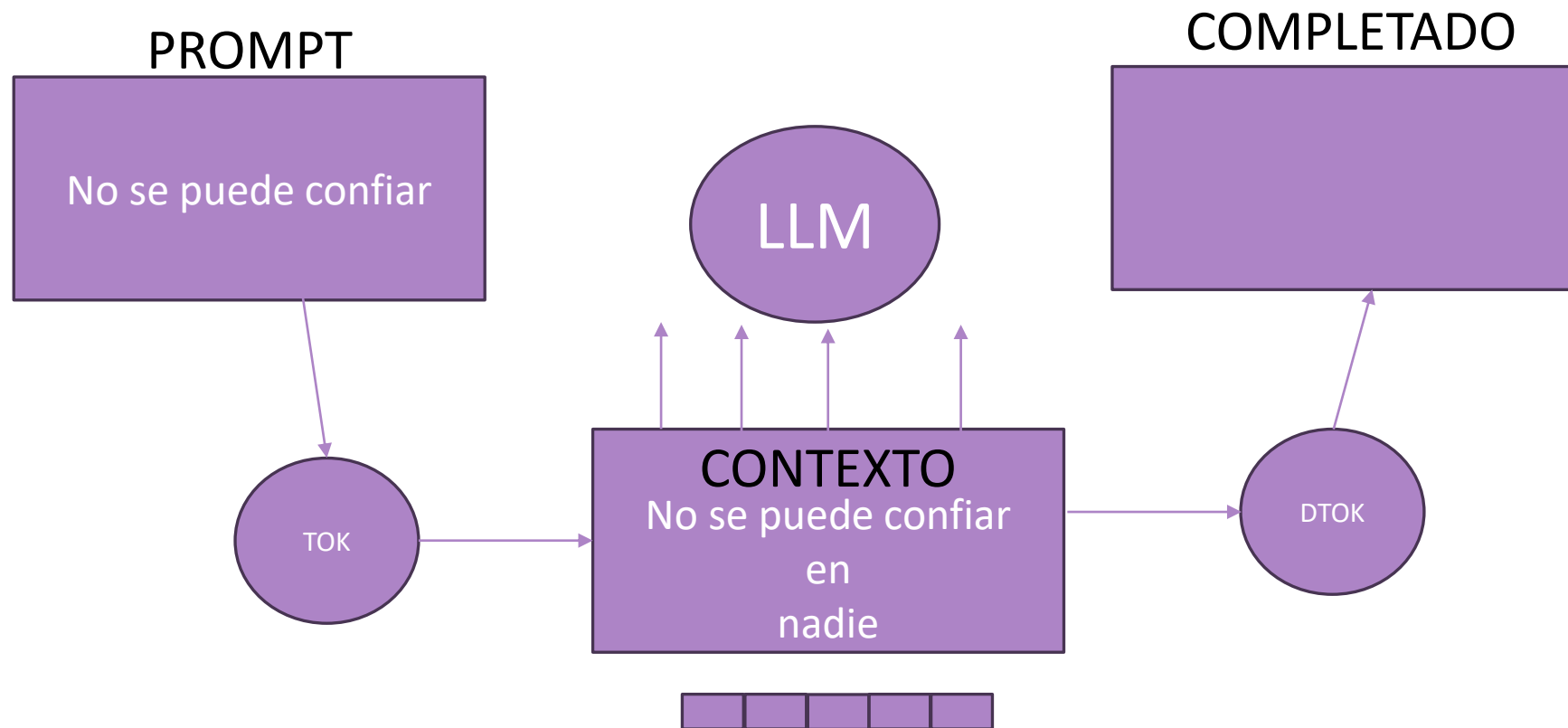
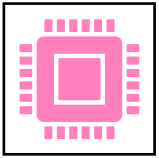




1. PROMPTING Y PROMT ENGINEERING



¿Qué es un prompt?



En la IA, el prompting se refiere a proporcionar una entrada inicial o un ejemplo para un modelo de lenguaje o de aprendizaje automático, con el fin de guiar su predicción o generación de respuesta.



Es la forma en que los usuarios interactúan con modelos de lenguaje como por ejemplo ChatGPT para obtener respuestas o generar contenido.

¿Cuándo usamos prompting?



Cuando necesitas respuestas rápidas y generales: Si solo necesitas información básica o una respuesta general sobre un tema, un prompt simple puede ser suficiente.



Cuando buscas inspiración creativa: Si deseas que el modelo genere contenido creativo, como historias o poesía, puedes dar un prompt amplio y permitir que el modelo sea creativo en su respuesta.



Cuando prefieres simplicidad: Para interacciones sencillas y rápidas, o cuando no tienes la necesidad de ajustar los detalles de la respuesta, un prompt básico puede ser adecuado.



escribeme un cuento



Por supuesto, aquí tienes un cuento:



Título: El Viaje del Pequeño Explorador

Había una vez un pequeño explorador llamado Leo. Desde que era un niño, Leo soñaba con descubrir lugares lejanos y desconocidos. Su habitación estaba llena de mapas, brújulas y objetos extraños que había recolectado en sus expediciones imaginarias por la casa de su abuela.

Un día, mientras exploraba el ático de su abuela en busca de tesoros olvidados, Leo encontró un viejo diario de viaje que pertenecía a su abuelo. Las páginas amarillentas estaban llenas de relatos sobre aventuras en tierras lejanas y misterios por descubrir. Leo estaba emocionado.

Regenerate



El cielo es



El cielo es azul durante el día cuando el sol está brillando y la atmósfera de la Tierra dispersa la luz solar de manera que vemos el color azul. Durante la noche, el cielo puede verse oscuro o lleno de estrellas, dependiendo de la ubicación y las condiciones atmosféricas. El color y aspecto del cielo también pueden variar durante el amanecer y el atardecer, creando hermosos tonos de naranja, rosa y rojo a medida que el sol se eleva o se pone en el horizonte.



Regenerate

R

Escribe un cuento corto, con solo 2 personajes. Debe tener algo de humor pero a la vez debe contener alguna moraleja. Debe estar orientado a niños de 4 a 8 años. Usa lenguaje sencillo



Título: El Conejo y el Oso en la Gran Carrera

Había una vez un conejo llamado Rafa y un oso llamado Bruno, que vivían en el bosque. Rafa era muy rápido, siempre saltando y corriendo por todas partes. Bruno, por otro lado, era grande y lento, pero tenía un corazón amable y siempre estaba sonriendo.

Un soleado día de verano, Rafa se burló de Bruno por su lentitud. "¡Mira qué rápido puedo correr!", se jactó Rafa. "Nunca podrías ganarme en una carrera, Bruno."

Bruno, sin enfadarse, respondió: "Bueno, tal vez soy lento, pero creo que todos te

☐ Stop generating

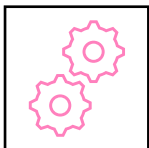
¿Qué es el prompt Engineering?



La ingeniería de prompt es una disciplina relativamente nueva para el desarrollo y la optimización de prompts para utilizar eficientemente modelos de lenguaje (ML) en una amplia variedad de aplicaciones y temas de investigación. Las habilidades de ingeniería de prompt ayudan a comprender mejor las capacidades y limitaciones de los grandes modelos de lenguaje (LLM).



