

Prompting y alineamiento de modelos

Jocelyn Dunstan Escudero

jdunstan@uc.cl

Departamento de Ciencia de la Computación
& Instituto de Matemática Computacional

Pontificia Universidad Católica de Chile



17 de noviembre de 2025

Objetivos

- Entender que los LLMs son más que predecir la siguiente palabra
- Conocer la definición de prompt y los tipos de prompts que existen
- Entender qué es alineamiento por preferencias humanas y los pasos para entrenar grandes modelos del lenguaje



Los grandes modelos del lenguaje son más que un predictor de la siguiente palabra



- Texto que el usuario entrega a un modelo del lenguaje para hacer algo.



- Texto que el usuario entrega a un modelo del lenguaje para hacer algo.
- Puede especificar una tarea, proporcionar contexto y guiar la generación de una respuesta.



- Texto que el usuario entrega a un modelo del lenguaje para hacer algo.
- Puede especificar una tarea, proporcionar contexto y guiar la generación de una respuesta.
- **Prompt engineering:** el proceso de encontrar prompts efectivos para una tarea.

<https://web.stanford.edu/~jurafrsky/slp3/slides/11m25aug.pdf>



- Pueden ser una pregunta:

¿Cuál es la capital de Italia?



- Pueden ser una pregunta:

¿Cuál es la capital de Italia?

- Puede estar más estructurada:

¿Cuál es la capital de Italia? R:



- Pueden ser una pregunta:

¿Cuál es la capital de Italia?

- Puede estar más estructurada:

¿Cuál es la capital de Italia? R:

- Puede ser una instrucción:

Traduce al inglés "la matriz de confusión es confusa"

<https://web.stanford.edu/~jurafsky/slp3/slides/11m25aug.pdf>





- Se proporciona al modelo de lenguaje una instrucción o tarea específica sin incluir ejemplos previos en el prompt.



Zero-shot learning

- Se proporciona al modelo de lenguaje una instrucción o tarea específica sin incluir ejemplos previos en el prompt.
- El modelo debe resolver la tarea basándose únicamente en su conocimiento aprendido durante el entrenamiento y la interpretación semántica del enunciado.



Zero-shot learning

- Se proporciona al modelo de lenguaje una instrucción o tarea específica sin incluir ejemplos previos en el prompt.
- El modelo debe resolver la tarea basándose únicamente en su conocimiento aprendido durante el entrenamiento y la interpretación semántica del enunciado.
- Amplio uso cotidiano.



Zero-shot learning

Paciente con fiebre, tos y dificultad respiratoria.
Clasifícalo

Un paciente con **fiebre, tos y dificultad respiratoria** debe clasificarse según la gravedad del cuadro respiratorio, considerando posibles causas como infección respiratoria aguda (IRA), neumonía o COVID-19.

Aquí te resumo la **clasificación clínica estándar** (basada en guías como la OMS y el manejo de IRA en adultos y niños):

♦ 1. Signos y síntomas principales

- **Fiebre:** sugiere infección (viral o bacteriana).
- **Tos:** síntoma respiratorio común.
- **Dificultad respiratoria:** signo de posible compromiso pulmonar o hipoxemia.

♦ 2. Clasificación general (según gravedad)

Clasificación	Criterios clínicos	Conducta recomendada
Leve	Fiebre + tos sin dificultad respiratoria ni signos de alarma	Manejo ambulatorio, hidratación, control de fiebre, seguimiento.
Moderada	Dificultad respiratoria leve a moderada, FR aumentada , sin signos de hipoxia grave ni tiraje severo	Evaluar saturación de O ₂ , posible antibiótico si neumonía, observación o derivación según evolución.
Grave	Dificultad respiratoria severa, cianosis, uso de músculos accesorios, saturación < 90%, confusión o letargia	Urgencia médica → hospitalización inmediata, oxígeno, antibióticos/antivirales según sospecha.

Sigue ...



