

# Deepseek R1 - Conversación

A: Hola, he oído mucho sobre los modelos DeepSeek-R1 y sus capacidades de razonamiento. ¿Puedes desglosarlo para mí?

B: ¡Claro! Empecemos por lo básico. DeepSeek-R1 es una serie de modelos desarrollados por DeepSeek-AI que se centran en mejorar las capacidades de razonamiento a través del aprendizaje por refuerzo (RL). Hay dos versiones principales: DeepSeek-R1-Zero y DeepSeek-R1.

A: ¿Cuál es la diferencia entre DeepSeek-R1-Zero y DeepSeek-R1?

B: DeepSeek-R1-Zero se entrena puramente a través de RL sin ningún ajuste fino supervisado (SFT). Demuestra fuertes capacidades de razonamiento, pero tiene problemas como mala legibilidad y mezcla de idiomas. DeepSeek-R1, por otro lado, incorpora un entrenamiento en múltiples etapas y datos de inicio frío antes de RL para abordar estos problemas y mejorar aún más el rendimiento.

A: Eso es interesante. ¿Cómo funciona el proceso de aprendizaje por refuerzo en estos modelos?

B: El proceso de RL implica usar un sistema de recompensas para guiar el aprendizaje del modelo. Para DeepSeek-R1-Zero, utilizan un sistema de recompensas basado en reglas que se centra en la precisión y el formato. El modelo aprende a generar un proceso de razonamiento seguido de la respuesta final, mejorando con el tiempo.

A: ¿Y qué hay de los datos de inicio frío en DeepSeek-R1? ¿Cómo ayuda eso?

B: Los datos de inicio frío proporcionan una pequeña cantidad de ejemplos de alta calidad y larga Cadena de Pensamiento (CoT) para ajustar el modelo base antes de RL. Esto ayuda a mejorar la legibilidad y a alinear el modelo con las preferencias humanas, haciendo que los procesos de razonamiento sean más coherentes y amigables para el usuario.

A: ¿Cómo aseguran que las respuestas del modelo sean precisas y bien formateadas?

B: Utilizan una combinación de recompensas de precisión y recompensas de formato. Las recompensas de precisión aseguran que las respuestas sean correctas, mientras que las recompensas de formato obligan al modelo a estructurar su proceso de pensamiento entre etiquetas específicas. Esto ayuda a mantener la consistencia y la legibilidad.

A: ¿Qué tipo de benchmarks han utilizado para evaluar estos modelos?

B: Han evaluado los modelos en una variedad de benchmarks, incluyendo AIME 2024, MATH-500, GPQA Diamond, Codeforces, y más. Estos benchmarks cubren tareas de matemáticas, codificación y razonamiento general, proporcionando una evaluación exhaustiva de las capacidades de los modelos.

A: ¿Cómo se desempeña DeepSeek-R1 en comparación con otros modelos como la serie o1 de OpenAI?

B: DeepSeek-R1 logra un rendimiento comparable a OpenAI-o1-1217 en tareas de razonamiento. Por ejemplo, obtiene un 79.8% de Pass@1 en AIME 2024 y un 97.3% en MATH-500, igualando o incluso superando a los modelos de OpenAI en algunos casos.

A: Eso es impresionante. ¿Y el proceso de destilación? ¿Cómo funciona eso?

B: La destilación implica transferir las capacidades de razonamiento de modelos más grandes como DeepSeek-R1 a modelos más pequeños y eficientes. Ajustan finos modelos de código abierto como Qwen y Llama utilizando los datos generados por DeepSeek-R1, resultando en modelos más pequeños que se desempeñan excepcionalmente bien.

A: ¿Cuáles son los beneficios de la destilación sobre el RL directo en modelos más pequeños?

B: La destilación es más económica y efectiva. Los modelos más pequeños entrenados directamente a través de un gran RL pueden no alcanzar el mismo rendimiento que aquellos destilados de modelos más grandes. La destilación aprovecha los patrones de razonamiento avanzados descubiertos por los modelos más grandes, lo que lleva a un mejor rendimiento en modelos más pequeños.

A: ¿Hay algún compromiso o limitación con el enfoque de destilación?

B: Una limitación es que los modelos destilados aún pueden requerir más RL para alcanzar su máximo potencial. Aunque la destilación mejora significativamente el rendimiento, aplicar RL a estos modelos puede dar aún mejores resultados. Sin embargo, esto requiere recursos computacionales adicionales.

