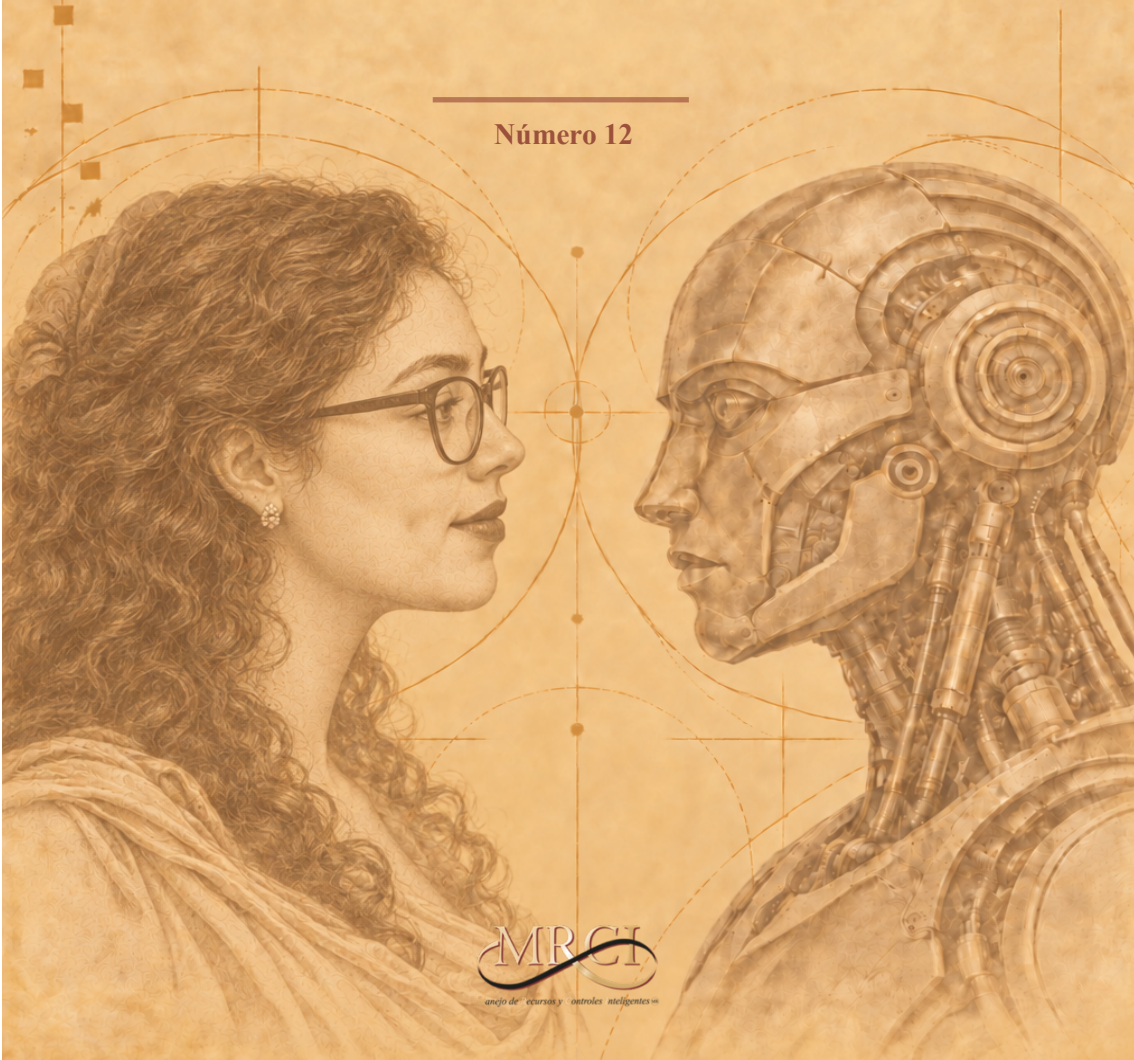
COLECCIÓN

MITOS Y REALIDADES DE LA IA

Entre la promesa y la realidad

TEMPERATURA Y MUESTREO Y POR QUÉ UNA RESPUESTA PUEDE CAMBIAR

Número 12



MRCI

anejo de ecursos y ontroles nteligentes

DR. JOSÉ ANTONIO PÉREZ RAMOS

TEMPERATURA Y MUESTREO Y POR QUÉ UNA RESPUESTA PUEDE CAMBIAR

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*.

TEMPERATURA Y MUESTREO Y POR QUÉ UNA RESPUESTA PUEDE CAMBIAR

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: Temperatura y muestreo y por qué una respuesta puede cambiar.

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; TEMPERATURA Y MUESTREO Y POR QUÉ UNA RESPUESTA PUEDE CAMBIAR, por cualquier medio, sin autorización escrita del autor.

PRINTED IN MEXICO
IMPRESO EN MÉXICO

ÍNDICE

LA PREGUNTA Y EL ALCANCE..... - 7 -

DÓNDE INTERVIENE LA SELECCIÓN ..... - 9 -

QUÉ TRANSFORMA LA TEMPERATURA..... - 11 -

UN EXPERIMENTO ARITMÉTICO REPRODUCIBLE ..... - 14 -

TOP K Y EL NÚMERO DE CANDIDATOS..... - 19 -

TOP P Y LA MASA ACUMULADA ..... - 21 -

EL ORDEN DE LAS OPERACIONES IMPORTA ..... - 24 -

ELEGIR UN TOKEN NO ES ELEGIR LA MEJOR RESPUESTA... - 27 -

CONCENTRACIÓN Y VERDAD SON PREGUNTAS DIFERENTES . - 29 -

REPETICIÓN DE EJECUCIONES Y SEMILLAS ..... - 32 -

COMPARAR CONFIGURACIONES..... - 35 -

TRES SITUACIONES DE USO ..... - 39 -

MÁS ALLÁ DE LOS CONTROLES MÁS CONOCIDOS..... - 41 -

MITOS QUE CONVIENE ABANDONAR ..... - 43 -

CONCLUSIÓN..... - 46 -

REFERENCIAS..... - 47 -



La pregunta y el alcance

¿Por qué una respuesta puede cambiar al modificar un solo número en la configuración de generación y por qué ese cambio no garantiza una mejora? La pregunta tiene una dimensión matemática y otra práctica: podemos describir con precisión qué sucede a una distribución cuando aplicamos determinada transformación, pero necesitamos evidencia adicional para conocer si el texto resultante sirve mejor a una persona, conserva los hechos o satisface una restricción.

