COLECCIÓN

MITOS Y REALIDADES DE LA IA

Entre la promesa y la realidad

TRANSFORMERS Y ATENCIÓN COMO MECANISMOS DE PROCESAMIENTO DE SECUENCIAS

Número 8



MRCI

anejo de ecursos y ontroles nteligentes



DR. JOSÉ ANTONIO PÉREZ RAMOS

TRANSFORMERS Y ATENCIÓN COMO MECANISMOS DE PROCESAMIENTO DE SECUENCIAS



SOBRE EL AUTOR

Doctor en Ciencias de lo Fiscal por el Instituto de Especialización para Ejecutivos. Maestro en Fiscal por la Universidad Autónoma "Benito Juárez" de Oaxaca. Maestro en Estudios Avanzados en Derechos Humanos por la Global Open University. Licenciado en Contaduría Pública por la Universidad Autónoma "Benito Juárez" de Oaxaca y Licenciado en Filosofía por la Universidad Nacionalista de México.

Socio fundador y director general de la Firma Manejo de Recursos y Controles Inteligentes (Firma MRCI). Fiscalista del año 2009 por la revista *Defensa Fiscal*. Doctor honoris causa por la Asamblea General del Claustro Doctoral Iberoamericano (2016). Doctor honoris causa por el Instituto Mexicano de Líderes de Excelencia (2017). Doctor honoris causa por 1Million Startups, Latinomics, Leaderships Forum y la Fundación Humanist World (2018). Doctor honoris causa por el Claustro de Doctores de la AISOL (American International School of Law) (2019).

Autor de diversas obras, entre las que destacan *Ius mutantur o del derecho que cambia* (Tirant Lo Blanch, 2025), *Futurología* (MRCI, 2025) y *Hasta que la muerte nos separe* (MRCI, 2025). Ha participado como coordinador en *Cuestiones tributarias. Problemas y controversias en el México actual* (Tirant Lo Blanch, 2023), *El costo de la justicia* (Apex Iuris, 2019), *Remuneraciones estratégicas inteligentes* (MRCI, 2015) y *Retos de la justicia tributaria* (Tirant Lo Blanch-AMDF, 2025).

Titular de *Pie de Página*, programa para la divulgación de los derechos humanos en México, a través de la plataforma Legal Cloud. Expositor en foros nacionales e internacionales. Colaborador en diversas revistas jurídicas, entre las que destacan *Defensa Fiscal* y *Legal Cloud*.



TRANSFORMERS Y ATENCIÓN COMO MECANISMOS DE PROCESAMIENTO DE SECUENCIAS

DR. JOSÉ ANTONIO PÉREZ RAMOS

PRIMERA EDICIÓN, ATEMPORAL

Derechos reservados, propiedad de
Dr. José Antonio Pérez Ramos

Comentarios y opiniones: investigacion@mrci.com.mx

Título original: Transformers y atención como mecanismos de procesamiento de secuencias.

Colección mitos y realidades de la inteligencia artificial.

Entre la promesa y la realidad.

Autor: Dr. José Antonio Pérez Ramos.

Queda prohibida la reproducción total y parcial de esta obra denominada; TRANSFORMERS Y ATENCIÓN COMO MECANISMOS DE PROCESAMIENTO DE SECUENCIAS, por cualquier medio, sin autorización escrita del autor.

PRINTED IN MEXICO
IMPRESO EN MÉXICO

ÍNDICE

QUÉ EXPLICA REALMENTE LA ATENCIÓN ..... - 7 -
LA ATENCIÓN ANTES DEL TRANSFORMER ..... - 9 -
CONSULTAS CLAVES Y VALORES ..... - 11 -
UN CÁLCULO QUE PUEDE COMPROBARSE..... - 15 -
CUANDO LOS PESOS CAMBIAN Y LA SALIDA NO..... - 19 -
QUÉ POSICIONES PUEDEN COMUNICARSE ..... - 20 -
EL ORDEN NECESITA REPRESENTACIÓN ..... - 24 -
UNA CAPA CONTIENE MÁS QUE ATENCIÓN ..... - 26 -
PARALELISMO Y COSTO COMPUTACIONAL ..... - 30 -
PESOS DE ATENCIÓN Y EXPLICACIÓN..... - 34 -
QUÉ DEBERÍA ACOMPAÑAR A UNA VISUALIZACIÓN..... - 36 -
CONTEXTO DISPONIBLE Y CONTEXTO UTILIZADO ..... - 38 -
TRES CASOS PARA COMPROBAR RELACIONES..... - 40 -
UNA EVALUACIÓN PEQUEÑA PERO INFORMATIVA ..... - 44 -
MITOS QUE MERECEN UNA RESPUESTA PRECISA ..... - 48 -
CONSECUENCIAS PARA EL USO RESPONSABLE ..... - 51 -
CONCLUSIÓN ..... - 53 -
REFERENCIAS..... - 54 -

Qué explica realmente la atención

Cuando un sistema relaciona un pronombre con una persona mencionada varias líneas antes, resulta tentador decir que prestó atención al antecedente correcto; esa expresión puede describir informalmente un resultado, pero deja sin resolver cómo se representó la entrada, qué conexiones participaron y si la relación se mantendría al cambiar el documento.

El problema central de este artículo es precisar qué significa atención en un Transformer y hasta dónde permite explicar el procesamiento de secuencias; una descripción útil debe mostrar operaciones concretas sin convertir un vocabulario matemático en una afirmación sobre conciencia, intención o comprensión humana.

La tesis sostiene que la atención es una combinación de representaciones dependiente de la entrada y condicionada por parámetros aprendidos, inserta en una arquitectura mayor; su valor explicativo depende de la pregunta que se formule y de la evidencia disponible, mientras un diagrama o una visualización de pesos sólo cubren una parte del problema.

Conviene distinguir arquitectura, ejecución y desempeño: la arquitectura establece operaciones y conexiones posibles, la ejecución produce cantidades para una entrada particular y el desempeño expresa si la salida cumple una tarea; confundir esos niveles conduce a inferencias como suponer que toda información accesible fue utilizada correctamente o que una operación eficiente vuelve confiable cualquier respuesta.