A: ¿Y el proceso de autoevolución en DeepSeek-R1-Zero? ¿Cómo funciona eso?

B: El proceso de autoevolución en DeepSeek-R1-Zero es fascinante. El modelo aprende naturalmente a resolver tareas de razonamiento cada vez más complejas aprovechando el cómputo extendido en tiempo de prueba. Esto lleva al surgimiento de comportamientos sofisticados como la reflexión y enfoques alternativos de resolución de problemas.

A: ¿Puedes dar un ejemplo de cómo evolucionan las capacidades de razonamiento del modelo con el tiempo?

B: ¡Claro! Por ejemplo, la longitud promedio de la respuesta del modelo aumenta con el tiempo, lo que indica que aprende a pasar más tiempo pensando y refinando sus soluciones. Esto lleva a un mejor rendimiento en benchmarks como AIME 2024, donde la puntuación de Pass@1 mejora del 15.6% al 71.0%.

A: ¿Y el ‘momento eureka’ mencionado en el papel? ¿Qué es eso?

B: El ‘momento eureka’ se refiere a un punto durante el entrenamiento en el que el modelo aprende a reevaluar su enfoque inicial a un problema, lo que lleva a mejoras significativas en sus capacidades de razonamiento. Es un testimonio de la capacidad del modelo para desarrollar estrategias de resolución de problemas avanzadas de manera autónoma.

A: ¿Cómo manejan el problema de la mezcla de idiomas en los modelos?

B: Para abordar la mezcla de idiomas, introducen una recompensa de consistencia lingüística durante el entrenamiento de RL. Esta recompensa alinea al modelo con las preferencias humanas, haciendo que las respuestas sean más legibles y coherentes. Aunque esto degrada ligeramente el rendimiento, mejora la experiencia del usuario en general.

A: ¿Cuáles son algunos de los intentos fallidos que mencionan en el papel?

B: Experimentaron con modelos de recompensa de proceso (PRM) y Búsqueda de Árbol de Monte Carlo (MCTS), pero ambos enfoques enfrentaron desafíos. PRM sufrió de hackeo de recompensas y problemas de

escalabilidad, mientras que MCTS luchó con el espacio de búsqueda exponencialmente más grande en la generación de tokens.

A: ¿Cuáles son las direcciones futuras para DeepSeek-R1?

B: Planean mejorar las capacidades generales, abordar la mezcla de idiomas, mejorar la ingeniería de prompts y mejorar el rendimiento en tareas de ingeniería de software. También tienen como objetivo explorar más el potencial de la destilación e investigar el uso de largas CoT para diversas tareas.

A: ¿Cómo planean mejorar las capacidades generales?

B: Planean aprovechar las largas CoT para mejorar tareas como la llamada de funciones, conversaciones de múltiples vueltas, roles complejos y salida json. Esto ayudará a hacer que el modelo sea más versátil y capaz de manejar una gama más amplia de tareas.

A: ¿Y el problema de la mezcla de idiomas? ¿Cómo planean abordarlo?

B: Planean optimizar el modelo para múltiples idiomas, asegurando que no predomine el inglés para el razonamiento y las respuestas al manejar consultas en otros idiomas. Esto hará que el modelo sea más accesible y útil para una audiencia global.

A: ¿Cómo planean mejorar la ingeniería de prompts?

B: Recomiendan a los usuarios describir directamente el problema y especificar el formato de salida utilizando un entorno de cero disparos. Este enfoque ha demostrado ser más efectivo que el prompting de pocos disparos, que puede degradar el rendimiento del modelo.

A: ¿Cuáles son los desafíos que enfrentan con las tareas de ingeniería de software?

B: Los largos tiempos de evaluación afectan la eficiencia del proceso de RL, haciendo que sea difícil aplicar un gran RL extensivamente en tareas de ingeniería de software. Planean implementar muestreo de rechazo en datos de ingeniería de software o incorporar evaluaciones asincrónicas para mejorar la eficiencia.

A: ¿Cómo aseguran que las respuestas del modelo sean útiles y seguras?

B: Implementan una etapa secundaria de aprendizaje por refuerzo dirigida a mejorar la utilidad y seguridad del modelo. Esto implica usar una combinación de señales de recompensa y distribuciones de prompts diversas para alinear el modelo con las preferencias humanas y mitigar riesgos potenciales.

