Capítulo 6

Parte 2: Ingeniería de prompts

Prompts e ingeniería de prompts

- Un **prompt** es una instrucción o pregunta que se le proporciona al modelo basado en PLN para indicarle el tipo de respuesta que se espera.
- **Prompt engineering**, se refiere a la habilidad de diseñar y optimizar prompts.
  - Esto no solo implica diseñar y refinar los prompts para obtener los resultados deseados de manera eficiente y precisa,
  - Sino también ajustar y probar diferentes enfoques en los modelos de PLN para mejorar su efectividad y precisión.

Prompts e ingeniería de prompts

- El **ajuste y prueba de diferentes enfoques** dentro de prompt engineering, son las modificaciones y optimizaciones específicas para mejorar la respuesta del modelo. Esto puede incluir:
  - **Ajustar los parámetros** del modelo (ya lo veremos).
  - **Incorporar más datos de entrenamiento relevantes**.
    - **Por ejemplo**: incluir datos adicionales en el prompt, agregar información de contexto adicional en el prompt, dar ejemplos concretos de como quiero las respuestas, más información sobre la tarea o el objetivo que deseo lograr con el prompt.
  - **Refinar la estructura** de los prompts para que el modelo entienda mejor la intención.
  - **Experimentar con diferentes formulaciones** de prompts para ver cuál produce las mejores respuestas.

Prompts e ingeniería de prompts

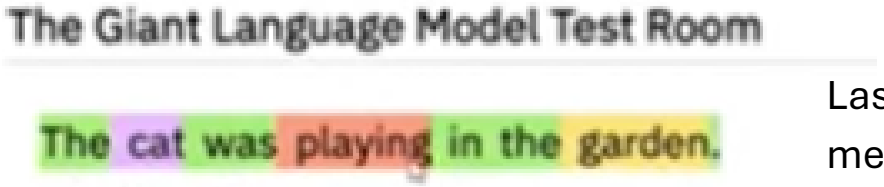
- Un buen ingeniero de prompts debe tener una mezcla de **habilidades técnicas y creativas**:
  - **Conocimiento en PLN**: Entender cómo funcionan los modelos de Procesamiento del Lenguaje Natural.
  - **Creatividad**: Ser capaz de pensar fuera de lo común para formular prompts efectivos.
  - **Análisis de datos**: Evaluar y ajustar prompts basándose en resultados y métricas de rendimiento.
  - **Resolución de problemas**: Identificar y solucionar problemas en la generación de respuestas.
  - **Comunicación clara**: Formular prompts y objetivos de manera precisa y comprensible.
  - **Experimentación**: Probar diferentes enfoques y ajustar parámetros para optimizar resultados.
  - **Comprensión de contexto**: Captar la intención y contexto detrás de las preguntas para generar respuestas precisas.

Conceptos básicos

- Los modelos de texto actuales procesan el texto que reciben como un input, dividiéndolo en **tokens**. A esto le llamamos **tokenización**.
  - Un token es una secuencia de caracteres que se trata como una unidad.
- Los tokens son importantes porque el precio de uso de la API de los modelos de texto se hace en tokens.
- La máxima cantidad de palabras que puede gestionar un modelo de IA de texto se mide en tokens.
  - Por ejemplo, hay una versión de GPT 4 que puede generar 32K tokens.

Conceptos básicos

- Los modelos de texto no entienden el prompt que le damos como entiende un ser humano, pero si es capaz de continuar la frase en muchos casos mejor a como lo haría un ser humano;
    - porque es capaz de predecir cual es la secuencia de palabras más probable según su entrenamiento.
- Los modelos de texto actuales son máquinas matemáticas de probabilidad y lenguaje.
- La tecnología que se usa para detectar los textos que han sido realizados por IA se basan en este funcionamiento de los modelos de texto.
    - Una herramienta analiza la probabilidad de cada palabra según el contexto para determinar si dicho texto está generado por un modelo de IA.
    - Cada palabra de un texto de input se colorea según su probabilidad.



Las de color verde son las más probables de todas. Luego un poco menos probable las que están en amarillo. Las coloreadas en rojo son poco probables. Las más improbables de todas son las de color morado.

Conceptos básicos

- Cuando en un texto la densidad de color rojo y color morado es baja, significa que este texto ha sido realizado por una IA.
- La IA usa muchas más palabras probables que un ser humano cuando escribe. Cuando escribe un ser humano utiliza muchas más palabras improbables.