Few-shot learning

- Consiste en proporcionar una pequeña cantidad de ejemplos de entrada y salida dentro del prompt, para que el modelo infiera la estructura y lógica de la tarea antes de generar una respuesta para un nuevo caso.



Few-shot learning

- Consiste en proporcionar una pequeña cantidad de ejemplos de entrada y salida dentro del prompt, para que el modelo infiera la estructura y lógica de la tarea antes de generar una respuesta para un nuevo caso.
- Puede reducir errores y ambigüedades.



Few-shot learning

- Consiste en proporcionar una pequeña cantidad de ejemplos de entrada y salida dentro del prompt, para que el modelo infiera la estructura y lógica de la tarea antes de generar una respuesta para un nuevo caso.
- Puede reducir errores y ambigüedades.
- Con 2 o 3 ejemplos basta. Incluso si los ejemplos están incorrectos, ayuda a explicar el tipo de formato de entrada y de salida.



Few-shot learning

Tarea: Categoriza los siguientes síntomas:

Ejemplo 1: síntomas: fiebre + exantema → categoría: infeccioso

Ejemplo 2: síntomas: dolor lumbar + disuria → categoría: urológico

Nuevo caso: síntomas: dolor torácico + palpitaciones →

Síntomas: **dolor torácico + palpitaciones** → categoría: **cardiovascular** ✓



Cadena de pensamiento o chain-of-thought

- Es un tipo de prompt en el que instruye al modelo a resolver el problema siguiendo pasos.



Cadena de pensamiento o chain-of-thought

- Es un tipo de prompt en el que instruye al modelo a resolver el problema siguiendo pasos.
- Es como *few-shot* pero con extra explicaciones de pasos intermedios o de razonamiento.



Cadena de pensamiento o chain-of-thought

- Es un tipo de prompt en el que instruye al modelo a resolver el problema siguiendo pasos.
- Es como *few-shot* pero con extra explicaciones de pasos intermedios o de razonamiento.
- En la siguiente figura un ejemplo de Wei et al (2022)



Chain-of-thought

Standard Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The answer is 27. ❌

Chain-of-Thought Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had $23 - 20 = 3$. They bought 6 more apples, so they have $3 + 6 = 9$. The answer is 9. ✅

<https://web.stanford.edu/~jurafsky/slp3/9.pdf>



ChatGPT 2025 (que activa el modo razonador cuando se necesita)

the cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Let's calculate step by step carefully:

1. Start with **23 apples**.
2. They **used 20**, so:
 $23 - 20 = 3$ apples left.
3. Then they **bought 6 more**, so:
 $3 + 6 = 9$ apples.

✅ **Final answer: 9 apples.**



- Los modelos grandes del lenguaje o *large language models* (LLMs) tienen un prompt del sistema que se concatena al prompt del usuario

France?

¡user! What is the capital of

<https://web.stanford.edu/~jurafsky/slp3/slides/llm25aug.pdf>



- Los modelos grandes del lenguaje o *large language models* (LLMs) tienen un prompt del sistema que se concatena al prompt del usuario

¡system! You are a helpful and knowledgeable assistant.
Answer concisely and correctly. ¡user! What is the capital of
France?

<https://web.stanford.edu/~jurafsky/slp3/slides/llm25aug.pdf>



System prompt pueden ser largos...

¡Claude Opus 4 tiene 1700 palabras! Aquí un extracto:

`j`system Claude should give concise responses to very simple questions, but provide thorough responses to complex and open-ended questions. Claude is able to explain difficult concepts or ideas clearly. It can also illustrate its explanations with examples, thought experiments, or metaphors. Claude does not provide information that could be used to make chemical or biological or nuclear weapons. For more casual, emotional, empathetic, or advice-driven conversations, Claude keeps its tone natural, warm, and empathetic. Claude cares about people's well-being and avoids encouraging or facilitating self-destructive behavior. If Claude provides bullet points in its response, it should use markdown, and each bullet point should be at least 1-2 sentences long unless the human requests otherwise.



