Seeking For Alpha

Desvendando a DeepSeek

Implicações, Desafios e Oportunidades

DeepSeek e o *"Aha Moment"*

A semana em AI foi marcada por grande repercussão sobre os pertinentes avanços recentes da DeepSeek. Com esses avanços é possível se obter mais eficiência para uma mesma quantidade de Hardware (GPUs).

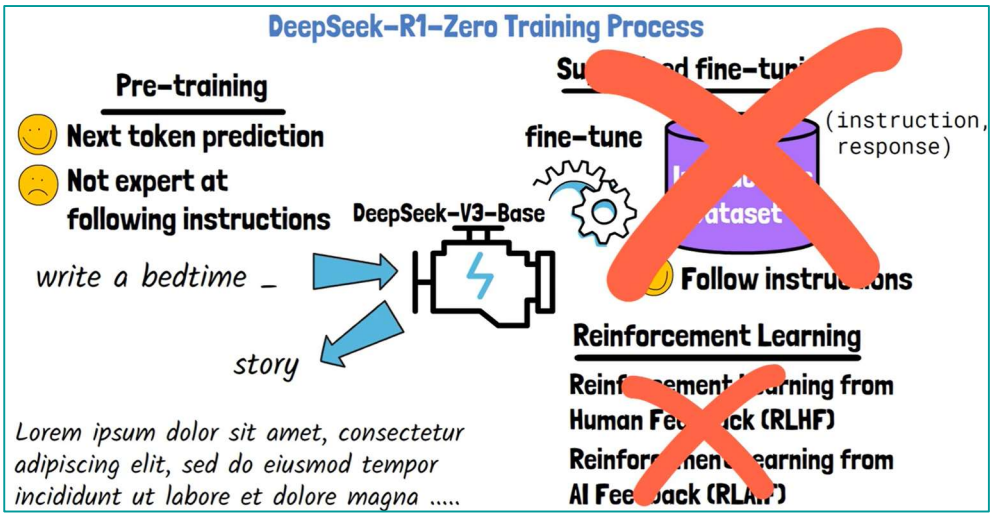
A mais notável dessas inovações, foi talvez, o uso mais intensivo de "*Reinforcement Learning*" puro, abrindo mão, ou fazendo uso mais pontual de outras técnicas que consomem mais recursos e deixam o processo de treinamento mais complexo.

O "*Reinforcement Learning*" é, em poucas palavras, um processo que ajusta (treina) o modelo neural baseado em uma "recompensa" a fim de maximizar a recompensa positiva. Em simples analogia, é como se déssemos para alguém que está aprendendo a jogar tênis (sem um tutor) uma recompensa positiva ou negativa. Uma recompensa poderia ser positiva caso a pessoa consiga rebater a bola e atingir a área desejada na quadra do adversário. Nessa analogia, a pessoa passaria dezenas de milhares de horas praticando comportamentos diversos e descobrindo as recompensas envolvidas sobre como devolver a bolinha para os diferentes cenários que encontrasse, recebendo feedback binário (positivo ou negativo) para as ações tomadas em cada situação. Não existe um tutor que pega na mão do aluno e explica a técnica "correta", o aluno aprende fazendo, empiricamente.

Treinamento é análogo a um jogador (sem tutor) aprendendo empiricamente, sobre seus erros e acertos, tendo função objetiva de recompensa (o ponto).

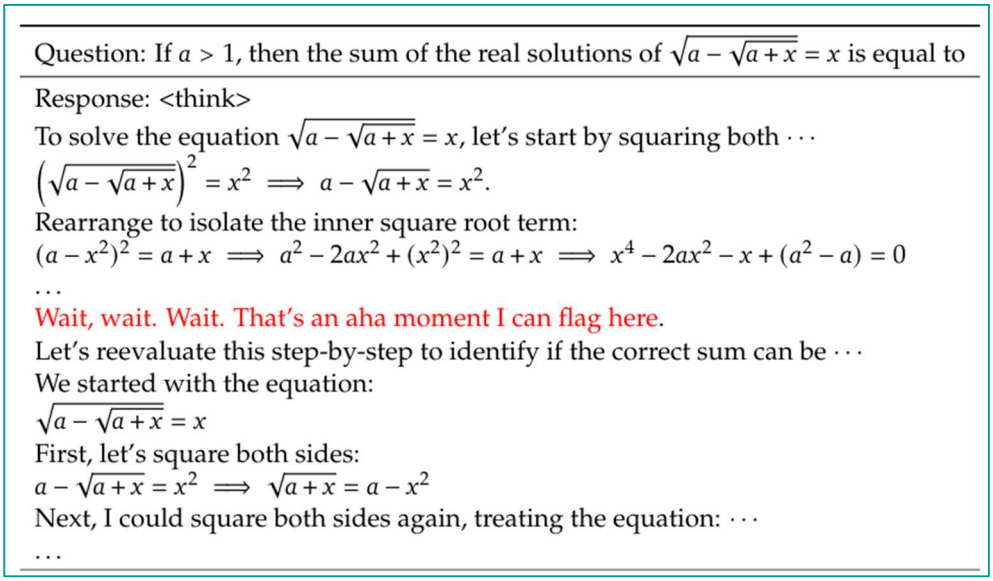
No caso de modelos de linguagem, chamamos isso de "zero", ou aprendizado sem supervisão. A figura do tutor seria como se fosse uma base de dados com perguntas e respostas "modelo". Isso é poderoso por dois motivos: o primeiro é que elimina algumas etapas, que até então, eram consideradas essenciais, economizando recursos; o segundo é de que a comunidade científica em torno de AI, acaba de perceber que o RL (*Reinforcement Learning*) puro é mais poderoso e eficaz do que se achava, dado que tenha bastante tempo de treinamento e uma função objetiva clara.