La contribución de este artículo consiste en hacer visible ese recorrido mediante operaciones pequeñas que cualquier lector puede reconstruir; el propósito no es recomendar un valor universal de temperatura, sino proporcionar un método para interpretar los controles, reconocer sus interacciones y formular comparaciones que permitan decidir con fundamento dentro de un entorno identificado.

La investigación es documental y explicativa, con corte al 3 de octubre de 2026; se consultaron artículos originales sobre generación y muestreo, así como documentación oficial de Transformers en la versión 4.57.1 y la página de



reproducibilidad de PyTorch disponible en esa fecha, cuyas indicaciones se atribuyen a esos recursos concretos y no a todas las aplicaciones de inteligencia artificial.

El experimento numérico que se presenta es una construcción hipotética ejecutada mediante cálculo, sin consultas experimentales a un modelo de lenguaje; sus resultados demuestran consecuencias de las reglas definidas y no permiten estimar el rendimiento de un producto, afirmar una tasa real de errores ni trasladar los valores elegidos a sistemas que empleen otras distribuciones o procedimientos.

Quedan fuera del desarrollo la explicación general del determinismo, las causas completas de la información falsa, el diseño integral de instrucciones y la calibración de la confianza; esos asuntos requieren tratamientos propios, mientras aquí se examina la etapa que transforma puntuaciones en selecciones y la manera de evaluar sus efectos dentro de una tarea acotada.



Dónde interviene la selección

Para estudiar la generación secuencial conviene observar un solo paso: el sistema dispone de un contexto y produce puntuaciones para los posibles tokens siguientes, después una regla decide cuál se incorpora al texto; el token elegido modifica el contexto del paso posterior, de modo que una variación inicial puede conducir a una continuación diferente sin que haya cambiado el conjunto de parámetros aprendidos durante el entrenamiento.

Un token es una unidad de la representación que utiliza el modelo y no tiene por qué coincidir con una palabra completa; puede representar un fragmento, un signo u otro elemento, por lo que el ejemplo con letras de este artículo funciona como una simplificación de candidatos y no como la descripción literal de un vocabulario comercial.

Las puntuaciones anteriores a la normalización suelen denominarse logits; para convertirlas en probabilidades se utiliza frecuentemente una función softmax, que transforma valores reales en cantidades positivas cuya suma es uno, aunque las restricciones de generación pueden excluir candidatos y asignarles probabilidad cero antes de la selección.



Es importante distinguir tres objetos: las puntuaciones que entrega el modelo, las probabilidades obtenidas tras aplicar las transformaciones y el token efectivamente seleccionado; dos ejecuciones pueden compartir los primeros objetos y diferir en la selección cuando existe muestreo, mientras un cambio de configuración puede modificar la propia distribución desde la que se extrae el resultado.

La documentación versionada de Transformers distingue generación voraz, muestreo multinomial y búsqueda por haces mediante combinaciones de configuración; esa distinción confirma que temperatura y estrategia de selección no son términos intercambiables, puesto que un argumento puede carecer de efecto dentro de una ruta de ejecución que no utiliza el procedimiento al que pertenece (Hugging Face, s.f.-a).

En una aplicación concreta también importa qué controles están expuestos y cuáles se aplican realmente; que una interfaz muestre una casilla o acepte un campo no demuestra por sí solo que ese valor gobierne cada componente de una respuesta, por lo que la interpretación técnica debe apoyarse en documentación correspondiente



al modelo, a la biblioteca y a la ruta de ejecución observada.

Qué transforma la temperatura

Sea z_i la puntuación de un candidato i y T una temperatura estrictamente positiva; la regla usual calcula q_i(T) = exp(z_i / T) / suma_j exp(z_j / T), donde el denominador suma sobre los candidatos considerados y permite que las probabilidades resultantes sumen uno, como se describe en la formulación de muestreo de Holtzman et al. (2020).

Si ya conocemos las probabilidades p_i obtenidas con T = 1, podemos expresar la misma transformación como q_i(T) = p_i^(1/T) / suma_j p_j^(1/T); esta segunda forma permite hacer los cálculos sin conocer los logits originales, siempre que se mantenga el mismo conjunto de candidatos y no se introduzcan otras transformaciones entre ambas representaciones.

La temperatura no debe aplicarse dividiendo cada probabilidad entre T y deteniéndose allí, porque la suma dejaría de ser uno; tampoco basta dividir y normalizar

después, ya que ese factor común se cancelaría y recuperaríamos exactamente la distribución inicial, mientras la operación correcta modifica potencias relativas antes de volver a normalizar.

La relación entre dos candidatos muestra el efecto con claridad: q_i(T) / q_j(T) = (p_i / p_j)^(1/T); si el primero tenía el doble de probabilidad que el segundo, con T = 0.5 pasa a tener cuatro veces su probabilidad, mientras con T = 2 la relación se reduce a la raíz cuadrada de dos, aproximadamente 1.414.