El playground de OpenAI y sus parámetros

- El **playground de OpenAI** permite a los usuarios experimentar con los modelos GPT de OpenAI (p.ej: escribiendo prompts).
  - Permite elegir el modelo de lenguaje a usar.
  - Permite además ajustar varios parámetros del modelo como la longitud de la secuencia de salida, temperatura, etc.
  - Esta plataforma es más apropiada para trabajar con IAs de texto que ChatGPT.
- Estudiaremos ahora los parámetros más importantes del playground de OpenAI.
- La temperatura sirve para controlar la diversidad en las respuestas. Los valores de este parámetro van de 0 a 1.
  - Con una temperatura alta la diversidad en las respuestas es mayor, por lo que el modelo será más creativo e impredecible en sus respuestas, pero también, con más alta probabilidad de error.
  - La temperatura baja hace que la diversidad de respuestas del modelo sea mucho menor, por lo que el modelo se hace mucho más predecible y determinista, pero también es más fiable en sus respuestas.

El playground de OpenAI y sus parámetros

- Con el parámetro TopP vamos a poder constreñir las diferentes opciones de texto barajadas por el modelo. Por lo tanto, vamos a poder limitar las respuestas irrelevantes o erróneas.
  - Con un TopP alto el modelo va a considerar más opciones, aunque tengan probabilidad baja. Eso va a producir respuestas más creativas, pero también es probable que sean respuestas más irrelevantes o incluso incorrectas.
  - Con un TopP bajo el modelo va a considerar menos opciones; solo las de probabilidad más alta. Esto va a producir respuestas menos creativas, pero también es más probable que sean respuestas relevantes y respuestas correctas.
- Los parámetros **frequency penalty** y **presence penalty** sirven para reducir la probabilidad de que el modelo repita varias veces lo mismo en su texto generado.

El playground de OpenAI y sus parámetros

- **Frequency Penalty**: Este parámetro penaliza la aparición repetida de tokens (palabras o frases) en la respuesta.
  - Si un token ya ha aparecido varias veces, este parámetro aumenta la penalización para reducir la repetición.
- **Presence Penalty**: Este parámetro penaliza la aparición de cualquier token que ya haya sido utilizado en la respuesta.
  - A diferencia del Frequency Penalty, no importa cuántas veces haya aparecido el token, solo que haya aparecido al menos una vez1

Las diferentes partes de un prompt

- Vamos a dividir un prompt en 3 elementos fundamentales:
  - Prompt = acción + modificadores + datos
- **Las acciones:**
  - A través de la acción vamos a establecer qué queremos que realice el modelo de texto.
  - Normalmente son verbos y órdenes sencillas; por ejemplo, escribe, resume, traduce, categoriza.
  - Las acciones pueden ser **instrucciones directas** (p.ej: resume, traduce, define, analiza, etc.) o **preguntas guiadas** p.ej: ¿qué opinas sobre …?
  - En un mismo prompt puede haber varias acciones.
    - Pero tendrás que plantearte en esa circunstancia si es mejor un prompt con dos acciones o realizar dos prompt, cada uno con una acción.

Las diferentes partes de un prompt

- **Modificadores:**
  - La columna vertebral de un prompt son los modificadores.
  - Son elementos que agregamos a la acción para constreñir los resultados y dar forma a la respuesta del modelo.
  - Cuantos más modificadores uses, más vas a moldear la respuesta del modelo de texto.
- Hay diferentes **tipos de modificadores** que se pueden usar en un prompt.
  - El **modificador tipo rol**: sirve para constreñir la personalidad que adopta la inteligencia artificial. La IA usará unas relaciones entre tokens propias de ese rol o esa personalidad.
    - Por ejemplo: perspectiva profesional (p.ej. Médico, abogado), perspectiva educativa (p.ej. Maestro, profesor, coach.), perspectiva técnica (p.ej: como técnico , ingeniero), perspectiva creativa/literaria (p.ej. Escritor, poeta.)
  - El **modificador asunto o tema** sirve para constreñir el tema sobre el que queremos que la IA realice una acción.

Las diferentes partes de un prompt

- **Tipos de modificadores (cont.)**:
  - El modificador **tipología de texto** sirve para constreñir el tipo de texto
    - P.ej: un post (para redes sociales), un correo electrónico, blogs, mensajes de texto (SMS), anuncios publicitarios, presentaciones.
    - P-ej: describir, narrar, exponer, argumentar, dar instrucciones o pasos, persuadir, informar.
  - **Modificador longitud de texto**: constriñe la cantidad de texto que queremos que responda el modelo.
  - **Modificador audiencia objetivo**: constriñe el texto según a quien va dirigida la respuesta del modelo.
    - P.ej: niños, jóvenes o adultos; principiantes, intermedios o expertos; estudiantes, profesores o investigadores; profesionales de distintos tipos).