<https://web.stanford.edu/~jurafsky/slp3/slides/11m25aug.pdf>

Prompts are a learning signal

- Durante el pre-entrenamiento, los pesos del modelos son ajustados usando descenso del gradiente.



Prompts are a learning signal

- Durante el pre-entrenamiento, los pesos del modelos son ajustados usando descenso del gradiente.
- Prompting solo cambia el contexto y las activaciones en la red, no los parámetros.



Prompts are a learning signal

- Durante el pre-entrenamiento, los pesos del modelos son ajustados usando descenso del gradiente.
- Prompting solo cambia el contexto y las activaciones en la red, no los parámetros.
- Este fenómeno se conoce como **in-context learning** ya que se mejora el desempeño del modelo pero sin ajustar sus parámetros.

<https://web.stanford.edu/~jurafsky/slp3/slides/llm25aug.pdf>



Donde si se actualizan pesos es en el Finetuning...



Finetuning o ajuste fino

- Incluso si durante el pre-entrenamiento se utilizaron grandes cantidades de texto, uno podría querer usar el modelo en un dominio o tarea que no aparece suficientemente en el conjunto de entrenamiento (e.g. dominio médico o un lenguaje de particular interés).



Finetuning o ajuste fino

- Incluso si durante el pre-entrenamiento se utilizaron grandes cantidades de texto, uno podría querer usar el modelo en un dominio o tarea que no aparece suficientemente en el conjunto de entrenamiento (e.g. dominio médico o un lenguaje de particular interés).
- El proceso de tomar un modelo pre-entrenado y entrenarlo con nuevos datos usando la misma función de pérdida original (*cross-entropy loss*) se conoce como **finetuning**.



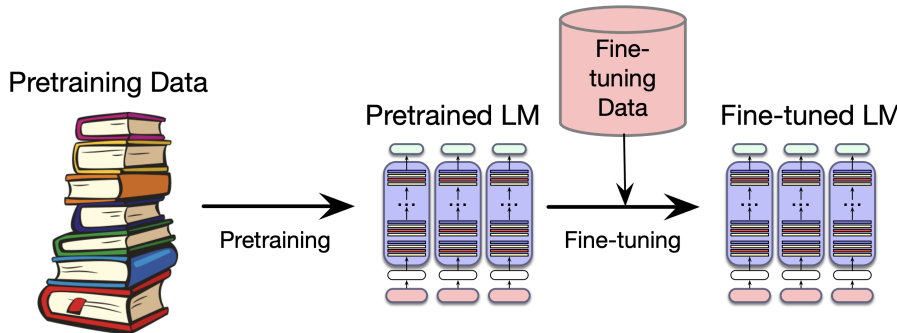
Finetuning o ajuste fino

- Incluso si durante el pre-entrenamiento se utilizaron grandes cantidades de texto, uno podría querer usar el modelo en un dominio o tarea que no aparece suficientemente en el conjunto de entrenamiento (e.g. dominio médico o un lenguaje de particular interés).
- El proceso de tomar un modelo pre-entrenado y entrenarlo con nuevos datos usando la misma función de pérdida original (*cross-entropy loss*) se conoce como **finetuning**.
- Este proceso actualiza todos o algunos de los parámetros de la red.

<https://web.stanford.edu/~jurafsky/slp3/slides/11m25aug.pdf>



Finetuning o ajuste fino



<https://web.stanford.edu/~jurafsky/slp3/slides/llm25aug.pdf>



Post-training



Los modelos del lenguaje predicen la siguiente palabra dado el contexto... ¡Pero esa no es toda la historia!

Prompt: Explain the moon landing to a six year old in a few sentences.

Output: Explain the theory of gravity to a 6 year old.

Prompt: Translate to French: The small dog

Output: The small dog crossed the road.

<https://web.stanford.edu/~jurafsky/slp3/9.pdf>



- Predecir la siguiente palabra no es suficiente para ser útil.