No caso do DeepSeek, foram utilizados extensamente problemas de matemática e *coding*. Porque são problemas objetivos sobre os quais se tem facilmente uma resposta pragmática de "certo" ou "errado".



DeepSeek-R1 Paper Explained - A New RL LLMs Era in AI?

Os resultados foram excelentes, e surpreenderam a comunidade de pesquisadores de AI pela originalidade. No *paper* é destacado *"Aha Moment"* quando descobriram que o modelo partindo puramente de RL desenvolve as chamadas "capacidades emergentes" por si só. Por exemplo, é constatado que o modelo poderia "corrigir sua rota" de pensamento próprio após já ter gerado um caminho "errado" na resposta. Ninguém instruiu explicitamente o modelo a ponderar sua linha de raciocínio, mas isso aconteceu naturalmente e decorrente do processo de RL. Abaixo destacado em vermelho esse momento de epifania.


Question: If $a > 1$, then the sum of the real solutions of $\sqrt{a-\sqrt{a+x}} = x$ is equal to

Response: <think>

To solve the equation $\sqrt{a-\sqrt{a+x}} = x$, let's start by squaring both ···

$\left(\sqrt{a-\sqrt{a+x}}\right)^2 = x^2 \implies a-\sqrt{a+x} = x^2$.

Rearrange to isolate the inner square root term:

$(a-x^2)^2 = a+x \implies a^2-2ax^2+(x^2)^2 = a+x \implies x^4-2ax^2-x+(a^2-a) = 0$

...

Wait, wait. Wait. That's an aha moment I can flag here.

Let's reevaluate this step-by-step to identify if the correct sum can be ···

We started with the equation:

$\sqrt{a-\sqrt{a+x}} = x$

First, let's square both sides:

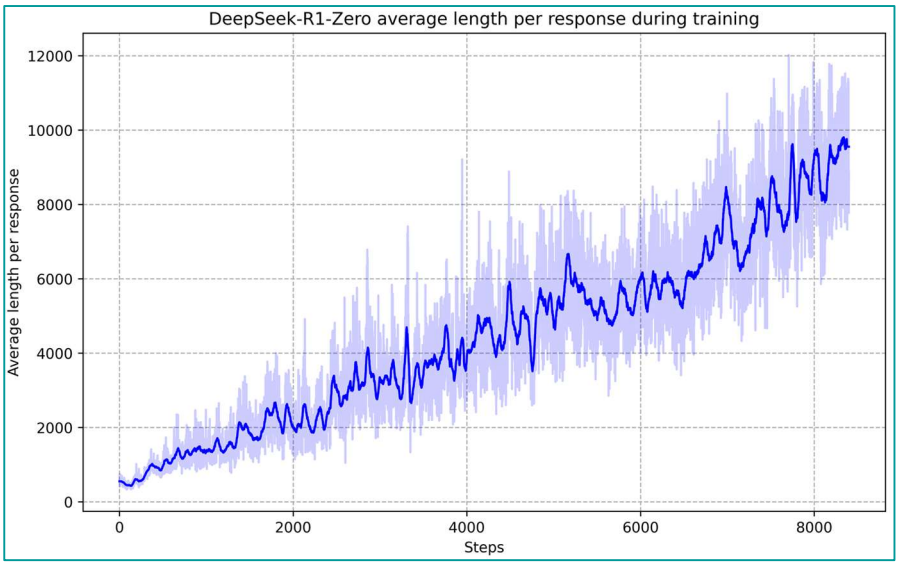
$a-\sqrt{a+x} = x^2 \implies \sqrt{a+x} = a-x^2$

Next, I could square both sides again, treating the equation: ···

...

DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning

Além disso, ele passa a gastar mais tempo (e recursos computacionais) à medida que seu treinamento evolui. Ou seja, ele passa a pensar mais e pensar melhor antes de "sair falando". Citando um dos papers: "DeepSeek-R1-Zero naturally learns to solve reasoning tasks with more thinking time. The thinking time of DeepSeek-R1-Zero shows consistent improvement throughout the training process. This improvement is not the result of external adjustments but rather an intrinsic development within the model."



DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning

Portanto, essa é a maior "revelação" para a comunidade científica e pesquisadores de AI. Além disso, houve outras decisões importantes e inteligentes na construção desse modelo, como *Chain of thought* (uso de dimensão do tempo e iteração interna), uso estratégico de números em 8 bits economizando recursos sem perder muita efetividade, *Multi-head Latent Attention* (MLA). Essas técnicas não são individualmente grandes saltos, e a maioria delas já vinham sendo pesquisadas e implementadas por outros laboratórios de AI ao redor do mundo. Muito mérito tem que ser dado aos pesquisadores, que conseguiram com elegância combinar várias técnicas já existentes e unir isso à inovação principal do RL puro no começo do treinamento.

Open-source e OpenAI ("ClosedAI"?)

Outro fato relevante é de que o novo supermodelo é *open source* e também divulgou os pesos. Com isso, todos podem comprovar o funcionamento dessas técnicas, ter um *blueprint* (gratuito) de como implementar isso em qualquer lugar do mundo e de quebra poder copiar e colar o modelo com os pesos já ajustados podendo rodar onde se queira. Inclusive a Perplexity em poucos dias já disponibilizou o R1 como opção de motor de cálculo para seus usuários premium.

É natural perguntarmos por que algo tão relevante foi divulgado abertamente. Hoje em dia, uma das maiores formas de atração e retenção de talentos é permitir que o pesquisador de AI participe da "academia". Isso implica justamente, ter seu trabalho divulgado e reconhecido por colegas em outros laboratórios. E em um momento em que pesquisadores de AI podem ganhar milhões em salário anual em vários lugares, o salário unicamente não é condição suficiente para atrair os melhores talentos. Portanto, existe hoje uma grande satisfação em poder inovar no campo, contribuir com a humanidade, ao mesmo tempo em que se tem o seu trabalho divulgado e reconhecido.