En consecuencia, una temperatura positiva conserva el orden de las probabilidades, salvo empates o efectos numéricos de una implementación; bajar T refuerza la ventaja relativa de los candidatos preferidos y subirla reduce esa ventaja, pero ninguna de esas operaciones introduce información externa que permita saber si el candidato favorecido expresa un hecho verdadero.

El valor T = 1 deja intacta la distribución a la que se aplica esta transformación; no significa que toda la aplicación funcione sin modificaciones, pues podrían existir filtros, restricciones, penalizaciones u otras etapas activas que



intervengan antes o después y cambien la distribución efectiva de selección.

Cuando T se aproxima a cero por valores positivos, la masa se concentra en el máximo si este es único; el caso T = 0 no se sustituye directamente en la fórmula, porque dividir entre cero no está definido, y la aceptación de un cero en un producto debe interpretarse mediante la regla especial que su implementación documente, nunca como una operación ordinaria de softmax.

Si varios candidatos comparten exactamente el máximo, el límite matemático distribuye la masa entre ellos por igual bajo esta transformación; una elección voraz necesita además una regla de desempate, lo que impide identificar sin condiciones el límite de temperatura con una única salida determinada en todos los casos.

En el extremo opuesto, al crecer T sin límite y conservar logits finitos sobre un vocabulario finito, la distribución se aproxima a la uniforme; esto describe una propiedad de la fórmula, no una recomendación de funcionamiento, pues hacer equiprobables candidatos semánticamente muy



distintos puede disminuir la utilidad del texto y no convierte esa dispersión en creatividad valiosa.

Un experimento aritmético reproducible

Consideremos seis candidatos, A, B, C, D, E y F, con probabilidades iniciales 0.40, 0.25, 0.15, 0.10, 0.06 y 0.04; suman exactamente uno y no representan mediciones de ningún modelo, sino una distribución construida para comparar operaciones manteniendo todos los demás elementos bajo control.

Con T = 0.5 se elevan las probabilidades al cuadrado: 0.1600, 0.0625, 0.0225, 0.0100, 0.0036 y 0.0016; la suma es 0.2602, de modo que A recibe 0.1600 / 0.2602, aproximadamente 0.614912, y las demás cantidades se obtienen mediante la misma división.

Con T = 2 se utiliza la raíz cuadrada de cada probabilidad y después se divide entre la suma de esas raíces; A baja aproximadamente a 0.277280 y F sube a 0.087684, lo cual ilustra que el cambio favorece relativamente a los candidatos menos probables sin alterar el orden A, B, C, D, E, F.

Tabla 1. *Distribución hipotética después de aplicar temperatura*

Candidato	T = 0.5	T = 1	T = 2
A	0.614912	0.400000	0.277280
B	0.240200	0.250000	0.219209
C	0.086472	0.150000	0.169798
D	0.038432	0.100000	0.138640
E	0.013836	0.060000	0.107390
F	0.006149	0.040000	0.087684

Nota. Elaboración propia mediante q_i(T) = p_i^(1/T) / suma_j p_j^(1/T); las cifras se redondean de manera independiente a seis decimales y sus sumas visibles pueden diferir levemente de uno.

La tabla también deshace una intuición engañosa: elevar la temperatura no aumenta la probabilidad de cada alternativa que no ocupa el primer lugar; B disminuye de 0.25 a aproximadamente 0.219209, porque el

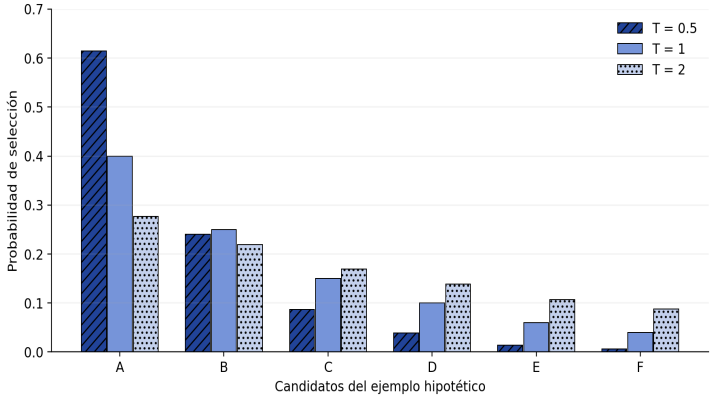


aplanamiento redistribuye masa de acuerdo con toda la distribución y no mediante una regla que premie automáticamente a todos los candidatos distintos del máximo.

Podemos describir la concentración mediante la entropía H = menos suma_i q_i log2(q_i), tomando como cero la contribución de una probabilidad nula; en este ejemplo los resultados son 1.542333 bits para T = 0.5, 2.200797 para T = 1 y 2.476282 para T = 2, cifras que cuantifican dispersión probabilística y no calidad conceptual.

La cantidad 2^H expresa un número efectivo de alternativas equiprobables y resulta aproximadamente 2.913, 4.597 y 5.565, respectivamente; es una comparación de concentración, por lo que un valor de 2.913 no significa que existan literalmente esa cantidad de tokens, sino que la incertidumbre medida equivale a la de un conjunto uniforme de ese tamaño efectivo.

Figura 1. *Concentración de la probabilidad según la temperatura*



Nota. Elaboración propia a partir de los valores sin redondear de la tabla 1; tramas y rótulos distinguen las series además del color, y ninguna barra representa una observación empírica de un modelo.

Para reproducir la selección sin depender de un generador pseudoaleatorio, fijemos seis números: 0.05, 0.28, 0.52, 0.69, 0.83 y 0.95; para cada uno recorremos las probabilidades acumuladas de A a F y elegimos el primer

candidato cuya suma acumulada sea estrictamente mayor que el número fijado.