Las diferentes partes de un prompt

- **Tipos de modificadores (cont.):**
  - **Modificador formato y estructura**: constriñe el formato y/o estructura del texto generado.
    - P.ej: estructura de lista, estructura de tabla, estructura de diálogo, estructura de código.
    - P.ej: formato XML, formato HTML, formato CSV.
  - **Modificador inserciones de texto**: sirve para localizar el punto donde queremos que la IA introduzca un determinado texto.

Las diferentes partes de un prompt

- **Ejemplo**: "Como si fueras un analista financiero (modificador de rol), escribe una entrada de blog (tipología de texto) sobre los beneficios del uso de blockchain en el sector financiero para un público de profesionales de negocios (audiencia objetivo). Asegúrate de utilizar una estructura de lista (estructura) y formatea los términos clave en negritas (formato). Explica detalladamente cómo la adopción de blockchain puede mejorar la transparencia y reducir costos operativos (tipología de texto: expositivo)."

Las diferentes partes de un prompt

- **Ejemplo**: "Como si fueras un científico ambiental (modificador de rol), escribe una presentación (tipología de texto) para estudiantes universitarios (audiencia objetivo) sobre los efectos del cambio climático en los ecosistemas marinos. Utiliza una estructura de lista (estructura) para destacar los puntos principales y elabora cada punto con descripciones detalladas (tipología de texto: descriptivo). Asegúrate de formatear las palabras clave en negritas y utiliza un formato HTML (formato) para que pueda ser fácilmente incrustada en una página web. Además, incluye datos actuales y citas de estudios recientes (contexto adicional) para respaldar tus argumentos."

Las diferentes partes de un prompt

- **Datos**
  - Podemos necesitar que la respuesta del modelo incluya un determinado dato.
    - Podemos aportar ciertos datos a la IA que no conoce, para que los tenga en cuenta en su respuesta.
    - Cuando queremos que la IA use unos datos específicos, tendremos que dárselos en el mismo prompt.
    - **Ejemplo**: Como si fueras un analista de mercado, escribe un informe expositivo sobre las tendencias actuales en la industria tecnológica para inversores intermedios . Asegúrate de incluir el dato de que 'el mercado global de IA se estima en 190 mil millones de dólares para 2025' (dato específico).

Las diferentes partes de un prompt

- **Datos** (**cont**.):
  - Estos datos pueden ir incluidos en línea con la parte principal del prompt, o ir aparte.
  - **Ejemplo**: "Como si fueras un economista, escribe un informe argumentativo sobre el impacto del comercio internacional en la economía global para un público de expertos en economía (audiencia objetivo). Asegúrate de utilizar una estructura de párrafo (estructura) y formatea las citas de estudios en negritas (formato)."
  - Datos adicionales:
    1. El comercio internacional representa aproximadamente el 30% del PIB global.
    2. Según la Organización Mundial del Comercio, el comercio global ha crecido un 3% anual en promedio desde 2000.
    3. El Banco Mundial indica que los países en desarrollo han incrementado su participación en el comercio global en un 50% en las últimas dos décadas.

Tipos de prompt

- Vamos a diferenciar entre los estándar prompt y los chain of thought prompt.
- **Standard prompt:** La IA no genera una cadena de pensamiento que le ayude a dar su respuesta.
  - Se trata de una respuesta directa, sin generar el proceso paso a paso que le ha llevado a dar dicha respuesta.
- **Cadena de pensamiento - CoT** : el modelo de IA genera una cadena de pensamiento que le ayudará a dar su respuesta; respuesta generando el paso a paso.
  - Lo podemos hacer directamente; por ejemplo: poniendo al final del prompt la frase: resuelve el problema paso a paso.
  - O sino podemos decir: resuelve el problema generando todos los pasos y soluciones intermedias.
  - El CoT es especialmente interesante em prompts de tipo lógico o matemático.

Tipos de prompt

- **Cadena de pensamiento (cont):**
  - También podemos indicar al modelo que genere ese CoT dándole ejemplos en el prompt.
  - Podemos aportar en un prompt uno o varios ejemplos del proceso lógico que debe seguir el modelo y éste lo reproducirá en la respuesta.
  - **Ejemplo 1:** Pedido Directo
    - Prompt: "Resuelve el siguiente problema de matemáticas paso a paso: Si tienes 15 manzanas y das 7 a tu amigo, ¿cuántas manzanas te quedan? Asegúrate de mostrar tu cadena de pensamiento y cómo llegas a la respuesta final."