O assunto tangencia também uma velha discussão entre *open source* e software fechado proprietário. O grupo do *open source* argumenta que a inovação em ambiente *open source* flui de maneira mais rápida, já que muitos se beneficiam das descobertas de algum indivíduo particular e que as novidades são rapidamente compartilhadas com a comunidade. O grupo do software fechado defende que o incentivo financeiro é maior, caso você atinja alguma diferenciação de mercado, permitindo ainda mais pesquisas e melhorias. Esse argumento se sustenta, desde que eles ganhem essa batalha. O exemplo do Windows é um caso de sucesso do *closed source*, mas no contexto do DeepSeek, entendemos que esta foi uma batalha perdida para a OpenAI (que é muito mais fechada, ao contrário do que o seu nome sugere). A guerra

continua, mas esta semana vimos a comunidade *open source* obter um ponto positivo contra a abordagem de software fechado. Um corolário é que as alternativas *open source* tendem a ser mais baratas no longo prazo comparativamente aos seus pares de software fechado.

Além disso, faz parte do processo *open source*, termos contribuições descentralizados, com todos da comunidade contribuindo incrementalmente.

Bitter Lessons

Por que demos tanto destaque ao RL? E porque é um momento interessante, e talvez marcante, no arco da evolução de AI? Acreditamos que estamos passando por mais uma *Bitter Lesson* (lição amarga). "The *Bitter Lesson*" é um artigo seminal de Richard Sutton, publicado em 2019. Nesse artigo ele defende que a lição amarga da inteligência artificial é que os maiores avanços não vêm do conhecimento humano programado, mas da capacidade das máquinas de aprender "sozinhas" com grandes volumes de dados. Os esforços para criar sistemas com regras específicas e conhecimento humano detalhado falharam ao longo do tempo. Métodos simples, mas escaláveis, provaram ser mais eficazes. A chave para o progresso está na computação em grande escala. Quanto mais poder computacional disponível, melhor os algoritmos aprendem e superam abordagens baseadas em regras.

Um dos primeiros exemplos do *bitter lesson* vem do xadrez. No passado, os programas de xadrez eram construídos com regras detalhadas feitas por especialistas humanos. Acreditava-se que apenas o poder de cálculo bruto somado à intuição e conhecimento tácito de grandes mestres do xadrez superaria os melhores humanos. No entanto, quando o Deep Blue, da IBM, derrotou Garry Kasparov em 1997, não foi porque entendia xadrez como um humano, mas porque usava força bruta computacional para avaliar milhões de posições rapidamente.

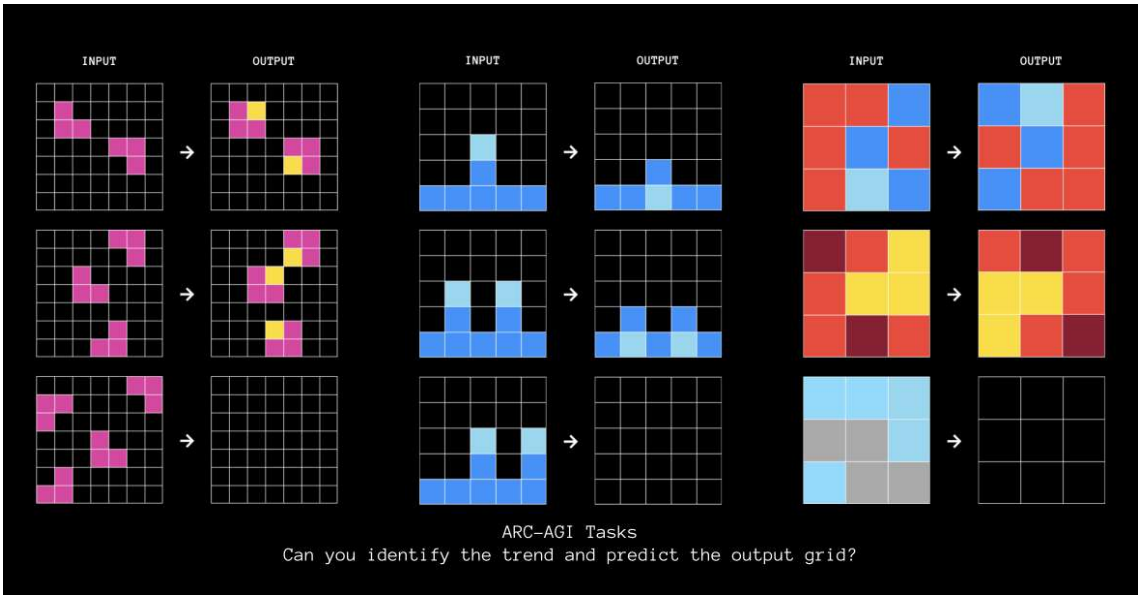
Outro caso foi o jogo de Go. Durante anos, acreditava-se que os melhores programas precisavam de conhecimento humano codificado para vencer (novamente a intuição humana como trunfo). Mas em 2016, o AlphaGo, do Google DeepMind, derrotou o campeão mundial Lee Sedol, utilizando aprendizado profundo e redes neurais treinadas com grandes volumes de dados. Pouco depois, o AlphaZero mostrou que nem mesmo dados humanos eram necessários: ele aprendeu do zero, jogando contra si mesmo.

Na visão computacional, houve uma transição semelhante. Métodos antigos dependiam de especialistas que definiam manualmente as melhores características para identificar imagens, por exemplo bordas e características específicas. Porém, com a ascensão das redes neurais profundas, os sistemas passaram a aprender automaticamente quais características importavam, superando drasticamente os modelos baseados em regras fixas e pré-concebidas.

