DeepSeek

As implicações para o futuro da AI

astercapital

Nas últimas semanas foram lançados o **DeepSeek-V3** (V3) e o **DeepSeek-R1** (R1), dois novos modelos chineses de inteligência artificial que têm gerado discussões sobre seu impacto no futuro da AI. A discussão estava presa dentro da comunidade de AI, porém, com o lançamento do aplicativo oficial da **DeepSeek** para iOS e Android, a discussão ganhou uma dimensão global.

Entendemos que grandes modelos de linguagem (LLMs) estão se commoditizando, tanto os generalistas quanto os especialistas, e acreditamos que o valor está nas camadas construídas em cima desses modelos. Os modelos lançados pela **DeepSeek** são um resultado esperado desse movimento e entendemos que não mudam os caminhos atuais e existentes de AI.

Os modelos do **DeepSeek** possuem uma otimização de hardware impressionante, principalmente para ajudar no desenvolvimento de outros modelos, tarefas lógicas, como revisão de códigos de programação, e também atividades do dia-a-dia, como, por exemplo, traduções.

Encontramos em nossas análises evidências que a **DeepSeek** tem muito trabalho pela frente para conseguir atuar em grande escala, talvez um trade-off de toda otimização construída.

Com um desempenho superior em alguns aspectos em relação a modelos de empresas como OpenAI, Claude, LLaMA e X.AI, o **DeepSeek-V3** e **R1** também se destacam por custos de treinamento até 20 vezes menores.

Ou seja, é possível que o lançamento de um produto inferior em alguns aspectos, mas exponencialmente mais barato, impacte empresas como a Microsoft, por exemplo, a ponto de transformar o mercado e afetar a demanda por inteligência artificial nos próximos anos?

Nossa análise sugere que, por enquanto, esse não é o caso. O produto apresenta um custo exponencialmente mais barato, mas em linha com o movimento de queda que já está ocorrendo. Ainda não encontramos evidências para revisar as nossas estimativas de investimento em inteligência artificial, nem uma redução na necessidade de infraestrutura já contratada para os próximos dois anos.

Esses modelos trazem, de fato, uma discussão maior sobre *open source* vs *closed source* e uma clara evidência de que continuaremos vivendo otimizações nos modelos de AI.

Os dois modelos **DeepSeek-V3** e **R1** são muito bons, melhores em alguns aspectos em relação a outros principais LLMs de mercado, embora com objetivos diferentes. Apesar de terem custos de treinamento significativamente

menores, parecem ser menos eficientes em algumas tarefas mais complexas que consideramos críticas para que empresas construam aplicações - como chatbots e agentes co-pilotos - e, portanto, uma ameaça menos relevante para os casos de uso mais populares do que se sugere inicialmente do ponto de vista competitivo. Fato é que esses dois modelos e seu aplicativo encontraram um espaço no mercado de AI.

Abaixo compartilhamos nossas análises e conclusões com base nos documentos técnicos da **DeepSeek**; em *benchmarks* realizados por especialistas em AI que foram divulgados ao longo das últimas semanas; e discussões de algumas das principais comunidades de AI.

DeepSeek

A **DeepSeek** nasceu como uma iniciativa dentro da High-Flyer, um *hedge fund* chinês fundado em fevereiro de 2016. Originalmente, a **DeepSeek** operava como um laboratório de pesquisa em inteligência artificial dentro da High-Flyer. No entanto, em maio de 2023, a empresa foi desmembrada, tornando-se uma organização independente.

No final do ano passado, a **DeepSeek** lançou seu novo modelo, o **DeepSeek-V3,** com licenciamento *open source* e técnicas inovadoras que reduzem os custos de treinamento e operação.

Um dos principais impactos sentidos pela comunidade de AI foi o fato dos modelos **R1** e **V3** mostrarem como utilizam o seu *Chain of Thought* (*CoT)*, uma técnica que instrui o modelo a "*mostrar como ele chega na resposta*" por meio de uma sequência de perguntas e palavras *(tokens)* intermediárias, ao invés de simplesmente mostrar a resposta final para uma questão.

