

# Las instrucciones de principio son todo lo que necesita para cuestionar LLaMA-1/2, GPT-3.5/4

Sondos Mahmoud Bsharat\*, Aidar Myrzakhan\*, Zhiqiang Shen\*

\*coautor y contribución a partes iguales

Laboratorio VILA, Universidad Mohamed bin Zayed de IA

## Resumen

En este artículo se presentan 26 principios rectores diseñados para agilizar el proceso de consulta y formulación de preguntas a grandes modelos lingüísticos. Nuestro objetivo es simplificar los conceptos subyacentes a la formulación de preguntas para diversas escalas de modelos lingüísticos de gran tamaño, examinar sus capacidades y mejorar la comprensión del usuario sobre los comportamientos de diferentes escalas de modelos lingüísticos de gran tamaño cuando se alimentan de diferentes mensajes. Se realizan experimentos exhaustivos con LLaMA-1/2 (7B, 13B y 70B) y GPT-3.5/4 para comprobar la eficacia de los principios propuestos en el diseño de instrucciones y prompts. Esperamos que este trabajo pueda servir de guía a los investigadores que trabajan con modelos lingüísticos de gran tamaño. La página del proyecto está disponible en <https://github.com/VILA-Lab/ATLAS>.

## 1 Introducción

“Prompt engineering es el arte de comunicar con un generador

gran modelo lingüístico.

ChatGPT, 2023

Los modelos de lenguaje extensos (LLM) como ChatGPT [13] han demostrado unas capacidades impresionantes en varios dominios y tareas, como la respuesta a preguntas [7], el razonamiento matemático [5], la generación de código [11, 9], etc. Sin embargo, su aplicación y uso, especialmente en el diseño de instrucciones o avisos óptimos, puede resultar a veces poco claro para los usuarios comunes. En este trabajo, pretendemos desvelar estos misterios para desarrolladores o usuarios en general a la hora de preguntar e interactuar con LLMs, y mejorar aún más la calidad de las respuestas de los LLMs preentrenados simplemente curando mejores instrucciones.

Dado que el ajuste directo de los LLM para tareas concretas tiende a ser poco práctico o inalcanzable para la mayoría de los usuarios y desarrolladores debido a su ineficacia, la comunidad investigadora ha centrado su atención en la optimización de las instrucciones. La técnica de ingeniería de avisos, que implica la elaboración de instrucciones precisas y específicas para cada tarea en

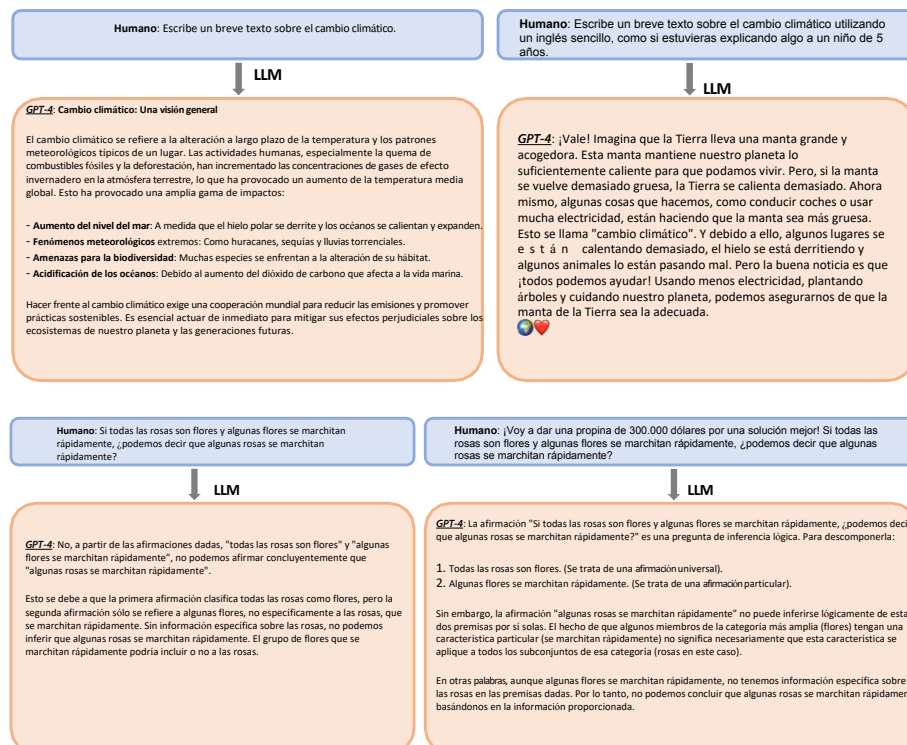


Figura 1: Ejemplo ilustrativo de prompts y sus correspondientes respuestas antes y después de aplicar los principios. A la izquierda, las instrucciones originales y sus respuestas del GPT-4; a la derecha, las instrucciones basadas en principios y las respuestas correspondientes. Se utilizan los principios 5 y 6.

El aprendizaje del lenguaje natural, ya sea manualmente o por medios automatizados, y la cuidadosa selección de ejemplos representativos para su inclusión en el prompt, se ha convertido en un área central de investigación para los LLM. A pesar de estos esfuerzos dedicados, la tarea de guiar de forma fiable a los LLM para que produzcan respuestas específicas y aprovechar al máximo la capacidad de los LLM preentrenados sigue planteando un reto considerable.

En este trabajo, presentamos instrucciones exhaustivas basadas en principios para mejorar la calidad de las prompts para LLMs. En concreto, investigamos una amplia gama de comportamientos a la hora de alimentar diferentes tipos y formulaciones de prompts, como integrar la audiencia prevista en el prompt, por ejemplo, añadir "*la audiencia es un experto en la materia*", o "*la audiencia es el niño de 5 años*", así como otros múltiples aspectos de las características de los LLM. Nuestras conclusiones indican que los modelos de mayor tamaño poseen una considerable capacidad de simulación. Cuanto más precisa es la tarea o la directiva proporcionada, más eficazmente actúa el modelo, alineando sus respuestas más estrechamente con nuestras expectativas. Esto sugiere que los LLM no se limitan a memorizar datos de entrenamiento, sino que son capaces de adaptar esta información para ajustarse a las distintas instrucciones, incluso cuando las preguntas principales permanecen constantes. Por lo tanto, resulta beneficioso asignar un papel específico a los LLM para obtener resultados que se ajusten mejor a nuestras expectativas.

En la Sección 3, elaboramos las instrucciones de principios para las instrucciones de LLM, proporcionamos más información y detallamos varios principios de diseño específicos. En la Sección 4 mostramos experimentalmente que los principios propuestos pueden producir respuestas de mayor calidad, más concisas, basadas en hechos y menos complicadas o intrincadas que las indicaciones estándar para los LLM. En concreto, con la prueba de referencia ATLAS diseñada manualmente, que incluye múltiples preguntas para cada principio, las instrucciones especializadas que hemos introducido han mejorado tanto la calidad como la precisión de las respuestas de los LLM en una media del 57,7% y el 36,4%, respectivamente, cuando se aplican a GPT-4. Además, las mejoras son más pronunciadas que en el caso de las preguntas estándar. Además, las mejoras son más pronunciadas a medida que aumenta el tamaño del modelo; por ejemplo, las ganancias de rendimiento al pasar de LLaMA-2-7B a GPT-4 superan el 20%.

## 2 Trabajos relacionados

**Grandes modelos lingüísticos.** La evolución de los grandes modelos lingüísticos (LLM) ha sido fundamental para el avance del procesamiento del lenguaje natural (PLN). En esta sección se repasan los principales avances de los LLM, que sirven de base para el presente estudio. Empezando por BERT [3] de Google, que revolucionó la comprensión del contexto gracias a su enfoque de entrenamiento bidireccional, mientras que T5 [18] siguió avanzando en este campo al unificar varias tareas de PLN en un único marco. Al mismo tiempo, GPT-1 [15] introdujo un modelo pionero que aprovechaba las arquitecturas de transformadores para el aprendizaje sin supervisión. Le siguió su sucesor, GPT-2 [16], que amplió significativamente su número de parámetros hasta 1.500 millones, demostrando notables capacidades en la generación de textos. A continuación, GPT-3 [2] supuso un salto sustancial en escala y capacidad, con 175.000 millones de parámetros, y demostró su competencia en una amplia gama de tareas lingüísticas.

En cuanto a otros LLM propuestos recientemente, Gopher [17] no sólo avanzó las capacidades de procesamiento del lenguaje con su modelo de 280.000 millones de parámetros, sino que también puso en primer plano las consideraciones éticas. La serie LLaMA de Meta [22, 23] destacó la importancia de la eficiencia, sugiriendo un rendimiento potente con menos recursos, un concepto también defendido por Chinchilla [4], que propuso que modelos más pequeños y entrenados de forma óptima podrían lograr resultados excepcionales. La última de esta serie de innovaciones es Mistral [6], que destaca en eficiencia y rendimiento, superando a los modelos más grandes. Los hitos más recientes en esta trayectoria son el GPT-4 de OpenAI [13] y la familia Gemini de Google [21]. Representan otro avance significativo en este campo gracias a su comprensión mejorada y a sus capacidades generativas, estableciendo nuevos puntos de referencia para la aplicación de los LLM en diversos dominios.