Es el proceso de diseñar y optimizar estos "prompts" para obtener resultados más precisos y relevantes de un modelo de lenguaje. Implica experimentar con diferentes formulaciones de las instrucciones, ajustar parámetros y evaluar los resultados para lograr los objetivos deseados.



La ingeniería de prompt no solo se trata de diseñar y desarrollar prompts. Abarca una amplia gama de habilidades y técnicas útiles para interactuar y desarrollar con LLM. Es una habilidad importante para interactuar, construir y comprender las capacidades de los LLM.



- En general, el prompting es una técnica muy útil para mejorar la calidad y precisión de los resultados obtenidos por los modelos de AI.
- Al proporcionar información adicional al modelo, podemos mejorar su comprensión del contexto y producir resultados más precisos y relevantes.

¿Cuándo aplicamos prompt engineering?



Cuando buscas respuestas específicas y detalladas: Si necesitas respuestas detalladas o específicas a preguntas complejas, es útil diseñar prompts detallados que incluyan contextos y detalles específicos.



Para optimizar la calidad de las respuestas: Cuando deseas mejorar la calidad y relevancia de las respuestas, el prompt engineering te permite ajustar y refinar tus prompts para obtener resultados más precisos.



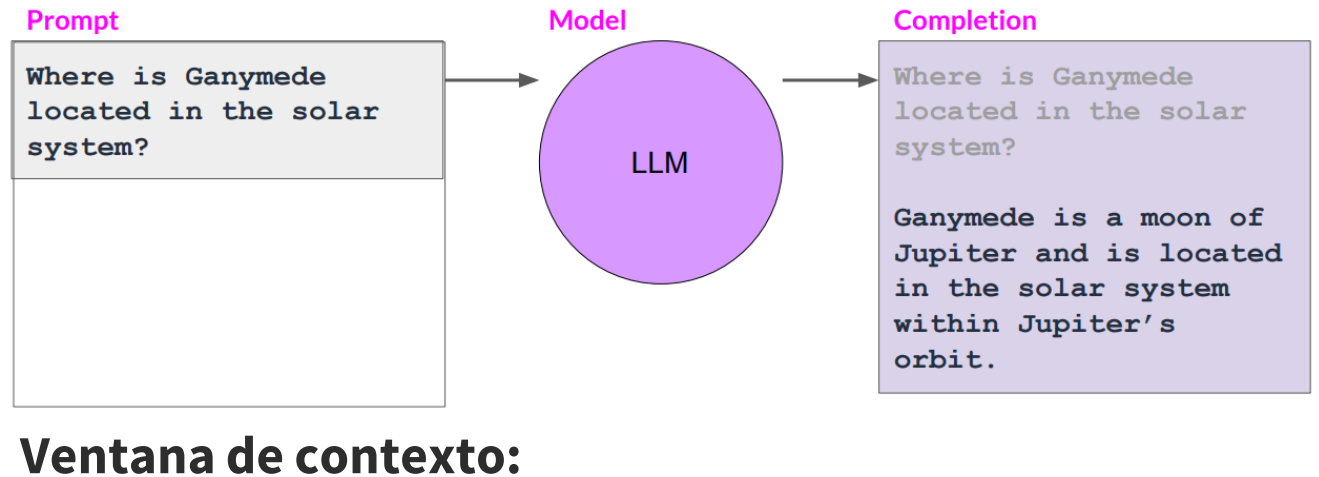
Cuando necesitas controlar el tono y estilo: Si deseas que el modelo responda en un tono o estilo particular, puedes lograrlo a través de instrucciones detalladas en el prompt.



Para tareas complejas o específicas: En situaciones en las que necesitas que el modelo realice tareas específicas, como resumir textos largos, traducir idiomas o generar código, el prompt engineering es esencial para guiar al modelo de manera efectiva.

Prompting & Prompt Engineering

- El texto de entrada que de damos al modelos se llama instruccion (**prompt**)
- El acto de generar texto se conoce como **inferencia**
- El texto de salida se conoce como **finalizacion (completion)**.
- La cantidad total de texto o la memoria que está disponible para usar para la inferencia se denomina **ventana de contexto**.





El poder de saber cómo usar el lenguaje

1. La instrucción

1. Instrucción directa, palabras clave, contextos y datos
2. cuanto más específico y relevante sea el contexto para la tarea que estás tratando de realizar, mejor será.

Prompt:

```
### Instrucción ###  
Traduce el texto a continuación al español:  
Texto: "Hello!"
```

Resultado:

¡Hola!

El poder de saber cómo usar el lenguaje

2. La especificidad

1. Cuanto más descriptiva y detallada sea el prompt, mejores serán los resultados.
2. Buen formato y un prompt descriptiva
3. Ejemplos en el prompt
4. Tener en cuenta la longitud, los detalles deben ser relevantes y contribuir a la tarea en cuestión.

Prompt:

```
Extrae los nombres de lugares del siguiente texto.  
Formato deseado:  
Lugar: <lista_separada_por_comas_de_nombres_de_empresa>  
Input: "Aunque estos avances son alentadores para los investigadores, aún hay mucho misterio.  
'A menudo tenemos una caja negra entre el cerebro y el efecto que vemos en la periferia', dice  
Henrique Veiga-Fernandes, un neuroinmunólogo en el Centro Champalimaud para lo Desconocido en Lisboa.  
'Si queremos usarlo en el contexto terapéutico, necesitamos entender el mecanismo'.
```

Resultado:

Place: Centro Champalimaud para lo Desconocido, Lisboa



El poder de saber cómo usar el lenguaje

3. Evitar imprecisión

1. es fácil caer en la trampa de querer ser demasiado inteligente sobre los temas y crear descripciones imprecisas

Explica el concepto de ingeniería de promoción. Mantén la explicación corta, solo unas pocas frases y no seas demasiado descriptivo.

Usa 2-3 oraciones para explicar el concepto de ingeniería de prompt a un estudiante de secundaria.



El poder de saber cómo usar el lenguaje

4. Evitamos las negaciones

1. evitar decir qué no hacer, sino decir qué hacer en su lugar. Esto fomenta la especificidad y se centra en los detalles que conducen a buenas respuestas del modelo.

El siguiente es un agente que recomienda películas a un cliente.
NO PIDA INTERESES. NO PIDA INFORMACIÓN PERSONAL.
Cliente: Por favor, recomiéndeme una película basada en mis intereses.
Agente:

El siguiente es un agente que recomienda películas a un cliente. El agente es responsable de recomendar una película de las más populares a nivel mundial. Debe abstenerse de preguntar a los usuarios por sus preferencias y evitar pedir información personal. Si el agente no tiene una película para recomendar, debe responder "Lo siento, no pude encontrar una película para recomendar hoy".
Cliente: Por favor, recomiéndeme una película basada en mis intereses.
Agente:



Ventajas:

- Mayor precisión:

Un prompt relevante, significa un mejor trabajo por parte del modelo de IA que a su vez sólo significa una respuesta refinada simulada para la situación con precisión.

- Rentabilidad:

El número de interacciones necesarias para obtener una respuesta precisa y satisfactoria se reduce con una pregunta específica y bien diseñada.

Desventajas:

- Dificultad para determinar la especificidad:

Determinar el equilibrio adecuado entre especificidad y generalidad puede ser un reto, ya que una pregunta demasiado específica puede limitar la gama de respuestas generadas, mientras que una pregunta demasiado general puede producir respuestas irrelevantes.

