

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 genere, basado en un indicador de entrada y un modelo de inteligencia artificial generativa, una primera pluralidad de conjuntos de tokens, cada conjunto de tokens en la primera pluralidad de
15 conjuntos de tokens correspondiente a una primera porción de una respuesta candidata al indicador de entrada;
genere especulativamente, usando el modelo de inteligencia artificial
20 generativa, una segunda pluralidad de conjuntos de tokens, cada conjunto de tokens en la segunda pluralidad de conjuntos de tokens correspondiente a una segunda porción de la respuesta
25 candidata al indicador de entrada basado

en la primera pluralidad de conjuntos de tokens;
mientras se genera especulativamente la segunda pluralidad de conjuntos de tokens,
5 seleccione un conjunto de tokens de la primera pluralidad de conjuntos de tokens; y
emita el conjunto de tokens seleccionado de la primera pluralidad de tokens y un
10 conjunto de tokens asociado en la segunda pluralidad de tokens como una respuesta al indicador de entrada.

2. El sistema de procesamiento de la reivindicación 1, en donde para seleccionar el conjunto de
15 tokens de la primera pluralidad de conjuntos de tokens, el uno o más procesadores se configuran para hacer que el sistema de procesamiento seleccione una secuencia más larga de tokens aceptados de la primera pluralidad de conjuntos de tokens.

20 3. El sistema de procesamiento de la reivindicación 1, en donde los conjuntos de tokens en la segunda pluralidad de conjuntos de tokens incluyen protección responsable de un número de tokens en el conjunto de tokens seleccionado de la primera pluralidad
25 de conjuntos de tokens.

4. El sistema de procesamiento de la reivindicación 1, en donde la primera pluralidad de conjuntos de tokens se representa como una estructura de datos de árbol y en donde un nodo raíz de la estructura de datos de árbol
5 corresponde al indicador de entrada.

5. El sistema de procesamiento de la reivindicación 4, en donde una profundidad de la estructura de datos de árbol corresponde a un número máximo de tokens generados por un único pase a través del modelo de
10 inteligencia artificial generativa.

6. El sistema de procesamiento de la reivindicación 4, en donde se establece un tamaño máximo de la estructura de datos de árbol basado en una métrica de complejidad computacional asociada con generar un
15 conjunto de tokens por el modelo de inteligencia artificial generativa.

7. El sistema de procesamiento de la reivindicación 4, en donde para seleccionar el conjunto de tokens de la primera pluralidad de conjuntos de tokens, el
20 uno o más procesadores se configuran para hacer que el sistema de procesamiento:

rechace un primer token en un primer nivel de la estructura de datos de árbol que representa la primera pluralidad de conjuntos
25 de tokens;

genere una distribución de probabilidad ajustada basada en el rechazo del primer token;

5 descarte o ignore tokens hijos del primer token 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 8. 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:

15 rechace cada conjunto de tokens en la primera pluralidad de conjuntos de tokens generados por el modelo de inteligencia artificial generativa; y

20 muestree, usando el modelo de inteligencia artificial generativa, un token basado en una distribución objetivo que excluye las probabilidades asociadas con cada conjunto de tokens en la primera pluralidad de conjuntos de tokens, en donde el conjunto de tokens
25 seleccionado de la primera pluralidad de

conjuntos de tokens comprende el token
muestreado.

9. El sistema de procesamiento de la
reivindicación 1, en donde el modelo de inteligencia
5 artificial generativa comprende un modelo de inteligencia
artificial generativa entrenado para generar múltiples
tokens en respuesta al indicador de entrada basado en las
incrustaciones de indicadores de pronósticos.

10. El sistema de procesamiento de la
10 reivindicación 1, en donde el modelo de inteligencia
artificial generativa comprende un modelo que incluye una
o más capas no autorregresivas y una o más
capas autorregresivas.

11. El sistema de procesamiento de la
15 reivindicación 10, en donde la una o más capas
autorregresivas comprenden una o más capas en una parte
superior de una pila de capas que representan el modelo de
inteligencia artificial generativa.

12. El sistema de procesamiento de la
20 reivindicación 10, en donde la una o más capas
autorregresivas comprenden una o más capas en una parte
inferior de una pila de capas que representan el modelo de
inteligencia artificial generativa.