Tipos de prompt

Cadena de pensamiento (cont):

- **Ejemplo 2**: Si tienes 30 libros y das 10 a tu amigo, luego compras 5 más, y finalmente pierdes 3, ¿cuántos libros te quedan?
  - Paso a paso:
  1. Número total de libros = 30.
  2. Libros que das a tu amigo = 10.
  3. Resta 10 de 30: 30 - 10 = 20 libros.
  4. Libros que compras = 5.
  5. Suma 5 a los 20 libros: 20 + 5 = 25 libros.
  6. Libros que pierdes = 3.
  7. Resta 3 de 25: 25 - 3 = 22 libros.
  - Ahora, resuelve este problema similar usando la misma cadena de pensamiento: Si tienes 50 caramelos y das 15 a tu amigo, luego compras 20 más, y finalmente comes 10, ¿cuántos caramelos te quedan? Asegúrate de seguir los pasos y mostrar tu cadena de pensamiento."

Entrenamiento de un modelo de lenguaje para una tarea específica

- Podemos **entrenar a la IA de texto** para que nos dé una respuesta más acorde con nuestras necesidades.
  - Podemos hacerlo mediante el agregado de **ejemplos etiquetados** en un prompt con el objetivo de que el modelo de IA aprenda de ellos extrayendo patrones y características y consiga aportar de esta manera una respuesta más acorde a nuestras necesidades.
- **Zero shot prompt** es un prompt sin muestras o ejemplos. Se usa cuando el modelo de IA puede realizar la tarea encomendada sin la necesidad de un entrenamiento.

Entrenamiento de un modelo de lenguaje para una tarea específica

- **One shot prompt**: prompt con un sola muestra o ejemplo. El único ejemplo en el prompt es para que la IA aprenda.
- **Few-shot prompt**: prompt con unas pocas muestras o ejemplos. Se añaden ejemplos etiquetados en el prompt para que la IA aprenda de ellos.
- El one shot prompt y el few showt prompt se usarán para pedir a la IA tareas más complejas en las cuales necesitamos aportar un entrenamiento a la IA y lo incluiremos con ejemplos etiquetados en el mismo prompt.

Entrenamiento de un modelo de lenguaje para una tarea específica

- **Zero Shot Prompt Example:**
- **User Prompt**: "Calculate the sum of 8, 15, and 23."
- **Chatbot Response**: "The sum of 8, 15, and 23 is 46."
- **One Shot Prompt Example:**
- **User Prompt**: "Calculate the sum of 8, 15, and 23.
  - Response: 8+15+23 = 46
  - Calculate the sum of 9, 24, 87
  - **Chatbot Response**: Response: 9 + 24 + 87 = 120

Entrenamiento de un modelo de lenguaje para una tarea específica

- Ahora vemos un ejemplo donde quiero que el modelo cambie la estructura de una frase.
- Quiero que primero me ponga el tiempo, a continuación, el lugar y en tercer lugar la voz pasiva.
  - Esto es prácticamente imposible en un zero shot prompt.
- **Prompt**: "Cambia la estructura de la frase de voz activa a voz pasiva. La nueva estructura debe ser : primero el tiempo, luego el lugar, y finalmente la voz pasiva."
- **Ejemplo 1**:
- **Input**: "El chef preparó una deliciosa comida en el restaurante ayer."
- **Resultado**: "Ayer, en el restaurante, una deliciosa comida fue preparada por el chef."
- **Ejemplo 2**:
- **Input**: "Los niños jugaron al fútbol en el parque esta mañana."
- **Resultado**: "Esta mañana, en el parque, el fútbol fue jugado por los niños."

Entrenamiento de un modelo de lenguaje para una tarea específica

- **Ejemplo 3:**
- **Input**:
  "Los científicos descubrieron una nueva especie en la selva la semana pasada."
- **Resultado**: "La semana pasada, en la selva, una nueva especie fue descubierta po r los científicos."
- **Input:** "El equipo lanzó un nuevo producto en la conferencia el mes pasado.“
- **Respuesta del chatbot:**
- **Resultado**: "El mes pasado, en la conferencia, un nuevo producto fue lanzado por el equipo."