Con T = 1 los intervalos son [0, 0.40), [0.40, 0.65), [0.65, 0.80), [0.80, 0.90), [0.90, 0.96) y [0.96, 1); los seis números seleccionan A, A, B, C, D y E, mientras con T = 0.5 seleccionan A, A, A, B, B y D y con T = 2 producen A, B, C, D, E y F.

Esta comparación utiliza los mismos números como entradas de tres transformaciones y no pretende que seis elecciones permitan estimar frecuencias poblacionales; los valores se prefijaron para mostrar el mecanismo, no se obtuvieron de un ensayo aleatorio ni acreditan que una temperatura genere siempre determinado repertorio de respuestas.

Tampoco se está generando una oración: cada extracción vuelve a la misma distribución inicial y corresponde a un paso aislado, mientras una secuencia real recalcula sus probabilidades condicionadas por el texto anterior; reconocer esa diferencia evita confundir un experimento estático sobre seis categorías con una evaluación de comportamiento lingüístico.



Top k y el número de candidatos

El muestreo top-k conserva los k candidatos de mayor probabilidad y elimina los demás antes de normalizar; la idea controla cuántas opciones pueden competir en el paso de selección, mientras temperatura modifica las relaciones probabilísticas entre opciones sin necesidad de excluirlas, por lo que ambos controles actúan sobre aspectos distintos.

En nuestro ejemplo, k = 3 conserva A, B y C, cuya masa original conjunta es 0.80; después de dividir por esa suma obtenemos 0.50, 0.3125 y 0.1875, de manera que A sigue siendo la opción principal, pero no se elige necesariamente si se realiza una extracción aleatoria entre los tres candidatos restantes.

Con k = 1 y un máximo único solo queda A, por lo que el muestreo deja de variar en ese paso; con k igual al número total de candidatos no se elimina ninguno, siempre dentro de esta definición matemática, pues las bibliotecas pueden usar valores especiales para desactivar filtros y esos convenios requieren documentación específica.



Fan et al. (2018) emplearon un esquema top-k para generar historias y describieron en su sección 5.4 la selección entre los diez candidatos más probables de cada paso; ese valor pertenece a sus experimentos históricos y no constituye una pauta general para aplicaciones actuales, lenguas diferentes o tareas que exijan precisión factual.

El principal límite de un k fijo puede observarse comparando dos contextos hipotéticos: en uno, A concentra 0.95 y los restantes candidatos apenas compiten; en otro, diez alternativas reciben 0.10 cada una, por lo que conservar tres candidatos tiene consecuencias muy diferentes aunque el parámetro numérico sea idéntico.

En la primera situación, ampliar el conjunto puede introducir opciones poco plausibles; en la segunda, recortarlo a tres elimina siete alternativas que la distribución consideraba igualmente posibles, de modo que k describe el tamaño permitido y no la fracción de probabilidad preservada ni una medida de suficiencia informativa.



Los empates plantean otra decisión de implementación: cuando varios candidatos ocupan el punto de corte con idéntica puntuación, una regla puede desempatar por identificador o conservar todas las opciones empatadas; por ello conviene separar la definición ideal de k candidatos de los detalles de un código concreto y registrar cómo resuelve los bordes cuando la prueba depende de ellos.

Top p y la masa acumulada

El muestreo top-p, también llamado de núcleo, ordena los candidatos por probabilidad y conserva el conjunto inicial más pequeño cuya masa acumulada alcanza o supera un umbral p; después normaliza dentro de ese conjunto, lo que permite variar el número de opciones conservadas en función de la forma de la distribución (Holtzman et al., 2020).

Con p = 0.80, nuestra distribución base acumula 0.40 en A, 0.65 en A y B, y 0.80 al incorporar C; por tanto conserva tres candidatos y produce las mismas probabilidades que top-k con k = 3 en este caso particular, coincidencia que no



convierte ambos métodos en equivalentes para cualquier contexto.

Si el umbral aumenta a 0.85, hace falta incluir D y la masa preservada pasa a 0.90; el método no recorta una fracción del token D para detenerse exactamente en 0.85, porque los candidatos son unidades discretas, de modo que el umbral representa una condición de cobertura y no siempre el valor exacto de la masa conservada.

Con p = 0.60 se conservan A y B, que suman 0.65; después de normalizar sus probabilidades son aproximadamente 0.615385 y 0.384615, lo que muestra que bajar el umbral no equivale a elegir siempre A y que la masa asignada a un candidato puede aumentar al excluir otros.

Un umbral de 0.80 tampoco significa que exista 80 % de probabilidad de que la respuesta sea verdadera; el número se refiere a masa acumulada bajo una distribución de candidatos y no a una evaluación de hechos, mientras una afirmación sobre exactitud requeriría definir respuestas correctas y observar resultados frente a un criterio externo.



Además, el núcleo puede contener un único candidato cuando su probabilidad ya supera el umbral; esta situación vuelve determinista la selección local aunque la ruta se denomine muestreo, por lo que activar una opción aleatoria no demuestra que haya variación efectiva en todos los pasos.

Hewitt et al. (2022) analizan el recorte como una aproximación para retirar parte del suavizado de la distribución y muestran que top-p puede excluir continuaciones plausibles cuando unos pocos candidatos concentran mucha masa; su explicación es un modelo analítico con supuestos y evidencia experimental delimitados, no una prueba de que cada token descartado sea incorrecto ni una descalificación universal del método.

La lectura práctica es que excluir la cola puede ser útil y, a la vez, perder opciones valiosas; un término especializado poco frecuente, una formulación infrecuente pero correcta o una variante cultural pueden quedar fuera por su posición probabilística, sin que esa exclusión esté apoyada en un juicio directo sobre su pertinencia.