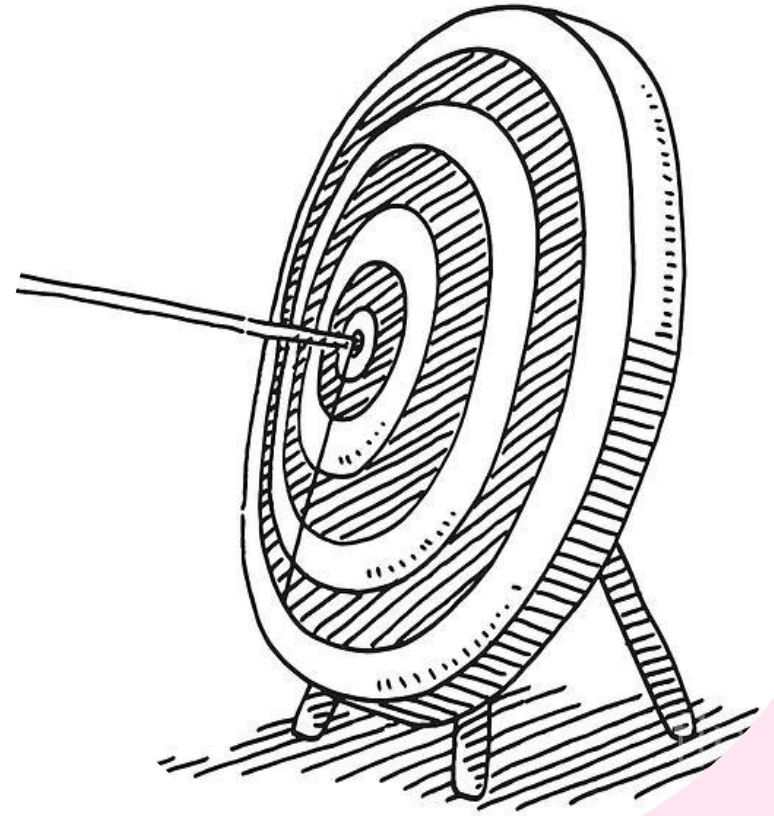


No hay una fórmula mágica,
sí hay recomendaciones

Prueba-ensayo para cada
caso

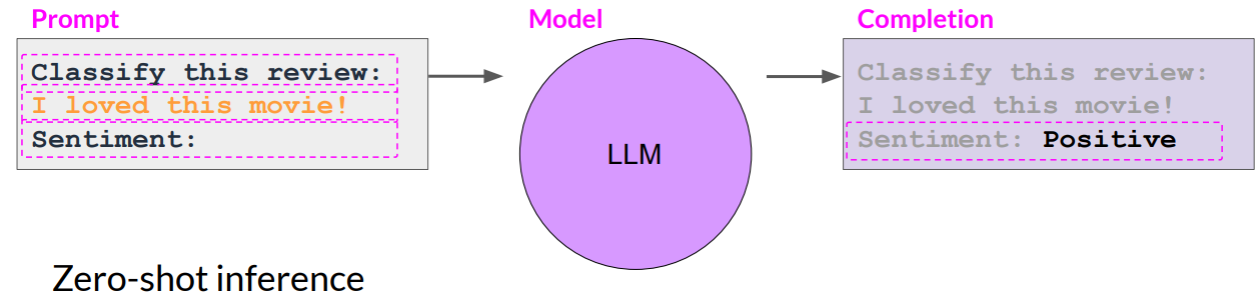
TÉCNICAS

1. Zero-shot inference
2. One-shot inference
3. Few-shot inference

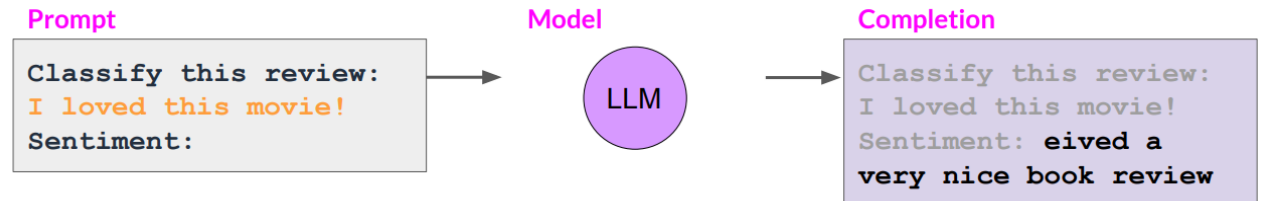


● In-Context Learning (ICL) – Zero-Shot Inference

- Una estrategia para conseguir que el modelo produzca mejores resultados es incluir ejemplos de la tarea que quieres que el modelo lleve a cabo dentro del prompt.
- Proporcionar ejemplos dentro de la ventana de contexto se denomina aprendizaje en contexto (**in-context learning**).

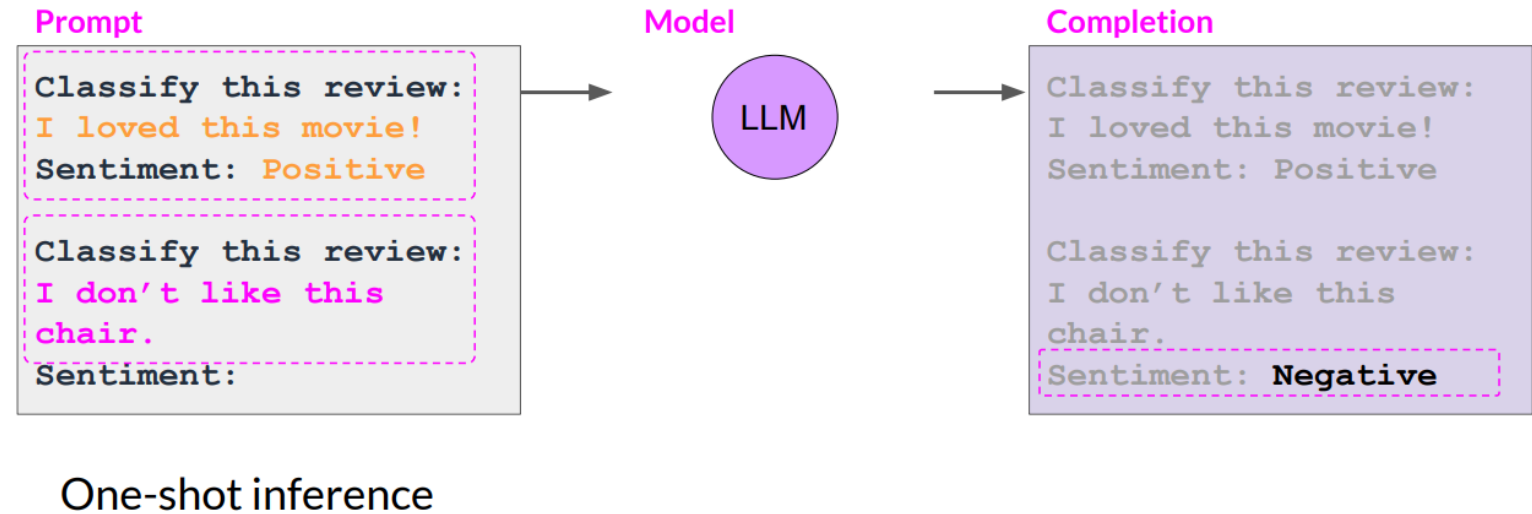


En cambio, los modelos pequeños pueden tener dificultades (GPT-2):