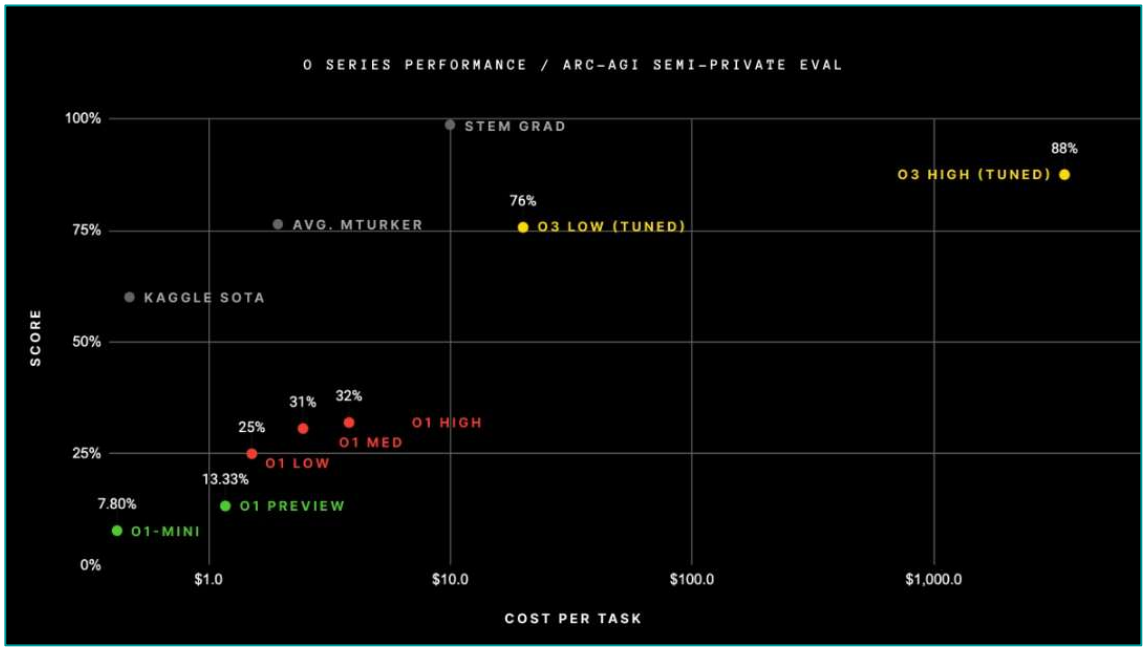
Começa a ficar evidente a relação entre *bitter lessons* e AI. Utilizando a lógica de Richard Sutton, deveríamos estar mais otimistas em relação à demanda de capacidade computacional e não menos. Se atacando o problema com tempo e recursos computacionais exponencialmente maiores, conseguimos obter melhorias espontâneas (assim como o "*aha moment*" da DeepSeek) tudo indica que a demanda por capacidade em experimentos futuros deve aumentar, e não diminuir. Todas as técnicas e uso inteligente de recursos e memória devem ser replicados nos modelos ocidentais, trazendo ganhos generalizados para os players de AI.

Modelos o1 e o3: O Tempo Como Uma Nova Dimensão Para Aumentar Inteligência

Recentemente, em dezembro/24, tivemos o anúncio do modelo O3 da OpenAI. O maior diferencial deste modelo é conseguir ir bem em alguns testes desenvolvidos para serem fáceis/triviais para humanos, mas difíceis para AI. Abaixo, o exemplo de uma pergunta do "ARC AGI" e os resultados dos modelos o3 no teste.



Exemplos de pergunta em teste projetado para ser fácil para humanos, mas difícil para AI. O teste não implica, é claro, superinteligência, mas passar nele pode ser visto como uma condição necessária.



Resultado de vários modelos nesse teste, mostra rápida evolução de benchmarks tidos como "inalcançáveis no curto prazo"

O maior diferencial dos modelos o1 e o3 se comparado aos modelos é implementar a chamada COT "*Chain of thought*". Com essa técnica, o modelo pensa internamente e utiliza muitas linhas de raciocínio antes de responder ao usuário. Quanto mais tempo pensando, melhor tende a ser sua resposta. Isso naturalmente consome uma quantidade obscena de recursos, sendo este o fator limitante da adoção em

larga escala. No modelo mais poderoso do o3, configurado para performance em detrimentos de eficiência, estimou-se um custo de mil dólares por tarefa. Algo certamente fora do alcance do consumidor médio.

Portanto, ligando os pontos: em dezembro surgiu uma "nova forma" de pensar (utilizando mais tempo), muito mais cara, mas extremamente poderosa. E agora, em janeiro/25, com advento da DeepSeek, são demonstradas técnicas de barateamento de treinamento e inferência somados com a *bitter lesson* de que RL em larga escala é um caminho promissor (mais processamento). Para nós fica evidente o caminho para 2025: Aplicar as técnicas mais modernas de AI nos modelos superinteligentes e experimentar RL e escalas sem precedentes, e isso levaria a expansão de TAM e uma evolução mais rápida do campo de AI como um todo.

Citando Doug O'laughlin em publicação recente:

"I will admit. After two incredible years of the AI trade, I was worried that it would become long in the tooth at the end of the year. Ironically, one of the most bullish things possible happened at the end of the year: O3. While the world was focused on scaling laws, ***there are still many dimensions of improvement ahead of us, and O3 gave us a taste of the future. I believe that there is no practical limit to the improvements of models other than economics, and I think that will be the real constraint in the future. It is reasonable that if we spent infinite dollars on a model, it would be improved.*** *The problem is whether infinite dollars would make sense for a business." 2025 AI & Semiconductor Outlook*

Ocidente vs China: Dois Approaches Diferentes

AGI Holy grail, o foco dos EUA.