El orden de las operaciones importa

Temperatura y filtros no son interruptores independientes cuando se aplican en secuencia; cada etapa recibe una distribución ya modificada, por lo que dos configuraciones con los mismos números pueden producir distribuciones diferentes si cambian las operaciones previas, el orden de aplicación o la regla empleada para volver a normalizar.

Apliquemos primero T = 0.5 y después top-p con p = 0.80; A acumula aproximadamente 0.614912 y A junto con B llega a 0.855111, de modo que el núcleo contiene dos candidatos y la distribución final es aproximadamente 0.719101 para A y 0.280899 para B.

Invirtamos ahora el orden: top-p con p = 0.80 aplicado a la distribución base conserva A, B y C, después T = 0.5 eleva sus probabilidades normalizadas al cuadrado y produce aproximadamente 0.653061, 0.255102 y 0.091837; en este segundo procedimiento C permanece disponible, lo cual establece una diferencia de soporte y no solo de redondeo.

La composición de top-k y top-p también puede depender del orden: si primero conservamos los tres mejores

candidatos y después aplicamos p = 0.80 sobre su distribución renormalizada, A y B ya acumulan 0.8125 y C queda fuera; si primero aplicamos top-p a la distribución original y después k = 3, los tres candidatos permanecen.

Tabla 2. *Resultados de cambiar el orden en el ejemplo hipotético*

Secuencia definida	Candidatos finales	Probabilidades finales
T = 0.5 y después p = 0.80	A y B	A 0.719101; B 0.280899
p = 0.80 y después T = 0.5	A, B y C	A 0.653061; B 0.255102; C 0.091837
k = 3 y después p = 0.80	A y B	A 0.615385; B 0.384615
p = 0.80 y después k = 3	A, B y C	A 0.500000; B 0.312500; C 0.187500

Nota. Elaboración propia; filtros aplicados a la distribución renormalizada de la etapa anterior y cifras redondeadas; las secuencias no describen una implementación concreta.



Existe una excepción instructiva: temperatura positiva y top-k, con un corte idéntico y sin ambigüedad por empates, conservan el mismo orden de candidatos; al aplicar ambas operaciones y normalizar correctamente, la distribución final dentro de los k seleccionados coincide, porque el factor de normalización intermedio se cancela al elevar todas las probabilidades a la misma potencia.

La excepción permite evitar una exageración frecuente: no todo orden altera necesariamente el resultado; hay que comprobar si una operación depende del rango, de la masa acumulada, de un umbral absoluto o de otra propiedad, y después analizar qué transformaciones preservan esa propiedad en el caso considerado.

En particular, top-p depende de las sumas acumuladas, que la temperatura modifica; por eso en la distribución propuesta el mismo umbral de 0.80 conserva dos candidatos con T = 0.5, tres con T = 1 y cuatro con T = 2, aunque el orden de preferencia permanezca intacto.

La documentación de utilidades de Transformers 4.57.1 describe clases para temperatura, top-k y top-p, y advierte que temperatura necesita una ruta de muestreo para



ejercer el efecto señalado; consultar esa documentación es más preciso que suponer que todo producto que usa los mismos nombres ejecuta una composición idéntica (Hugging Face, s.f.-b).

Elegir un token no es elegir la mejor respuesta

La selección voraz elige el candidato con mayor puntuación en cada paso; es una regla local, por lo que no garantiza que la secuencia completa tenga la mayor probabilidad posible ni que alcance el mejor resultado según una tarea, una diferencia que puede demostrarse sin recurrir a ningún modelo real.

Imaginemos un árbol de dos pasos con A de probabilidad 0.60 y B de probabilidad 0.40; después de A, su mejor continuación tiene probabilidad condicional 0.51, mientras después de B una continuación alcanza 0.99, de modo que el camino voraz A seguido de su mejor opción tiene probabilidad 0.306 y el camino que inicia en B alcanza 0.396.

Este ejemplo hipotético muestra que maximizar cada paso por separado no equivale a maximizar el producto



completo; una búsqueda por haces conserva varias trayectorias parciales para explorar alternativas, pero el criterio de probabilidad de una trayectoria sigue siendo distinto de la corrección factual, la claridad o el cumplimiento de una finalidad profesional.

Holtzman et al. (2020) compararon estrategias en generación abierta de texto y encontraron problemas de repetición y diversidad asociados a ciertas elecciones de decodificación; sus resultados ayudan a cuestionar que la secuencia más probable sea siempre la más deseable, pero no autorizan a trasladar su clasificación experimental a todas las tareas ni a declarar un ganador permanente entre métodos.

La probabilidad de un texto también depende de la longitud y de cómo se representa la terminación; comparar secuencias completas exige describir las reglas de finalización y cualquier normalización, mientras confundir una respuesta breve con una respuesta mejor solo porque termina antes puede esconder una pérdida de información necesaria para la tarea.



Un cambio en la primera palabra puede alterar el registro, el tema elegido o la estructura del resto de la respuesta; esa propagación no implica que cada token posterior sea aleatorio de manera independiente, sino que las probabilidades de los pasos siguientes se calculan sobre contextos que ya son diferentes.

Por esta razón, no basta contar la coincidencia exacta entre dos textos para decidir cuál configuración es más estable en lo que importa; dos redacciones distintas pueden conservar los mismos hechos, mientras dos textos casi idénticos pueden diferir justamente en una fecha, una negación o una condición que cambia su significado práctico.