El método empleado es una revisión documental explicativa de fuentes primarias, complementada por cálculos y casos hipotéticos reproducibles; no se entrenó un modelo, no se midió una aplicación comercial y no se realizó una revisión sistemática exhaustiva, por lo que las conclusiones no constituyen una clasificación del mercado ni una estimación del comportamiento de todos los sistemas existentes.

El corte documental corresponde al 2 de octubre de 2026, hora de Oaxaca; se consultan versiones identificadas de trabajos publicados entre 2015 y 2024 y una revisión de prepublicación de 2023, sin presentar sus resultados históricos como características verificadas de productos disponibles en esa fecha de corte.

El artículo se concentra en la circulación y transformación de información entre posiciones; la división del texto en tokens, el ajuste de parámetros durante el entrenamiento y las estrategias de selección de la siguiente unidad se desarrollan en otros números de la colección, aunque aquí se introducen las distinciones necesarias para evitar errores de interpretación.

La atención antes del Transformer

La atención no nació con el Transformer: Bahdanau et al. (2015) propusieron para traducción un mecanismo que construía un contexto diferente para cada paso del decodificador a partir de una suma ponderada de representaciones de la entrada; el modelo conservaba componentes recurrentes y aprendía conjuntamente la alineación y la traducción, como muestran sus ecuaciones de la sección 3.1.

El antecedente importa porque separa una operación de la arquitectura que la utiliza; puede existir atención dentro de una red recurrente y puede diseñarse un Transformer con distintas disposiciones de sus componentes, de modo que

encontrar la palabra atención en la documentación no identifica por sí sola todo el sistema.

En el esquema recurrente de ese trabajo, el estado de un paso dependía del estado anterior; el mecanismo añadido permitía consultar representaciones de diferentes partes de la entrada, en lugar de exigir que una única representación de tamaño fijo cargara con toda la información necesaria para generar la traducción (Bahdanau et al., 2015).

La propuesta de Vaswani et al. (2017) reorganizó el procesamiento de secuencias alrededor de capas de atención y transformaciones por posición, con codificador y decodificador; la eliminación de recurrencia alineada con las posiciones facilitó paralelizar cálculos durante el procesamiento de secuencias conocidas, sin convertir la generación autorregresiva en una producción completa y simultánea de la salida.

Para apreciar la diferencia puede imaginarse una tarea de revisión documental: un procedimiento obliga a que cada elemento reciba únicamente el resumen acumulado por el anterior, mientras otro permite combinar directamente



representaciones de varias posiciones; esta comparación describe rutas de comunicación y no demuestra que alguna de ellas comprenda, recuerde o juzgue como una persona.

La reducción de la distancia entre posiciones dentro del grafo de cómputo tampoco resuelve por sí sola el problema de aprender relaciones útiles; una conexión disponible puede recibir un peso poco conveniente, transportar información redundante o quedar compensada por otras operaciones, por lo que la posibilidad arquitectónica y el resultado efectivo deben mantenerse separados.

El nombre Transformer designa una familia de diseños y no una lista universal de dimensiones, capas o cabezas; reproducir los números de una publicación histórica como si fueran especificaciones de cualquier modelo posterior impediría reconocer qué cambia realmente entre sistemas y qué permanece como principio de construcción.

Consultas claves y valores

Una representación es un conjunto ordenado de números que participa en el cálculo; en una secuencia textual puede

asociarse a un token, aunque después de varias transformaciones ya incorpore información procedente de otras posiciones y no deba interpretarse como el significado aislado de una palabra.

La atención por producto punto escalado compara consultas con claves y utiliza las compatibilidades resultantes para ponderar valores; Vaswani et al. (2017, sección 3.2) describen cómo esas cantidades se obtienen mediante proyecciones aprendidas y cómo las salidas de distintas cabezas se combinan posteriormente.

Consulta, clave y valor son nombres funcionales, no objetos de una base de datos que necesariamente contengan una pregunta escrita, una etiqueta legible y su respuesta; una misma representación puede originar cantidades distintas según la transformación aplicada, del mismo modo que una misma lista de datos admite varias operaciones sin dejar de ser la entrada de partida.

La consulta permite calcular qué claves son compatibles con ella, mientras el valor aporta el contenido numérico que finalmente se combina; distinguir ambos papeles evita afirmar que el peso asignado a una posición describe toda

la información que esa posición transporta o todo el efecto que tendrá en la salida, como sintetiza la Tabla 1.

Tabla 1. *Funciones que conviene separar al leer una explicación de atención*

Componente	**Operación o función**	**Lo que no acredita por sí solo**
Consulta	Participa en la comparación con las claves disponibles	Una intención o pregunta consciente
Clave	Ofrece una representación con la cual calcular compatibilidad	La verdad del contenido asociado
Valor	Aporta el vector que se incorpora a la combinación	Una frase literal recuperada de una fuente
Peso	Determina la proporción de un valor en una combinación concreta	Su importancia causal para toda la respuesta

Componente	Operación o función	Lo que no acredita por sí solo
Máscara	Restringe qué posiciones pueden participar	Una evaluación del contenido permitido

Nota. Elaboración propia a partir de las funciones descritas por Vaswani et al. (2017, sección 3.2); la tercera columna explicita límites de interpretación, no resultados experimentales adicionales.

El producto punto multiplica componentes correspondientes de dos vectores y suma los resultados; si las coordenadas se han aprendido para una tarea, su magnitud puede servir para calcular compatibilidad, pero no tiene por qué coincidir con una semejanza que una persona reconocería al leer dos palabras fuera de contexto.

Los puntajes se escalan y después se normalizan mediante softmax, operación que convierte los puntajes de una fila en pesos positivos cuya suma es uno cuando las entradas son finitas; al excluir posiciones con una



máscara, esas posiciones reciben peso nulo en la formulación ideal y las restantes se normalizan dentro del conjunto permitido.

Decir que esos pesos forman una distribución normalizada no significa que expresen la probabilidad de que cada fragmento sea verdadero; la interpretación de una cantidad depende del evento y del procedimiento que la definen, de modo que trasladar el vocabulario probabilístico al terreno de la confiabilidad factual exige una justificación que esta operación no proporciona.