Essa técnica já é amplamente utilizada, porém, foi a primeira vez que um modelo grande e impactante, como o **R1** e o **V3,** apresentou o seu funcionamento de forma detalhada.

Além disso, ao estudar a documentação técnica dos modelos, encontramos outros aspectos que tornam o **V3** e o **R1** tão diferentes:

> **i) Multi-Token Prediction (MTP):** Tradicionalmente, os modelos de AI utilizam técnicas que preveem um *token* (unidade de dados processada por AI) de maneira serial. No entanto, a **DeepSeek** pode prever várias palavras simultaneamente, reduzindo o número de etapas necessárias para produzir o texto completo. Ao processar blocos maiores de *tokens*, o MTP faz melhor uso de recursos computacionais como GPUs.

ii) FP8 Mixed Precision Training: Outra técnica focada em eficiência computacional e de custo de memória. Os modelos **V3** e **R1** usaram pontos flutuantes de 8 bits, ao invés de 16 ou 32 bits como a maioria dos modelos atuais. Essa característica acompanhada de estabilidade e técnicas para mitigar a perda de precisão, permitem que os modelos funcionem em infra estruturas menores.

iii) Uso otimizado de hardware: Devido a restrições políticas e comerciais, a NVIDIA não consegue exportar seus chips mais avançados para a China. Para contornar isso, o **V3** e o **R1** foram ajustados para funcionarem em máquinas com chips mais simples (A800 e H800), o que também ajuda a reduzir o custo de operação.

Se isso for verdade, os modelos da **DeepSeek** são realmente mais econômicos tanto no treinamento quanto na inferência? **A resposta é: sim, mas apenas para alguns usos específicos**.

Comparações

O LMSYS (*Large Model Systems Organization*) é uma organização de pesquisa dedicada ao estudo e avaliação de modelos de linguagem de grande escala e uma de suas principais iniciativas é a Chatbot Arena, uma plataforma que permite comparações diretas entre diferentes LLMs por meio de batalhas anônimas e aleatórias, nas quais os usuários interagem com dois modelos lado a lado e votam em suas respostas preferidas.

A Chatbot Arena utiliza um sistema de classificação baseado no modelo de Bradley-Terry, que é semelhante ao sistema de classificação Elo usado em xadrez e outros jogos competitivos. Com isso, é capaz de avaliar o desempenho desses modelos, gerando uma pontuação "Arena".

A tabela de classificação mais atualizada apresenta o modelo **R1** em 4º lugar na classificação geral e em 1º lugar no *ranking* de modelos *open source*.

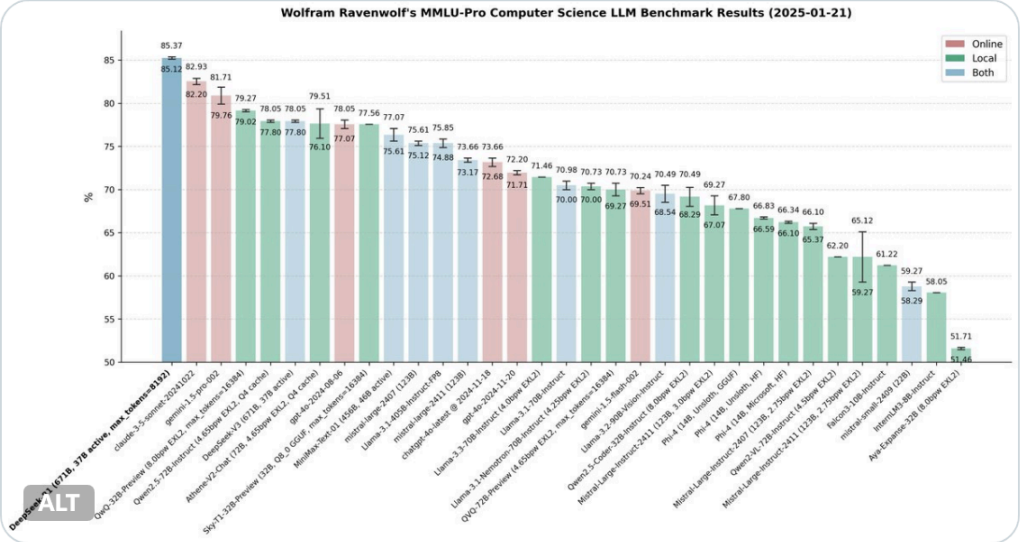
Wolfram Ravnwolf, consultor de AI, postou em seu X os resultados de testes com os modelos **DeepSeek V3** e **R1** apresentando bom desempenho em diversos *benchmarks*, mas é importante entender o contexto dos testes realizados.