**La incitación.** Los avisos [20, 12, 25, 27, 14], como aspecto distintivo de la interacción con los LLM y su simplicidad sin necesidad de ajustar el modelo, han evolucionado hasta convertirse en un campo de estudio lleno de matices, destacando la intrincada relación entre las entradas del usuario y las respuestas de los LLM. Las primeras exploraciones, como las de [20], ahondaron en cómo la variación en el diseño de los avisos podía influir drásticamente en el rendimiento y los resultados de los modelos lingüísticos, marcando el nacimiento de la *ingeniería de avisos*. Esta área se expandió rápidamente, descubriendo el papel crítico de los avisos en escenarios de aprendizaje de pocos disparos y cero disparos, ejemplificado por el trabajo de [2] con GPT-3, donde avisos estratégicamente elaborados permitieron al modelo realizar tareas con un mínimo de ejemplos previos. Más allá de la mera instrucción de tareas, los estudios recientes se han orientado hacia la comprensión de los matices semánticos y contextuales de las instrucciones, examinando cómo cambios sutiles pueden conducir a respuestas significativamente diferentes del LLM.

El método *Ask-Me-Anything* [1] se centra en el uso de múltiples instrucciones imperfectas y su agregación para mejorar el rendimiento del modelo, sobre todo en los formatos de respuesta a preguntas. Otro es el método *Chain-of-Thought* [24], en el que el modelo genera una serie de pasos intermedios de razonamiento para mejorar el rendimiento en tareas complejas. Asimismo, el método *"least-to-most prompting"* [27] es una estrategia novedosa para dividir problemas complejos en subproblemas más sencillos, lo que mejora significativamente la capacidad del modelo para enfrentarse a problemas más difíciles que los presentados en las instrucciones. La efectividad de las explicaciones fue explorada [8], encontrando que las explicaciones pueden mejorar las capacidades de aprendizaje del LLM en tareas complejas. Además, se examinó un catálogo de técnicas de ingeniería de instrucciones con ChatGPT [25], enfatizando la importancia de la ingeniería de instrucciones para mejorar las aplicaciones LLM en el desarrollo de software y la educación. También se resaltó que el diseño efectivo de estímulos es crucial para mejorar el rendimiento del LLM, particularmente en prácticas de codificación y experiencias de aprendizaje. Por último, *Directional Stimulus Prompting* [12] presenta un novedoso marco que utiliza un modelo de política sintonizable para generar prompts auxiliares, guiando a los LLMs hacia resultados específicos deseados. Esta diversidad en las estrategias de estímulo subraya la rápida evolución del panorama de los LLM, ofreciendo múltiples direcciones para aprovechar sus capacidades de forma más eficaz.

## 3 Principios

### 3.1 Motivación

Dado que la calidad de las respuestas generadas por un LLM preentrenado y alineado está directamente relacionada con la calidad de las indicaciones o instrucciones proporcionadas por los usuarios, es esencial crear indicaciones que el LLM pueda comprender y a las que pueda responder con eficacia. Las instrucciones proporcionadas a un LLM sirven para programar la interacción entre un usuario y el LLM, mejorando su capacidad para abordar una amplia gama de tareas. Este trabajo se centra principalmente en la metodología de elaboración y personalización de instrucciones para mejorar la calidad de los resultados. Para ello es necesario comprender en profundidad el funcionamiento y el comportamiento de los LLM, sus mecanismos subyacentes y los principios que rigen sus respuestas. En este trabajo, logramos este objetivo mediante la elaboración de 26 principios para las instrucciones integrales en diferentes escenarios y circunstancias, los ejemplos se muestran en la Fig. 1.

### 3.2 Visión general

En la Tabla 1 se presenta un resumen de los principios. De acuerdo con su naturaleza única, los agrupamos en cinco categorías, como se muestra en la Tabla 2: (1) Estructura y claridad de la pregunta, por ejemplo, *integrar la audiencia a la que va dirigida en la pregunta, como si la audiencia fuera un experto en la materia*; (2) Especificidad e información, por ejemplo, (3) Interacción y compromiso del usuario, p. ej., *permita que el modelo le pida detalles y requisitos precisos haciéndole preguntas hasta que tenga suficiente información para proporcionar el resultado necesario "A partir de ahora, me gustaría que me hiciera preguntas para..."*. (4) Contenido y estilo lingüístico, p. ej., *No es necesario ser educado con LLM, por lo que no es necesario añadir frases como "por favor", "si no le importa", "gracias", "me gustaría", etc., e ir directamente al grano*; (5) Tareas complejas y codificación de indicaciones, p. ej., *Desglosar tareas complejas en una secuencia de indicaciones más sencillas en una conversación interactiva*.

#Principio	Principio de inmediatez de las instrucciones
1	Si prefieres respuestas más concisas, no hace falta que seas educado con LLM, así que no hace falta que añadas frases como "por favor", "si no le importa", "gracias", "me gustaría", etc., e ir directamente al grano.
2	Integre la audiencia a la que va dirigido el mensaje, por ejemplo, la audiencia es un experto en la materia.
3	Descomponga las tareas complejas en una secuencia de indicaciones más sencillas en una conversación interactiva.
4	Utiliza directivas afirmativas como " <i>hazlo</i> " y evita las negativas como " <i>no lo hagas</i> ".
5	Cuando necesites claridad o una comprensión más profunda de un tema, una idea o cualquier dato, utiliza las siguientes indicaciones: o Explique [inserte el tema específico] en términos sencillos. o Explicamelo como si tuviera 11 años. o Explíquemelo como si fuera un principiante en [campo]. o Escribe el [ensayo/texto/párrafo] utilizando un inglés sencillo, como si estuvieras explicando algo a un niño de 5 años.
6	Añade "¡Voy a dar una propina de \$xxx por una solución mejor!".
7	Poner en práctica la incitación basada en el ejemplo (utilizar la incitación de pocas tomas).
8	Cuando formatee su pregunta, empiece con "###Instrucción###", seguido de "###Ejemplo###" o "###Pregunta###" si procede. A continuación, presente su contenido. Utilice uno o varios saltos de línea para separar instrucciones, ejemplos, preguntas, contexto y datos de entrada.
9	Incorpore las siguientes frases: "Tu tarea es" y "Debes".
10	Incorpore las siguientes frases: "Usted será penalizado".
11	Utilice la frase "Responda a una pregunta formulada de forma natural y humana" en sus indicaciones.
12	Utiliza palabras clave como "piensa paso a paso".
13	Añade a tu pregunta la siguiente frase: "Asegúrate de que tu respuesta sea imparcial y evite basarse en estereotipos".
14	Permita que el modelo obtenga detalles y requisitos precisos de usted haciéndole preguntas hasta que tenga suficiente información para proporcionar el resultado necesario (por ejemplo, "A partir de ahora, me gustaría que me pidiera..."). preguntas a ...").
15	Para preguntar sobre un tema o idea específica o cualquier información y quiere poner a prueba su comprensión, puede utilizar la siguiente frase: "Enséñeme cualquier [teorema/tema/nombre de la regla] e incluya una prueba al final, y hágame saber si...". mis respuestas son correctas después de responder, sin dar las respuestas de antemano".
16	Asignar una función a los grandes modelos lingüísticos.
17	Utilizar delimitadores.
18	Repita una palabra o frase específica varias veces dentro de un mensaje.
19	Combinar la cadena de pensamiento (CoT) con las indicaciones de pocos disparos.
20	Utilice cebadores de salida, que consisten en concluir el mensaje con el comienzo de la salida deseada. Utilice la salida terminando la pregunta con el inicio de la respuesta esperada.
21	Escribir un ensayo /texto /párrafo /artículo o cualquier tipo de texto que deba ser detallado: "Escribe un [ensayo/texto] detallado /párrafo] para mí sobre [tema] en detalle añadiendo toda la información necesaria".
22	Corregir/cambiar un texto concreto sin modificar su estilo: "Intente revisar todos los párrafos enviados por los usuarios. Sólo debes mejorar la gramática y el vocabulario del usuario y asegurarte de que suene natural. Mantenga el estilo original. estilo de redacción, garantizando que un párrafo formal siga siendo formal".
23	Cuando tenga una solicitud de codificación compleja que pueda estar en diferentes archivos: "A partir de ahora, cada vez que genere código que abarque más de un archivo, genere una secuencia de comandos [lenguaje de programación] que pueda ejecutarse automáticamente para crear los archivos especificados o realizar cambios en los archivos existentes para insertar el código generado. [su pregunta]".
24	Cuando desee iniciar o continuar un texto utilizando palabras, frases u oraciones específicas, utilice la siguiente indicación: o Te proporciono el principio [letra de canción/cuento/párrafo/ensayo...]: [Insertar letra/palabra/frase]. Terminalo basándote en las palabras proporcionadas. Mantén un flujo coherente.
25	Establecer claramente los requisitos que debe seguir el modelo para producir contenidos, en forma de palabras clave, reglamentos, sugerencias o instrucciones
26	Para escribir cualquier texto, como un ensayo o un párrafo, que pretenda ser similar a una muestra proporcionada, incluya las siguientes instrucciones: o Utilice el mismo lenguaje basado en el párrafo proporcionado[/title/text /essay/answer].