La combinación resultante es un vector y normalmente no es todavía una palabra ni una respuesta; para llegar a una predicción intervienen otras transformaciones y una salida apropiada para la tarea, de manera que un ejemplo de atención debe indicar dónde termina su alcance y qué operaciones ha dejado fuera.

Un cálculo que puede comprobarse

Construyamos un ejemplo hipotético con una consulta de dos componentes, q = (1, 0), dos claves, k1 = (1, 0) y k2 = (0, 1), y dos valores, v1 = (2, 0) y v2 = (0, 2); los números



se eligen exclusivamente para facilitar el cálculo y no provienen de una ejecución de un modelo de lenguaje.

El producto punto entre q y k1 es 1, mientras el producto entre q y k2 es 0; al dividir ambos por la raíz cuadrada de la dimensión de las claves, que en este caso es 2, los puntajes escalados quedan aproximadamente en 0.707107 y 0.

Para obtener el primer peso se divide la exponencial de 0.707107 entre la suma de esa exponencial y la de 0; el resultado es aproximadamente 0.669762, mientras el segundo peso es 0.330238, de modo que la suma vale uno salvo diferencias de redondeo.

La primera componente de la salida se obtiene multiplicando 0.669762 por 2 y sumando 0.330238 por 0; la segunda se calcula con los valores intercambiados, por lo que la combinación produce aproximadamente (1.339523, 0.660477), calculada con los pesos completos antes de redondear, como resume la Tabla 2.

Tabla 2. *Resultados del ejemplo hipotético de dos posiciones*

Magnitud	Posición 1	Posición 2
Producto punto	1	0
Puntaje escalado	0.707107	0.000000
Peso normalizado	0.669762	0.330238
Aporte a componente 1	1.339523	0.000000
Aporte a componente 2	0.000000	0.660477

Nota. Elaboración propia; los resultados se calcularon sin redondeo intermedio y se muestran con seis decimales, por lo que recomputar desde los valores impresos puede producir diferencias en el último decimal.

El resultado no selecciona exclusivamente el primer valor: conserva aportes de ambos, aunque en proporciones diferentes; describir la atención como una búsqueda que siempre devuelve una única ficha perdería precisamente

esta propiedad de combinación, salvo que se estudie una variante diseñada con otro criterio.

Ahora cambiemos la consulta por (0, 1) y mantengamos las claves y los valores; los puntajes y los pesos intercambian sus posiciones, por lo que la salida pasa a (0.660477, 1.339523), mostrando cómo una modificación de la consulta puede cambiar la combinación sin alterar el conjunto de valores disponibles.

La magnitud de la consulta también importa: multiplicarla por una constante positiva modifica los puntajes relativos que recibe softmax y puede concentrar la distribución, aun cuando la dirección del vector permanezca igual; por ello no conviene reducir toda compatibilidad a la idea intuitiva de cercanía angular entre dos elementos.

Esta observación ayuda a entender la función del escalamiento, pero no convierte el factor utilizado en una regla infalible de aprendizaje; su justificación en el diseño original se relaciona con el comportamiento de productos punto y gradientes bajo determinados supuestos, mientras cada arquitectura conserva decisiones adicionales que afectan su estabilidad (Vaswani et al., 2017, sección 3.2.1).



Introduzcamos después una restricción que impida utilizar la segunda clave; el peso de esa posición pasa a cero y la única posición permitida recibe peso uno, de modo que la salida resulta (2, 0), sin necesidad de cambiar los valores ni afirmar que la información excluida carezca de importancia semántica.

El ejemplo contiene una precaución de implementación: debe existir al menos una posición válida para aplicar esta normalización de la manera descrita; una fila completamente bloqueada exige un tratamiento específico y no autoriza dividir entre una suma nula o interpretar cualquier resultado numérico como una distribución significativa.

Cuando los pesos cambian y la salida no

Supongamos ahora que ambos valores son idénticos e iguales a (3, 5), manteniendo dos pesos que suman uno; cualquier reparto de esos pesos produce la misma salida (3, 5), por lo que una modificación visible del mapa de atención no implica necesariamente una modificación de la combinación.



También pueden coincidir salidas con valores distintos: para los valores escalares 2, 4 y 10, los pesos 0.5, 0.3 y 0.2 producen 4.2, y los pesos 0.2, 0.7 y 0.1 producen igualmente 4.2; este contraejemplo aritmético demuestra falta de unicidad en esa operación reducida, sin pretender reproducir un experimento sobre una red completa.

En cambio, aplicar los pesos 0.2, 0.3 y 0.5 a esos mismos valores produce 6.6; la atención puede modificar la salida de forma importante, pero la dirección y magnitud del cambio dependen de los valores y de las operaciones posteriores, no únicamente del tamaño visual de un peso.

Estos cálculos permiten formular una pregunta más precisa que «a qué palabra miró»: qué representación se ponderó, con qué coeficiente, en qué capa y con qué efecto después de las transformaciones restantes; responderla requiere observar más información que una lista de colores sobre el texto.

Qué posiciones pueden comunicarse

En autoatención, las consultas, claves y valores proceden de representaciones de una misma secuencia; en atención



cruzada, las consultas proceden de una secuencia o conjunto de representaciones y las claves y valores de otro, como sucede entre el decodificador y la salida del codificador en la arquitectura de 2017 (Vaswani et al., 2017, sección 3.2.3).

La diferencia puede visualizarse mediante un caso hipotético de traducción: relacionar dos posiciones del texto de entrada corresponde a una operación dentro de esa secuencia, mientras combinar una representación de la salida parcial con representaciones de la entrada implica una relación entre conjuntos; ninguna de las dos operaciones tiene por qué reproducir una correspondencia palabra por palabra.

En la autoatención causal se impide que una posición utilice información de posiciones posteriores; para predecir la siguiente unidad durante el entrenamiento, la secuencia objetivo se desplaza de forma coherente con la máscara, de modo que la respuesta que se busca predecir no esté disponible en la entrada que la predice.

La palabra causal en este contexto describe una restricción de acceso dentro de una secuencia ordenada;



no significa que el modelo identifique causas del mundo, pruebe relaciones causales entre hechos o produzca una explicación científica de por qué ocurrió un acontecimiento.