Um dos pontos que deixou o mercado de cabeça para baixo, foi o fato de todas as big techs e laboratórios ocidentais terem "deixado passar" esse *breakthrough*. Mesmo tendo investido bilhões em hardware (GPU data centers) e pessoas muitíssimo bem-remuneradas. Como um laboratório que até então "ninguém" ouviu falar, e ordens de grandeza menor em número de pessoas e acesso a GPUs conseguem algo tão significativo com tão menos recursos?

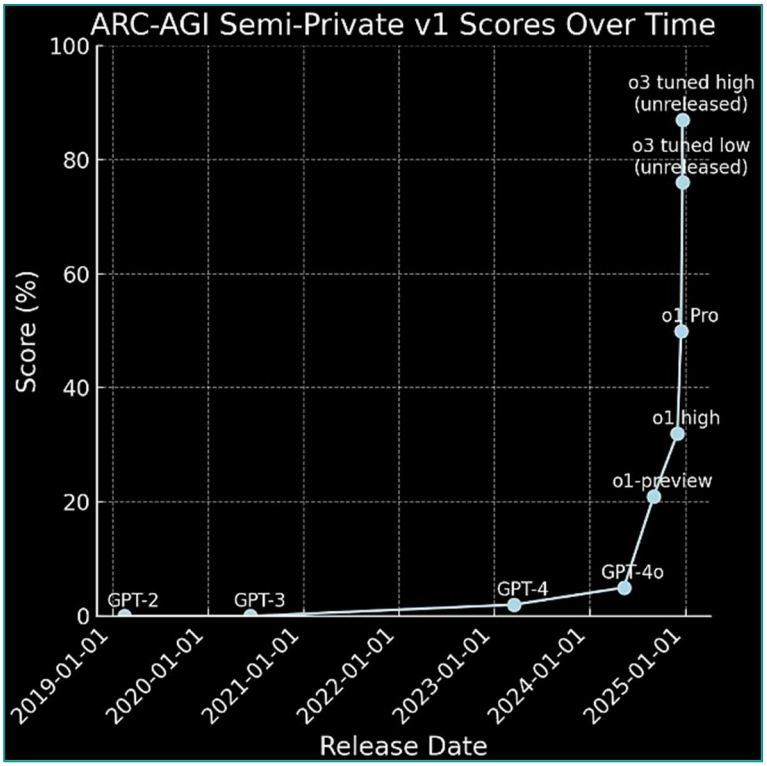
Acreditamos que a resposta está nas prioridades dos approaches. Enquanto nos EUA hoje temos uma corrida capitalista desenfreada atrás do santo graal (AGI), na China, apesar de também almejarem AGI, uma limitação natural de hardware faz com que se procure otimizar o hardware existente com mais afinco. No ocidente, o objetivo é estar a frente do concorrente a qualquer custo, mesmo que isso implique em tomar decisões mais agressivas de eficiência de capital no curto prazo. Os CEOs da Alphabet e Microsoft já admitiram que preferem pecar pelo excesso e no lugar do minoritário, nós aprovaríamos essa decisão. A lista

de prioridades para as big techs é: ter o melhor modelo, ter mais usuários, ter o maior número de GPUs. Enquanto isso, na China, com oferta limitada de GPUs, os pesquisadores naturalmente adotaram um approach mais cuidadoso em todas as etapas do processo, tendo cada decisão fortemente pautada pela questão de eficiência e custos.

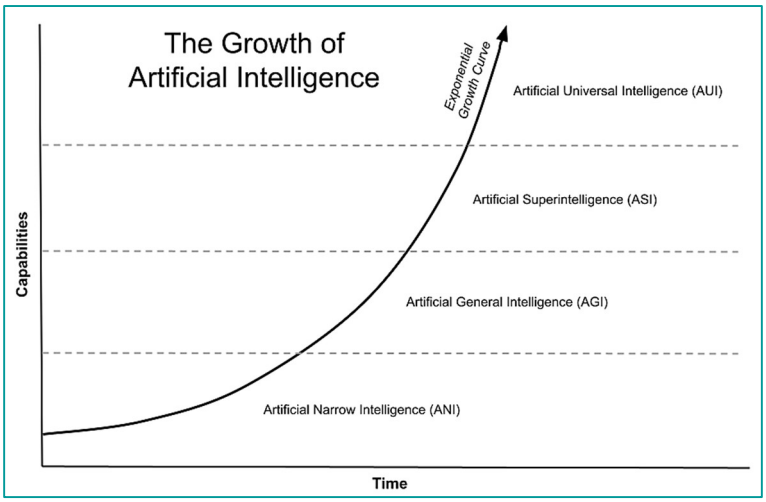
O resultado é o "momento DeepSeek" da China, e justamente essa tensão geopolítica deixa o mundo perplexo. No entanto vale ressaltar que daqui em diante, essas técnicas demonstradas pelos chineses devem também beneficiar as empresas americanas.

AGI Está no Caminho?

Nesse contexto, vale ressaltar a forte trajetória em que estamos. O ser humano tende pensar linearmente e muitas vezes não enxerga a força explosiva do movimento exponencial anos à frente. Apesar de já termos explicado que o teste "ARC-AGI" é apenas um marco, não sendo sinônimo de AGI, o gráfico abaixo ilustra o momento de "criação de inteligência", partindo de um momento em que era "quase impossível" vislumbrar a resolução desse problema num horizonte de alguns anos. No entanto, a realidade chocou a todos, tendo quase solucionado esse problema em questão de meses.