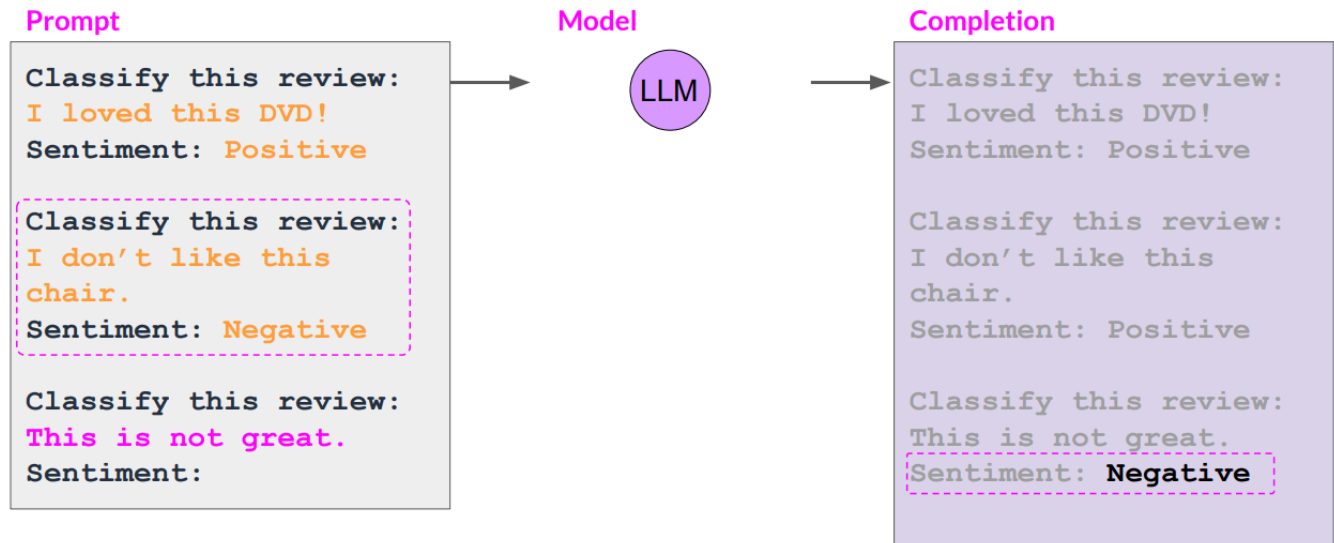
● In-Context Learning (ICL) – One-Shot Inference

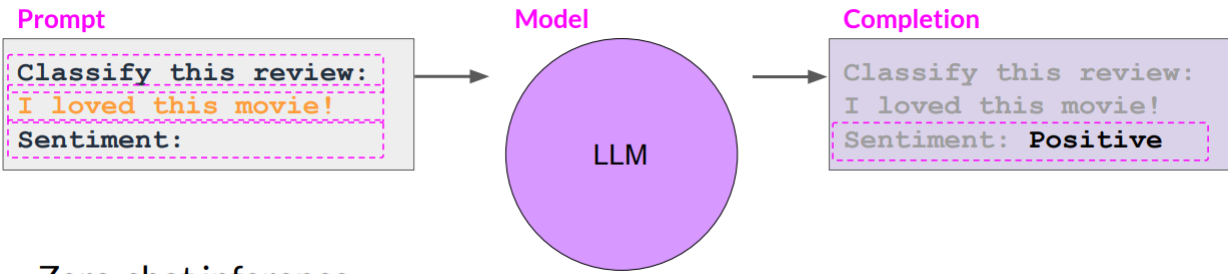
- A diferencia del zero-shot prompt
- La inclusión de un solo ejemplo se conoce como **one-shot inference**



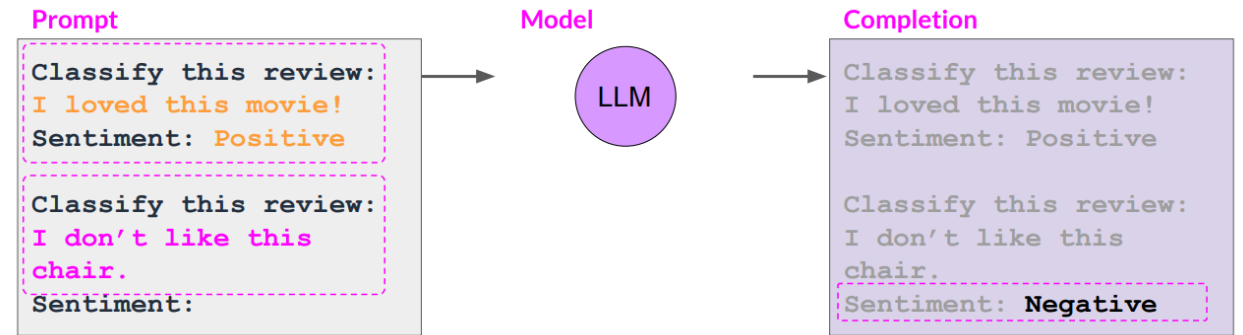
● In-Context Learning (ICL) – Few-Shot Inference

- A veces, un solo ejemplo no es suficiente para que el modelo aprenda lo que quieres que haga.
- Por lo tanto, se puede ampliar la idea de dar un solo ejemplo para incluir múltiples ejemplos.
- Es lo que se conoce como **few-shot inference**.

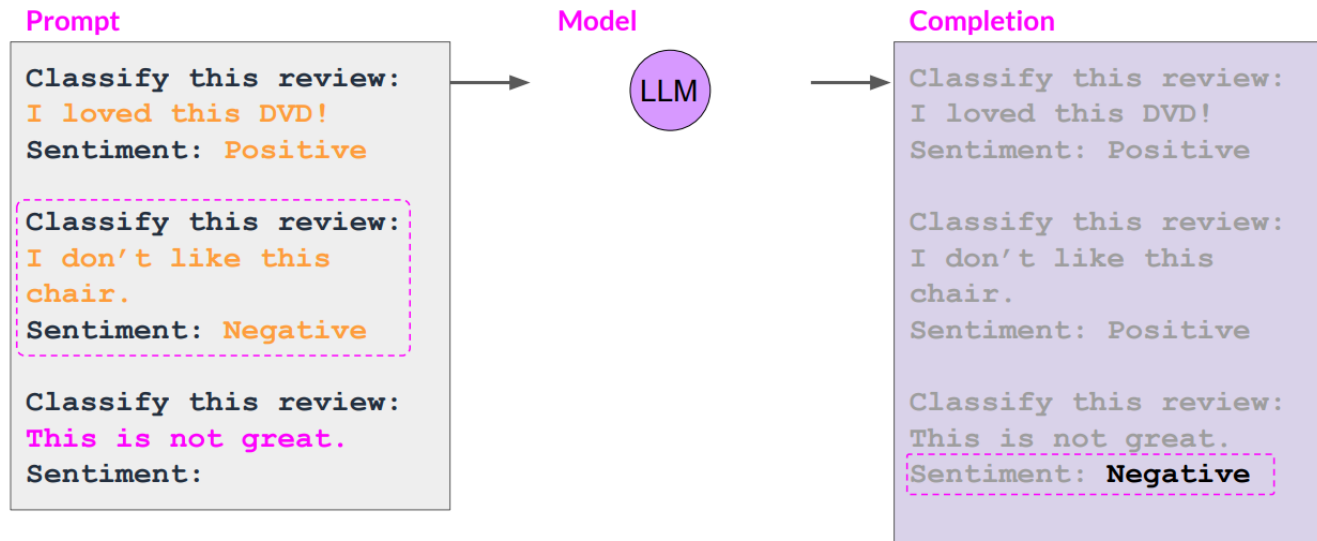




Zero-shot inference



One-shot inference



Limitations of in-context learning

Classify this review:

I loved this movie!