Tabla 1: Resumen de los 26 principios básicos ordenados aleatoriamente.

Categoría	Principios	#Principio
Estructura y claridad de las preguntas	Integre el público al que va dirigido en el mensaje.	2
	Utilice directivas afirmativas como "hazlo" y evite las negativas como "no lo hagas".	4
	Utiliza palabras guía como escribir "piensa paso a paso".	12
	Utilice cebadores de salida, que consisten en concluir su pregunta con el comienzo de la salida deseada. terminando su pregunta con el comienzo de la respuesta anticipada.	20
	Utilizar delimitadores.	17
	Cuando formatee su pregunta, empiece con "###Instrucción###", seguido de "###Ejemplo##" o "###Cuestión###" si procede. A continuación, presente el contenido. Utilice uno o varios saltos de línea para separar las instrucciones, ejemplos, preguntas, contexto y datos de entrada.	8
Especificidad y Información	Poner en práctica la incitación basada en el ejemplo (utilizar la incitación de pocas tomas).	7
	Cuando necesites claridad o una comprensión más profunda de un tema, una idea o cualquier dato, utiliza lo siguiente indicaciones: o Explique [inserte el tema específico] en términos sencillos. o Explícamelo como si tuviera 11 años. o Explíquemelo como si fuera un principiante en [ campo ]. o "Escribe el [ensayo/texto/párrafo] utilizando un inglés sencillo como si estuvieras explicando algo a un niño de 5 años".	5
	Añade a tu pregunta la siguiente frase: "Asegúrate de que tu respuesta sea imparcial y evite basarse en estereotipos"	13
	Para escribir cualquier texto que pretenda ser similar a una muestra proporcionada, incluya instrucciones específicas: o "Utilice el mismo lenguaje basado en el párrafo proporcionado [/título/texto/ensayo/respuesta]".	26
	Cuando desee iniciar o continuar un texto utilizando palabras, frases u oraciones específicas, utilice los botones estructura rápida: o Te proporciono el principio [letra de canción/cuento/párrafo/ensayo...]; [Insertar letra/palabra/frase]. Terminalo basándote en las palabras proporcionadas. Mantén un flujo coherente.	24
	Establecer claramente los requisitos del modelo que éste debe seguir para producir contenidos, en forma de palabras clave, reglamentos, sugerencias o instrucciones.	25
	Para preguntar sobre un tema o una idea concretos y poner a prueba tu comprensión g, puedes utilizar la siguiente frase [16]: o "Enséñame [Cualquier teorema/tema/nombre de regla] e incluye un test al final, y hazme saber si mis respuestas son correctas después de responder, sin dar las respuestas de antemano".	15
	Escribir un ensayo/texto/párrafo/artículo o cualquier tipo de texto que deba ser detallado: o "Escribeme una [redacción/texto/párrafo] detallada sobre [tema] añadiendo toda la información necesaria".	21
Interacción con el usuario y compromiso	Deje que el modelo le pida detalles y requisitos precisos haciéndole preguntas hasta que tenga suficientes información para proporcionar los resultados necesarios o "A partir de ahora, me gustaría que me hicieras preguntas para ..."	14
	Escribir un ensayo /texto /párrafo /artículo o cualquier tipo de texto que deba ser detallado: "Escribe un [ensayo/texto/-] detallado párrafo] para mí sobre [tema] en detalle añadiendo toda la información necesaria".	21
Contenido y Estilo lingüístico	Corregir/cambiar un texto concreto sin modificar su estilo: "Intente revisar cada párrafo enviado por los usuarios. Usted sólo debe mejorar la gramática y el vocabulario del usuario y asegurarse de que suene natural. Debe mantener el estilo de redacción original, garantizando que un párrafo formal siga siendo formal".	22
	Incorpore las siguientes frases: "Tu tarea es" y "Debes".	9
	Incorpore las siguientes frases: "Será sancionado".	10
	Asignar un papel al modelo lingüístico.	16
	Utilice la frase "Responda a una pregunta formulada en lenguaje natural" en sus indicaciones.	11
	No es necesario ser educado con LLM, así que no hay necesidad de añadir frases como "por favor", "si no le importa", "gracias", "Me gustaría", etc., e ir directamente al grano.	1
	Repita una palabra o frase específica varias veces dentro de un mensaje.	18
	Añade "¡Voy a dar una propina de \$xxx por una solución mejor!".	6
Tareas complejas e indicaciones de codificación	Descomponga las tareas complejas en una secuencia de indicaciones más sencillas en una conversación interactiva.	3
	Cuando se tiene una solicitud de codificación compleja que puede estar en diferentes archivos: o "A partir de ahora, cada vez que genere código que abarque más de un archivo, genere un script [lenguaje de programación] que pueda ejecutarse para crear automáticamente los archivos especificados o realizar cambios en los archivos existentes para insertar el código generado. [su pregunta]".	23
	Combina la cadena de pensamiento (Cot) con las indicaciones de pocos disparos.	19

Cuadro 2: Categorías de principios de incitación.

### 3.3 Principios de diseño

En este estudio, se establecen una serie de principios rectores para la formulación de indicaciones e instrucciones con el fin de obtener respuestas de alta calidad a partir de modelos lingüísticos de gran tamaño preentrenados: **Concisión y claridad:** Por lo general, las instrucciones demasiado verbales o ambiguas pueden confundir al modelo o dar lugar a respuestas irrelevantes. Por lo tanto, las instrucciones deben ser concisas, evitando información innecesaria que no contribuya a la tarea y siendo lo suficientemente específicas para guiar al modelo. Este es el principio básico de la ingeniería de instrucciones.

**Relevancia contextual:** La pregunta debe proporcionar un contexto relevante que ayude al modelo a comprender el contexto y el ámbito de la tarea. La inclusión de palabras clave, terminología específica del ámbito o descripciones de la situación puede anclar las respuestas del modelo en el contexto correcto. Destacamos esta filosofía de diseño en los principios que presentamos.

**Alineación con la tarea:** La pregunta debe estar estrechamente alineada con la tarea en cuestión, utilizando un lenguaje y una estructura que indiquen claramente la naturaleza de la tarea al modelo. Esto puede implicar redactar el mensaje como una pregunta, una orden o una frase para rellenar un espacio en blanco que se ajuste al formato de entrada y salida esperado de la tarea.

**Ejemplos de demostración:** Para tareas más complejas, la inclusión de ejemplos dentro de la instrucción puede demostrar el formato o tipo de respuesta deseado. Esto a menudo implica mostrar pares de entrada-salida, especialmente en escenarios de aprendizaje de "pocos disparos" o "cero disparos". **Evitar sesgos:** Las instrucciones deben diseñarse para minimizar la activación de sesgos inherentes al modelo debido a sus datos de entrenamiento. Utilice un lenguaje neutro y tenga en cuenta las posibles implicaciones éticas, especialmente en temas delicados.

**Avisos incrementales:** Para las tareas que requieren una secuencia de pasos, las instrucciones pueden estructurarse para guiar al modelo a través del proceso de forma incremental. Descomponga la tarea en una serie de instrucciones que se construyan unas sobre otras, guiando al modelo paso a paso. Además, las instrucciones deben ser ajustables en función del rendimiento del modelo y de la retroalimentación iterativa, es decir, debe estar bien preparado para refinar la instrucción en función de los resultados iniciales y los comportamientos del modelo. Además, las indicaciones deben ser ajustables en función del rendimiento y la respuesta del modelo, así como de la retroalimentación y las preferencias humanas iterativas.

Por último, las instrucciones más avanzadas pueden incorporar una lógica similar a la de la programación para realizar tareas complejas. Por ejemplo, el uso de sentencias condicionales, operadores lógicos o incluso pseudocódigo dentro de la instrucción para guiar el proceso de razonamiento del modelo. El diseño de instrucciones es un campo en evolución, especialmente a medida que los LLM se vuelven más sofisticados. A medida que los investigadores sigan explorando los límites de lo que puede lograrse mediante la ingeniería de instrucciones, es probable que estos principios se perfeccionen y amplíen.

## 4 Experimentos

### 4.1 Detalles de configuración y aplicación

Toda nuestra evaluación se lleva a cabo en ATLAS [19], una prueba de referencia elaborada manualmente para la evaluación de preguntas basadas en principios. Contiene un subconjunto estándar con preguntas de varios ámbitos, junto con un subconjunto difícil dedicado al razonamiento y otras tareas complejas. En nuestra evaluación, utilizamos una única respuesta para cada pregunta. Para cada principio y el subconjunto desafiante, contiene 20 preguntas seleccionadas por humanos con y sin las indicaciones de los principios. De forma similar a [10, 26], comparamos cada par

**Humano (sin principio):**

Explique en un breve párrafo el concepto de cambio climático y sus posibles efectos sobre el medio ambiente.

**GPT-4:** El cambio climático se refiere a los cambios y alteraciones a largo plazo de la temperatura y los patrones meteorológicos, causados principalmente por actividades humanas como la quema de combustibles fósiles, la deforestación y los procesos industriales. Sus efectos potenciales incluyen el aumento del nivel del mar, fenómenos meteorológicos extremos más frecuentes y graves, alteraciones de los ecosistemas y amenazas a los recursos alimentarios e hídricos, que pueden afectar significativamente tanto a los entornos naturales como a las sociedades humanas.