Nos experimentos apresentados por Ravnwolf, que indicam que a **DeepSeek** se destaca em relação a outros modelos, as questões testadas eram, em sua maioria, matemáticas ou lógicas, que exigiam respostas diretas e únicas.

Wolfram Ravenwolf 🐺🐦‍⬛
@WolframRvnwlf

Yesterday was a historic day, and you all know why: The coronation of a new king... **DeepSeek-R1**! We finally got Sonnet at home, local o1 even. There is no moat, China is the GOAT - and DeepSeek is the real Open AI! 🚀

Seriously, look at that score! (Will add more variants soon.)



9:01 PM · Jan 21, 2025 · **19.4K** Views

Fonte: https://x.com/WolframRvnwlf/status/1881854573352001712

O resultado é impressionante, principalmente quando comparado o custo de treinamento do modelo que, segundo a **DeepSeek**, está 20x menor que o custo dos principais modelos do mercado. É importante ressaltar que no documento técnico da **DeepSeek,** parte dos custos com pesquisas prévias e algumas etapas do processo de treinamento não foram incluídas - **e essas etapas não incluídas podem ter custos significativos.**

Essa eficiência, no entanto, pode gerar algumas preocupações entre os investidores, principalmente ao perceberem que modelos como os da **DeepSeek** não dependem dos chips avançados da NVIDIA e têm a capacidade de competir com os principais modelos do mercado. Essas preocupações podem também existir ao notarem que o modelo do Google gemini-1.5-pro-002 está em 3º lugar no *ranking* da imagem, rodando com os chips TPU *(Tensor Processor Unit)*, que são criados pelo próprio Google.

Isso levantou dúvidas sobre a real necessidade dos chips mais modernos da NVIDIA, sua dominância e a precisão das estimativas que indicam uma demanda crescente por eles.

Modelos como a **DeepSeek**, com custos reduzidos, poderiam oferecer uma alternativa eficiente e mais acessível, provocando uma reavaliação dos investimentos de longo prazo em AI.