A: ¿Cuáles son algunas de las tendencias emergentes en el aprendizaje por refuerzo para LLMs?

B: Algunas tendencias emergentes incluyen el uso de modelos de recompensa más avanzados, explorar nuevos algoritmos de RL e integrar RL con otras técnicas de entrenamiento como la destilación. También hay un creciente interés en hacer que el RL sea más eficiente y escalable para modelos más grandes.

A: ¿Cómo comparan el rendimiento de los modelos destilados con otros modelos comparables?

B: Comparan los modelos destilados con otros modelos como GPT-4o-0513, Claude-3.5-Sonnet-1022 y QwQ-32B-Preview en varios benchmarks. Los modelos destilados, como DeepSeek-R1-Distill-Qwen-7B, superan a estos modelos en todos los frentes, demostrando la efectividad del enfoque de destilación.

A: ¿Cuáles son algunos de los puntos clave del papel de DeepSeek-R1?

B: Los puntos clave incluyen el potencial de RL para mejorar las capacidades de razonamiento en LLMs, la efectividad de la destilación para transferir estas capacidades a modelos más pequeños y la importancia de abordar problemas como la mezcla de idiomas y la sensibilidad a los prompts. El papel también destaca la necesidad de más investigación para hacer que el RL sea más eficiente y escalable.

A: ¿Cómo aseguran que las respuestas del modelo sean precisas y bien formateadas?

B: Utilizan una combinación de recompensas de precisión y recompensas de formato. Las recompensas de precisión aseguran que las respuestas sean correctas, mientras que las recompensas de formato obligan al modelo a estructurar su proceso de pensamiento entre etiquetas específicas. Esto ayuda a mantener la consistencia y la legibilidad.

A: ¿Qué tipo de benchmarks han utilizado para evaluar estos modelos?

B: Han evaluado los modelos en una variedad de benchmarks, incluyendo AIME 2024, MATH-500, GPQA Diamond, Codeforces, y más. Estos benchmarks cubren tareas de matemáticas, codificación y razonamiento general, proporcionando una evaluación exhaustiva de las capacidades de los modelos.

A: ¿Cómo se desempeña DeepSeek-R1 en comparación con otros modelos como la serie o1 de OpenAI?

B: DeepSeek-R1 logra un rendimiento comparable a OpenAI-o1-1217 en tareas de razonamiento. Por ejemplo, obtiene un 79.8% de Pass@1 en AIME 2024 y un 97.3% en MATH-500, igualando o incluso superando a los modelos de OpenAI en algunos casos.

A: Eso es impresionante. ¿Y el proceso de destilación? ¿Cómo funciona eso?

B: La destilación implica transferir las capacidades de razonamiento de modelos más grandes como DeepSeek-R1 a modelos más pequeños y eficientes. Ajustan finos modelos de código abierto como Qwen y Llama utilizando los datos generados por DeepSeek-R1, resultando en modelos más pequeños que se desempeñan excepcionalmente bien.

A: ¿Cuáles son los beneficios de la destilación sobre el RL directo en modelos más pequeños?

B: La destilación es más económica y efectiva. Los modelos más pequeños entrenados directamente a través de un gran RL pueden no alcanzar el mismo rendimiento que aquellos destilados de modelos más grandes. La destilación aprovecha los patrones de razonamiento avanzados descubiertos por los modelos más grandes, lo que lleva a un mejor rendimiento en modelos más pequeños.

A: ¿Hay algún compromiso o limitación con el enfoque de destilación?

B: Una limitación es que los modelos destilados aún pueden requerir más RL para alcanzar su máximo potencial. Aunque la destilación mejora significativamente el rendimiento, aplicar RL a estos modelos puede dar aún mejores resultados. Sin embargo, esto requiere recursos computacionales adicionales.

A: ¿Y el proceso de autoevolución en DeepSeek-R1-Zero? ¿Cómo funciona eso?

B: El proceso de autoevolución en DeepSeek-R1-Zero es fascinante. El modelo aprende naturalmente a resolver tareas de razonamiento cada vez más complejas aprovechando el cómputo extendido en tiempo de prueba. Esto lleva al surgimiento de comportamientos sofisticados como la reflexión y enfoques alternativos de resolución de problemas.