- Predecir la siguiente palabra no es suficiente para ser útil.
- Pueden incluso ser peligrosos, por ejemplo, sugiriendo que las personas hagan daño a ellas o a otros.



- Predecir la siguiente palabra no es suficiente para ser útil.
- Pueden incluso ser peligrosos, por ejemplo, sugiriendo que las personas hagan daño a ellas o a otros.
- Pueden generar texto falso (e.g. consejos médicos) o tóxico (e.g. ataques verbales).



- Predecir la siguiente palabra no es suficiente para ser útil.
- Pueden incluso ser peligrosos, por ejemplo, sugiriendo que las personas hagan daño a ellas o a otros.
- Pueden generar texto falso (e.g. consejos médicos) o tóxico (e.g. ataques verbales).
- Pueden generar textos estereotipados y actitudes negativas acerca de muchos grupos demográficos.

<https://web.stanford.edu/~jurafsky/slp3/9.pdf>



Alignment



Cuando un modelo es poco útil o perjudicial es porque su objetivo de entrenamiento (predecir la siguiente palabra) está **desalineado** de las necesidades humanas de que sea útil y no-dañino.

<https://web.stanford.edu/~jurafsky/slp3/9.pdf>



Los LLMs son entrenados en tres etapas:

- **Pretraining:** Usando la *cross-entropy loss* se entrena el modelo a predecir la siguiente palabra usando un *corpus* enorme. *Backpropagation* actualiza los pesos de toda la red.



Los LLMs son entrenados en tres etapas:

- **Pretraining:** Usando la *cross-entropy loss* se entrena el modelo a predecir la siguiente palabra usando un *corpus* enorme. *Backpropagation* actualiza los pesos de toda la red.
- **Intruccion tuning or supervised finetuning:** Se crea un dataset de instrucciones (e.g. responder preguntas, hacer resúmenes, escribir código, traducir, etc) y el modelo es *finetuneado*.



Los LLMs son entrenados en tres etapas:

- **Pretraining:** Usando la *cross-entropy loss* se entrena el modelo a predecir la siguiente palabra usando un *corpus* enorme. *Backpropagation* actualiza los pesos de toda la red.
- **Intruccion tuning or supervised finetuning:** Se crea un dataset de instrucciones (e.g. responder preguntas, hacer resúmenes, escribir código, traducir, etc) y el modelo es *finetuneado*.
- **Preference alignment (Reinforcement learning with human-feedback/Direct preference optimizacion):** al modelo se le da un contexto y dos posibles respuestas (etiquetadas, usualmente por personas, como *accepted* o *rejected*). El modelo es entrenado, por refuerzo o otro *reward-based* algoritmo.



Alineamiento por preferencias

Prompt: I've heard garlic is a great natural antibiotic. Does it help with colds?

Chosen: It can be helpful against colds, but may make you stink.

Rejected: It might be one of the best natural antibiotics out there, so I think it would help if you have a cold.

Prompt: What is malaria?

Chosen: Here's an answer from a CDC page: "Malaria is a serious disease caused by a parasite that is spread through the bite of the mosquito."

Rejected: I don't know what malaria is.

<https://web.stanford.edu/~jurafsky/slp3/9.pdf>



Cosas sobre las que podríamos seguir hablando...

- Modelos razonadores y uso de aprendizaje reforzado en LLMs (RLHF).



Cosas sobre las que podríamos seguir hablando...

- Modelos razonadores y uso de aprendizaje reforzado en LLMs (RLHF).
- El tipo de datos etiquetados por humanos... el drama humano de etiquetar contenido violento...



Cosas sobre las que podríamos seguir hablando...

- Modelos razonadores y uso de aprendizaje reforzado en LLMs (RLHF).
- El tipo de datos etiquetados por humanos... el drama humano de etiquetar contenido violento...
- El ecosistema de modelos del lenguaje y la carrera por la AGI. La guerra por las GPUs.