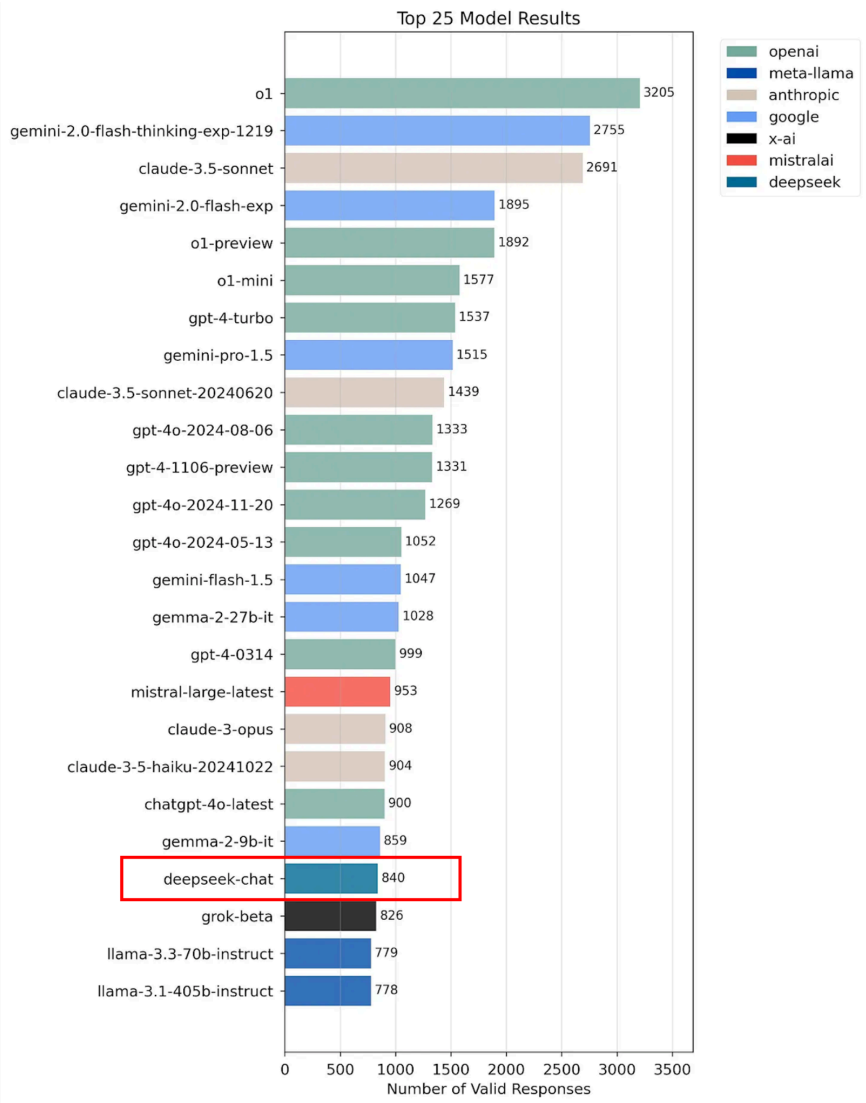
Até que ponto a troca entre eficiência e complexidade impacta a capacidade dos modelos DeepSeek V3 e R1 de escalar para aplicações mais exigentes? Será que, ao priorizar o custo, o modelo compromete sua aplicabilidade em contextos mais amplos que demandam criatividade na interação com o humano? Ou mesmo suacapacidade de dentro da mesma pergunta, orientada pelo usuário, gerar múltiplas respostas diferentes?

Custo x Complexidade

Um *benchmark* privado e importante onde o **V3** e o **R1** tiveram baixa performance foi o AidanBench. Este teste avalia a capacidade de grandes modelos de linguagem (LLMs) gerarem ideias criativas, mantendo coerência com o contexto e seguindo instruções de maneira confiável em cenários mais abertos e próximos ao mundo real:

> *"AidanBench avalia grandes modelos de linguagem (LLMs) com base em sua capacidade de gerar ideias novas em resposta a perguntas abertas, com foco em criatividade, confiabilidade, atenção ao contexto e seguimento de instruções.* ***Diferentemente de benchmarks com respostas definitivas, o AidanBench avalia os modelos em tarefas mais abertas e próximas de cenários do mundo real****. Ao testar vários LLMs de última geração, ele mostra uma baixa correlação com benchmarks existentes, oferecendo uma visão mais detalhada sobre seu desempenho em cenários abertos."*
>
> Fonte: https://openreview.net/pdf?id=fz969ahcvJ (tradução livre)

Abaixo o resultado do último teste feito:



Fonte:
https://x.com/aidan_mclau/status/1872444303974543859?s=46&t=niG-Z9FrMrJD0ib4JDlS9g

O AidanBench não avalia o desempenho geral de um modelo de linguagem, mas sim sua habilidade de gerar respostas variadas e manter a coerência com o contexto. Nos testes, o **DeepSeek** mostrou um problema de repetição nas respostas, ou seja, ele não consegue gerar muitas variações para a mesma pergunta ou se aprofundar em um tema como demandado em interações com chatbots ou agentes co-pilotos. Além disso, o modelo fica mais lento com o tempo, demorando mais para gerar respostas.

