procesamiento:

rechace un primer token en un primer nivel de una estructura de datos de árbol que representa la pluralidad de conjuntos de tokens;
genere una distribución de probabilidad ajustada basada en el rechazo del primer token; descarte o ignore, de la estructura de datos de árbol, los tokens hijos del primer token en niveles más profundos que el primer nivel de la estructura de datos de árbol; y
determine si aceptar o rechazar un segundo token en el primer nivel de la estructura de datos de árbol basado en la distribución de probabilidad ajustada.

10. El sistema de procesamiento de la reivindicación 1, en donde para seleccionar el conjunto de tokens de la pluralidad de conjuntos de tokens, el uno o más procesadores se configuran para hacer que el sistema de procesamiento:

rechace cada conjunto de tokens generados por el primer modelo de inteligencia artificial generativa; y
muestree, usando el segundo modelo de inteligencia artificial generativa, un token basado en una distribución objetivo que excluye las probabilidades asociadas con cada conjunto de tokens generado por el primer modelo de inteligencia artificial generativa, en donde el conjunto de tokens seleccionado comprende el token muestreado.

11. El sistema de procesamiento de la reivindicación 1, en donde:

el primer modelo de inteligencia artificial generativa corresponde a un modelo de borrador en una canalización de decodificación especulativa, y
el segundo modelo de inteligencia artificial generativa corresponde a un modelo objetivo en la canalización de decodificación especulativa.

12. Un método implementado por procesador, en donde comprende:

recibir una pluralidad de conjuntos de tokens generados basado en un indicador de entrada y un primer modelo de inteligencia artificial

generativa, cada conjunto de tokens en la pluralidad de conjuntos de tokens comprende una secuencia de tokens correspondiente a una respuesta candidata al indicador de entrada;

5 seleccionar, usando un segundo modelo de inteligencia artificial generativa y ajuste recursivo de una distribución objetivo asociada con la pluralidad recibida de conjuntos de tokens, un conjunto de tokens

10 de la pluralidad de conjuntos de tokens; y emitir el conjunto de tokens seleccionado como una respuesta al indicador de entrada.

13. El método de la reivindicación 12, en donde la pluralidad de conjuntos de tokens se organiza en una

15 estructura de datos de árbol.

14. El método de la reivindicación 13, en donde un nodo raíz de la estructura de datos de árbol corresponde al indicador de entrada y en donde cada ruta a través de la estructura de datos de árbol corresponde a una secuencia

20 diferente de tokens correspondiente a la respuesta candidata al indicador de entrada.

15. El método de la reivindicación 13, en donde: la estructura de datos de árbol incluye una pluralidad de niveles, cada nivel

25 correspondiente a un token en la secuencia de tokens, y

un número de tokens en un nivel particular en la estructura de datos de árbol está basado en un factor de ramificación asociado con un nivel inmediatamente anterior al nivel particular en la estructura de datos de árbol.

16. El método de la reivindicación 13, en donde una profundidad de la estructura de datos de árbol corresponde a un parámetro que define un número máximo de tokens generados por un único pase a través del primer modelo de inteligencia artificial generativa.

17. El método de la reivindicación 12, en donde un tamaño de cada conjunto de tokens está basado en una métrica de complejidad computacional asociada con generar un conjunto de tokens objetivo por el segundo modelo de inteligencia artificial generativa.

18. El método de la reivindicación 12, en donde el ajuste recursivo de la distribución objetivo comprende:

- determinar si aceptar o rechazar un primer token en un conjunto de tokens de la pluralidad de conjuntos de tokens; y
- ajustar una distribución de probabilidad usada para verificar un segundo token en el conjunto de tokens posterior al primer token basado en la determinación de si aceptar o rechazar el primer token.

19. El método de la reivindicación 18, en donde ajustar la distribución de probabilidad comprende restar un valor de probabilidad asociado con el primer token de la distribución de probabilidad basado en determinar rechazar
5 el primer token.

20. El método de la reivindicación 12, en donde la selección del conjunto de tokens de la pluralidad de conjuntos de tokens comprende:

rechazar un primer token en un primer nivel de
10 una estructura de datos de árbol que representa la pluralidad de conjuntos de tokens;
emitir una distribución de probabilidad ajustada basada en el rechazo del primer token;
descartar o ignorar, de la estructura de
15 datos de árbol, los tokens hijos del primer token en niveles más profundos que el primer nivel de la estructura de datos de árbol; y
determinar si aceptar o rechazar un segundo token en el primer nivel de la estructura de
20 datos de árbol basado en la distribución de probabilidad ajustada.

21. El método de la reivindicación 12, en donde la selección del conjunto de tokens de la pluralidad de conjuntos de tokens comprende:

rechazar cada conjunto de tokens generados por el primer modelo de inteligencia artificial generativa; y
muestrear, usando el segundo modelo de
5 inteligencia artificial generativa, un token basado en una distribución objetivo que excluye las probabilidades asociadas con cada conjunto de tokens generado por el primer modelo de inteligencia artificial
10 generativa, en donde el conjunto de tokens seleccionado comprende el token muestreado.

22. El método de la reivindicación 12, en donde:
el primer modelo de inteligencia artificial generativa corresponde a un modelo de
15 borrador en una canalización de decodificación especulativa, y
el segundo modelo de inteligencia artificial generativa corresponde a un modelo objetivo en la canalización de decodificación
20 especulativa.