Sentiment: Positive

Classify this review:

I don't like this chair.

Sentiment: Negative

Classify this review:

This sofa is so ugly.

Sentiment: Negative

Classify this review:

Who would use this product?

Sentiment:

Context Window

Even with
multiple
examples

- El aprendizaje en contexto puede no funcionar en modelos más pequeños (LLM)
- Especificar ejemplos ocupa espacio en la ventana de contexto
- En su lugar, prueba a realizar **fine-tuning**



Zero-Shot, One-shot, Few-Shots

Escribe un ejemplo de zero-shot, one shot y few shots para una tarea de:

- Clasificación (ej: clasificación de un texto según el lenguaje)
- NER



Clasificación: Zero-Shot, One-shot, Few-Shots

0

Detect the language of the following text:
Hello, how are you?

1

Detect the language of the following text: Hello how are you?
Example:
Salut comment Ça va?
Language: French

FEW

Detect the language of the following text: Hello, How are you?
Example:
Salut comment Ça va?
Language: French

Detect the language of the following text: How are you?
Example:
Hola como estas?
Language: Spanish

Detect the language of the following text: How are you?
Example:
Guzal
Language: Turkish



Traducción: Zero-Shot, One-shot, Few-Shots

0

Extract the entities: My name is Rocio, i live in Vitoria, i work for Hiberus.

Entities:

- PER:
- LOC:
- ORG:

1

Extract the entities: My name is Rocio, i live in Vitoria, i work for Hiberus.

Entities:

- PER: Rocio
- LOC: Vitoria
- ORG: Hiberus

FEW

Extract the entities: My name is Rocio, i live in Vitoria, i work for Hiberus.

Entities:

- PER: Rocio
- LOC: Vitoria
- ORG: Hiberus

Extract the entities: My name is Bert , i live in HuggingFace, i work for Transformer.

Entities:

- PER: Bert
- LOC: HF
- ORG: Transformer

Extract the entities: My name is Pablo, i live in London, i work for DataCompany.

Entities:

- PER: Pablo
- LOC: London
- ORG: DataCompany



System message

Es la forma de proporcionar al modelo cierto **contexto**, **instrucciones** u otra información relevante para la tarea o escenario que se pretende lograr. Puede ayudar al modelo a **entender el propósito**, el **alcance** y las **limitaciones** de sus respuestas, así como el **tono** y el **estilo** que se espera del mismo.

- Definir el perfil, las capacidades y las limitaciones del modelo para su escenario
- Definir el formato de salida del modelo
- Proporcionar comportamientos adicionales

The image features a central black rectangle with a thin white border. To the left of this rectangle is a large light green circle containing two white zigzag lines. Above the rectangle's top right corner is a pink circle with a horizontal line through its center. Below the rectangle's bottom left corner is a small light green circle. In the bottom right corner, there is a large pink shape and four parallel white diagonal lines. The text "GENERATIVE CONFIGURATION" is centered within the black rectangle in a white, bold, sans-serif font.

GENERATIVE CONFIGURATION

Generative Configuration

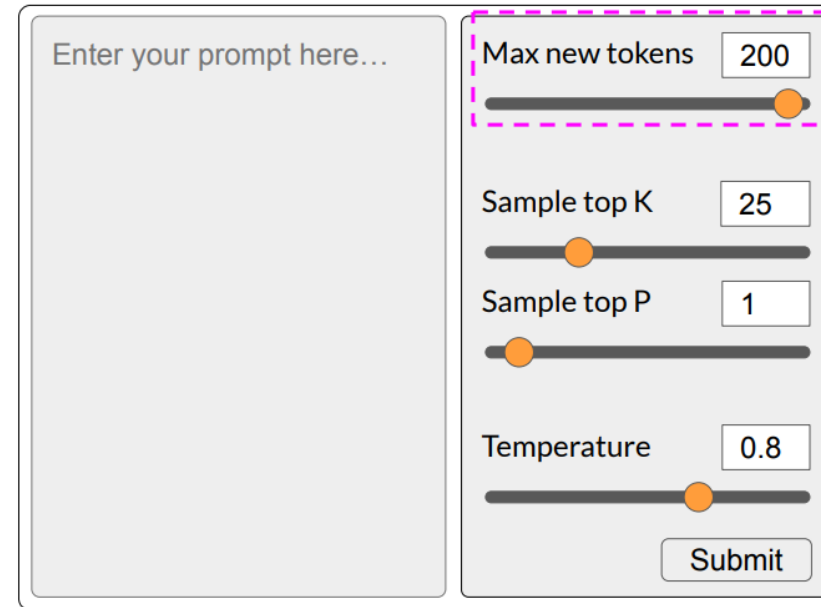
La **configuración generativa** son parámetros de configuración para la inferencia que podemos utilizar **para influir en el modo** en el que el modelo toma la decisión final sobre la generación o predicción de la palabra siguiente.

The image shows a user interface for configuring a generative model. On the left is a large text input area with the placeholder "Enter your prompt here...". To the right is a settings panel. The settings panel contains four parameters, each with a slider and a numeric input field: "Max new tokens" (value 200), "Sample top K" (value 25), "Sample top P" (value 1), and "Temperature" (value 0.8). These four parameters are grouped together by a dashed pink rectangular border. Below this group is a "Submit" button. A bracket on the right side of the settings panel points to the dashed pink box and is labeled "Inference configuration parameters".

Parameter	Value
Max new tokens	200
Sample top K	25
Sample top P	1
Temperature	0.8

Generative Configuration – max new tokens

- **Max new tokens** es probablemente el más sencillo de estos parámetros.
- Lo utilizamos para limitar el número de tokens que generará el modelo.



The image shows a configuration interface for a generative model. On the left is a large text input area with the placeholder "Enter your prompt here...". On the right is a sidebar containing four sliders, each with a corresponding numerical input box and a "Submit" button at the bottom. The sliders are for "Max new tokens" (set to 200), "Sample top K" (set to 25), "Sample top P" (set to 1), and "Temperature" (set to 0.8). The "Max new tokens" slider and its input box are highlighted with a dashed pink rectangular border.