Una máscara triangular sencilla permite que la primera posición consulte la primera, que la segunda consulte las dos primeras y que la tercera consulte las tres primeras; el patrón ilustra quién puede aportar información en cada fila, mientras los pesos concretos se calculan después dentro de esas conexiones permitidas.

En una atención bidireccional sin esa restricción, cada posición puede utilizar otras ubicadas a ambos lados, siempre que las demás máscaras lo permitan; esto resulta compatible con tareas donde toda la entrada está disponible, como clasificar un texto ya recibido, aunque no sea el mismo planteamiento que generar una secuencia desconocida paso a paso.

BERT ofrece un ejemplo documentado de codificador Transformer bidireccional: Devlin et al. (2019, sección 3) describen esa arquitectura y su preentrenamiento con predicción de tokens seleccionados, además de representaciones posicionales aprendidas; el caso



muestra que Transformer no es sinónimo de generador autorregresivo ni de chatbot.

La máscara causal y el enmascaramiento de tokens de BERT cumplen funciones distintas: la primera restringe conexiones de atención, mientras el segundo forma parte de la construcción de una tarea de predicción durante el preentrenamiento; compartir la palabra máscara no vuelve intercambiables los procedimientos.

También existen posiciones de relleno para agrupar secuencias de distinta longitud en un mismo lote; impedir que ese relleno participe como contenido útil es otra función de una máscara, por lo que una explicación técnica debe aclarar qué restricción se aplica y qué problema pretende evitar.

Para quien revisa una aplicación, esta distinción se traduce en preguntas verificables: si el sistema analiza un documento completo o una transmisión parcial, si dispone de toda la información al producir cada salida y si se han omitido partes por límites de entrada; una respuesta errónea puede deberse a restricciones de acceso antes de involucrar la calidad de las relaciones aprendidas.



El orden necesita representación

Las frases «El comité supervisa al equipo» y «El equipo supervisa al comité» contienen casi el mismo material léxico, pero asignan funciones diferentes; cualquier procedimiento que ignore esas relaciones podría confundirlas, incluso si identifica correctamente todas las palabras que aparecen.

Una operación de autoatención sin información posicional ni otra señal de orden trata de manera equivalente una permutación coherente de sus entradas y salidas; esta propiedad matemática permite comprender por qué el contenido de los vectores, por sí solo, no ofrece automáticamente una descripción de quién apareció antes o después.

El diseño original añadió codificaciones posicionales sinusoidales a las representaciones de entrada y comparó esa opción con representaciones aprendidas; la información de posición se incorporaba así al cálculo sin reconstruir una cadena recurrente que obligara a procesar cada elemento después del anterior (Vaswani et al., 2017, sección 3.5).



Representar una posición absoluta y representar una distancia relativa son problemas relacionados pero distintos; saber que un elemento ocupa el lugar veinte no expresa directamente que está dos posiciones después de otro, aunque una arquitectura pueda aprender a aprovechar ambas clases de información.

Su et al. (2023, secciones 3.1 y 3.2) describen RoPE, una incorporación de posición mediante rotaciones de consultas y claves que introduce una dependencia relativa en su producto punto; aquí se cita la versión cinco de la prepublicación consultada y no se mezclan sus datos con los de una publicación posterior.

La diferencia entre sumar una codificación y transformar consultas o claves por rotación muestra que el orden no tiene una única implementación; reconocer esa variedad evita presentar un dibujo introductorio como especificación exhaustiva de todos los modelos que emplean atención.

Una representación posicional calculable para una secuencia larga no demuestra que el modelo haya aprendido a trabajar bien en esa longitud; la fórmula puede seguir produciendo números cuando cambian las

distancias y, aun así, la distribución de situaciones puede diferir de la utilizada durante el aprendizaje.

En una tarea hipotética de cronología, una fecha situada antes de otra en el documento no tiene por qué ser anterior en el calendario; el sistema debe distinguir orden de aparición y relación temporal descrita, por lo que la presencia de una codificación posicional no sustituye la evaluación de la conclusión.

El mismo problema aparece cuando una cláusula posterior corrige una descripción anterior: no basta con preferir lo último por su ubicación ni con conservar lo primero por haber aparecido antes, porque la relación depende del contenido y del alcance de la corrección; la arquitectura ofrece recursos para representar esas relaciones, mientras su aplicación requiere comprobarse.

Una capa contiene más que atención

La salida de una cabeza de atención no agota el procesamiento del bloque; en la arquitectura original intervienen varias cabezas, proyecciones de salida, conexiones residuales, normalización y redes de

transformación por posición, cuya combinación determina qué representación pasa al bloque siguiente (Vaswani et al., 2017, secciones 3.1 a 3.3).

Una cabeza es una instancia del mecanismo con sus propias proyecciones de consultas, claves y valores; permite calcular relaciones diferentes dentro de un mismo bloque, pero no debe describirse automáticamente como un especialista humano al que se asignó de antemano gramática, memoria o lógica.

Si una inspección encuentra un patrón asociado con cierta relación lingüística, todavía debe determinar su estabilidad entre entradas y su efecto al intervenir el modelo; bautizar una cabeza con una función a partir de un ejemplo atractivo puede convertir una observación local en una explicación más amplia que la evidencia.

Las salidas de las cabezas se combinan y proyectan, por lo que tampoco existe una correspondencia necesaria entre el número de cabezas y el número de asuntos que el sistema puede atender correctamente; aumentar un componente arquitectónico no ofrece, por sí solo, una medida de calidad para una tarea profesional.

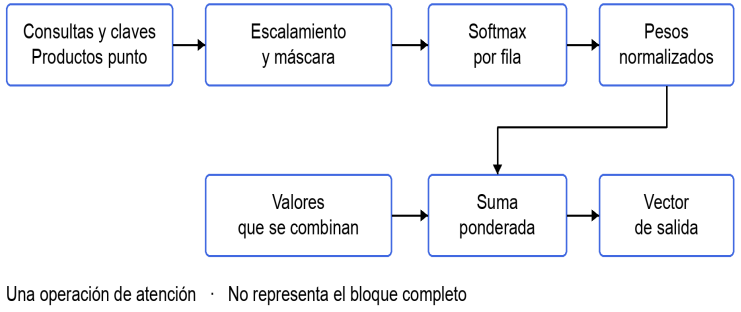


Una conexión residual conserva una ruta por la que la representación de entrada se suma a una transformación del bloque; como consecuencia, un mapa de atención no describe toda la información que puede llegar al siguiente punto de cómputo, incluso cuando el mapa se haya extraído correctamente.

