Engineer

Modelo de lenguaje de gran tamaño

Red neuronal entrenada en texto que predice el próximo token

Qué es

Un modelo de lenguaje de gran tamaño es una red neuronal entrenada con una cantidad enorme de texto para hacer una sola cosa: dado el texto escrito hasta el momento, asignar una probabilidad a cada token que podría venir después. Un token es un fragmento corto de texto, a menudo una palabra completa y a veces parte de una palabra o un único signo de puntuación. Escribir una respuesta consiste en repetir ese paso en un bucle. El software elige un token según esas probabilidades, lo añade al texto y vuelve a preguntar, hasta que el modelo produce un token que indica que debe detenerse o se alcanza un límite de longitud. El perfil de NIST para la IA generativa, publicado en julio de 2024, lo resume en una línea: «Los LLM predicen el siguiente token o palabra de una oración o frase». Las siglas inglesas LLM, que utiliza el propio perfil, corresponden a Large Language Model y designan el mismo concepto.

La arquitectura de los modelos de los últimos años es el Transformer, descrito en un artículo de 2017 sobre traducción automática. Prescindió de las capas recurrentes y convolucionales de los modelos anteriores y se basó únicamente en la atención, un mecanismo que permite que cada posición del texto consulte directamente todas las demás. Eso facilitó mucho distribuir el entrenamiento entre numerosos procesadores, y el modelo del artículo se entrenó en mucho menos tiempo que sus predecesores.

«De gran tamaño» se refiere al número de parámetros, los pesos que la red aprende durante el entrenamiento. Un artículo de 2020 describió un modelo con 175 mil millones de parámetros, diez veces el número de parámetros de cualquier modelo de lenguaje anterior que no fuera disperso, y mostró que, a esa escala, un modelo podía abordar una tarea nueva a partir de unos pocos ejemplos resueltos incluidos en su entrada, sin modificar en absoluto sus pesos.

Un modelo entrenado solo para continuar texto hace exactamente eso: continúa el texto, que no es lo mismo que hacer lo que se le pide. Un artículo de 2022 describió el paso que convierte ese modelo en un asistente. Unas personas escribieron las respuestas que querían para un conjunto de peticiones y el modelo se sometió a un ajuste fino con ellas; después ordenaron por preferencia las respuestas del propio modelo, que siguió entrenándose a partir de esas clasificaciones, mediante un método llamado aprendizaje por refuerzo con retroalimentación humana. En las evaluaciones del artículo, las personas prefirieron las respuestas de un modelo de 1,3 mil millones de parámetros entrenado así a las del modelo original, que tenía cien veces esa cantidad de parámetros.

Utilice el concepto en una decisión

Aplique «¿Qué es un modelo de lenguaje de gran tamaño?» a una decisión real. Podemos ayudarle a relacionar el concepto con el sistema que está construyendo o comprando. Un agente conectado a los sistemas que utiliza, probado con casos definidos por sus especialistas, supervisado por su personal y sin depender de un único modelo.

Explore nuestros servicios Póngase en contacto →

Compruebe los supuestos detrás del término

Utilice la explicación y los trabajos enlazados para formular preguntas sobre su proyecto. Podemos examinar los supuestos con usted antes de elegir una herramienta. Motor de 268 controles para el repositorio — 268 controles de commit y 146 pruebas de los propios controles

Motor de 268 controles para el repositorio Póngase en contacto →

Por qué importa

Un modelo de lenguaje asume trabajos que resultan difíciles de cubrir con reglas fijas: clasificar un correo de texto libre en la carpeta correcta, extraer los campos de un albarán escaneado, redactar una respuesta en el idioma del cliente o resumir un hilo largo. El software puede preparar ahora un borrador de trabajos que antes requerían que una persona leyera cada elemento, y una persona puede revisarlo.