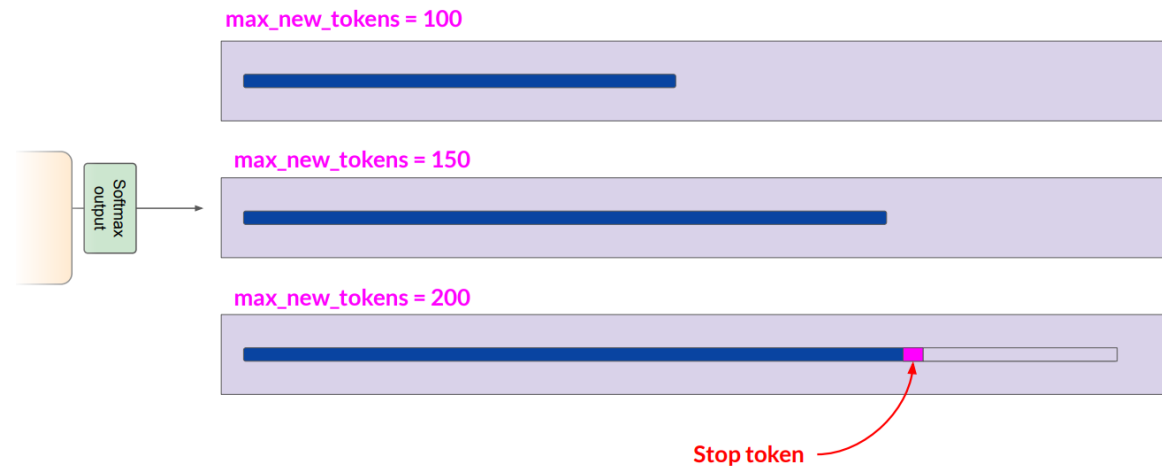
Parameter	Value
Max new tokens	200
Sample top K	25
Sample top P	1
Temperature	0.8

Max new tokens

<https://towardsdatascience.com/why-openai-api-is-more-expensive-for-non-english-languages-553da4a1eccc>

<https://huggingface.co/spaces/JacobLinCool/tiktoken-calculator>

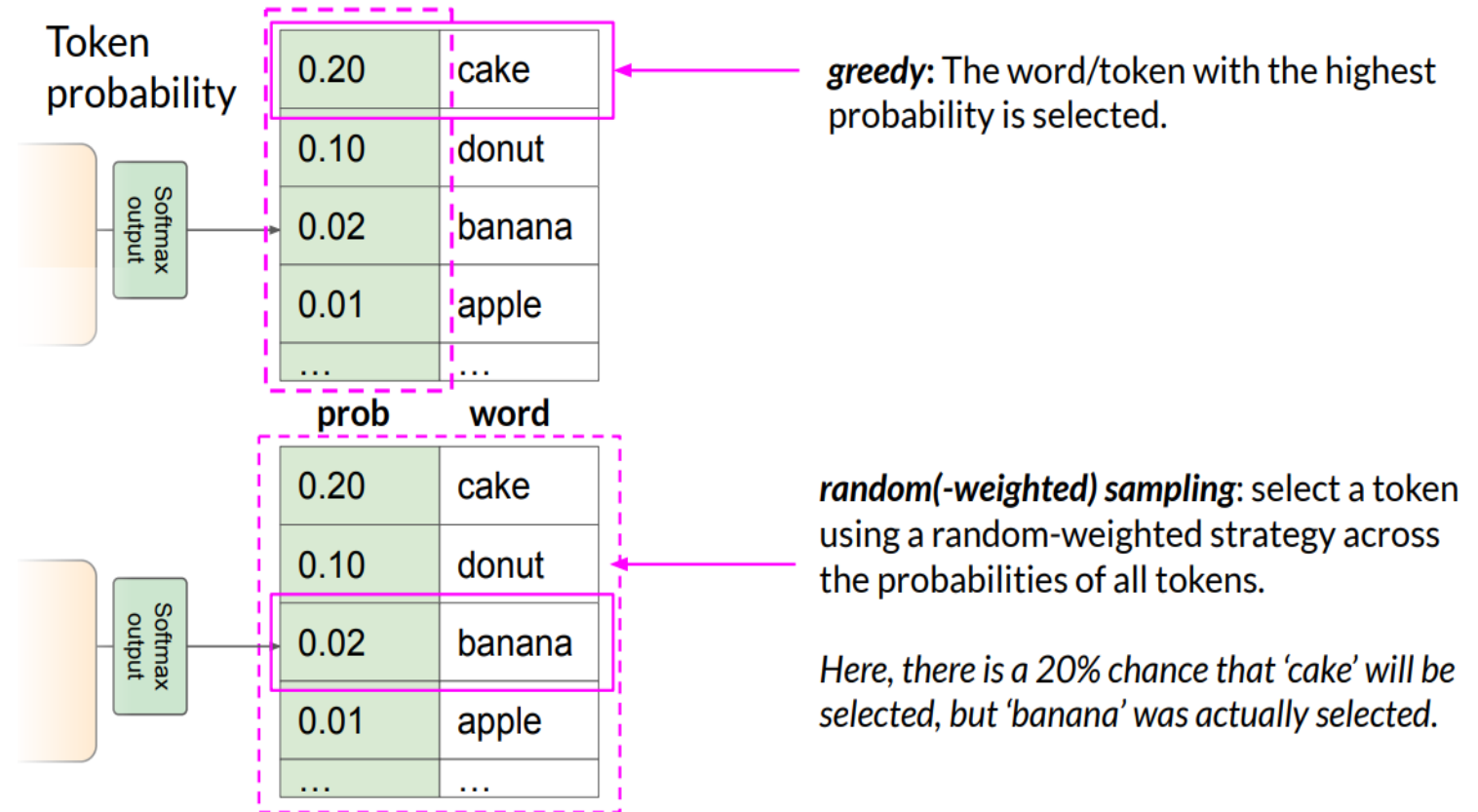
Generative Configuration – max new tokens



- Es como poner un tope al número de veces que el modelo pasará por el proceso de selección. Ejemplos de un máximo de nuevos tokens fijado en 100, 150 o 200.
- OJO, en el ejemplo de 200 la finalización se hace antes, esto se debe a que se alcanzó otra condición de parada, como la predicción del modelo y el token de fin de secuencia.

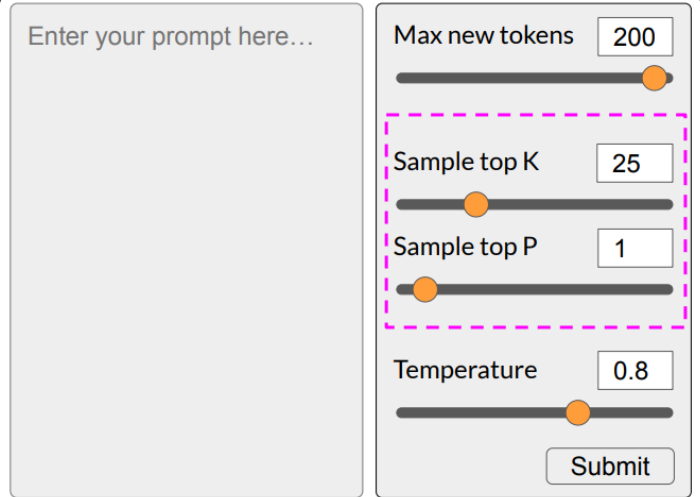
Generative Configuration – Do sample Greedy vs. Random Sampling