La red aplicada por posición transforma el vector de cada lugar con parámetros compartidos entre posiciones dentro de esa capa; aunque no mezcle directamente posiciones en ese paso, opera sobre representaciones que ya pueden incorporar información combinada anteriormente, por lo que llamarla independiente del contexto sería equívoco.

La normalización modifica las representaciones conforme a una regla de cálculo y contribuye a organizar la dinámica de la red; su ubicación y detalles pertenecen al diseño concreto, de modo que describir el orden de operaciones de 2017 no debe convertirse en una afirmación universal sobre todas las variantes posteriores; la Figura 1 aísla el recorrido de una operación de atención.

Figura 1. *Recorrido reducido de una operación de atención*



Nota. Diagrama de elaboración propia basado en las operaciones descritas por Vaswani et al. (2017, sección 3.2); representa una sola operación y omite proyecciones iniciales, múltiples cabezas, conexiones residuales y otros componentes del bloque.

Apilar bloques permite transformar repetidamente las representaciones; una conexión de una capa tardía puede combinar información que ya recorrió varias posiciones, de modo que leer su columna como si contuviera únicamente la palabra original borraría parte del historial de procesamiento.



Esta observación tiene una consecuencia práctica para la divulgación: un buen diagrama debe especificar su nivel de detalle y sus omisiones; mostrar una operación simplificada es útil, siempre que el lector no tenga que adivinar si observa una cabeza, un bloque, todo el modelo o una aplicación que además consulta herramientas externas.

Paralelismo y costo computacional

Durante el procesamiento de una secuencia conocida, varias consultas pueden organizarse en matrices y calcularse conjuntamente; la máscara causal permite restringir las dependencias aunque las operaciones se ejecuten de manera paralela, por lo que paralelismo físico y disponibilidad lógica de información no significan lo mismo.

Una analogía limitada ayuda a verlo: varias personas pueden resolver simultáneamente ejercicios impresos con información diferente en cada hoja, sin que la simultaneidad autorice a consultar respuestas reservadas; en el cálculo de atención, la restricción se expresa



mediante máscaras y estructura de datos, no mediante una decisión moral del sistema.

En generación autorregresiva, la siguiente unidad depende del prefijo disponible y la unidad seleccionada amplía ese prefijo; el procedimiento habitual conserva por ello una dependencia entre pasos de salida, aunque dentro de cada paso se paralelicen numerosas operaciones y existan técnicas para reducir trabajo repetido.

El procesamiento inicial de una entrada extensa y la producción sucesiva de una respuesta pueden presentar costos distintos; para evaluar una aplicación conviene medir por separado el tiempo hasta que empieza a responder y el tiempo que necesita para completar la salida, en lugar de atribuir ambos a una sola palabra como velocidad.

La atención densa compara cada consulta con todas las claves permitidas de la secuencia; si se consideran todas las parejas de una secuencia de longitud n, la matriz correspondiente contiene n por n posiciones, aunque una implementación no necesite almacenar cada matriz intermedia completa.



Con mil posiciones hay un millón de parejas ordenadas, con dos mil hay cuatro millones y con ocho mil hay sesenta y cuatro millones; estas cantidades son un cálculo ilustrativo de tamaño de matriz, no una medición de memoria total, energía, tiempo de ejecución ni dinero cobrado por un proveedor.

El crecimiento cuadrático de esa parte del cálculo no permite inferir que todo el costo del modelo se multiplique siempre por cuatro al duplicar la longitud; también intervienen anchura de las representaciones, número de capas, proyecciones, tamaño del lote, organización de memoria y características de la implementación.

Dao et al. (2022, secciones 2.2 y 3.1) muestran con FlashAttention que puede reorganizarse el cálculo exacto de atención mediante bloques y recomputación para reducir transferencias de memoria; la propuesta evita materializar grandes matrices intermedias en memoria de alto ancho de banda, sin sustituir por ello la operación densa por una regla semántica diferente.

La palabra exacto se refiere aquí a la operación matemática que se pretende calcular, con las



precauciones propias de la aritmética numérica; no significa que las respuestas sean verdaderas, que dos equipos produzcan necesariamente los mismos bits ni que desaparezca el crecimiento cuadrático de todas las operaciones aritméticas.

Comparar eficiencia exige conservar una tarea y una calidad de referencia, además de identificar equipo, longitud y configuración; una reducción de accesos a memoria puede ser decisiva en un entorno y tener otra importancia relativa en uno distinto, por lo que no se trasladan al presente las aceleraciones de una publicación como promesas universales.

Otra decisión consiste en limitar las conexiones, por ejemplo a vecindades locales; eso modifica qué posiciones pueden comunicarse directamente y puede exigir rutas adicionales para relaciones distantes, de modo que ahorrar trabajo mediante restricciones no equivale a ejecutar de forma más eficiente exactamente la misma atención densa.

Para un usuario, la pregunta útil es si la aplicación conserva la calidad necesaria con el costo y la demora

aceptables; conocer la arquitectura ayuda a interpretar los compromisos, pero una elección responsable necesita observar el resultado del flujo completo y no únicamente una operación interna aislada.

Pesos de atención y explicación

Un mapa de atención representa coeficientes de una operación concreta, habitualmente mediante intensidades de color; ofrece una observación valiosa del cálculo si se conoce cómo fue extraído, pero la conclusión de que explica una respuesta requiere precisar qué se entiende por explicación.

Puede buscarse una descripción de la arquitectura, una indicación de fragmentos relevantes para una persona o una identificación causal de componentes que cambiaron una predicción; esos objetivos no son equivalentes, aunque una misma visualización pueda resultar útil como punto de partida para varios de ellos.

Jain y Wallace (2019) compararon pesos de atención con medidas de importancia basadas en gradientes y eliminación de elementos, además de explorar



distribuciones alternativas que conservaran predicciones; sus resultados cuestionaron la interpretación automática de esos pesos como explicación en los modelos y tareas estudiados, principalmente con codificadores recurrentes bidireccionales.