13. Un método implementado por procesador, en
25 donde comprende:

generar, basado en un indicador de entrada y un modelo de inteligencia artificial generativa, una primera pluralidad de conjuntos de tokens, cada conjunto de tokens en la primera pluralidad de conjuntos de tokens correspondiente a una primera porción de una respuesta candidata al indicador de entrada;

generar especulativamente, usando el modelo de inteligencia artificial generativa, una segunda pluralidad de conjuntos de tokens, cada conjunto de tokens en la segunda pluralidad de conjuntos de tokens correspondiente a una segunda porción de la respuesta candidata al indicador de entrada basado en la primera pluralidad de conjuntos de tokens;

mientras que se genera especulativamente la segunda pluralidad de conjuntos de tokens, seleccionar un conjunto de tokens de la primera pluralidad de conjuntos de tokens; y emitir el conjunto de tokens seleccionado de la primera pluralidad de tokens y un conjunto de tokens asociado en la segunda pluralidad de tokens como una respuesta al indicador de entrada.

14. El método de la reivindicación 13, en donde seleccionar el conjunto de tokens de la primera pluralidad de conjuntos de tokens comprende seleccionar una secuencia más larga de tokens aceptados de la primera pluralidad de
5 conjuntos de tokens.

15. El método de la reivindicación 13, en donde los conjuntos de tokens en la segunda pluralidad de conjuntos de tokens incluyen protección responsable de un número de tokens en el conjunto de tokens seleccionado de
10 la primera pluralidad de conjuntos de tokens.

16. El método de la reivindicación 13, en donde la primera pluralidad de conjuntos de tokens se representan como una estructura de datos de árbol y en donde un nodo raíz de la estructura de datos de árbol corresponde al
15 indicador de entrada.

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

18. El método de la reivindicación 16, en donde se establece un tamaño máximo de la estructura de datos de árbol basado en una métrica de complejidad computacional asociada con generar un conjunto de tokens por el modelo
25 de inteligencia artificial generativa.

19. El método de la reivindicación 16, en donde la selección del conjunto de tokens de la primera pluralidad de conjuntos de tokens comprende:

5 rechazar un primer token en un primer nivel de la estructura de datos de árbol que representa la primera pluralidad de conjuntos de tokens;

emitir una distribución de probabilidad ajustada basada en el rechazo del

10 primer token;

descartar o ignorar tokens hijos del primer token de la estructura de datos de árbol; y

determinar si aceptar o rechazar un segundo token en el primer nivel de la estructura de

15 datos de árbol basado en la distribución de probabilidad ajustada.

20. El método de la reivindicación 13, en donde la selección del conjunto de tokens de la primera pluralidad de conjuntos de tokens comprende:

20 rechazar cada conjunto de tokens en la primera pluralidad de conjuntos de tokens generados por el modelo de inteligencia artificial generativa; y

muestrear, usando el modelo de inteligencia

25 artificial generativa, un token basado en una distribución objetivo que excluye las

probabilidades asociadas con cada conjunto de tokens en la primera pluralidad de conjuntos de tokens, en donde el conjunto de tokens seleccionado de la primera pluralidad de conjuntos de tokens comprende el token muestreado.

21. El método de la reivindicación 13, en donde el modelo de inteligencia artificial generativa comprende un modelo de inteligencia artificial generativa entrenado para generar múltiples tokens en respuesta al indicador de entrada basado en las incrustaciones de indicadores de pronósticos.

22. El método de la reivindicación 13, en donde el modelo de inteligencia artificial generativa comprende un modelo que incluye una o más capas no autorregresivas y una o más capas autorregresivas.

23. El método de la reivindicación 22, en donde la una o más capas autorregresivas comprenden una o más capas en una parte superior de una pila de capas que representan el modelo de inteligencia artificial generativa.

24. El método de la reivindicación 22, en donde la una o más capas autorregresivas comprenden una o más capas en una parte inferior de una pila de capas que representan el modelo de inteligencia artificial generativa.

25. Un sistema de procesamiento, en donde comprende:

5 medios para generar, basados en un indicador de entrada y un modelo de inteligencia artificial generativa, una primera pluralidad de conjuntos de tokens, cada conjunto de tokens en la primera pluralidad de conjuntos de tokens correspondiente a una primera porción de una respuesta candidata al
10 indicador de entrada;

medios para generar especulativamente, usando el modelo de inteligencia artificial generativa, una segunda pluralidad de conjuntos de tokens, cada conjunto de tokens
15 en la segunda pluralidad de conjuntos de tokens correspondiente a una segunda porción de la respuesta candidata al indicador de entrada basado en la primera pluralidad de conjuntos de tokens;

20 medios para seleccionar, mientras se genera especulativamente la segunda pluralidad de conjuntos de tokens, un conjunto de tokens de la primera pluralidad de conjuntos de tokens; y

25 medios para emitir el conjunto de tokens seleccionado de la primera pluralidad de

tokens y un conjunto de tokens asociado en la segunda pluralidad de tokens como una respuesta al indicador de entrada.