A: ¿Puedes dar un ejemplo de cómo evolucionan las capacidades de razonamiento del modelo con el tiempo?

B: ¡Claro! Por ejemplo, la longitud promedio de la respuesta del modelo aumenta con el tiempo, lo que indica que aprende a pasar más tiempo pensando y refinando sus soluciones. Esto lleva a un mejor rendimiento en benchmarks como AIME 2024, donde la puntuación de Pass@1 mejora del 15.6% al 71.0%.

A: ¿Y el ‘momento eureka’ mencionado en el papel? ¿Qué es eso?

B: El ‘momento eureka’ se refiere a un punto durante el entrenamiento en el que el modelo aprende a reevaluar su enfoque inicial a un problema, lo que lleva a mejoras significativas en sus capacidades de razonamiento. Es un testimonio de la capacidad del modelo para desarrollar estrategias de resolución de problemas avanzadas de manera autónoma.

A: ¿Cómo manejan el problema de la mezcla de idiomas en los modelos?

B: Para abordar la mezcla de idiomas, introducen una recompensa de consistencia lingüística durante el entrenamiento de RL. Esta recompensa alinea al modelo con las preferencias humanas, haciendo que las respuestas sean más legibles y coherentes. Aunque esto degrada ligeramente el rendimiento, mejora la experiencia del usuario en general.

A: ¿Cuáles son algunos de los intentos fallidos que mencionan en el papel?

B: Experimentaron con modelos de recompensa de proceso (PRM) y Búsqueda de Árbol de Monte Carlo (MCTS), pero ambos enfoques enfrentaron desafíos. PRM sufrió de hackeo de recompensas y problemas de escalabilidad, mientras que MCTS luchó con el espacio de búsqueda exponencialmente más grande en la generación de tokens.

A: ¿Cuáles son las direcciones futuras para DeepSeek-R1?

B: Planean mejorar las capacidades generales, abordar la mezcla de idiomas, mejorar la ingeniería de prompts y mejorar el rendimiento en tareas de ingeniería de software. También tienen como objetivo explorar más el potencial de la destilación e investigar el uso de largas CoT para diversas tareas.

A: ¿Cómo planean mejorar las capacidades generales?

B: Planean aprovechar las largas CoT para mejorar tareas como la llamada de funciones, conversaciones de múltiples vueltas, roles complejos y salida json. Esto ayudará a hacer que el modelo sea más versátil y capaz de manejar una gama más amplia de tareas.

A: ¿Y el problema de la mezcla de idiomas? ¿Cómo planean abordarlo?

B: Planean optimizar el modelo para múltiples idiomas, asegurando que no predomine el inglés para el razonamiento y las respuestas al manejar consultas en otros idiomas. Esto hará que el modelo sea más accesible y útil para una audiencia global.

A: ¿Cómo planean mejorar la ingeniería de prompts?

B: Recomiendan a los usuarios describir directamente el problema y especificar el formato de salida utilizando un entorno de cero disparos. Este enfoque ha demostrado ser más efectivo que el prompting de pocos disparos, que puede degradar el rendimiento del modelo.

A: ¿Cuáles son los desafíos que enfrentan con las tareas de ingeniería de software?

B: Los largos tiempos de evaluación afectan la eficiencia del proceso de RL, haciendo que sea difícil aplicar un gran RL extensivamente en tareas de ingeniería de software. Planean implementar muestreo de rechazo en datos de ingeniería de software o incorporar evaluaciones asincrónicas para mejorar la eficiencia.

A: ¿Cómo aseguran que las respuestas del modelo sean útiles y seguras?

B: Implementan una etapa secundaria de aprendizaje por refuerzo dirigida a mejorar la utilidad y seguridad del modelo. Esto implica usar una combinación de señales de recompensa y distribuciones de prompts diversas para alinear el modelo con las preferencias humanas y mitigar riesgos potenciales.

A: ¿Cuáles son algunas de las tendencias emergentes en el aprendizaje por refuerzo para LLMs?

B: Algunas tendencias emergentes incluyen el uso de modelos de recompensa más avanzados, explorar nuevos algoritmos de RL e integrar RL con otras técnicas de entrenamiento como la destilación. También hay un creciente interés en hacer que el RL sea más eficiente y escalable para modelos más grandes.