Concentración y verdad son preguntas diferentes

Una distribución puede concentrar gran parte de su masa en un candidato equivocado; si el mecanismo de temperatura refuerza esa preferencia, obtendremos mayor consistencia de selección sin haber corregido el error, lo que demuestra por qué reducir la variación no constituye una comprobación independiente del contenido.



Para verlo de forma mínima, asignemos a tres etiquetas hipotéticas las probabilidades 0.60, 0.30 y 0.10, y supongamos que una clave externa identifica como correcta únicamente la segunda; con T = 0.5, la probabilidad de elegir esa etiqueta pasa de 0.30 a 0.09 / 0.46, aproximadamente 0.195652, mientras la elección voraz selecciona siempre la etiqueta incorrecta que ocupa el primer lugar.

La clave de corrección no se deduce de las probabilidades, sino que se estipula externamente para construir el contraejemplo; si otra tarea tuviera como correcta la primera etiqueta, la misma transformación aumentaría la probabilidad de acierto, de modo que el resultado depende de cómo se relaciona la distribución del sistema con la respuesta exigida.

La demostración tampoco afirma que bajar la temperatura perjudique siempre la exactitud; establece algo más acotado y suficiente para la práctica: el parámetro, considerado por sí solo, no determina la verdad y debe evaluarse frente a referencias o verificadores que puedan distinguir un acierto de una equivocación.

La diversidad necesita la misma separación conceptual: aumentar el número de formas de expresar una idea puede favorecer un trabajo creativo, pero introducir hechos incompatibles, alternativas irrelevantes o contradicciones no constituye una mejora por el solo hecho de ampliar el repertorio de salidas.

Una aplicación de extracción podría requerir que diferentes ejecuciones devuelvan exactamente el mismo código leído en un documento; una aplicación de ideación podría valorar alternativas conceptualmente distintas, por lo que el criterio de éxito debe fijarse antes de mover controles y no reconstruirse después para justificar el resultado que parezca más atractivo.

También conviene distinguir repetición legítima de repetición defectuosa: conservar nombres técnicos, etiquetas de un esquema o referencias necesarias puede ser indispensable; una penalización indiscriminada contra tokens ya presentes podría dificultar esa fidelidad, razón por la cual los mecanismos contra repetición merecen una prueba específica en vez de tratarse como correctores universales de calidad.



Repetición de ejecuciones y semillas

Una semilla fija el estado inicial de un generador pseudoaleatorio dentro de un procedimiento determinado; para reproducir una extracción se necesita además que coincidan la distribución, el algoritmo de muestreo, el orden de consumo de números y las operaciones que influyen en el resultado, de modo que registrar solo la semilla conserva una parte del experimento y deja otras sin identificar.

La documentación oficial de PyTorch advierte que no se garantizan resultados completamente reproducibles entre versiones, cambios de código o plataformas, ni necesariamente entre CPU y GPU aun con semillas iguales; también distingue el control de fuentes de aleatoriedad de la selección de operaciones deterministas, advertencias consultadas en la página actualizada el 14 de mayo de 2026 (PyTorch, 2026).

En consecuencia, una prueba local debería conservar los identificadores de modelo y tokenizador, la versión de las bibliotecas, la configuración efectiva, los datos de entrada y las características relevantes del entorno; si alguna parte no puede observarse porque pertenece a un servicio

externo, esa ausencia limita la reconstrucción posible y debe registrarse en lugar de reemplazarse por una promesa de identidad absoluta.

Reproducir bytes y reproducir una conclusión son objetivos diferentes: lo primero exige que el archivo resultante sea idéntico, mientras lo segundo puede admitir variaciones de forma siempre que una evaluación repetida sostenga el mismo resultado; decidir cuál de ambos se necesita evita exigir identidad textual cuando basta estabilidad semántica o aceptar equivalencia aparente cuando una diferencia exacta tiene consecuencias.

Para depurar un problema es útil empezar con una entrada conservada sin cambios y desactivar fuentes de variación que la implementación permita controlar; si aparece una diferencia, conviene localizar la primera etapa en que los registros divergen, pues atribuir toda discrepancia a temperatura puede ocultar cambios en recuperación documental, instrucciones, transformaciones previas o recursos externos.

La repetición no debe transformarse en una falsa encuesta sobre la verdad: si una configuración produce la misma



afirmación veinte veces, eso prueba únicamente una regularidad dentro del procedimiento observado; una salida estable puede compartir el mismo error de origen y carecer de evidencia externa, incluso cuando su redacción parezca especialmente segura.

En el experimento estático de seis candidatos, la probabilidad de que dos extracciones independientes coincidan es suma_i p_i^2, igual a 0.2602 para T = 1 y aproximadamente 0.444996 para T = 0.5; esta comparación cuantifica coincidencia local bajo supuestos explícitos y no estima la probabilidad de que dos respuestas completas sean iguales.

Asimismo, si F mantiene probabilidad 0.04 en veinte extracciones independientes de la misma distribución, la probabilidad de observarlo al menos una vez es 1 − 0.96^20, aproximadamente 0.557998; extrapolar ese cálculo a veinte respuestas de un servicio requeriría justificar tanto la independencia como la constancia del evento y de su probabilidad.



Comparar configuraciones

Una evaluación útil empieza por definir el producto esperado: extraer un dato, proponer alternativas, elaborar un resumen fiel o producir una salida estructurada son tareas distintas; antes de elegir métricas conviene especificar qué cuenta como éxito, qué errores invalidan una respuesta y cuáles pueden corregirse con un esfuerzo razonable.

El conjunto de prueba debe contener entradas representativas y casos difíciles, sin limitarse a ejemplos que la configuración inicial ya resuelva bien; resulta útil separar una porción destinada al ajuste y otra reservada para comprobar la elección final, porque evaluar repetidamente sobre los mismos casos puede favorecer decisiones adaptadas a peculiaridades del conjunto conocido.