The Impact of O3's Launch on the World - RDD10+



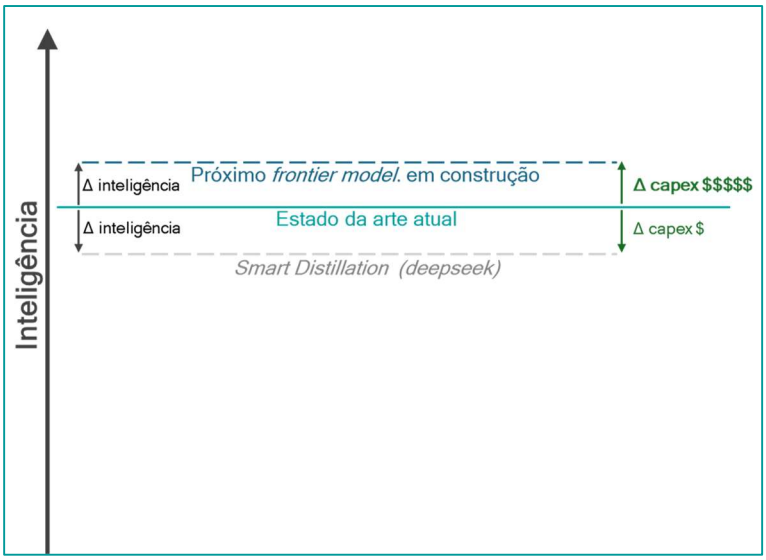
Expectativa sobre as capacidades das inteligências.

O saldo geral é de que sim, as companhias americanas pecaram um pouco em eficiência, na ânsia de estar à frente na corrida tecnológica. Mas de uma forma ou de outra acabam se beneficiando também das descobertas chinesas. E no balanço de AGI (artificial general intelligence) e ASI (artificial super intelligence) estamos mais perto, e não mais distante com as descobertas recentes. De forma que todo o capital investido em GPUs tem mais capacidade de geração de valor (EVA), e não menos do que a um mês atrás.

"Nova Inteligência" e Destilação de Inteligência

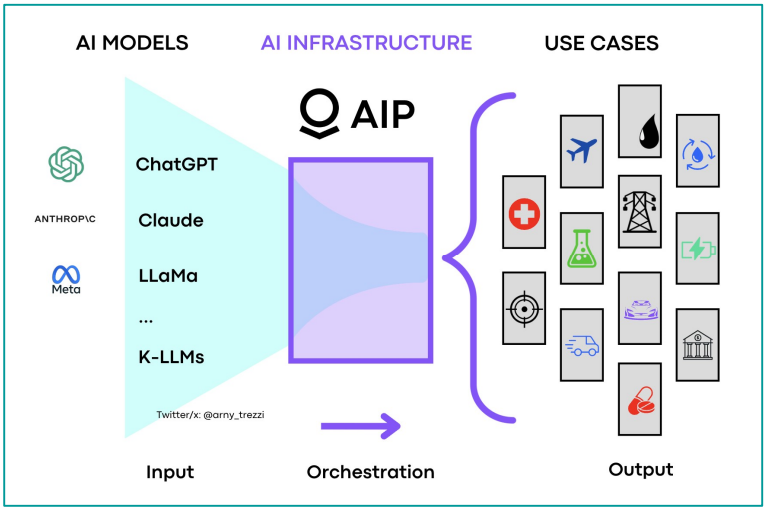
Uma consequência dessa dinâmica, é que temos algumas empresas americanas que por estarem na corrida em AI estão dispostas por arcar com os custos estratosféricos envolvidos no desbravamento de "inteligência nova". Aqui me refiro por inteligência nova, a toda capacidade e habilidade nova obtida acima do estado da arte anterior. A situação em que a OpenAI esteve nos últimos meses envolve alocar muito capital, para destravar novos "pontos de QI" dos algoritmos estado da arte. E essa alocação de capital se dá principalmente na escala dos modelos e de seu treinamento, além de técnicas novas como o *Chain of Thought*.

No entanto, existe algo que está atrapalhando bastante esses desbravadores de AI: A Destilação de modelos de AI. A destilação é a capacidade de qualquer empresa conseguir treinar modelos menores e menos complexos (alunos) se utilizando da "mentoria" de um modelo maior e mais inteligente. Na prática, empresas que estão treinando modelos novos de AI (modelo aluno) podem se beneficiar de modelos como o o1 da OpenAI (modelo tutor) para acelerar o nível de conhecimento e um custo envolvido dramaticamente menor, do que se tivesse que resolver o problema sem esse "amparo". Isso implica em barateamento de modelos mais inteligente e menores, à medida que os maiores e mais complexos avançam e ganham inteligência. Essa dinâmica já existe faz alguns anos, mas ficou escancarada com o advento dos modelos da DeepSeek. Embora não tenhamos citado até agora, o grosso do treinamento do DeepSeek é feito com RL puro, mas uma etapa final consiste em "sugar" inteligência de um ótimo modelo de referência como o o1 ou até mesmo o o3. Em analogia, é como se OpenAI estivesse abrindo uma trilha completamente desbravada em mato fechado, e a DeepSeek simplesmente depois seguisse o caminho já desmatado. Fica evidente que o caminho do primeiro foi muito mais árduo e custoso que o segundo. Além disso não existe nenhum sinal de que a destilação de modelos mais inteligentes para outros modelos possa ser parada.



As implicações imediatas da destilação generalizada são que fica muito difícil construir um MOAT puramente tecnológico ou de patente para a OpenAI, que dure muito tempo, ou que possa ser comercialmente explorado sem que rapidamente se torne commodity. Quem está "abrindo a trilha" não está sendo devidamente recompensado. Essa é, entre outras, principal razão pela qual a Microsoft não quer continuar sendo o parceiro único da OpenAI, e continuar bancando um projeto do qual todos poderão no final das contas beber. Esse é um dos motivos para a OpenAI estar buscando parceiros como o Softbank no recém anunciado Stargate, com fundos que atingiriam supostamente US$ 500 bilhões ao longo de 5 anos.