La propiedad que lo hace útil también dificulta incorporarlo a una aplicación. Su salida suele obtenerse mediante muestreo, por lo que una misma entrada puede producir respuestas distintas en ejecuciones diferentes, y se expresa con fluidez tanto si acierta como si no. NIST llama confabulación, más conocida como alucinación, a una salida incorrecta afirmada con seguridad, y la describe como un resultado natural del diseño de estos modelos: producen texto que coincide con las estadísticas del texto con el que se entrenaron, y eso es a menudo correcto y a veces no. NIST también señala que un modelo puede inventar el razonamiento o las citas que parecen justificar una respuesta. Una respuesta incorrecta tiene

exactamente el mismo aspecto que una correcta, así que la única manera de saber con qué frecuencia acierta un modelo en su trabajo es medirlo con ese trabajo.

Tampoco se trata de quedarse con un único modelo. Aparecen modelos nuevos con regularidad, dos pueden comportarse de forma muy distinta ante la misma entrada y el proveedor del modelo al que llama una aplicación decide cuándo cambia. Un equipo que basa su software en un modelo de lenguaje depende de una pieza que no controla; el trabajo que perdura es todo lo que lo rodea: los casos que lo evalúan, las herramientas a las que puede llamar y las reglas sobre lo que puede hacer por sí solo.

El perfil de NIST identifica doce ámbitos de riesgo de la IA generativa, entre ellos la confabulación, la privacidad de los datos, la seguridad de la información y la propiedad intelectual. Está dirigido a organizaciones que gestionan riesgos, más que a desarrolladores, pero sirve como lista de comprobación para un equipo que decide a qué puede acceder una función basada en un modelo.

Cómo se aplica en la práctica

Una función basada en un modelo de lenguaje empieza por casos escritos, no por una instrucción o prompt. Cada caso contiene una entrada real y la respuesta que aceptaría una persona que conoce el trabajo, conservadas en un archivo bajo control de versiones junto al código:

```
# cases/delivery-notes.yaml
- id: note-0142
  input: scans/note-0142.txt
  expect:
    supplier: "Example Freight"
    reference: "DN-58213"
    pallets: 6
- id: note-0157
  input: scans/note-0157.txt
  expect:
    supplier: "Sample Haulage"
    reference: "DN-59002"
    pallets: 14
```

El modelo que se ocupa de cada tipo de trabajo se elige mediante la configuración, de modo que un modelo pequeño y rápido puede extraer datos mientras uno mayor redacta las respuestas que requieren criterio:

```
# routing.yaml
jobs:
  extract-delivery-note:
    model: small-fast
    cases: cases/delivery-notes.yaml
  draft-claim-reply:
    model: large
    cases: cases/claim-replies.yaml
```

Cada cambio se ejecuta contra todos los casos antes de publicarse, tanto si se cambia una instrucción como una herramienta o un modelo. El resultado es una puntuación por caso, conservada a lo largo del tiempo para que una corrección de un caso no estropee otro sin que se detecte. Cuando se publica un modelo nuevo, probarlo consiste en cambiar una línea de `routing.yaml` y ejecutar los casos; la decisión de adoptarlo se basa en las puntuaciones, no en el anuncio.

Así describe este sitio sus agentes de IA a medida. La elección del modelo es configuración, no código: uno pequeño y rápido para clasificar y extraer datos y otro grande para lo que exige criterio. Un encargo comienza con treinta a cincuenta ejemplos reales escritos y sus respuestas correctas, y cada cambio, ya sea de una instrucción, una herramienta o el modelo, se ejecuta contra todos ellos antes de publicarse. La página también indica que el modelo se puede sustituir y que el código, los casos y los datos pertenecen al cliente, para que el agente siga siendo suyo cuando cambie el modelo que utiliza.

Dónde surgen los problemas

Respuestas expresadas con seguridad, pero incorrectas. Un modelo al que se pregunta por una cláusula contractual que no tiene delante puede describirla de todos modos, con el mismo tono seguro que utiliza para una respuesta correcta, e incluso citar una fuente inexistente. Proporcione al modelo los documentos en los que debe basar sus respuestas, pídale que los cite y compruebe mediante código que cada cita apunta a un documento y un pasaje reales. Mida la tasa de error con casos escritos y remita a una persona las respuestas con consecuencias antes de que se actúe a partir de ellas.