Wiegreffe y Pinter (2019) objetaron aspectos de esa interpretación y de la libertad concedida a las distribuciones alternativas; propusieron controles con pesos uniformes, variación entre inicializaciones, un modelo diagnóstico y adversarios entrenados de forma consistente, trabajando con clasificación binaria y arquitecturas LSTM, un tipo de red recurrente.

El contraste no se resuelve eligiendo el título más contundente: ambos trabajos obligan a examinar el vínculo entre un coeficiente y la predicción, mientras discrepan sobre qué intervenciones y criterios permiten valorar su utilidad explicativa; ninguno constituye una auditoría general de todos los Transformers ni de cada mapa producido por aplicaciones posteriores.

Si dos configuraciones de atención preservan una predicción, puede demostrarse que esa salida no identifica



de forma única la distribución utilizada; para sostener además que una explicación particular carece de valor, debe justificarse qué propiedad explicativa se exigía y por qué las configuraciones comparadas son pertinentes para esa exigencia.

En sentido inverso, que un patrón resulte estable o útil para un diagnóstico tampoco garantiza que sea una explicación causal completa; puede aportar evidencia parcial sobre un mecanismo y dejar fuera otras rutas, interacciones o representaciones que también contribuyen a la salida.

Los ejemplos aritméticos anteriores permiten comprender esta cautela sin resolver por anticipado la controversia experimental: si los valores coinciden, los pesos pueden variar sin cambiar la combinación, mientras valores diferentes pueden producir cambios importantes; sólo observar los coeficientes omite una parte necesaria del cálculo.

Qué debería acompañar a una visualización

Una visualización informativa debería identificar el modelo y su versión, la entrada, el bloque, la cabeza y la posición

de consulta; también debería aclarar si muestra pesos originales, promedios entre cabezas o alguna transformación adicional, porque promediar puede ocultar diferencias relevantes y ya constituye una decisión analítica.

La escala de color requiere atención propia: un tono intenso puede significar valores altos dentro de una fila, dentro de toda la figura o después de una normalización gráfica; comparar imágenes con escalas distintas como si expresaran magnitudes directamente equivalentes introduce una inferencia que el gráfico quizá no permita.

También importa cómo se agrupan tokens para mostrar palabras: sumar, promediar o tomar el máximo entre fragmentos produce resúmenes diferentes; una interfaz legible puede resultar útil para una persona y, al mismo tiempo, ocultar que la unidad visual ya no coincide con la unidad exacta de cálculo.

Una prueba causal debe explicitar la intervención y la salida medida; modificar una entrada, fijar activaciones o alterar pesos son operaciones distintas, cada una con supuestos propios, por lo que observar un cambio no basta



para atribuirlo a una única causa si se alteraron simultáneamente contenido, posición y estructura.

La estabilidad tampoco debe confundirse con corrección: un sistema puede repetir consistentemente una respuesta equivocada, y un mapa puede mantenerse parecido mientras cambia una negación decisiva; la revisión del contenido necesita un criterio externo a la apariencia del gráfico.

Contexto disponible y contexto utilizado

Una ventana de contexto indica una capacidad de entrada bajo determinadas condiciones, pero no mide directamente cuántas relaciones complejas pueden resolverse de forma fiable; disponer de espacio para incluir un documento es distinto de extraer su información pertinente y aplicarla correctamente.

Liu et al. (2024) estudiaron preguntas sobre varios documentos y recuperación de pares clave y valor, variando la ubicación de la información relevante; observaron sensibilidad a la posición en varios modelos evaluados y, en distintas condiciones, mejor desempeño



cuando esa información aparecía al principio o al final que en el centro.

El alcance de ese resultado es histórico y experimental: las tareas, versiones y configuraciones están descritas en el estudio y no autorizan afirmar que cualquier modelo de 2026 falla siempre en el centro; su contribución para este análisis consiste en mostrar que longitud nominal y uso efectivo requieren evaluaciones diferentes.

El estudio tampoco identifica por sí solo una única causa interna de todos los fallos de contexto; observar una curva de desempeño al cambiar posiciones caracteriza el comportamiento y permite contrastar hipótesis, mientras atribuirla a una cabeza específica, a la posición o a otro componente exige pruebas adicionales.

Una aplicación puede añadir además recuperación de documentos, selección de fragmentos, resúmenes y herramientas de búsqueda; cuando una respuesta omite información, primero conviene comprobar qué recibió realmente el modelo, porque el documento visible para la persona y la entrada procesada no siempre son el mismo objeto.



Si una excepción desaparece durante la extracción de un PDF, aumentar la ventana de contexto no la recupera automáticamente; si la excepción se recibió pero se aplicó a una persona distinta, el problema ya no se resuelve simplemente demostrando que su texto estaba presente, de modo que el diagnóstico debe seguir el recorrido de la información.

De manera análoga, una referencia correcta puede acompañar una conclusión incorrecta: citar el párrafo que contiene una condición no prueba que se haya interpretado su alcance; por eso una evaluación documental debe separar localización del fragmento, reproducción fiel, relación con el caso y conclusión final.

Tres casos para comprobar relaciones

Una excepción que cambia la respuesta

Consideremos un reglamento ficticio de préstamo de equipo: la regla general permite retirar una cámara por dos días, pero una cláusula autoriza cinco días para proyectos previamente registrados; el caso consultado corresponde a un proyecto registrado y la respuesta correcta, dentro de



este escenario inventado, debe aplicar el plazo de cinco días.

La tarea no requiere decidir si una norma real es válida ni ofrecer asesoría jurídica, sino conservar una relación lógica definida en el propio ejemplo; una salida que cite ambas cláusulas y concluya dos días habría recuperado información pertinente sin aplicar correctamente la excepción.

Podrían prepararse tres versiones que ubiquen la cláusula al principio, en el centro o al final, conservando el resto del contenido y su significado; el propósito sería observar si la ubicación cambia la respuesta, sin anticipar que ocurrirá un fallo ni atribuir sus posibles causas a una única parte del modelo.

Debe incluirse también un caso de proyecto no registrado, cuya respuesta seguiría la regla de dos días; si sólo se evaluaran excepciones aplicables, un sistema que siempre respondiera cinco días podría aparentar éxito sin discriminar la condición relevante.