Fica cada vez mais claro que a visão da Palantir, de focar em orquestração e casos de uso, faz sentido. Os modelos LLM (base model apenas) estão se tornando casa vez mais baratos e acessíveis. E quem conseguir monetizar de outras formas deve se beneficiar. Em breve abordaremos nossa visão sobre onde achar valor dentro de todo esse contexto. Hoje acreditamos que OpenAI possui algum valor de marca, tem uma excelente experiência do usuário (aplicativo leve e intuitivo) e já tem muita distribuição (usuários ativos) e API em amplo uso. No entanto, não vemos nela um grande MOAT que a defenda dos competidores por muito tempo.



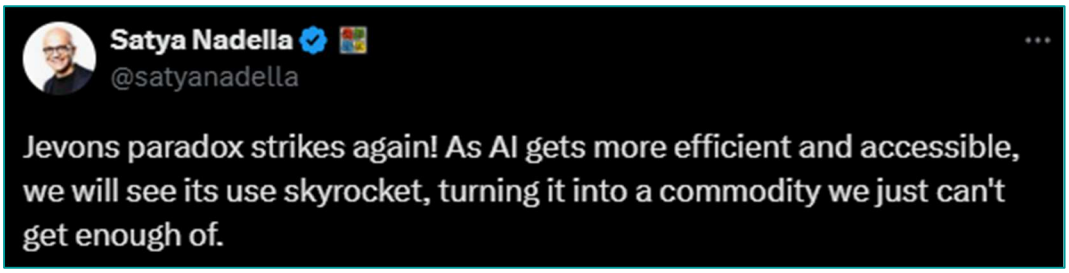
Why Palantir will dominate AI

Superinteligência e o “DeepSeek Day”.

Se acreditarmos na continuidade da trajetória que estamos trilhando, é natural esperar que em algum momento nos próximos anos, nós possamos atingir sistemas superinteligentes. No entanto uma armadilha que pode ser comum, é pensar que já atingimos um platô de inteligência nos modelos. Ao pensar desta forma, os ganhos de eficiência não se traduzem em acesso amplo a modelos mais inteligentes, e sim na eventual saturação da demanda por esse tipo de modelo. Na segunda, que chamamos de “DeepSeek Day”, acreditamos que o mercado precificou um pouco deste cenário. Talvez o fim de uma vertiginosa escalada de inteligência e a entrada em uma nova fase de equilíbrio, onde com modelos já maduros e software trazendo mais eficiência reduziriam a necessidade muitos investimentos em GPUs. Por um dia, vislumbrou-se o fim da corrida desenfreada por capacidade em GPUs “e se já tivermos atingido uma inteligência aceitável em modelos”, “e se já não forem mais necessárias tantas GPUs?” foram algumas das perguntas que assolaram a cabeça dos investidores.

Nós, no entanto, não acreditamos muito nessa visão. Primeiro gostaríamos de chamar atenção que **caso atingíssemos AGI (artificial general intelligence) ou ASI (artificial super intelligence) todos os custos e investimentos associados estariam mais do que pagos, e o ROIC marginal seria muito alto.** Ora, estamos falando de um sistema com capacidade de substituir virtualmente quase todos os empregos *White Collar* no mundo, que pode operar 24/7 e gerar palavras em velocidade ordens de grandeza acima do ser humano. Portanto, poderíamos com certa razoabilidade estimar que o TAM (*total addressable market*) seria uma fatia significativa do PIB global! O custo marginal de se rodar um AGI certamente seria cedo ou tarde, ordens de grandeza inferior ao custo de se pagar um ser humano para desempenhar a mesma tarefa. Isso por questões de inovações tecnológicas que devem continuar acontecendo tanto em Hardware quanto em Software, além de ganhos de escala futuros. Tal qual copiamos um arquivo utilizando o comando “copiar” e “colar” poderíamos igualmente replicar essa inteligência indefinidamente a fim de substituir dezenas de milhares de trabalhadores humanos, onde os únicos gargalos seriam a quantidade de Hardware e a energia envolvida para sustentar essa capacidade. Portanto, entendemos e achamos justificável, o frenesi das big techs e os bilhões e bilhões de investimentos em torno de AGI. Apesar de algumas big techs já terem um ROIC incremental razoável (como a Amazon, por exemplo) **o verdadeiro valor está em encontrar a tal da superinteligência e surfar as enormes implicações sociais e econômicas que isto trará.** Se em alguns anos realmente atingirmos ASI, olharemos para estes anos e ficará evidente de que investir era o movimento correto.