Instrucciones ocultas en los datos. Una aplicación que permite al modelo leer correos, páginas web o archivos subidos permite a los autores de ese texto dirigirse al modelo. Un artículo de 2023 lo demostró en aplicaciones reales: las instrucciones insertadas en contenido que el modelo probablemente leería hicieron que filtrara datos y cambiara el comportamiento de la aplicación. El perfil de NIST incluye la inyección indirecta de instrucciones entre los riesgos de seguridad de la IA generativa. Trate todo lo que lee el modelo como entrada no fiable, aplique los permisos en las herramientas en lugar de en la instrucción, conceda a las herramientas únicamente los derechos de la persona en cuyo nombre actúa el modelo y exija la aprobación de una persona para cualquier acción que no se pueda deshacer.

Un modelo que cambia sin que usted lo controle. Una versión nueva o un proveedor distinto pueden responder de otro modo a la misma entrada, y una respuesta que pasaba las pruebas el mes pasado falla hoy sin cambios en su propio código. Indique una versión exacta del modelo siempre que el proveedor la ofrezca y ejecute el archivo completo de casos con cada cambio, incluido un cambio de modelo, antes de que llegue a los usuarios.

Una salida que el código no puede analizar. Un modelo al que se pide JSON devuelve JSON seguido de una frase amable, omite un campo o escribe un número como cadena. Describa la estructura esperada en JSON Schema, compruebe cada respuesta contra ella y trate cualquier respuesta que no la cumpla como un error, para reintentar o remitir a una persona, nunca como datos.

Datos personales en instrucciones y registros. Todo lo que se envía a un modelo alojado sale de sus propios sistemas, y un registro de instrucciones y respuestas conserva datos personales mucho después de terminar la petición. NIST añade que los modelos pueden repetir material sensible de sus datos de entrenamiento, un problema que denomina memorización de datos. Envíe solo los campos que requiere cada trabajo, someta los registros de instrucciones a las mismas reglas de conservación y acceso que sus datos y lea las condiciones del proveedor sobre conservación y entrenamiento antes de que le llegue cualquier dato personal.

Lecturas adicionales

- NIST AI 600-1: perfil de inteligencia artificial generativa — el complemento de NIST a su marco de gestión de riesgos de IA, publicado en julio de 2024: doce riesgos de la IA generativa, incluida la confabulación, y acciones sugeridas para cada uno.
- La atención es todo lo que se necesita — el artículo de 2017 que presentó el Transformer, la arquitectura de los modelos de lenguaje actuales.
- Los modelos de lenguaje aprenden con pocos ejemplos — el artículo de 2020 sobre un modelo de 175 mil millones de parámetros que aborda tareas nuevas a partir de ejemplos incluidos en su entrada.
- Entrenamiento de modelos de lenguaje para seguir instrucciones con retroalimentación humana — ajuste fino con respuestas escritas y salidas ordenadas por preferencia: el paso de predictor de texto a asistente.
- No es lo que había aceptado: comprometer aplicaciones reales con LLM mediante inyección indirecta de instrucciones — instrucciones insertadas en datos que lee un modelo, demostradas en aplicaciones reales.

Glosario

Elija según sus restricciones reales

Describa el flujo de trabajo, las interfaces existentes y los usuarios que necesitan el cambio. Podemos definir una entrega comprobable y un traspaso que permita a su equipo mantener lo entregado.

Vea cuánto cuesta Póngase en contacto →

Haga práctica la terminología

Pregunte cómo se aplica «¿Qué es un modelo de lenguaje de gran tamaño?» a su organización. Basta un ejemplo concreto de su sistema actual para iniciar la conversación.

Explore nuestros servicios

welcome@engineer.company +45 50 10 61 84 Copenhague, Dinamarca

Llévese esta página

Tarjeta de contacto

Creado con esmero y cariño por D.

CVR 45969703 · Tagensvej 105C, 3. th, 2200 København N

Derechos de autor © 2025 – actualidad · Engineer ApS · Todos los derechos reservados.

Impreso desde https://engineer.company/es/glossary/large-language-model/