Una variante adicional podría declarar que el registro ocurrió después de la solicitud y dejar sin especificar si basta con registrarse posteriormente; en ese caso la evaluación debería admitir que falta información, porque obligar al sistema a elegir uno de los plazos convertiría la propia prueba en una fuente de certeza injustificada.

Dos antecedentes y una ambigüedad real

Imaginemos la frase «Marina habló con Elena cuando ella terminó el informe»; sin información adicional, el pronombre puede admitir más de una interpretación, por lo que una respuesta que elige a una persona con absoluta seguridad no es necesariamente mejor que otra que reconoce la ambigüedad.

Para construir un caso inequívoco puede añadirse una segunda oración que indique quién había redactado y concluido el informe; después se cambiarían nombres y posiciones, manteniendo las relaciones, con objeto de comprobar si la respuesta conserva el vínculo y no depende únicamente de la proximidad superficial de un nombre.

La evaluación debe distinguir una entrada ambigua de un error del sistema: si el documento no permite decidir, la salida adecuada puede describir alternativas y solicitar el dato faltante; adjudicar una respuesta única después de ver lo que produjo el modelo sesgaría el análisis.

Un mapa que concentre peso en Elena no resuelve la ambigüedad del texto original; muestra una propiedad de una operación, mientras la determinación de qué permite concluir el documento requiere criterios lingüísticos y contextuales que no desaparecen por visualizar un coeficiente.

Una corrección que conserva el hecho anterior

Supongamos que un registro ficticio señala primero una reunión para el martes y después añade una corrección explícita que la traslada al jueves; la pregunta por la fecha vigente debe responder jueves, mientras la pregunta por la fecha anunciada originalmente debe responder martes.

Este caso obliga a relacionar contenido, orden y alcance de la pregunta; responder siempre con la última fecha puede acertar una consulta y fallar la otra, por lo que una



prueba que no distinga ambos objetivos no permite saber qué relación utilizó el sistema.

También conviene separar una corrección de una mera mención: un párrafo posterior que recuerde una reunión del jueves anterior no modifica necesariamente el anuncio del martes; intercambiar estos casos ayuda a detectar una regla superficial de escoger la fecha más reciente en el texto.

Los tres escenarios comparten una enseñanza práctica: la calidad depende de conservar relaciones y condiciones, no de repetir fragmentos llamativos; sus números, nombres y reglas son construcciones didácticas y no se atribuyen a personas, organizaciones ni resultados medidos.

Una evaluación pequeña pero informativa

Puede construirse una prueba propuesta con tres posiciones de la cláusula, dos longitudes de documento, dos formulaciones equivalentes y dos condiciones de aplicabilidad; el producto de esas dimensiones genera veinticuatro casos por escenario base, cantidad que

organiza el diseño y no representa una muestra ya ejecutada.

Las formulaciones equivalentes deben revisarse antes de comenzar, porque cambiar una negación, un plazo o una condición alteraría la respuesta correcta; mover contenido también exige conservar referencias como «la cláusula anterior», cuyo significado podría cambiar al reordenar párrafos; la Tabla 3 reúne los criterios propuestos.

Tabla 3. *Aspectos que registrar en la prueba propuesta*

Aspecto	Comprobación	Error que ayuda a distinguir
Entrada efectiva	Conservar el texto realmente procesado	Omisión previa a la inferencia
Recuperación	Localizar los fragmentos pertinentes	Falta de acceso o selección útil
Aplicación	Relacionar condición y caso correctamente	Cita adecuada con conclusión errónea

Aspecto	Comprobación	Error que ayuda a distinguir
Incertidumbre	Reconocer cuando falta un dato decisivo	Elección injustificadamente segura
Estabilidad	Comparar variantes con significado conservado	Sensibilidad a posición o formulación

Nota. Propuesta de elaboración propia; la comparación por posición se inspira en la pregunta experimental de Liu et al. (2024), mientras las condiciones y criterios de esta tabla corresponden a los casos hipotéticos del artículo.

Antes de ejecutar, debe establecerse una clave de evaluación que defina las respuestas aceptables y el tratamiento de la ambigüedad; conviene separar acierto de conclusión, calidad de los fragmentos citados y explicación de la condición, en lugar de ocultar todos los fallos dentro de una puntuación única.



La prueba debe conservar el nombre y versión del modelo, las instrucciones, la configuración de generación y cualquier transformación de los documentos; si una aplicación no permite fijar alguno de esos elementos, esa limitación debe acompañar la interpretación de los resultados.

Repetir ejecuciones puede revelar variación de salida, aunque tampoco garantiza cobertura de todos los comportamientos posibles; la cantidad de repeticiones debería responder al uso previsto y a la variabilidad observada, sin confundir una coincidencia ocasional con estabilidad suficiente para una decisión importante.

El análisis por grupos resulta más informativo que un promedio aislado: un sistema puede acertar la mayoría de casos breves y fallar todos aquellos donde la excepción aparece en el centro de un documento largo; ese patrón cambia la forma de utilizarlo aunque la cifra agregada parezca aceptable.

Si se modifica una instrucción para mejorar los resultados, debe conservarse un conjunto distinto de casos para comprobar la mejora; evaluar únicamente sobre los



ejemplos que guiaron la modificación puede ocultar que se adaptó la solución a un conjunto reducido y no a la relación general que se pretendía probar.

Cuando se detecte un fallo, conviene empezar por una reproducción mínima que mantenga la condición decisiva; retirar contenido puede facilitar el diagnóstico, pero debe hacerse con cuidado para no eliminar precisamente la dificultad que originó el error y luego afirmar que el problema quedó resuelto.

Las correcciones posibles pertenecen a distintos niveles: reparar la extracción, seleccionar mejor documentos, aclarar la pregunta, separar una tarea compleja o introducir revisión humana; la evidencia del caso debe orientar la elección, porque modificar el prompt no repara automáticamente un componente que nunca entregó la información necesaria.

Mitos que merecen una respuesta precisa

«La atención permite mirar todo al mismo tiempo» mezcla paralelismo con acceso y uso; algunas operaciones procesan varias posiciones conjuntamente, pero las



máscaras restringen conexiones, la generación puede conservar dependencias secuenciales y la calidad depende de qué relaciones se aprovechan realmente.

