Texto Simplificado e Traduzido:

GSM-Symbolic: explorando os desafios do raciocínio matemático em grandes modelos de linguagem

Visão geral

O progresso recente em modelos de linguagem grande (LLMs) levou a um aumento do interesse em sua capacidade de realizar raciocínio lógico, especialmente em matemática. O benchmark GSM8K é uma ferramenta comum usada para avaliar o raciocínio matemático desses modelos com perguntas típicas do que um aluno do ensino fundamental pode encontrar. Embora os LLMs tenham mostrado um desempenho aprimorado nesse benchmark, ainda não está claro se isso indica um verdadeiro avanço no raciocínio matemático. Essa incerteza lança dúvidas sobre a confiabilidade das métricas de desempenho atuais.

Em resposta, conduzimos um estudo abrangente envolvendo vários modelos abertos e proprietários líderes. Para resolver as deficiências de avaliação existentes, desenvolvemos o GSM-Symbolic, um benchmark aprimorado usando modelos simbólicos, que permite a geração de perguntas variadas. Esse novo benchmark fornece avaliações controladas, oferecendo insights críticos e métricas mais confiáveis para avaliar as habilidades de raciocínio dos modelos.

Nossas descobertas indicam variações significativas no desempenho dos LLMs em diferentes versões da mesma pergunta. Especificamente, quando apenas os números em uma pergunta mudam, o desempenho dos modelos cai no benchmark simbólico GSM. Além disso, notamos que, à medida que a complexidade das perguntas (medida pelo número de cláusulas) aumenta, o desempenho dos modelos diminui significativamente. Nossa hipótese é que isso ocorre porque os LLMs atuais não realizam um raciocínio lógico genuíno. Em vez disso, eles imitam as etapas de raciocínio observadas em seus dados de treinamento. Quando uma cláusula irrelevante é adicionada a uma pergunta, o desempenho pode cair em até 65%, ressaltando uma limitação fundamental no discernimento de informações relevantes para resolver problemas.

Introdução

Os LLMs demonstraram capacidades impressionantes em várias áreas, incluindo processamento de linguagem natural e tarefas criativas. Seu potencial para lidar com tarefas complexas de raciocínio em áreas como codificação e matemática atraiu atenção significativa. No entanto, continua sendo uma questão crítica se esses modelos podem realmente realizar o raciocínio lógico. Uma análise mais detalhada revela limitações substanciais, com pesquisas sugerindo que o raciocínio em LLMs tem mais a ver com correspondência probabilística de padrões do que com raciocínio lógico estruturado.

O raciocínio matemático é vital para a resolução de problemas em muitas aplicações científicas e práticas. O conjunto de dados GSM8K é amplamente usado para testar as habilidades de raciocínio matemático dos LLMs. No entanto, embora o GSM8K forneça questões matemáticas simples com soluções detalhadas, ele oferece apenas uma única métrica em um conjunto fixo de perguntas, limitando a compreensão do raciocínio dos modelos. Também existe o risco de

contaminação dos dados, dada a popularidade do conjunto de dados. A natureza estática do GSM8K não permite que experimentos controlados explorem as limitações do modelo sob condições e níveis de dificuldade variados.

Portanto, é necessária uma estrutura de avaliação mais flexível e adaptável, que possa gerar diversos formatos de perguntas e ajustar os níveis de complexidade para investigar minuciosamente os pontos fortes e fracos do raciocínio dos LLMs.

Contribuições

- **GSM-Symbolic Benchmark: ** Introduzimos um benchmark aprimorado usando modelos simbólicos para criar diversas variantes de perguntas GSM8K. Isso permite uma avaliação diferenciada e confiável dos LLMs em diferentes configurações, indo além das métricas de precisão de um único ponto. Nosso estudo em grande escala fornece informações significativas sobre o comportamento dos LLMs em tarefas de raciocínio matemático.

- **Variabilidade de desempenho: ** Destacamos a confiabilidade questionável dos resultados atuais do GSM8K, demonstrando que o desempenho do LLM pode variar inesperadamente em diferentes instâncias da mesma pergunta. Declínios de desempenho no GSM Symbolic sugerem uma possível contaminação de dados.

- **Sensibilidade a mudanças numéricas: ** Mostramos que, embora os LLMs sejam mais estáveis com mudanças superficiais, como nomes próprios, eles são altamente sensíveis a alterações numéricas. A degradação do desempenho aumenta com a complexidade das perguntas, indicando dificuldades com maior dificuldade.

- **Conjunto de dados GSM-Noop: ** Apresentamos esse conjunto de dados para demonstrar quedas significativas de desempenho (até 65%) em modelos de última geração quando informações irrelevantes, mas aparentemente relevantes, são adicionadas aos problemas. Isso expõe uma falha crítica na capacidade dos modelos de discernir informações relevantes devido à sua confiança na correspondência de padrões em vez do raciocínio formal.

Conclusão

No geral, nosso trabalho fornece uma compreensão abrangente das limitações do LLMS no raciocínio matemático, enfatizando a necessidade de metodologias de avaliação mais confiáveis e uma maior exploração das capacidades de raciocínio do LLM. Nossos resultados sugerem limitações significativas na capacidade dos LLMs de realizar um raciocínio matemático genuíno. Esses modelos apresentam alta variação no desempenho em diferentes instâncias de perguntas, quedas significativas de desempenho com pequenos aumentos de complexidade e sensibilidade a informações irrelevantes. Isso indica que o raciocínio do LLMS pode ser mais parecido com uma combinação sofisticada de padrões do que com o raciocínio lógico real. No futuro, mais pesquisas são necessárias para desenvolver modelos de IA que possam realizar o raciocínio formal, avançando além do reconhecimento de padrões para obter habilidades robustas e generalizáveis de resolução de problemas.

Métricas do Texto Original:

Índice de Flesch Reading Ease: 55.34

Grau de Flesch-Kincaid: 9.50
Índice SMOG: 11.90
Índice de Coleman-Liau: 14.56
Índice ARI: 14.90
Pontuação de Dale-Chall: 8.11

Métricas do Texto Simplificado:

Índice de Flesch Reading Ease: 26.61
Grau de Flesch-Kincaid: 14.30
Índice SMOG: 16.00
Índice de Coleman-Liau: 19.03
Índice ARI: 18.60
Pontuação de Dale-Chall: 9.70