Texto Simplificado e Traduzido:

GSM-Symbolic: investigando as restrições do raciocínio matemático em grandes modelos de linguagem

Resumo

O recente progresso em modelos de linguagem grande (LLMs) gerou um interesse significativo em suas capacidades de raciocínio formal, particularmente no domínio da matemática. O benchmark GSM8K é amplamente utilizado para avaliar as capacidades de raciocínio matemático dos modelos em questões do ensino fundamental. Embora tenha havido um notável aprimoramento no desempenho dos LLMs no GSM8K recentemente, o grau em que essas melhorias refletem avanços genuínos no raciocínio matemático permanece incerto, questionando assim a validade das métricas relatadas. Em resposta a essas preocupações, realizamos um extenso estudo de vários modelos abertos e fechados de última geração. Para abordar as limitações das avaliações atuais, propomos o GSM-Symbolic, um benchmark aprimorado derivado de modelos simbólicos, permitindo a criação de uma variedade diversificada de perguntas. O GSM-Symbolic permite avaliações mais controladas, gerando insights críticos e métricas mais confiáveis para avaliar as capacidades de raciocínio dos modelos. Nossas descobertas indicam que os LLMs apresentam uma variabilidade considerável ao responder a diferentes instâncias da mesma pergunta. Especificamente, o desempenho de todos os modelos se deteriora quando somente os valores numéricos da questão são modificados no benchmark simbólico GSM. Além disso, examinamos a fragilidade do raciocínio matemático nesses modelos e mostramos que seu desempenho diminui significativamente à medida que o número de cláusulas em uma pergunta aumenta. Nossa hipótese é que esse declínio decorre da limitação atual da incapacidade dos LLMs de realizar um raciocínio lógico autêntico; em vez disso, eles tentam imitar as etapas de raciocínio observadas em seus dados de treinamento. Quando uma cláusula adicional percebida como relevante para a pergunta é introduzida, observamos quedas substanciais de desempenho (até 65%) em todos os modelos de última geração, mesmo que a cláusula adicionada não contribua para a cadeia de raciocínio necessária para chegar à resposta final. No geral, nossa pesquisa oferece uma compreensão mais diferenciada das capacidades e limitações dos LLMs no raciocínio matemático.

Introdução

Os LLMs demonstraram um potencial notável em vários domínios, incluindo processamento de linguagem natural, resposta a perguntas e tarefas criativas. Seu potencial para realizar tarefas complexas de raciocínio, particularmente em codificação e matemática, atraiu considerável interesse de pesquisadores e profissionais. No entanto, se os LLMs atuais são realmente capazes de um raciocínio lógico genuíno continua sendo um foco significativo de pesquisa. Embora alguns estudos destaquem capacidades impressionantes, um exame mais detalhado revela limitações substanciais. A literatura sugere que o processo de raciocínio em LLMs é uma correspondência probabilística de padrões em vez de um raciocínio formal. Embora os LLMs possam combinar padrões de raciocínio mais abstratos, eles ficam aquém do verdadeiro raciocínio lógico. Pequenas mudanças nos tokens de entrada podem alterar drasticamente as saídas do modelo, indicando um forte viés de token e sugerindo que esses modelos são altamente sensíveis e frágeis. Além disso, em tarefas que exigem a seleção correta de vários

tokens, a probabilidade de chegar a uma resposta precisa diminui exponencialmente com o número de tokens ou etapas envolvidas, ressaltando sua falta de confiabilidade inerente em cenários complexos de raciocínio.

O raciocínio matemático é uma habilidade cognitiva crucial que apoia a resolução de problemas em várias aplicações científicas e práticas. Consequentemente, a capacidade dos LLMs de realizar tarefas de raciocínio matemático com eficácia é vital para o avanço da inteligência artificial e de suas aplicações no mundo real. O conjunto de dados GSM8K se tornou uma referência popular para avaliar as capacidades de raciocínio matemático dos LLMs. Embora inclua questões matemáticas simples com soluções detalhadas, o que o torna adequado para técnicas como a solicitação de Cadeia de Pensamento (CoT), ele fornece apenas uma única métrica em um conjunto fixo de perguntas. Essa limitação restringe insights abrangentes sobre o raciocínio matemático dos modelos. Além disso, a popularidade e a prevalência do GSM8K podem aumentar o risco de contaminação inadvertida de dados. Finalmente, a natureza estática do GSM8K não permite que experimentos controláveis entendam as limitações do modelo, como comportamento sob condições variadas ou mudanças nos aspectos da questão e nos níveis de dificuldade.

