Deepseek se tornou viral.

O laboratório chinês da AI Deepseek invadiu a consciência convencional nesta semana depois que seu aplicativo de chatbot chegou ao topo dos gráficos da Apple App Store. Os modelos de IA da Deepseek, que foram treinados usando técnicas com eficiência de computação, levaram os analistas de Wall Street-e os tecnólogos-a questionar se os EUA podem manter sua liderança na corrida de IA e se a demanda por chips de IA sustentará.

Mas de onde veio o Deepseek e como se elevou à fama internacional tão rapidamente?

O que você vai ler:

- As origens comerciais de Deepseek
- Modelos fortes de Deepseek
- Uma abordagem disruptiva

As origens comerciais de Deepseek

A Deepseek é apoiada pela High-Flyer Capital Management, um fundo de hedge quantitativo chinês que usa a IA para informar suas decisões comerciais.

O entusiasta da IA, Liang Wenfeng, co-fundou o High-Flyer em 2015. Wenfeng, que teria começado a se envolver em negociações enquanto um estudante da Universidade de Zhejiang, lançou a gestão de capital de alto voo como um fundo de hedge em 2019, focado no desenvolvimento e implantação de algoritmos de IA.

Em 2023, o High-Flyer começou a Deepseek como um laboratório dedicado à pesquisa de ferramentas de IA separadas de seus negócios financeiros. Com o High-Flyer como um de seus investidores, o laboratório girou em sua própria empresa, também chamada Deepseek.

Desde o primeiro dia, a Deepseek construiu seus próprios clusters de data center para treinamento de modelos. Mas, como outras empresas de IA na China, a Deepseek foi afetada pelas proibições de exportação dos EUA em hardware. Para treinar um de seus modelos mais recentes, a empresa foi forçada a usar o NVIDIA H800 Chips, uma versão menos poderosa de um chip, o H100, disponível para empresas americanas.

Diz -se que a equipe técnica de Deepseek distorce Young. A empresa teria recruta agressivamente o doutorado em pesquisadores de IA das principais universidades chinesas. A Deepseek também contrata pessoas sem nenhum histórico de ciência da computação para ajudar sua tecnologia a entender melhor uma ampla gama de assuntos, de acordo com o New York Times.

memória virtual

Modelos fortes de Deepseek

A Deepseek apresentou seu primeiro conjunto de modelos-Deepseek Coder, Deepseek LLM e Deepseek Chat-em novembro de 2023. Mas não foi até a primavera passada, quando a startup lançou sua família de modelos de Deepseek-V2 da próxima geração, que a indústria da IA começou a prestar atenção.

Deepseek-V2, um sistema de análise de texto e imagem de uso geral, teve um bom desempenho em vários benchmarks de IA-e era muito mais barato de executar do que os modelos comparáveis na época. Forçou a concorrência doméstica de Deepseek, incluindo Bytedance e Alibaba, a reduzir os preços de uso de alguns de seus modelos e tornar os outros completamente gratuitos.

Deepseek-V3, lançado em dezembro de 2024, apenas adicionou à notoriedade de Deepseek.

De acordo com os testes internos de benchmark da Deepseek, o DeepSeek V3 supera modelos para downloads e disponíveis abertamente, como os modelos Llama da Meta e "fechado", que só podem ser acessados por meio de uma API, como o GPT-4O do OpenAI.

Igualmente impressionante é o modelo de "raciocínio" do Deepseek. Lançado em janeiro, o Deepseek afirma que o R1 é executado, bem como o modelo O1 do OpenAI em benchmarks - chave.

Sendo um modelo de raciocínio, o R1 efetivamente se chicando, o que ajuda a evitar algumas das armadilhas que normalmente disparam modelos. Os modelos de raciocínio demoram um pouco mais-geralmente segundos a minutos a mais-para chegar a soluções em comparação com um modelo típico de não-reamento. A vantagem é que eles tendem a ser mais confiáveis em domínios como física, ciência e matemática.

No entanto, há uma desvantagem para R1, Deepseek V3 e outros modelos de Deepseek. Sendo a IA desenvolvida em chinês, eles estão sujeitos a benchmarking pelo regulador da Internet da China para garantir que suas respostas "incorporem os principais valores socialistas". No aplicativo Chatbot da Deepseek, por exemplo, o R1 não responderá perguntas sobre a Praça Tiananmen ou a autonomia de Taiwan.

Uma abordagem disruptiva

Se o Deepseek tem um modelo de negócios, não está claro o que é esse modelo exatamente. A empresa prende seus produtos e serviços bem abaixo do valor de mercado – e distribui os outros gratuitamente.

A maneira como a Deepseek diz, os avanços da eficiência permitiram manter a extrema competitividade de custos. Alguns especialistas contestam os números que a empresa forneceu, no entanto.

Seja qual for o caso, os desenvolvedores adotaram os modelos da Deepseek, que não são de código aberto, pois a frase é comumente entendida, mas estão disponíveis sob licenças permissivas que permitem uso comercial. De acordo com Clem Delangue, o CEO de abraçar o rosto, uma das plataformas que hospedam os modelos de Deepseek, os desenvolvedores em abraçar o rosto criaram mais de 500 modelos “derivados” de R1 que acumularam 2,5 milhões de downloads combinados.

O sucesso de Deepseek contra rivais maiores e mais estabelecidos foi descrito como “AI elevando a IA” e inaugurando “uma nova era de bordo da AI”. O sucesso da empresa foi pelo menos em parte responsável por fazer com que o preço das ações da Nvidia caia em 18% na segunda -feira e por obter uma resposta pública do CEO da Openai, Sam Altman.

Quanto ao que o futuro de Deepseek pode ter, não está claro. Modelos aprimorados são um dado. Mas o governo dos EUA parece estar ficando cauteloso com o que considera influência estrangeira prejudicial.

O TechCrunch tem um boletim informativo focado na IA! Inscreva -se aqui para obtê -lo em sua caixa de entrada toda quarta -feira.

(tagstotranslate) ai