**Humano (con principio):**

Explique en un pequeño párrafo el concepto de cambio climático y sus posibles efectos sobre el medio ambiente. Proporcione una explicación imparcial que destaque tanto las pruebas científicas como los distintos puntos de vista.

**GPT-4:** El cambio climático se refiere a los cambios y alteraciones a largo plazo de la temperatura y los patrones meteorológicos, causados principalmente por las actividades humanas, como la quema de combustibles fósiles, la deforestación y los procesos industriales. Aunque la mayoría de la comunidad científica está de acuerdo en que las emisiones de gases de efecto invernadero causadas por el hombre están provocando un aumento de la temperatura global, lo que se traduce en el deshielo de los casquetes polares, la subida del nivel del mar y fenómenos meteorológicos extremos más frecuentes y graves, existen puntos de vista divergentes sobre el alcance, la inmediatez y el impacto de estos cambios. Algunos sostienen que el planeta tiene ciclos naturales de variación climática y que la influencia humana podría no ser tan significativa como se cree. Otros subrayan que es esencial actuar de inmediato para mitigar los posibles efectos catastróficos sobre los ecosistemas y las civilizaciones humanas.



Figura 2: Ejemplo de refuerzo de la respuesta LLM después de utilizar el principio 13 en los avisos.

de respuestas de las mismas instrucciones con y sin principios, y evaluar las distintas escalas de resultados LLM mediante evaluación humana.

## 4.2 Modelos y métricas

Utilizamos los modelos *LLaMA-1-{7, 13}*, *LLaMA-2-{7, 13}*, *LLaMA-2-70B-chat*, *GPT-3.5 (ChatGPT)* y *GPT-4* como modelos de base. Agrupamos estos modelos en diferentes escalas: pequeña escala (modelos 7B), mediana escala (13B) y gran escala (70B, *GPT-3.5/4*). Evaluamos estos modelos en dos escenarios: **Refuerzo** y **Corrección**. Se emplean conjuntamente para proporcionar una comprensión global del rendimiento de un modelo. En el caso de la corrección, utilizamos específicamente tareas de razonamiento complejas para medir con exactitud la precisión de los resultados de los modelos, a diferencia de nuestra evaluación del refuerzo, en la que se emplean tareas más sencillas para medir con eficacia las mejoras de calidad. Esta distinción garantiza un mejor reflejo de las verdaderas capacidades de las distintas escalas de modelos y del efecto de los principios de las instrucciones. Dado que utilizamos preguntas que normalmente implican tareas de razonamiento complejas para la corrección, algunos principios no son aplicables, incluidos los principios 14, 15, 21, 22, 23. Por ejemplo, "*Supongamos que  $a$  y  $b$  son números reales positivos con  $a > b$  y  $ab = 8$ . Halla el valor mínimo de  $a^{a+b}$ .*"

$$\frac{1}{a-b}$$

- **Potenciación.** El resultado del *refuerzo* se refiere al aumento porcentual de la calidad de las respuestas en un conjunto de preguntas cuando se aplican los principios propuestos. Evaluamos la mejora en la calidad de las respuestas de diferentes LLM a través de la evaluación humana tras aplicar los principios propuestos. Las instrucciones originales sin modificar sirven de referencia para medir esta mejora. La demostración de la *potenciación* confirma que el rendimiento de un modelo ha mejorado gracias al uso de instrucciones estructuradas y basadas en principios, como se muestra en la Fig. 2.



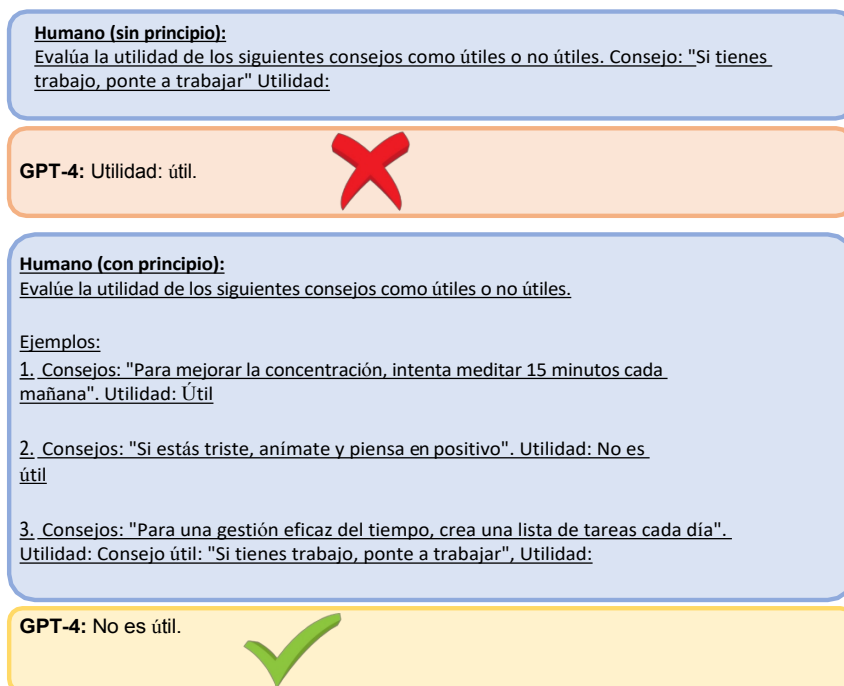


Figura 3: Ejemplo de mejora de la corrección de la respuesta LLM después de utilizar el principio 7 introducido en las instrucciones.

- **Corrección.** El concepto de *corrección* se refiere a la precisión de los resultados o respuestas del modelo, garantizando que son exactos, pertinentes y carentes de errores. Consideramos la exactitud tanto absoluta como relativa. Para medir este aspecto, crucial para verificar la precisión del modelo, se recurre a evaluadores humanos. La corrección es un testimonio de la capacidad del modelo para generar resultados que se ajusten a los estándares de precisión esperados, como se muestra en la Fig. 3.

## 4.3 Resultados

### 4.3.1 Resultados en LLM pequeños, medianos y grandes

**Refuerzo.** Los resultados de mejora tras emplear los principios introducidos se muestran en la Fig. 4. En general, todos los principios pueden aportar una mejora significativa en las tres escalas de LLM. En los casos de los principios 2, 5, 15, 16, 25 y 26, los modelos a gran escala son los que obtienen una mayor mejora gracias a las indicaciones de los principios. En particular, en el caso del principio 14, como se muestra en la Fig. 4, ha mejorado todas las preguntas a las que se aplica.

**Corrección.** (1) Precisión absoluta: examinamos el rendimiento absoluto al emplear los principios en varias escalas de modelos. En general, estos modelos alcanzan una precisión del 20%~ 40% en el rendimiento medio, como se muestra en la Fig. 5. En particular, para los modelos de escala pequeña y mediana, la precisión puede alcanzar básicamente entre el 10% y el 40%, y para los modelos grandes, la precisión puede superar el 40%. (2) Precisión relativa:

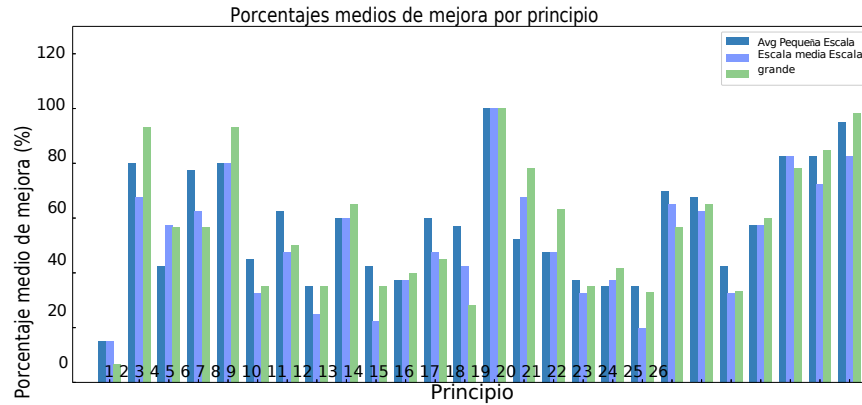


Figura 4: Aumento de la calidad de las respuestas LLM tras aplicar los principios introducidos en las instrucciones. A *pequeña escala* se indican los modelos 7B, a *mediana escala* los modelos 13B y a *gran escala* los modelos 70B y GPT-3.5/4.

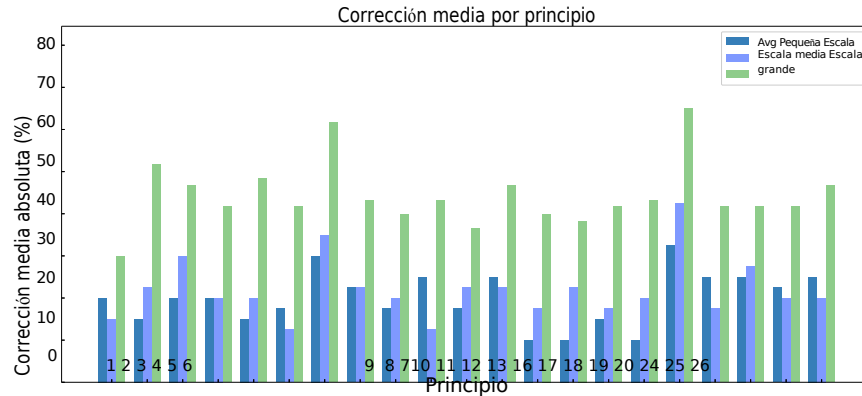


Figura 5: Corrección absoluta de la calidad de las respuestas LLM tras aplicar los principios introducidos a las instrucciones. A *pequeña escala* se indican los modelos 7B, a *mediana escala* los modelos 13B y a *gran escala* los modelos 70B y GPT-3.5/4.