«El peso más alto señala la razón de la respuesta» omite los valores combinados, las rutas residuales y las transformaciones posteriores; puede servir como hipótesis inicial, pero una afirmación causal requiere definir la intervención, el resultado y las alternativas pertinentes.

«Una ventana mayor significa mejor lectura» confunde capacidad nominal con desempeño; una evaluación debe comprobar recuperación y aplicación bajo longitudes, posiciones y tareas relevantes, sin convertir un resultado histórico o una prueba sencilla en una garantía para documentos futuros.

«Si el cálculo es exacto, la respuesta será exacta» traslada una propiedad algorítmica al contenido; una implementación puede calcular fielmente una operación de un modelo que produzca una conclusión incorrecta, del mismo modo que una operación aritmética bien ejecutada no corrige premisas falsas.



«Cada cabeza comprende un tema» atribuye una especialización humana sin demostrarla; las cabezas proporcionan subespacios y operaciones diferenciadas, pero asignarles funciones estables requiere investigación sobre el modelo concreto y no una interpretación sugerida por su nombre.

«Un Transformer es sólo atención» ignora componentes indispensables para transformar y transmitir representaciones; la atención es central en la organización de esa familia, pero una descripción del modelo debe incluir los demás componentes y distinguirlos de herramientas externas o funciones de la aplicación.

Estos errores no se corrigen reemplazando una metáfora optimista por otra despectiva; reducir el sistema a una calculadora de promedios también omite las representaciones aprendidas y la composición de muchas operaciones, por lo que una explicación rigurosa debe conservar tanto su capacidad técnica como sus límites demostrables.

Consecuencias para el uso responsable

Comprender atención permite formular mejores preguntas a quien ofrece una aplicación: qué parte del documento llega al modelo, qué longitud se probó realmente, qué ocurre con excepciones distantes y cómo se comprueba una conclusión; las respuestas útiles describen procedimientos y evidencia, no únicamente nombres de arquitecturas.

En tareas documentales, resulta razonable exigir que la salida identifique los fragmentos decisivos y explique cómo se relacionan con el caso; esa práctica facilita la revisión humana, aunque no convierte las citas ni la explicación generada en una observación directa de todos los procesos internos.

Cuando el costo de omitir una condición es alto, debe preservarse una vía independiente de comprobación; la supervisión necesita acceso al documento y criterios de decisión, porque revisar sólo una respuesta fluida o un mapa atractivo no ofrece suficiente base para detectar una excepción mal aplicada.

La transparencia también exige reconocer lo que no se conoce: si una aplicación no expone su arquitectura exacta o sus pesos de atención, puede evaluarse su comportamiento mediante pruebas controladas, pero no debe afirmarse que se observó un mecanismo interno que en realidad permaneció inaccesible.



Conclusión

La atención de un Transformer combina representaciones mediante compatibilidades calculadas para una entrada y bajo restricciones definidas; consultas, claves, valores, posición y máscaras permiten describir esa operación con precisión, mientras las demás partes del bloque explican por qué su salida no equivale todavía a una respuesta completa.

Los cálculos ilustrativos muestran que cambiar pesos puede cambiar la combinación, conservarla o afectar componentes distintos, según los valores disponibles; el contraste entre investigaciones de interpretabilidad confirma la necesidad de formular con cuidado qué explicación se busca y qué observaciones podrían sostenerla.

La consecuencia principal es distinguir acceso, procesamiento y uso correcto: una arquitectura puede permitir una conexión, una ejecución puede asignarle peso y una aplicación puede aun así llegar a una conclusión equivocada; evaluar relaciones, excepciones e incertidumbre ofrece una base más sólida que atribuir concentración humana al sistema.

La pregunta decisiva para una tarea concreta no es si el modelo presta atención en el sentido cotidiano, sino si conserva y aplica la información pertinente de manera comprobable bajo las condiciones de uso; responderla requiere evidencia del sistema y del documento, además de una revisión proporcional a las consecuencias del error.

Referencias

Bahdanau, D., Cho, K., y Bengio, Y. (2015). *Neural machine translation by jointly learning to align and translate* [Traducción automática neuronal mediante el aprendizaje conjunto de alineación y traducción] (Versión 5). arXiv. https://doi.org/10.48550/arXiv.1409.0473

Dao, T., Fu, D. Y., Ermon, S., Rudra, A., y Ré, C. (2022). *FlashAttention: Fast and memory-efficient exact attention with IO-awareness* [FlashAttention: Atención exacta rápida y eficiente en memoria mediante el conocimiento de las operaciones de entrada y salida] (Versión 2). arXiv. https://doi.org/10.48550/arXiv.2205.14135

Devlin, J., Chang, M.-W., Lee, K., y Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding [BERT: Preentrenamiento de Transformers bidireccionales profundos para la comprensión del lenguaje]. En J. Burstein, C. Doran, y T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Jain, S., y Wallace, B. C. (2019). Attention is not explanation [La atención no es explicación]. En J. Burstein, C. Doran, y T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 3543–3556). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1357

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., y Liang, P. (2024). Lost in the middle: How language models use long contexts [Perdidos en el centro: Cómo utilizan los modelos de lenguaje los contextos extensos]. *Transactions of the Association for Computational Linguistics, 12*, 157–173. https://doi.org/10.1162/tacl_a_00638

Su, J., Lu, Y., Pan, S., Murtadha, A., Wen, B., y Liu, Y. (2023). *RoFormer: Enhanced transformer with rotary position embedding* [RoFormer: Transformer mejorado con representación posicional rotatoria] (Versión 5). arXiv. https://doi.org/10.48550/arXiv.2104.09864

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., y Polosukhin, I. (2017). Attention is all you need [La atención es todo lo que necesitas]. En I. Guyon, U. von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, y R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 30). Curran Associates. https://papers.nips.cc/paper/7181-attention-is-all-you-need

Wiegreffe, S., y Pinter, Y. (2019). Attention is not not explanation [La atención no es no explicación]. En K. Inui, J. Jiang, V. Ng, y X. Wan (Eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 11–20). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1002