Alguns usuários que estão usando a versão *open source* em suas próprias estruturas também têm dificuldade em atingir a velocidade mínima necessária para gerar *tokens* rapidamente. Isso pode ser causado pelo fato de o modelo ter

sido treinado com uma grande quantidade de dados em chinês, o que dificulta a geração de conteúdo em inglês.

Outro ponto importante é que o **DeepSeek** apresentou falhas nas questões de segurança dos modelos. Os testes mostraram que estes modelos têm um desempenho abaixo do esperado em comparação com modelos similares, o que levanta questões sobre sua confiabilidade em ambientes com dados sensíveis, como aqueles encontrados nos setores de saúde e finanças.

Nenhum *benchmarking* deve ser levado a ferro e fogo, e, como os próprios autores comentam, novos testes estão sendo feitos e novos ajustes ao *benchmarking* estão acontecendo. Em algumas discussões, usuários comentam a necessidade de estudos adicionais antes de qualquer conclusão e avaliam se há falhas (*mode collapse*) em um problema no qual o modelo se fixa em produzir saídas limitadas, ignorando a diversidade dos dados reais. Isso poderia ser causado pelas estratégias econômicas usadas no processo de treinamento.

Conclusão

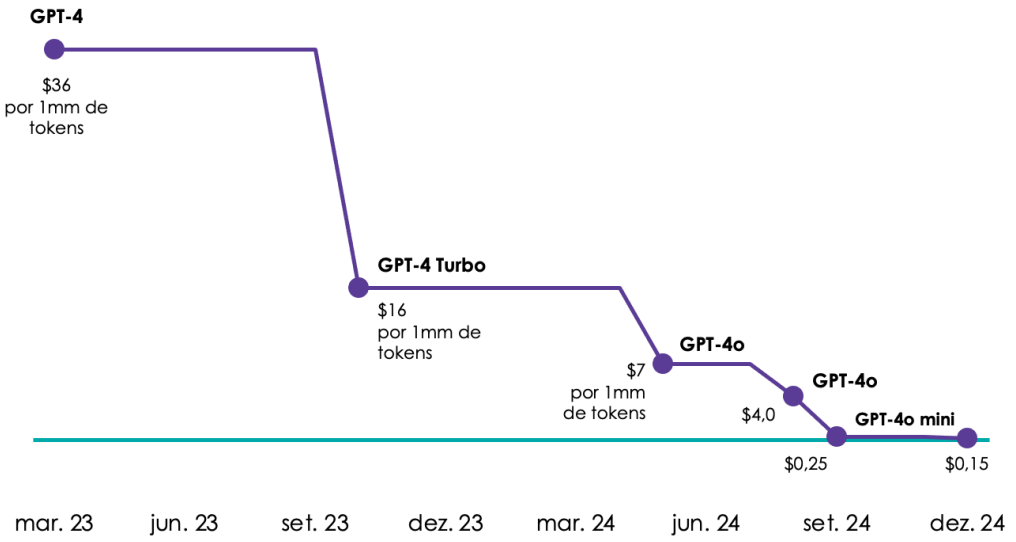
Nossos estudos e análises indicam que, apesar de o **DeepSeek V3** e o **R1** apresentarem bom desempenho a um custo significativamente mais baixo, não há evidências até o momento de que tenha o potencial disruptivo necessário para transformar completamente o caminho atual do mercado de forma tão impactante quanto sugerido.

O **V3** e o **R1** se destacam pelo custo reduzido e pelo uso de técnicas inovadoras, que permitem respostas rápidas e precisas, mas que já são conhecidas e estudadas por todas as empresas de AI. Ao mesmo tempo, apresentam limitações claras: tendem a gerar respostas mais curtas, perdem mais consistência do que outros modelos em interações repetidas e, até o momento, **demonstram queda de desempenho em ambientes de alta escala ao longo do tempo**.