Do sample:
False vs True



Generative Configuration – top-k and top-p

- **top p** y **top k** son técnicas de muestreo que podemos emplear para **limitar el muestreo aleatorio e incrementar la probabilidad de que el resultado sea sensato.**
- Para limitar las opciones a la vez que se permite cierta variabilidad es posible especificar un valor top k que indica al modelo a elegir solo entre los primeros k tokens con la probabilidad más alta

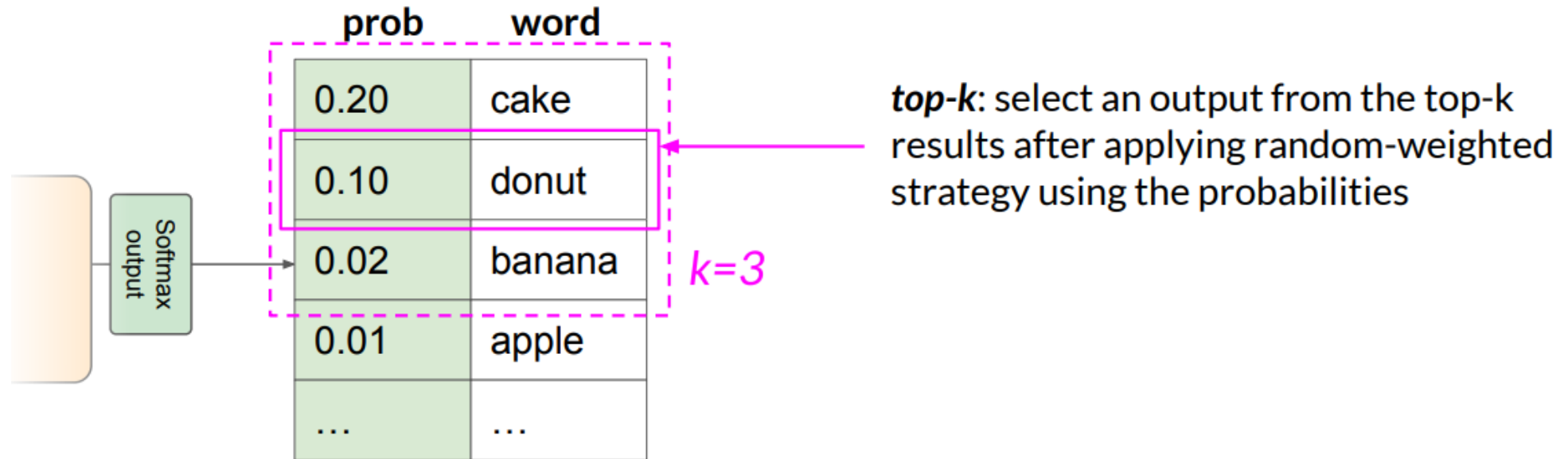


The image shows a user interface for configuring a generative model. On the left is a large text input area with the placeholder "Enter your prompt here...". On the right is a control panel with four sliders and a "Submit" button. The sliders are: "Max new tokens" (set to 200), "Sample top K" (set to 25), "Sample top P" (set to 1), and "Temperature" (set to 0.8). The "Sample top K" and "Sample top P" sliders are enclosed in a dashed pink box, highlighting the top-k and top-p sampling settings.

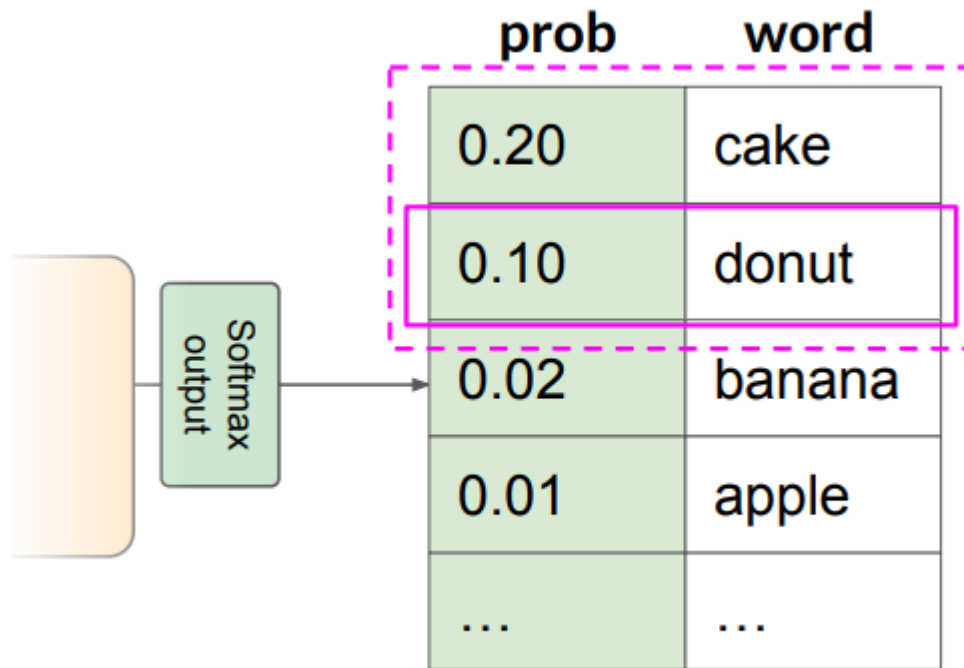
Parameter	Value
Max new tokens	200
Sample top K	25
Sample top P	1
Temperature	0.8