La primera comparación puede variar un control y conservar los demás, con el fin de interpretar su efecto; una vez identificadas regiones útiles, es necesario explorar interacciones relevantes, especialmente entre temperatura y top-p, ya que los cálculos anteriores muestran que el



tamaño del conjunto permitido puede cambiar con ambos parámetros a la vez.

Cada ejecución debería conservar la entrada exacta, la configuración efectiva y el resultado íntegro, junto con una identificación de versión; también conviene registrar si hubo truncamiento por longitud, una interrupción o un error de herramienta, porque esos eventos pueden explicar una respuesta incompleta sin que la causa sea la estrategia de muestreo.

La cantidad de repeticiones depende de la variabilidad observada, del costo de cada error y de la precisión que se necesite para comparar; no existe un número único que vuelva concluyente cualquier evaluación, pero una sola salida por configuración ofrece evidencia débil cuando el procedimiento admite resultados distintos ante la misma entrada.

Para una tarea con clave verificable puede medirse proporción de aciertos y tipos de error; para textos abiertos conviene combinar revisión del contenido con criterios sobre diversidad útil, coherencia y cumplimiento, sin permitir que una calificación estilística compense

automáticamente hechos incorrectos o requisitos incumplidos.

Tabla 3. *Criterios de comparación según la finalidad de la salida*

Finalidad	Evidencia principal	Riesgo de una lectura insuficiente
Extracción de campos	Coincidencia con la fuente y esquema válido	Confundir formato correcto con dato correcto
Resumen de un documento	Afirmaciones sustentadas y cobertura de puntos esenciales	Premiar fluidez pese a omisiones
Ideación	Alternativas pertinentes y diferencias conceptuales	Contar sinónimos como nuevas ideas
Explicación técnica	Exactitud de relaciones y ejemplos comprobables	Confundir extensión con profundidad
Producción repetida	Distribución de calidad y esfuerzo de revisión	Mostrar solo la mejor salida

Nota. Elaboración propia; la tabla propone criterios de evaluación y no presenta resultados observados ni un orden de preferencia entre configuraciones.

Cuando intervienen evaluadores humanos, conviene ocultarles los nombres de las configuraciones y presentar los textos en un orden que no favorezca sistemáticamente una de ellas; también es necesario definir los criterios antes de juzgar, registrar desacuerdos y separar la preferencia estética del cumplimiento de requisitos comprobables.

La comparación debería informar qué ocurre con la salida típica y con las salidas problemáticas; una configuración puede producir ocasionalmente un texto excelente y exigir mucha revisión en los demás casos, mientras otra ofrece resultados menos llamativos pero más previsibles, por lo que publicar solo el mejor ejemplo altera la decisión práctica.

El análisis final debe incluir el trabajo humano necesario para transformar la salida en un producto aceptable; si obtener diez variantes obliga a revisar diez errores distintos, el aumento de diversidad puede carecer de valor operativo, mientras una variación controlada entre propuestas pertinentes puede ahorrar tiempo en la fase de exploración.



Tres situaciones de uso

En un primer caso hipotético, una oficina necesita extraer de documentos el número de expediente, la fecha y el nombre de un trámite; la decisión razonable es evaluar configuraciones frente a campos comprobados, conservar un mecanismo de salida desconocida cuando el documento no contiene el dato y verificar el formato mediante un validador independiente.

Aquí la temperatura podría influir en la variación observada, pero no sustituye la lectura correcta del documento ni resuelve por sí misma un texto ilegible; si todas las configuraciones confunden un carácter, el siguiente paso útil consiste en investigar la entrada y su procesamiento, en lugar de seguir ajustando un parámetro que actúa después de esa pérdida de información.

En un segundo caso hipotético, una persona solicita conceptos para una campaña sobre ahorro de agua sin incorporar cifras no verificadas; la evaluación puede valorar diversidad de enfoques, adecuación al público y fidelidad a las restricciones, mientras descarta variantes que solo cambian adjetivos o inventan estadísticas para resultar persuasivas.



La comparación podría conservar una configuración como base y probar alternativas con mayor dispersión, pero no debe suponer que la temperatura por sí sola produce novedad conceptual; solicitar enfoques distintos, proporcionar restricciones útiles y revisar las propuestas son intervenciones separadas cuyos efectos conviene identificar para no atribuir todo cambio al muestreo.

En un tercer caso hipotético, un equipo prepara un resumen de un informe técnico y necesita conservar las salvedades del original; dos respuestas pueden parecer equivalentes hasta que una elimina la condición bajo la cual se obtuvo un resultado, por lo que la estabilidad importante reside en las proposiciones y sus límites, no en repetir exactamente las mismas palabras.

Si una configuración devuelve resúmenes variados pero conserva sistemáticamente datos y condiciones, la variación puede ser aceptable; si otra produce siempre el mismo resumen y omite la salvedad decisiva, su estabilidad textual no corrige el defecto, lo que vuelve a situar la verificación del contenido por encima de la apariencia de uniformidad.



Los tres casos muestran una secuencia de trabajo común: precisar la finalidad, conservar entradas comparables, definir una referencia de evaluación y observar varias salidas cuando exista variación; la elección final puede ser distinta en cada caso sin que ello constituya una contradicción, porque las funciones de utilidad y los errores tolerables también son diferentes.

Más allá de los controles más conocidos

La investigación sobre decodificación no termina en temperatura, top-k y top-p; Meister et al. (2023) proponen el muestreo localmente típico, que atiende a la cercanía entre la información de un candidato y la entropía condicional de la distribución, de modo que su selección no se reduce a tomar siempre los candidatos más probables.