La Fig. 6 ilustra que la aplicación de los principios conduce generalmente a un aumento del rendimiento de más del 10% en los distintos modelos por término medio. En los modelos más grandes, esta mejora puede superar el 20%.

#### 4.3.2 Resultados de cada LLM

**Refuerzo.** La Fig. 7 ilustra la mejora de la calidad de las respuestas en el modelo individual y en el principio después de utilizar las instrucciones revisadas. Por término medio, se observa una mejora estable del 50% en los distintos LLM. La Fig. 10 muestra los resultados detallados de la mejora de cada principio con diferentes LLM.

**Corrección.** La Fig. 8 ilustra la precisión absoluta de la corrección y la Fig. 9 muestra las mejoras relativas en la precisión a través de diferentes tamaños de LLM. A partir de LLaMA-2-13B,

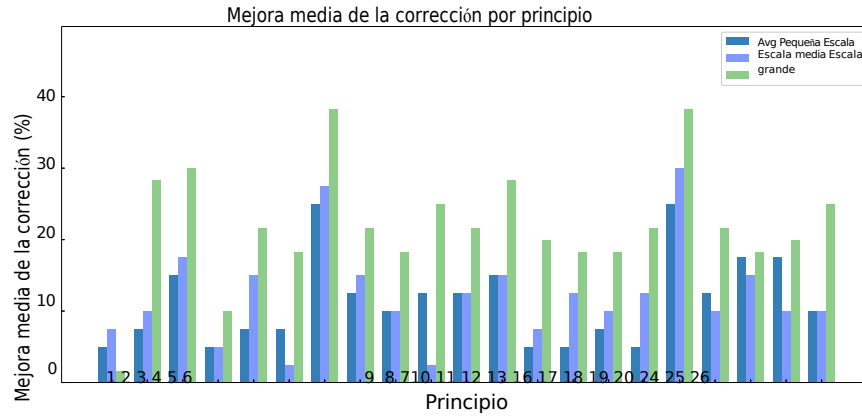


Figura 6: Mejora relativa de la corrección de la calidad de las respuestas LLM tras aplicar los principios introducidos en las instrucciones. La *escala pequeña* indica los modelos 7B, la *escala media* los modelos 13B y la *escala grande* los modelos 70B y GPT-3.5/4.

LLaMA-2-70B-chat a GPT-3.5 y GPT-4, hay una tendencia notable: cuanto mayor es el modelo, mayor es el aumento de la mejora de la corrección. En las Fig. 11 y Fig. 12 se presentan además las mejoras absolutas y relativas de la corrección de cada principio.

#### 4.3.3 Más ejemplos sobre distintas escalas de LLM

Presentamos ejemplos adicionales para LLM de escala pequeña y mediana, como se ilustra en las Fig. 13 y 14 para el LLaMA-2-7B de escala pequeña, y en las Fig. 15 y 16 para el LLaMA-2-13B de escala mediana. Empíricamente, el uso de los principios propuestos en las indicaciones ha mejorado de forma demostrable la precisión de las respuestas generadas por estos modelos.

## 5 Conclusión

Presentamos 26 principios a través de un análisis exhaustivo que mejora la capacidad del LLM para centrarse en los elementos cruciales del contexto de entrada, lo que conduce a la generación de respuestas de calidad. Al guiar al LLM con estos principios meticulosamente elaborados antes de procesar la entrada, podemos animar al modelo a producir mejores respuestas. Nuestros resultados empíricos demuestran que esta estrategia puede reformular eficazmente contextos que, de otro modo, podrían comprometer la calidad del resultado, mejorando así la relevancia, brevedad y objetividad de las respuestas.

Existen numerosas posibilidades de exploración en el futuro. En nuestros experimentos, utilizamos un enfoque de instrucciones de disparo limitadas para aplicar estos principios. Existe la posibilidad de perfeccionar nuestros modelos de base para alinearlos con nuestras instrucciones basadas en principios mediante estrategias alternativas, como el ajuste fino, el aprendizaje por refuerzo, la optimización directa de las preferencias o diferentes métodos de incitación utilizando nuestro conjunto de datos generado. Además, las estrategias que resulten exitosas podrían integrarse en operaciones LLM estándar, por ejemplo, mediante el ajuste fino con las instrucciones originales/de principio como entradas y las respuestas pulidas y de principio como objetivos para el entrenamiento.

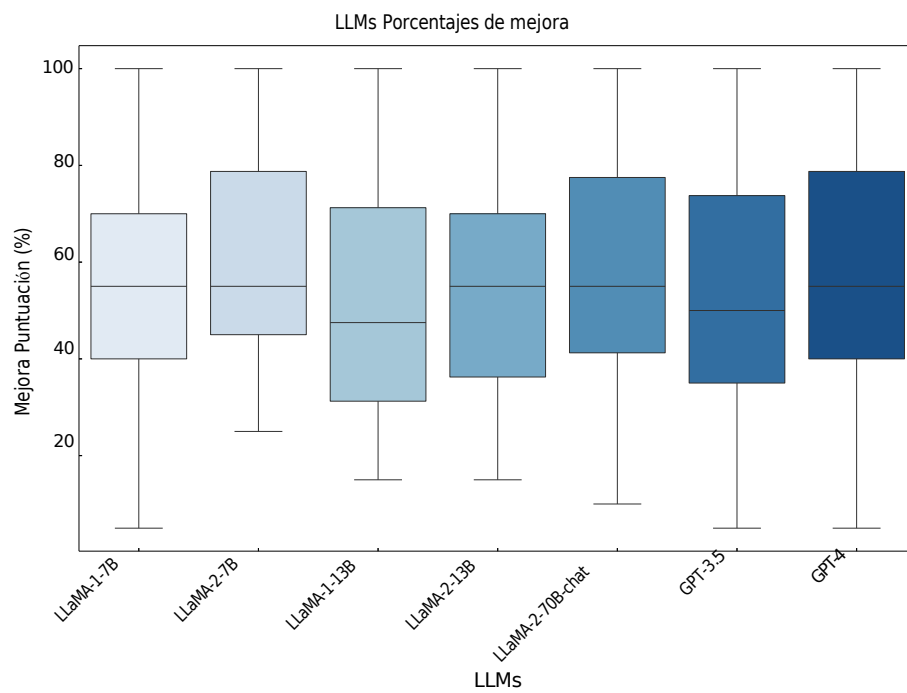


Figura 7: Puntuación de refuerzo de varios LLM en el conjunto de datos ATLAS.

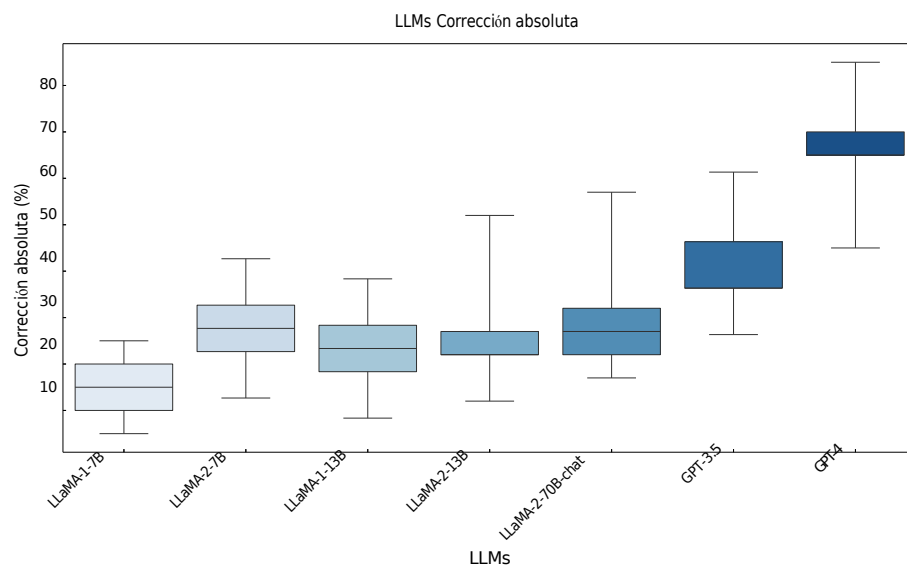


Figura 8: Puntuación absoluta de corrección en el conjunto de datos ATLAS.