Comparativamente, **modelos como o gpt-4o-latest oferecem respostas mais completas e de qualidade superior, mostrando-se mais adequados para tarefas complexas, mesmo sendo mais caras**. Por outro lado, o **DeepSeek V3** é bem posicionado para aplicações simples e de baixo custo, como perguntas e respostas objetivas, revisões básicas de códigos de programação e perguntas matemáticas.

O menor custo do **DeepSeek V3** e **R1** pode ser visto como uma tendência acelerada de redução nos custos totais observada antes mesmo do lançamento desses modelos, como mostra a figura abaixo.

Custo de processamento por 1 milhão de tokens no GPT



Fonte: Aster Capital e CB Insights

Em nossa avaliação, o **V3** e o **R1** ocupam um nicho importante de mercado, oferecendo uma solução viável para demandas operacionais mais simples.

Ainda assim, acreditamos que o futuro da inteligência artificial reside no desenvolvimento de agentes (*Agentic AI*) capazes de realizar tarefas mais complexas. Para que esses agentes prosperem, é necessário que os modelos subjacentes consigam manter coerência em interações prolongadas, lidar com janelas extensas de *tokens* e realizar inferências sofisticadas.

Em nossa visão, o **DeepSeek V3** é uma boa alternativa para cenários simples, mas não representa uma mudança de paradigma no mercado de que seguirá evoluindo com soluções cada vez mais avançadas, otimizadas e voltadas para aplicações exigentes e diversificadas.

Disclaimer

As informações contidas neste material expressam opiniões pessoais do gestor para fins meramente informativos e não constituem qualquer tipo de aconselhamento de investimentos, não devendo ser utilizadas para estar finalidade. As declarações contidas neste documento não devem ser interpretadas como fatos ou verdades absolutas. Nenhuma informação contida neste material constitui uma solicitação, oferta ou recomendação para compra ou venda de cotas de fundos de investimento, ou de quaisquer outros valores mobiliários. A Aster Capital não comercializa nem distribui cotas de fundos de investimento ou qualquer outro ativo financeiro. Aos investidores é recomendada a leitura cuidados do formulário de informações complementares, da lâmina de informações essenciais, se houver, e o regulamento antes de investir. Este material não é direcionado para quem se encontrar proibido por lei a acessar as informações nele contidas, as quais não devem ser usadas de qualquer forma contrária a qualquer lei de qualquer jurisdição. Fundos de Investimento não contam com a garantia de administrador do fundo, do gestor da carteira, de qualquer mecanismo de seguro ou, ainda, do Fundo Garantidos de Créditos - FGC. A rentabilidade divulgada já é líquida das taxas de administração, de performance e dos outros custos pertinentes ao fundo, mas não é líquida de impostos. Para avaliação da performance do fundo de investimentos, é recomendável uma análise de, no mínimo, 12 (doze) meses. A rentabilidade obtida no passado não representa garantia de rentabilidade futura. Fundos de investimento utilizam estratégias com derivativos como parte integrante de sua política de investimento. Tais estratégias, da forma como são adotadas, podem resultas em significativas perdas patrimoniais para seus cotistas, podendo, inclusive, acarretar tanto perdas superiores ao capital aplicado, quanto uma consequente obrigação do cotista de aportar recursos adicionais para cobrir o prejuízo do fundo. Fundos de investimento podem realizar aplicações em ativos financeiros no exterior. Os fundos podem ainda estar expostos a uma significativa concentração em ativos de poucos emissores, com riscos daí decorrentes. Não há garantia de que os fundos multimercados terão o tratamento tributário para fundos de longo prazo. A Aster Capital, seus administradores, sócios e funcionários não se responsabilizam pela publicação acidental de informações incorretas, e isentam-se de responsabilidade sobre quaisquer danos resultantes direta ou indiretamente da utilização das informações contidas neste material. O conteúdo deste material é para uso exclusivo de seu destinatário e não pode ser copiado, reproduzido, publicado, retransmitido ou distribuído, no todo ou em parte, por qualquer meio e modo, sem a prévia e expressa autorização, por escrito, da Aster Capital. A utilização das informações aqui contidas se dará exclusivamente por conta e risco do seu usuário.

astercapital