Hewitt et al. (2022), por su parte, desarrollan un umbral de recorte dependiente de la entropía y lo evalúan en tareas concretas; estas propuestas muestran que los controles habituales son decisiones de diseño entre varias posibles, mientras sus resultados deben interpretarse con los



modelos, lenguas y conjuntos de prueba utilizados en las investigaciones originales.

La consecuencia práctica no es acumular todos los controles disponibles, sino entender qué problema intenta resolver cada uno y cómo se relaciona con los demás; añadir filtros puede reducir el conjunto de alternativas hasta volver irrelevante un parámetro que parecía decisivo, además de dificultar la identificación de la causa cuando el resultado empeora.

Las restricciones de formato y las validaciones externas cumplen una función diferente: una salida puede cumplir una gramática y contener datos incorrectos, pero exigir un formato válido facilita que otro componente revise sus campos; por ello conviene distinguir el control sobre qué secuencias están permitidas del control sobre si las afirmaciones que esas secuencias expresan están justificadas.

Tampoco debe confundirse temperatura de generación con cualquier procedimiento que utilice el mismo término en calibración; aquí se modifica la distribución empleada para seleccionar continuaciones, mientras evaluar



probabilidades como indicadores de acierto plantea otra pregunta y exige datos etiquetados, una definición de corrección y un procedimiento específico de comprobación.

Mitos que conviene abandonar

La afirmación de que una temperatura baja produce necesariamente hechos verdaderos confunde concentración con corrección; el contraejemplo de las tres etiquetas demuestra que reforzar la opción preferida puede reducir el acierto cuando la respuesta correcta no ocupa el primer lugar, sin necesidad de atribuir al sistema intención alguna de engañar.

La idea de que una temperatura alta siempre genera mejores ideas confunde dispersión con utilidad; una salida distinta puede ser redundante, incoherente o incompatible con las restricciones, mientras una propuesta valiosa necesita pertinencia y evaluación, propiedades que no se obtienen mediante una transformación probabilística aislada.



La frase según la cual top-p conserva cierto porcentaje de las palabras confunde masa con cantidad; el número de candidatos depende de la distribución y puede cambiar entre pasos, por lo que el mismo umbral puede permitir una única alternativa o un conjunto amplio sin que el parámetro haya variado.

La creencia de que top-k y top-p son equivalentes porque coinciden en un ejemplo ignora el criterio de recorte; uno fija un número y el otro exige una cobertura de masa, mientras la tabla 2 muestra que incluso combinarlos requiere describir el orden y la normalización para conocer qué candidatos sobreviven.

La promesa de que una semilla garantiza resultados idénticos en cualquier entorno excede lo que documenta PyTorch (2026); una reconstrucción responsable identifica los elementos controlados, comprueba el resultado y declara las diferencias que no puede observar, sin convertir una conveniencia experimental en una garantía universal.

Finalmente, interpretar una respuesta diferente como prueba de que el modelo aprendió algo nuevo durante la



conversación confunde generación con actualización de parámetros; el muestreo puede cambiar una continuación sin reentrenamiento, mientras demostrar aprendizaje persistente exigiría evidencia sobre un mecanismo distinto al analizado en este artículo.



Conclusión

Temperatura, top-k y top-p modifican cómo se selecciona una continuación a partir de una distribución; conocer sus fórmulas permite anticipar concentración, exclusión de candidatos y dependencias entre operaciones, pero no permite deducir por sí solo si una respuesta será verdadera, útil o suficientemente estable para una finalidad concreta.

El experimento hipotético permitió comprobar tres resultados: la temperatura modifica relaciones probabilísticas sin cambiar su orden para valores positivos, el recorte por masa puede cambiar de tamaño al variar la concentración y determinadas composiciones dependen del orden; esas propiedades ofrecen un vocabulario preciso para investigar diferencias que de otro modo se atribuirían vagamente a creatividad o imprevisibilidad.

La elección exige pruebas comparables y criterios de aceptación; ninguna configuración de muestreo sustituye la verificación independiente de los hechos.

Referencias

Fan, A., Lewis, M., & Dauphin, Y. (2018). Hierarchical neural story generation [Generación neuronal jerárquica de historias]. En I. Gurevych y Y. Miyao (Eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 889–898).

Hewitt, J., Manning, C. D., y Liang, P. (2022). Truncation sampling as language model desmoothing [Muestreo por truncamiento como eliminación del suavizado de modelos de lenguaje]. En Y. Goldberg, Z. Kozareva, y Y. Zhang (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022* (pp. 3414–3427). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.findings-emnlp.249

Holtzman, A., Buys, J., Du, L., Forbes, M., y Choi, Y. (2020). *The curious case of neural text degeneration* [El curioso caso de la degeneración del texto neuronal] (versión 2) [Prepublicación]. arXiv. https://doi.org/10.48550/arXiv.1904.09751

Hugging Face. (s.f.-a). *Generation* [Generación]. Transformers, versión 4.57.1. https://huggingface.co/docs/transformers/v4.57.1/en/main_classes/text_generation

Hugging Face. (s.f.-b). *Utilities for generation* [Utilidades para la generación]. Transformers, versión 4.57.1. https://huggingface.co/docs/transformers/v4.57.1/en/internal/generation_utils

Meister, C., Pimentel, T., Wiher, G., y Cotterell, R. (2023). Locally typical sampling [Muestreo localmente típico]. *Transactions of the Association for Computational Linguistics, 11*, 102–121. https://doi.org/10.1162/tacl_a_00536

PyTorch. (2026, 14 de mayo). *Reproducibility* [Reproducibilidad]. PyTorch main documentation. Recuperado el 3 de octubre de 2026, de https://docs.pytorch.org/docs/main/notes/randomness.html