Para resolver essas limitações, é necessária uma estrutura de avaliação mais versátil e adaptável, que possa gerar diversas variantes de perguntas e ajustar os níveis de complexidade para explorar melhor a robustez e as habilidades de raciocínio dos LLMs. Isso facilitaria uma compreensão mais profunda dos pontos fortes e fracos desses modelos em tarefas de raciocínio matemático. Fazemos as seguintes contribuições:

- Apresentamos o GSM-Symbolic, um benchmark aprimorado que gera diversas variantes de perguntas do GSM8K usando modelos simbólicos, permitindo uma avaliação mais detalhada e confiável do desempenho dos LLMs em várias configurações, indo além das métricas de precisão de ponto único. Nosso estudo em grande escala sobre 25 modelos abertos e fechados de última geração fornece informações significativas sobre o comportamento dos LLMs em tarefas de raciocínio matemático.
- Questionamos a confiabilidade dos resultados relatados atualmente no GSM8K e demonstramos que o desempenho dos LLMs pode ser visto como uma distribuição com variação injustificada em diferentes instanciações da mesma pergunta. Mostramos que o desempenho de todos os modelos cai no GSM Symbolic, indicando uma possível contaminação de dados.
- Mostramos que os LLMs exibem mais robustez às mudanças em elementos superficiais, como nomes próprios, mas são muito sensíveis às mudanças nos valores numéricos. Mostramos que a degradação e a variância do desempenho aumentam à medida que o número de cláusulas aumenta, indicando que as capacidades de raciocínio dos LLMs enfrentam dificuldades com o aumento da complexidade.
- Finalmente, questionamos ainda mais as habilidades de raciocínio dos LLMs e apresentamos o conjunto de dados GSM-noop. Ao adicionar informações aparentemente relevantes, mas, em última análise, irrelevantes aos problemas, demonstramos quedas substanciais de desempenho (até 65%) em todos os modelos de última geração. Isso revela uma falha crítica na capacidade dos modelos de discernir informações relevantes para a resolução de problemas, provavelmente porque seu raciocínio não é formal no termo de senso comum e é baseado principalmente na correspondência de padrões. Mostramos que, mesmo quando fornecidos com vários exemplos da mesma pergunta ou exemplos contendo informações irrelevantes semelhantes, os LLMs lutam para superar os desafios colocados pelo GSM-noop. Isso sugere

problemas mais profundos em seus processos de raciocínio que exigem uma investigação mais aprofundada.

No geral, nosso trabalho fornece uma compreensão abrangente das limitações dos LLMs no raciocínio matemático. Nossos resultados enfatizam a necessidade de metodologias de avaliação mais confiáveis e pesquisas adicionais sobre as capacidades de raciocínio de grandes modelos de linguagem.

Trabalho relacionado: modelos de raciocínio e linguagem

O raciocínio lógico é uma característica crítica dos sistemas inteligentes. Avanços recentes em modelos de linguagem grande (LLMs) demonstraram um potencial significativo em vários domínios, mas suas habilidades de raciocínio permanecem incertas e inconsistentes. Muitos trabalhos investigaram se os LLMs são realmente capazes de raciocinar examinando como esses modelos resolvem tarefas que requerem raciocínio lógico. Uma direção interessante se concentra na modelagem da computação realizada por transformadores. Por exemplo, foram traçados paralelos entre componentes como módulos de atenção e feed-forward e primitivas computacionais simples. Pesquisas mostraram que os transformadores falham em generalizar em tarefas não regulares e indicaram que a memória estruturada (por exemplo, fita de memória) é necessária para lidar com tarefas complexas. Isso está relacionado à eficácia da solicitação da Cadeia de Pensamento (CoT) e do uso de blocos de rascunho para LLMs como memória adicional para cálculos intermediários. No geral, os resultados atuais sugerem que, embora a arquitetura do transformador tenha limitações e não tenha a expressividade necessária para resolver problemas em várias classes de complexidade, essas limitações podem ser atenuadas com memória adicional (por exemplo, rascunhos). No entanto, isso ainda exige a geração de grandes quantidades de tokens para resolver um problema. Embora esses trabalhos forneçam insights sobre a complexidade computacional teórica dos transformadores, na prática, ainda não está claro se esses LLMs podem realizar raciocínio lógico formal para resolver tarefas.

Há evidências substanciais que sugerem que o processo de raciocínio em LLMs não é formal, embora pareça que esses modelos entendam símbolos e possam trabalhar com eles de forma limitada. Em vez disso, os LLMs provavelmente executam um formulário

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: 23.87
Grau de Flesch-Kincaid: 15.40

Índice SMOG: 16.30
Índice de Coleman-Liau: 17.98
Índice ARI: 18.90
Pontuação de Dale-Chall: 9.33