Entrenamiento de un modelo de lenguaje para una tarea específica

- Se puede combinar el one shot prompt o el few shot prompt con la cadena de pensamiento.
- **Ejemplo**: Resuelva el ejercicio siguiendo el siguiente proceso:
  - Pregunta: (4+5) – 2 – (1+1)
  - Operaciones:
  - 9 – 2 – (1+1)
  - 9 – 2 – 2
  - Resultado: 5
  - Pregunta: (3 + 5) + 8 – (2+1)

Entrenamiento de un modelo de lenguaje para una tarea específica

- **Respuesta del chatbot:**
  - Operaciones:
  - 8 + 8 – (2+1)
  - 8 + 8 – 3
  - Resultado: 13

Recomendaciones para la creación de prompts

- **Evitar textos extensos:** evita crear prompts que generen textos muy extensos ya que siempre hay más probabilidad de que el modelo cree repeticiones y bucles.
  - Es preferible generar varios textos más pequeños donde la probabilidad de repeticiones y bucles es mucho menor.
  - Finalmente, uniremos todos esos textos.
  - **Ejemplo**: queremos crear un post para un blog; una opción es escribir un prompt que genere todo el texto de ese post (no recomendable); otra opción es en primer lugar crear un prompt para que el modelo genere los diferentes títulos o apartados del post;
    - ❑ después nosotros elegimos qué apartados nos interesan;
    - ❑ a continuación, ejecutar un prompt para generar el texto asociado a cada uno de los apartados, y
    - ❑ finalmente unir todos los textos generados.

Recomendaciones para la creación de prompts

- Escribe prompts que sean muy **específicos y detallados**.
  - Evita las ambigüedades
  - Crea el prompt, revísalo, ejecútalo y luego modifícalo para ir optimizándolo.
  - Cuanto más específico y detallado sea el prompt, menor será la probabilidad de que el modelo alucine.
- Puedes especificar en el prompt **que no alucine**, cuando tengas sospechas de que pueda llegar a hacerlo.
  - **Ejemplo**: en un prompt para crear la descripción de un producto de una tienda online, le podemos indicar al modelo en el prompt que no invente datos ni características. Usa solo los datos y características que te proporciono en el prompt.

Recomendaciones para la creación de prompts

- Es una buena idea **separar el prompt en sus diferentes partes**.
  - Separar acción, modificadores y datos.
  - Separar el prompt en sus diferentes partes nos va a permitir analizar el prompt mucho más fácilmente.
  - También nos va a permitir revisarlo con mayor facilidad y modificarlo o editarlo.
  - También nos va a permitir hacer más fácilmente su reutilización realizando modificaciones.

Recomendaciones para la creación de prompts

- **Visualiza los datos que el modelo no conoce y apórtaselos en el prompt**.
  - Debes tener claro en cada prompt si debes aportar al modelo datos necesarios que este desconoce.
    - ❑ Normalmente los va a desconocer porque dichos datos no estaban disponibles en el data set con el que fue entrenado;
    - ❑ aunque también puede ser por otros motivos.
  - ChatGPT no tiene datos actualizados por el momento al año presente.
    - ❑ Si no le damos los datos, es probable que el modelo los invente.

Recomendaciones para la creación de prompts

- **Ejemplo**: Imagina que queremos que el modelo genere un plan de empresa; tendremos que aportarle todos los datos necesarios de la empresa que él lógicamente no conoce.
  - Sino no podrá generar un plan de empresa de una manera correcta.
  - Seguramente, inventará muchos de los datos.
- Cuando veas que el modelo no contesta lo que quieres, dale ejemplos.
  - Usa el one shot o el few shot prompting.
  - Pero empieza siempre con el zero shot prompting, ya que es más sencillo y menos laborioso.
  - Y si no te funciona, pasa al one shot o al few shot prompting.

Recomendaciones para la creación de prompts

- **La integración de IAs de texto va a tener un costo**
  - **Ejemplo**, integrar una IA de texto en una herramienta o integrar una IA de texto en una página web o en una tienda online va a tener un costo.
  - Normalmente ese coste va a depender del número de tokens del prompt que se ejecuta y de la respuesta del modelo.
- **Ejemplo**: si queremos añadir un chatbot basado en ChatGPT en nuestra tienda online, esto va a tener un costo.
  - Tendremos que crear un prompt inicial que se ejecutará junto al texto que escriba el usuario; por lo tanto, ese prompt inicial tendrá un costo según el número de tokens y las veces que se ejecute.
  - Si el prompt es muy largo, entonces va a tener un mayor costo. Y si el prompt es más corto, va a tener un menor costo.
  - Por lo tanto, hay que acostumbrarse a parafrasear y simplificar los prompts en la manera de lo posible.