23. Un sistema de procesamiento, en donde
comprende:
medios para recibir una pluralidad de
conjuntos de tokens generados basado en un
25 indicador de entrada y un primer modelo de inteligencia artificial generativa, cada

conjunto de tokens en la pluralidad de
conjuntos de tokens comprende una secuencia de
tokens correspondiente a una respuesta
candidata al indicador de entrada;
5 medios para seleccionar, usando un segundo
modelo de inteligencia artificial generativa
y ajuste recursivo de una distribución
objetivo asociada con la pluralidad recibida
de conjuntos de tokens, un conjunto de tokens
10 de la pluralidad de conjuntos de tokens; y
medios para emitir el conjunto de tokens
seleccionado como una respuesta al indicador
de entrada.

24. El sistema de procesamiento de la
15 reivindicación 23, en donde la pluralidad de conjuntos de
tokens se organiza en una estructura de datos de árbol.

25. El sistema de procesamiento de la
reivindicación 24, en donde un nodo raíz de la estructura
de datos de árbol corresponde al indicador de entrada y en
20 donde cada ruta a través de la estructura de datos de árbol
corresponde a una secuencia diferente de tokens
correspondiente a la respuesta candidata al indicador
de entrada.

26. El sistema de procesamiento de la
25 reivindicación 24, en donde:

la estructura de datos de árbol incluye una pluralidad de niveles, cada nivel correspondiente a un token en la secuencia de tokens, y

5 un número de tokens en un nivel particular en la estructura de datos de árbol está basado en un factor de ramificación asociado con un nivel inmediatamente anterior al nivel particular en la estructura de datos de árbol.

10 **27.** El sistema de procesamiento de la reivindicación 24, en donde una profundidad de la estructura de datos de árbol corresponde a un parámetro que define un número máximo de tokens generados por un único pase a través del primer modelo de inteligencia
15 artificial generativa.

28. El sistema de procesamiento de la reivindicación 23, en donde un tamaño de cada conjunto de tokens está basado en una métrica de complejidad computacional asociada con generar un conjunto de tokens
20 objetivo por el segundo modelo de inteligencia artificial generativa.

29. El sistema de procesamiento de la reivindicación 23, en donde el ajuste recursivo de la distribución objetivo comprende:

determinar si aceptar o rechazar un primer token en un conjunto de tokens de la pluralidad de conjuntos de tokens; y
ajustar una distribución de probabilidad usada para verificar un segundo token en el conjunto de tokens posterior al primer token basado en la determinación de si aceptar o rechazar el primer token.

30. El sistema de procesamiento de la reivindicación 29, en donde ajustar la distribución de probabilidad comprende restar un valor de probabilidad asociado con el primer token de la distribución de probabilidad basado en determinar rechazar el primer token.

31. El sistema de procesamiento de la reivindicación 23, en donde los medios para seleccionar el conjunto de tokens de la pluralidad de conjuntos de tokens comprenden:

medios para rechazar un primer token en un primer nivel de una estructura de datos de árbol que representa la pluralidad de conjuntos de tokens;

medios para emitir una distribución de probabilidad ajustada basada en el rechazo del primer token;

medios para descartar o ignorar, de la estructura de datos de árbol, los tokens

hijos del primer token en niveles más profundos que el primer nivel de la estructura de datos de árbol; y medios para determinar si aceptar o rechazar un segundo token en el primer nivel de la estructura de datos de árbol basado en la distribución de probabilidad ajustada.

32. El sistema de procesamiento de la reivindicación 23, en donde los medios para seleccionar el conjunto de tokens de la pluralidad de conjuntos de tokens comprenden:

medios para rechazar cada conjunto de tokens generados por el primer modelo de inteligencia artificial generativa; y medios para muestrear, usando el segundo modelo de inteligencia artificial generativa, un token basado en una distribución objetivo que excluye las probabilidades asociadas con cada conjunto de tokens generado por el primer modelo de inteligencia artificial generativa, en donde el conjunto de tokens seleccionado comprende el token muestreado.

33. El sistema de procesamiento de la reivindicación 23, en donde:

el primer modelo de inteligencia artificial generativa corresponde a un modelo de

borrador en una canalización de
decodificación especulativa, y
el segundo modelo de inteligencia artificial
generativa corresponde a un modelo objetivo en
5 la canalización de decodificación
especulativa.

34. Un medio legible por computador no
transitorio que tiene instrucciones ejecutables
almacenadas en este que, cuando son ejecutadas por un
10 procesador, realizan una operación que comprende:

recibir una pluralidad de conjuntos de tokens
generados basado en un indicador de entrada y
un primer modelo de inteligencia artificial
generativa, cada conjunto de tokens en la
15 pluralidad de conjuntos de tokens comprende
una secuencia de tokens correspondiente a una
respuesta candidata al indicador de entrada;
seleccionar, usando un segundo modelo de
inteligencia artificial generativa y ajuste
20 recursivo de una distribución objetivo
asociada con la pluralidad recibida de
conjuntos de tokens, un conjunto de tokens
de la pluralidad de conjuntos de tokens; y
emitir el conjunto de tokens seleccionado
25 como una respuesta al indicador de entrada.