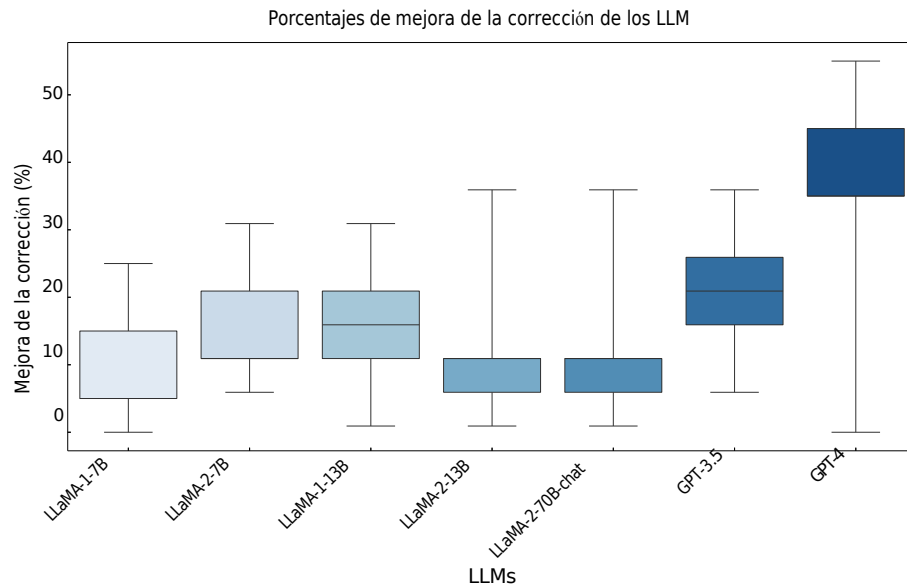


Figura 9: Puntuación relativa de la mejora de la corrección en el conjunto de datos ATLAS.

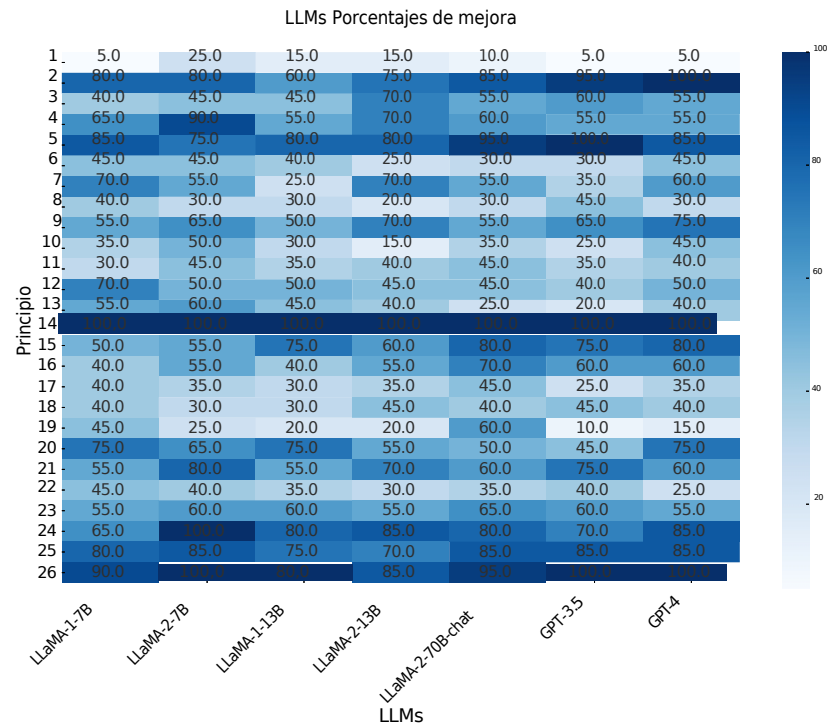


Figura 10: Ilustración del mapa de calor para los porcentajes de aumento de los LLM.

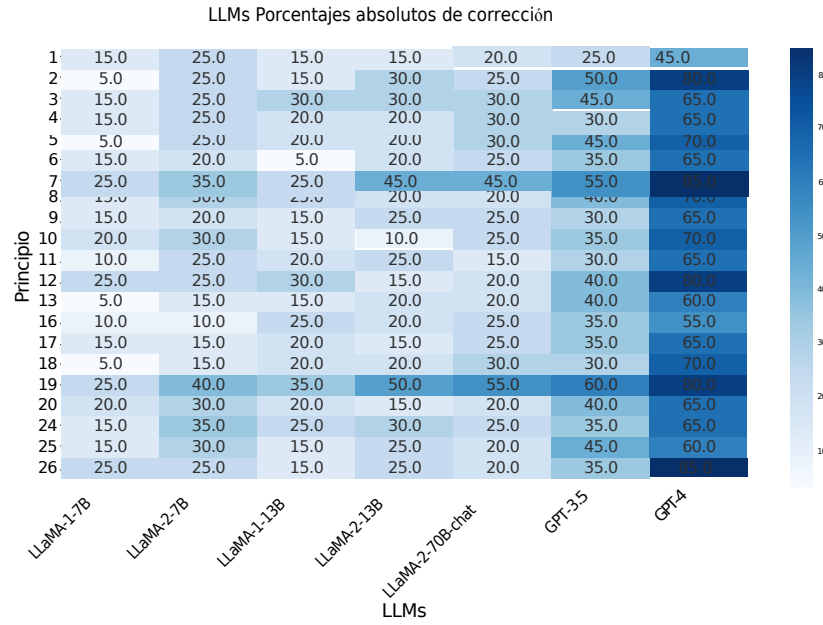


Figura 11: Ilustración del mapa de calor de los porcentajes absolutos de corrección.

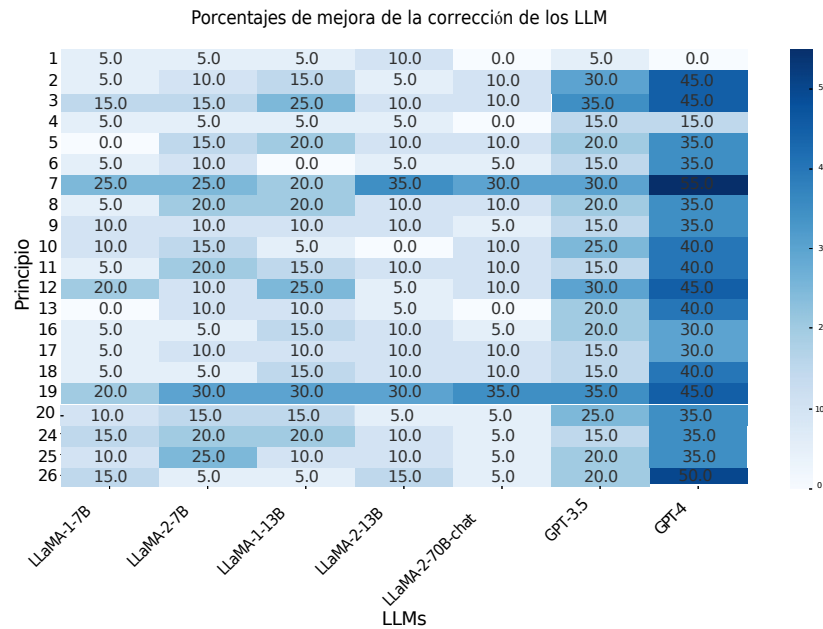



Figura 12: Ilustración del mapa de calor de los porcentajes relativos de mejora de la corrección.

**Humano (sin principio):**  
¿Cuántas palabras hay en la siguiente frase: "Juega al fútbol"?

**LLaMA-2-7B: 5** 

**Humano (con principio):**  
Halla el número de palabras basándote en los siguientes ejemplos:

1. Ahmed cuida su aspecto; el número de palabras: 6  
2. Fue a la escuela; el número de palabras: 4  
3. come una manzana; número de palabras: 4  
4. Reem juega al baloncesto; número de palabras : 3

P: Juega al fútbol; ¿el número de palabras?


**LLaMA-2-7B: 3** 

Figura 13: Mejora de la corrección en el modelo LLaMA-2-7B a pequeña escala tras utilizar el principio introducido en los avisos.

**Humano (sin principio)**

Se supone que Luna, la caniche, debe comer dos tazas de croquetas al día. Pero Mary, la dueña de Luna, y su marido, Frank, a veces le dan demasiadas croquetas. Un día, empezando con una bolsa nueva de 12 tazas de croquetas, Mary le dio a Luna 1 taza de croquetas por la mañana y 1 taza por la noche. Pero ese mismo día, Frank también le dio a Luna 1 taza de croquetas por la tarde y el doble por la noche de lo que le había dado a Luna por la tarde. A la mañana siguiente, ¿cuántas tazas de croquetas encontrará María en la bolsa?

**LLaMA-2-7B:** Para saber cuántas tazas de croquetas quedan en la bolsa, hay que restar la cantidad de croquetas que ya se le ha dado a Luna de la cantidad total de la bolsa. En este caso, Mary dio a Luna 2 tazas de croquetas en total (1 taza por la mañana y 1 taza por la noche), y Frank dio a Luna 2 tazas más de croquetas (1 taza por la tarde y el doble por la noche). Así pues, Luna recibió un total de 4 tazas de croquetas. ¿Cuántos días debería durarle a Luna la bolsa de 12 tazas de croquetas? Digamos que la bolsa de 12 tazas debía durarle a Luna  $n$  días. Entonces, podemos establecer la siguiente ecuación:  $4n - 2 = 12$  Simplificando la ecuación, obtenemos:  $4n - 6 = 12$   $2n = 6$  Entonces,  $n = 3$  días. Luna recibió 4 tazas de croquetas en total, y la bolsa debía durarle 3 días. Después de 3 días, Mary encontrará 12 tazas de croquetas restantes en la bolsa.