26. El sistema de procesamiento de la reivindicación 25, en donde seleccionar el conjunto de tokens de la primera pluralidad de conjuntos de tokens comprende medios para seleccionar una secuencia más larga de tokens aceptados de la primera pluralidad de conjuntos de tokens.

27. El sistema de procesamiento de la reivindicación 25, en donde los conjuntos de tokens en la segunda pluralidad de conjuntos de tokens incluyen protección responsable de un número de tokens en el conjunto de tokens seleccionado de la primera pluralidad de conjuntos de tokens.

28. El sistema de procesamiento de la reivindicación 25, en donde la primera pluralidad de conjuntos de tokens se representa como una estructura de datos de árbol y en donde un nodo raíz de la estructura de datos de árbol corresponde al indicador de entrada.

29. El sistema de procesamiento de la reivindicación 28, en donde una profundidad de la estructura de datos de árbol corresponde a un número máximo de tokens generados por un único pase a través del modelo de inteligencia artificial generativa.

30. El sistema de procesamiento de la reivindicación 28, en donde se establece un tamaño máximo de la estructura de datos de árbol basado en una métrica de complejidad computacional asociada con generar un conjunto de tokens por el modelo de inteligencia artificial generativa.

31. El sistema de procesamiento de la reivindicación 28, en donde los medios para seleccionar el conjunto de tokens de la primera pluralidad de conjuntos de tokens comprenden:

medios para rechazar un primer token en un primer nivel de la estructura de datos de árbol que representa la primera 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 tokens hijos del primer token 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 25, en donde los medios para seleccionar el

conjunto de tokens de la primera pluralidad de conjuntos de tokens comprenden:

5 medios para rechazar cada conjunto de tokens en la primera pluralidad de conjuntos de tokens generados por el modelo de inteligencia artificial generativa; y
10 medios para muestrear, usando el modelo de inteligencia artificial generativa, un token basado en una distribución objetivo que excluye las probabilidades asociadas con cada conjunto de tokens en la primera pluralidad de conjuntos de tokens, en donde el conjunto de tokens seleccionado de la primera pluralidad de conjuntos de tokens comprende
15 el token muestreado.

33. El sistema de procesamiento de la reivindicación 25, en donde el modelo de inteligencia artificial generativa comprende un modelo de inteligencia artificial generativa entrenado para generar múltiples
20 tokens en respuesta al indicador de entrada basado en las incrustaciones de indicadores de pronósticos.

34. El sistema de procesamiento de la reivindicación 25, en donde el modelo de inteligencia artificial generativa comprende un modelo que incluye una
25 o más capas no autorregresivas y una o más capas autorregresivas.

35. El sistema de procesamiento de la reivindicación 34, en donde la una o más capas autorregresivas comprenden una o más capas en una parte superior de una pila de capas que representan el modelo de
5 inteligencia artificial generativa.

36. El sistema de procesamiento de la reivindicación 34, en donde la una o más capas autorregresivas comprenden una o más capas en una parte inferior de una pila de capas que representan el modelo de
10 inteligencia artificial generativa.

37. Un medio legible por computadora no transitorio que tiene instrucciones ejecutables almacenadas en este que, cuando se ejecutan por uno o más procesadores, realizan una operación que comprende:

15 generar, basado en un indicador de entrada y un modelo de inteligencia artificial generativa, una primera pluralidad de conjuntos de tokens, cada conjunto de tokens en la primera pluralidad de conjuntos de tokens correspondiente a una primera porción
20 de una respuesta candidata al indicador de entrada;

generar especulativamente, usando el modelo de inteligencia artificial generativa, una
25 segunda pluralidad de conjuntos de tokens, cada conjunto de tokens en la segunda

pluralidad de conjuntos de tokens correspondiente a una segunda porción de la respuesta candidata al indicador de entrada basado en la primera pluralidad de conjuntos de tokens;

5 mientras que se genera especulativamente la segunda pluralidad de conjuntos de tokens, seleccionar un conjunto de tokens de la primera pluralidad de conjuntos de tokens; y

10 emitir el conjunto de tokens seleccionado de la primera pluralidad de tokens y un conjunto de tokens asociado en la segunda pluralidad de tokens como una respuesta al indicador de entrada.