Top-k and top-p sampling

Generative Configuration – top-k sampling



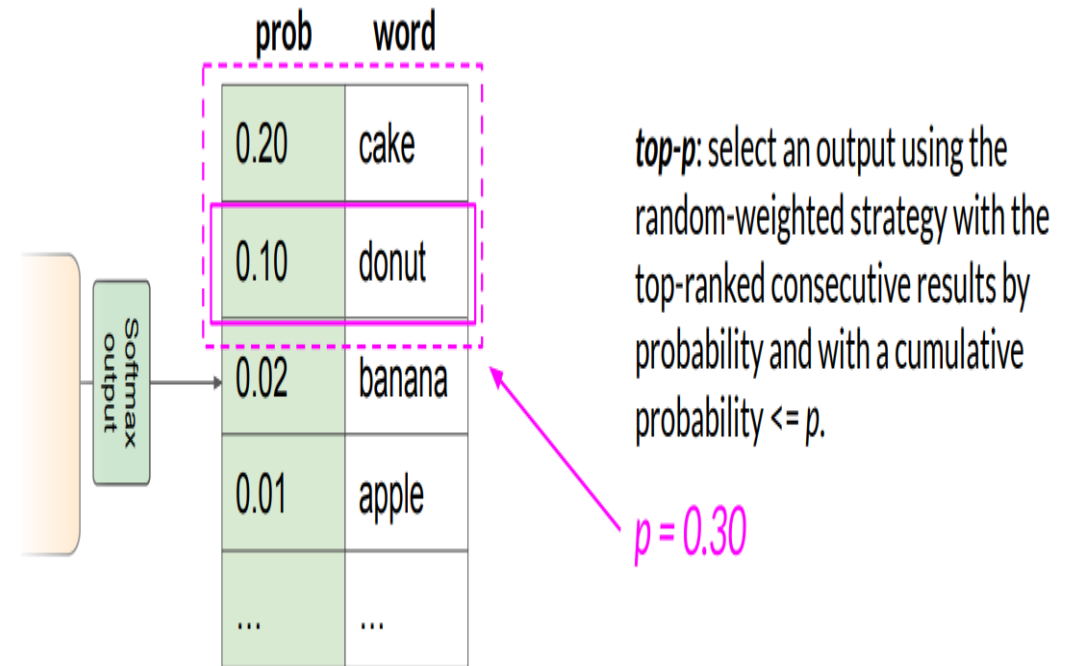
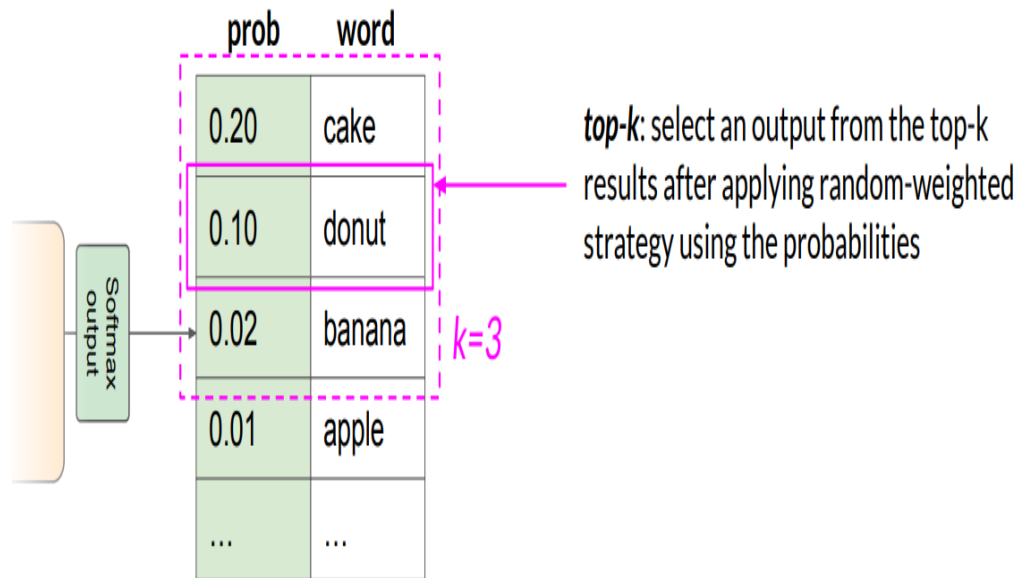
Generative Configuration – top-p sampling



top-p: select an output using the random-weighted strategy with the top-ranked consecutive results by probability and with a cumulative probability $\leq p$.

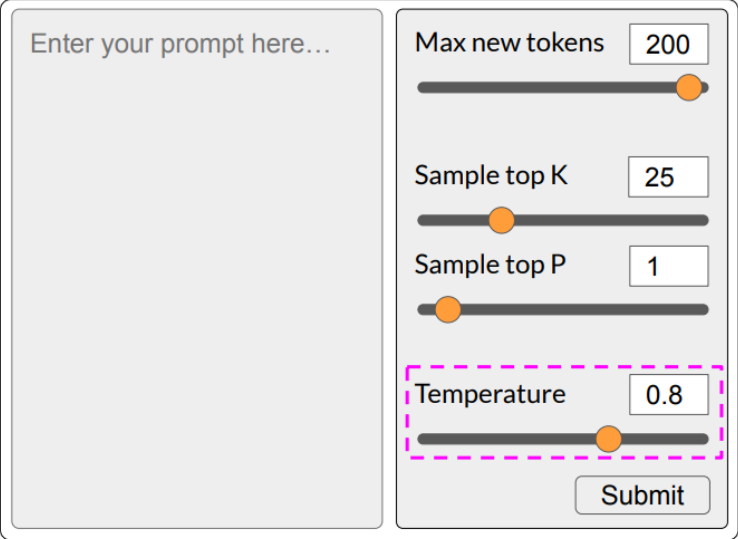
$p = 0.30$

Generative Configuration – top-p sampling



Generative Configuration – temperature

- La **temperatura** es un parámetro que permite controlar la aleatoriedad del resultado generado por el modelo
- Este parámetro influye en la forma de la distribución de probabilidad que el modelo calcula para el siguiente token



The image shows a user interface for configuring a generative model. On the left is a large text input area labeled "Enter your prompt here...". On the right is a settings panel with four sliders and a "Submit" button. The sliders are for "Max new tokens" (set to 200), "Sample top K" (set to 25), "Sample top P" (set to 1), and "Temperature" (set to 0.8). The "Temperature" slider and its label are enclosed in a dashed pink box. The "Submit" button is at the bottom right of the settings panel.

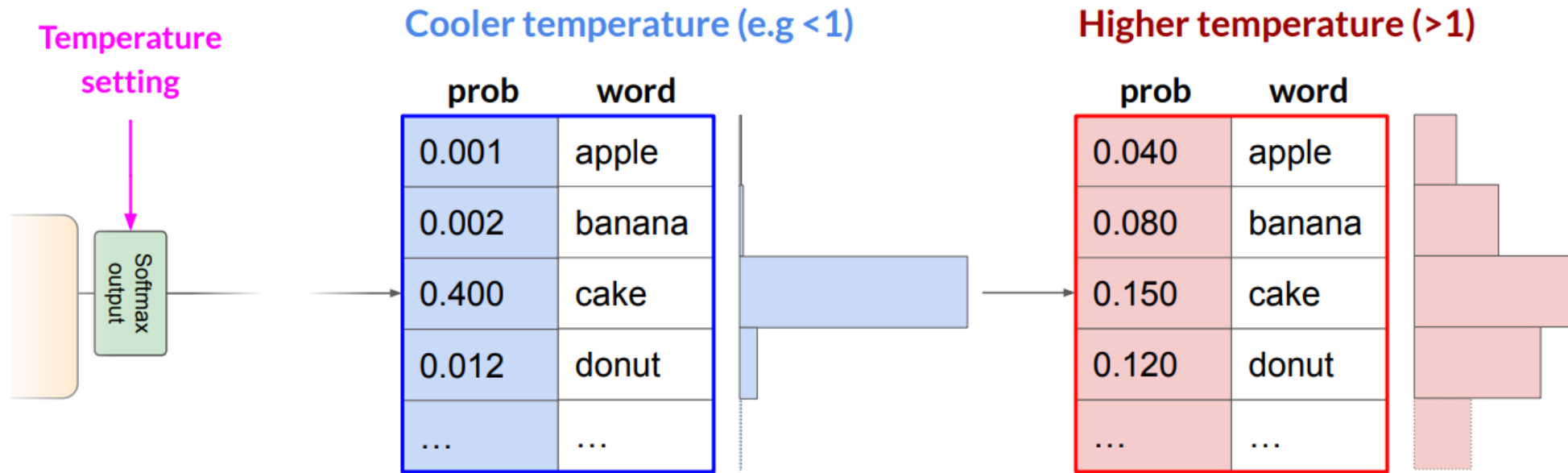
Parameter	Value
Max new tokens	200
Sample top K	25
Sample top P	1
Temperature	0.8

Temperature



- *Mayor temperatura implica mayor aleatoriedad*
- *Menor temperatura implica menor aleatoriedad.*

Generative Configuration – temperature





LAB