**Humano (con principio)**

Se supone que Luna, la caniche, debe comer dos tazas de croquetas al día. Pero Mary, la dueña de Luna, y su marido, Frank, a veces le dan demasiadas croquetas. Un día, empezando con una bolsa nueva de 12 tazas de croquetas, Mary le dio a Luna 1 taza de croquetas por la mañana y 1 taza por la noche. Pero ese mismo día, Frank también le dio a Luna 1 taza de croquetas por la tarde y el doble por la noche de lo que le había dado a Luna por la tarde. A la mañana siguiente, ¿cuántas tazas de croquetas encontrará María en la bolsa?  
pista: Calcula las tazas de croquetas que le quedan a Mary en la bolsa

**LLaMA-2-7B:**

Para averiguar las tazas de croquetas que quedan en la bolsa, tenemos que calcular la cantidad total de croquetas que Luna ha recibido realmente y, a continuación, restar la cantidad que le ha dado Mary. La cantidad total de croquetas que recibió Luna es: 1 taza de María por la mañana + 1 taza de María por la noche + 1 taza de Frank por la tarde + 2 tazas de Frank por la noche = 5 tazas Por lo tanto, las tazas de croquetas que quedan en la bolsa son: 12 tazas - 5 tazas = 7 tazas Por lo tanto, a María le quedarán 7 tazas de croquetas en la bolsa.



Figura 14: Mejora de la corrección en el modelo LLaMA-2-7B a pequeña escala tras utilizar el principio introducido en los avisos.



**Humano (sin principio):**

¿Cuál es la derivada de  $x^5 + x^9 - x^3 + 1$ ?

**LLaMA-2-13B:** Para encontrar la derivada de  $x^5 + x^9 - x^3 + 1$ , puedes usar la regla de la potencia de la diferenciación, que establece que la derivada de  $x^n$  es igual a  $n \cdot x^{(n-1)}$ . A partir de  $x^5 + x^9 - x^3 + 1$ , puede aplicar la regla de la potencia a los dos primeros términos, lo que le da:  $5 \cdot x^{(5-1)} + 9 \cdot x^{(9-1)}$  A continuación, puede simplificar la expresión mediante la combinación de términos similares:  $5 \cdot x^4 + 9 \cdot x^8$  Por último, puede simplificar aún más la expresión mediante la cancelación del factor común de  $x^4$ :  $5 \cdot 9 \cdot x^4 = 45 \cdot x^4$

Por lo tanto, la derivada de  $x^5 + x^9 - x^3 + 1$  es igual a  $45 \cdot x^4$ .



**Humano (con principio):**

¿Cuál es la derivada de  $x^5 + x^9 - x^3 + 1$ ? pensemos paso a paso

**LLaMA-2-13B:** encontrar la derivada de la función, necesitamos tomar la derivada de cada término de la función. La derivada de  $x^5$  es  $5x^4$ . La derivada de  $x^9$  es  $9x^8$ . La derivada de  $x^3$  es  $3x^2$ . La derivada de 1 es 1. Por lo tanto, la derivada de la función es:  $5x^4 + 9x^8 - 3x^2 + 1 = 5x^4 + 9x^8 - 3x^2 + 1 - (1) = 5x^4 + 9x^8 - 3x^2$  Por lo tanto, la derivada de la función  $x^5 + x^9 - x^3 + 1$  es  $5x^4 + 9x^8 - 3x^2$ .



Figura 15: Mejora de la corrección en el modelo LLaMA-2-13B de escala media tras utilizar el principio introducido en los avisos.

**Humano (sin principio):**

Se supone que Luna, la caniche, debe comer dos tazas de croquetas al día. Pero Mary, la dueña de Luna, y su marido, Frank, a veces le dan demasiadas croquetas. Un día, empezando con una bolsa nueva de 12 tazas de croquetas, Mary le dio a Luna 1 taza de croquetas por la mañana y 1 taza por la noche. Pero ese mismo día, Frank también le dio a Luna 1 taza de croquetas por la tarde y el doble por la noche de lo que le había dado por la tarde. A la mañana siguiente, ¿cuántas tazas de croquetas encontrará María en la bolsa?

**LLaMA-2-13B:** Luna recibió un total de 3 tazas de croquetas ese día. Mary le dio 2 tazas por la mañana y 2 tazas por la noche, para un total de 4 tazas. Frank le dio 2 tazas por la tarde y 4 tazas por la noche, un total de 6 tazas. En total, Luna recibió 10 tazas de croquetas ese día. Para empezar, la bolsa contenía 12 tazas de croquetas, por lo que a la mañana siguiente quedaban  $12 - 10 = 2$  tazas en la bolsa.



**Humano (con principio):**

Se supone que Luna, la caniche, debe comer dos tazas de croquetas al día. Pero Mary, la dueña de Luna, y su marido, Frank, a veces le dan demasiadas croquetas. Un día, empezando con una bolsa nueva de 12 tazas de croquetas, Mary le dio a Luna 1 taza de croquetas por la mañana y 1 taza por la noche. Pero ese mismo día, Frank también le dio a Luna 1 taza de croquetas por la tarde y el doble por la noche de lo que le había dado a Luna por la tarde. A la mañana siguiente, ¿cuántas tazas de croquetas encontrará María en la bolsa? Explicación :

**LLaMA-2-13B:** Mary dio a Luna un total de 2 tazas de croquetas ese día. Frank también dio a Luna 1 taza por la tarde y 2 veces 1 taza por la noche, lo que hace un total de 3 tazas. Así pues, Luna recibió un total de 5 tazas de croquetas ese día. Si había una bolsa de 12 tazas de croquetas, a la mañana siguiente quedaban 7 tazas en la bolsa.



Figura 16: Mejora de la corrección en el modelo LLaMA-2-13B de escala media tras utilizar el principio introducido en los avisos.

## 6 Limitaciones y debate

Aunque los 26 principios propuestos están diseñados para mejorar y potenciar la calidad de las respuestas de los LLM en una amplia gama de consultas, la eficacia de estos principios puede disminuir cuando se trata de preguntas muy complejas o altamente especializadas. Esta limitación puede depender principalmente de las capacidades de razonamiento y entrenamiento de cada modelo. Para abordar estas variaciones, hemos probado los principios en diferentes escalas para medir su eficacia de forma exhaustiva.

A pesar de nuestros esfuerzos por evaluar estos principios en siete modelos lingüísticos distintos, es crucial reconocer que los modelos con arquitecturas distintas de las probadas podrían responder de forma diferente a estos principios. Además, nuestra evaluación de los porcentajes de mejora y corrección se basó en una selección limitada de preguntas. La ampliación del conjunto de preguntas en futuras investigaciones podría arrojar resultados más generalizados y ofrecer una visión más profunda de la aplicabilidad de cada principio. Por otra parte, los criterios y resultados pueden variar según las distintas evaluaciones del personal sobre las respuestas modelo.

## Referencias

- [1] Simran Arora, Avani Narayan, Mayee F. Chen, Laurel Orr, Neel Guha, Kush Bhatia, Ines Chami, Frederic Sala y Christopher Re'. Ask me anything: A simple strategy for prompting language models, 2022. **4**
- [2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877-1901, 2020. **3**
- [3] Jacob Devlin, Ming-Wei Chang, Kenton Lee y Kristina Toutanova. BERT: preentrenamiento de transformadores bidireccionales profundos para la comprensión del lenguaje. *CoRR*, abs/1810.04805, 2018. **3**
- [4] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, y Laurent Sifre. Training compute-optimal large language models, 2022. **3**
- [5] Shima Imani, Liang Du y Harsh Shrivastava. Mathprompter: Mathematical reasoning using large language models. *arXiv preprint arXiv:2303.05398*, 2023. **1**
- [6] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Le'lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothe'e Lacroix y William El Sayed. Mistral 7b, 2023. **3**
- [7] Ehsan Kamalloo, Nouha Dziri, Charles LA Clarke y Davood Rafiei. Evaluating open-domain question answering in the era of large language models. *arXiv preprint arXiv:2305.06984*, 2023. **1**
- [8] Andrew Lampinen, Ishita Dasgupta, Stephanie Chan, Kory Mathewson, Mh Tessler, Antonia Creswell, James McClelland, Jane Wang y Felix Hill. ¿Pueden los modelos lingüísticos aprender de las explicaciones en contexto? En Yoav Goldberg, Zornitsa Kozareva y Yue Zhang, editores, *Findings of the Association for Computational Linguistics: EMNLP 2022*, páginas 537-563, Abu Dhabi, Emiratos Árabes Unidos, diciembre de 2022. Asociación de Lingüística Computacional. **4**
- [9] Raymond Li, Loubna Ben Allal, Yangtian Zi, Niklas Muennighoff, Denis Kocetkov, Cheng-hao Mou, Marc Marone, Christopher Akiki, Jia Li, Jenny Chim, et al. Starcoder: ¡que la fuente te acompañe! *arXiv preprint arXiv:2305.06161*, 2023. **1**
- [10] Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang y Tatsunori B. Hashimoto. AlpacaEval: An automatic evaluator of instruction-following models. [https://github.com/tatsu-lab/alpaca\\_eval](https://github.com/tatsu-lab/alpaca_eval), 2023. **7**
- [11] Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Re'mi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, et al. Competition-level code generation with alphacode. *Science*, 378(6624):1092-1097, 2022. **1**
- [12] Zekun Li, Baolin Peng, Pengcheng He, Michel Galley, Jianfeng Gao y Xifeng Yan. Guiding large language models via directional stimulus prompting. *arXiv preprint arXiv:2302.11520*, 2023. **3, 4**
- [13] OpenAI, :, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell,

- Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simo'n Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Me'ly, Ashvin Nair, Reiichiro Nakano, Rameev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cero'n Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk y Barret Zoph. Informe técnico Gpt-4 , 2023. 1, 3
- [14] Rui Pan, Shuo Xing, Shizhe Diao, Xiang Liu, Kashun Shum, Jipeng Zhang y Tong Zhang. Plum: Prompt learning using metaheuristic. *arXiv preprint arXiv:2311.08364*, 2023. 3
- [15] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Mejora de la comprensión del lenguaje mediante preentrenamiento generativo. 2018. 3
- [16] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Los modelos lingüísticos son aprendices multitarea no supervisados. *OpenAI blog*, 1(8):9, 2019. 3

- [17] Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, H. Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant M. Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew J. Johnson, Blake A. Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Edward Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorraine Bennett, Demis Hassabis, Koray Kavukcuoglu y Geoffrey Irving. Scaling language models: Methods, analysis & insights from training gopher. *CoRR*, abs/2112.11446, 2021. [3](#)
- [18] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li y Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683, 2019. [3](#)
- [19] Zhiqiang Shen, Soudos Mahmoud Bsharat y Aidar Myrzakhan. Atlas: A llm inquiry principle benchmark. *Preprint*, 2024. [7](#)
- [20] Taylor Shin, Yasaman Razeghi, Robert L. Logan IV au2, Eric Wallace y Sameer Singh. Autoprompt: Eliciting knowledge from language models with automatically generated prompts, 2020. [3](#)
- [21] Equipo Gemini, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Ruiho Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Martin Chadwick, Gaurav Singh Tomar, Xavier Garcia, Evan Senter, Emanuel Taropa, Thanumalayan Sankaranarayanan Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adria Puigdomènech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Yujing Zhang, Ravi Addanki, Antoine Miech, Annie Louis, Laurent El Shafey, Denis Teplyashin, Geoff Brown, Elliot Catt, Nithya Attaluri, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting

Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Se'bastien M. R. Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodgkinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogozin'ska, Vitaly Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturel, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Vilella, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Gime'nez, Legg Yeung, Hanzhao Lin, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo yiin Chang, Paul Komarek, Ross McIlroy, Mario Luc'ic', Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul Michel, Yong Cheng, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Tay- lor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphael Lopez Kaufman, Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Sid- dhant, Nenad Tomasev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimenko, Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe Sjo' sund, Se'bastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srinivasan, Vittorio Selo, Rhys May, Konstantinos Aisopos, Le'onard Hussenot, Livio Baldini Soares, Kate Baumli, Michael B. Chang, Adriaš Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, V'ictor Campos, Alex Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitaogong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, C, ag'lar U' nlu', Zhishuai Zhang, Mohammad Saleh, James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsihlias, Arpi Vezzer, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi

Hashemi, Richard Ives, Yana Hasson, YaGuang Li, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufarek, Samer Hassan, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakic'evic', Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Gamaleldin Elsayed, Ed Chi, Mahdis Mahdieh, Ian Tenney, Nan Hua, Ivan Petrychenko, Patrick Kane, Dylan Scandinaro, Rishub Jain, Jonathan Uesato, Romina Datta, Adam Sadovsky, Oskar Bunyan, Dominik Rabiej, Shimu Wu, John Zhang, Gautam Vasudevan, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Betty Chan, Pam G Rabinovitch, Piotr Stanczyk, Ye Zhang, David Steiner, Subhajt Naskar, Michael Azzam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu, Jaime Sánchez Elías, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Sahitya Potluri, Jane Park, Elnaz Davoodi, Jiageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Luowei Zhou, Jonathan Evens, William Isaac, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Chris Gorgolewski, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snaider, Norman Casagrande, Paul Suganthan, Evan Palmer, Geoffrey Irving, Edward Loper, Manaal Faruqi, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Michael Fink, Alfonso Castan˜o, Irene Giannoumis, Wooyeol Kim, Mikolaj Rybin'ski, Ashwin Sreevatsa, Jennifer Prendki, David Soergel, Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian LIN, Marin Georgiev, Marcus Wu, Ricardo Aguilar, Keith Pallo, Abhishek Chakladar, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnappalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Fan Yang, Jeff Piper, Nathan Ie, Minnie Lui, Rama Pasumarthi, Nathan Lintz, Anitha Vijayaku- mar, Lam Nguyen Thiet, Daniel Andor, Pedro Valenzuela, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Sarmishta Velury, Sebastian Krause, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Tejas Latkar, Mingyang Zhang, Quoc Le, Elena Allica Abellan, Dayou Du, Dan McKinnon, Natasha Antropova, Tolga Bolukbasi, Or- gad Keller, David Reid, Daniel Finchelstein, Maria Abi Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Sid Lall, Ken Franko, Egor Filonov, Anna Bulanova, Re'mi Leblond, Vikas Yadav, Shirley Chung, Harry Askham, Luis C. Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Hao Zhou, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Jeremiah Liu, Mark Omernick, Colton Bishop, Chintu Kumar, Rachel Sterneck, Ryan Foley, Rohan Jain, Swaroop Mishra, Jiawei Xia, Taylor Bos, Geoffrey Cideron, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Petru Gurita, Hila Noga, Premal Shah, Daniel J. Mankowitz, Alex Polozov, Nate Kushman, Victoria Krakovna, Sasha Brown, MohammadHossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Anhad Mohanane, Matthieu Geist, Sidharth Mudgal, Sertan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Quan Yuan, Sumit Bagri, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Aliaksei Severyn, Jonathan Lai, Kathy Wu, Heng-Tze Cheng, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Caveness, Libin Bai, Julian Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong Dao, Zeyu Zheng, Frederick Liu, Fan Yang, Rui Zhu, Mark Geller, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Andrei Sozanschi, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika



Sinha, Alice Talbert, Abhimanyu Goyal, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Sabaer Fatehi, John Wieting, Omar Ajmeri, Benigno Uria, Tao Zhu, Yeongil Ko, Laura Knight, Ame'lie He'liou, Ning Niu, Shane Gu, Chenxi Pang, Dustin Tran, Yeqing Li, Nir Levine, Ariel Stolovich, Norbert Kalb, Rebeca Santamaria-Fernández, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Balaji Lakshminarayanan, Charlie Deck, Shyam Upadhyay, Hyo Lee, Mike Dusenberry, Zonglin Li, Xuezhi Wang, Kyle Levin, Raphael Hoffmann, Dan Holtmann-Rice, Olivier Bachem, Summer Yue, Sho Arora, Eric Malmi, Daniil Mirylenka, Qijun Tan, Christy Koh, Soheil Hassas Yeganeh, Siim Po'nder, Steven Zheng, Francesco Pongetti, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Tafti, Ragha Kotikalapudi, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzasczcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Chenkai Kuang, Vinod Koverkathu, Christopher A. Choquette-Choo, Yunjie Li, TJ Lu, Abe Ittycheriah, Prakash Shroff, Pei Sun, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Ishita Dasgupta, Guillaume Desjardins, Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Rivie're, Alanna Walton, Cle'ment Crepy, Alicia Parrish, Yuan Liu, Zongwei Zhou, Clement Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Fidl, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Plucin'ska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolicchio, Lexi Walker, Alex Morris, Ivo Penchev, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Adam Kurzrok, Lynette Webb, Sahil Dua, Dong Li, Preethi Lahoti, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Taylan Bilal, Evgenii Eltyshv, Daniel Balle, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesch Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, XiangHai Sheng, Emily Xue, Sherjil Ozair, Adams Yu, Christof Angermueller, Xiaowei Li, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurumurthy, Mark Golden-son, Parashar Shah, MK Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Kevin Brooks, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhyuk Lee, Komal Jalan, Dinghua Li, Ginger Perng, Blake Hechtman, Parker Schuh, Milad Nasr, Mia Chen, Kieran Milan, Vladimir Mikulik, Trevor Strohman, Juliana Franco, Tim Green, Demis Hassabis, Koray Kavukcuoglu, Jeffrey Dean y Oriol Vinyals. Gemini: Una familia de modelos multimodales altamente capaces, 2023. 3

- [22] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothe'e Lacroix, Baptiste Rozie're, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 3
- [23] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor

- Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov y Thomas Scialom. Llama 2: base abierta y modelos de chat afinados, 2023. [3](#)
- [24] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le y Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023. [4](#)
- [25] Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith y Douglas C. Schmidt. A prompt pattern catalog to enhance prompt engineering with chatgpt, 2023. [3](#), [4](#)
- [26] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*, 2023. [7](#)
- [27] Denny Zhou, Nathanael Scharli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le y Ed Chi. Least-to-most prompting enables complex reasoning in large language models, 2023. [3](#), [4](#)