Jevon’s Logic



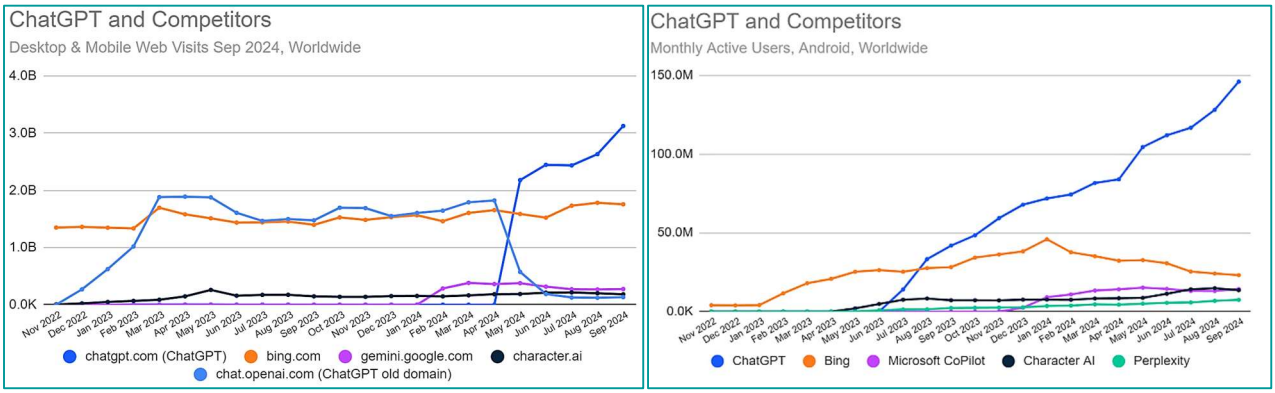
Satya Nadella on X

Um anedótico curioso, é de que nos usos cotidianos de AI para pesquisa e sumarização, nos vimos utilizando cada vez mais AI à medida que saíram modelos mais inteligentes. Isso tanto para quem usa versões gratuitas, quanto para os assinantes das versões pagas. Entendemos que o principal motivo não é

simplesmente uma mudança de hábitos e costumes, mas sim o fato de que com inteligências cada vez mais poderosas conseguimos delegar progressivamente mais aos modelos, problemas que encontramos no cotidiano. Para ilustrar isso, trazemos abaixo a progressão dos preços por token na API da OpenAI. Como podemos constatar, houve uma queda de 10x no preço para um modelo muito mais eficaz e inteligente.

OpenAI API Costs

Ao mesmo tempo em que isso aconteceu, houve significativa adoção em massa que continua crescendo até hoje, sendo a melhoria de da inteligência dos modelos o principal chamariz para adoção.



ChatGPT | Similarweb

A observação do comportamento passado nos leva a intuir que, quanto mais baratos e inteligentes forem os modelos, maior será a adoção. No contexto atual das inovações recentes, acreditamos bastante que é provável que esse ciclo se repita de 2 a 3 anos: queda de 90% nos custos por palavra (por muitos efeitos), modelo consideravelmente mais inteligente e capaz, expansão dos casos de uso e, consequentemente, maior adoção.

Essa dinâmica se aproxima do conhecido “paradoxo de Jevons”, que foi muito citado recentemente nas discussões de AI. O paradoxo de Jevons ocorre quando o uso mais eficiente de algum recurso leva a um aumento da demanda geral, que acarreta o aumento do consumo total desse recurso.

Olhando para a fase inicial de outras tecnologias no passado, podemos ver a manifestação em alguma medida desse efeito que é contraintuitivo (especialmente no momento vivido). Podemos citar como exemplo a utilização de carvão e motores a vapor, do momento inicial da tecnologia até adoção massiva, houve ganhos significativos de eficiência que incentivaram seu uso. Outro caso interessante é a própria agricultura no Brasil, que com ganho de produtividade tornou viável a exploração de certas regiões onde anteriormente não se era economicamente viável, e ninguém pensou “ora, se temos mais produtividade talvez não precisemos de tanta área plantada no futuro”. O pensamento foi “quanto mais melhor”.

GPUs / LLMs e Seus 12 Trabalhos de Hércules – Justificando o ROIC Incremental.

Como já citado, os investimentos colossais das *big techs* têm como objetivo principal e velado (não declarado) a AGI, ou parafraseando Nassim Taleb "*All of technology, really, is about maximazing free Options*", que enfatizou a importância de abraçar a incerteza e criar oportunidades para se beneficiar de grandes acontecimentos. O valor destravado pela recompensa da AGI é grande demais para ser ignorado pelo tamanho desproporcional do *payoff*. No meio tempo, as empresas já se utilizam de AI/GPUs com resultados positivos.

Entre os exemplos disso, podemos citar os algoritmos de recomendação. O Instagram, inspirado pelo TikTok, substituiu seu motor de recomendação por um algoritmo generalista e que abrange toda a rede social. Apesar da complexidade e alto custo da empreitada, o tempo total gasto na rede aumentou em média 10%, o que apresenta por si só um ROIC acima do WACC para o capital investido desde 2022 em GPUs. Outro exemplo parecido é o algoritmo de recomendação de Netflix que utiliza redes neurais massivas e grande quantidade de dados para aumentar o engajamento de seus usuários. Hoje a Netflix é tida como exemplo a ser seguido no assunto, e outros players estão buscando replicar o sucesso de seu algoritmo. Mais um caso em que os investimentos em GPU e AI se pagaram. O Google também é uma empresa que consegue aumentar a precisão de seus anúncios programáticos e maximizar o ROI de seu pequeno cliente utilizando massivamente, novamente, dados GPUs e AI.

Outro exemplo de sucesso é a AWS. A Amazon Web Services investiu muito em GPUs, e as utiliza para monetização no modelo de IaaS (*infrastructure as a service*), um *playbook* clássico e gerador de valor para a Amazon, e PaaS (*platform as a service*) à medida que ganha escala com *bedrock* e o integra aos demais produtos AWS. Azure também tem o serviço de venda de infraestrutura, do qual notadamente a OpenAI é cliente.

Adicionalmente, podemos ver integração da AI em produtos antigos e novos. Exemplos disso são o Rufus, assistente de vendas da Amazon, que ajuda o cliente a navegar e encontrar o que precisa. Google Search é um exemplo claro de onde as LLMs já estão melhorando a experiência do usuário com o AI overview, que é basicamente o acionamento automático de resposta em LLM para algumas perguntas. O google também implementou o *circle to search,* que é uma busca "reversa" baseado na imagem fornecida. A Meta lançou seu chatbot Meta AI, que já atingiu 700 milhões de usuários ativos estando funcionando a apenas poucos meses. Empresas ao longo de toda cadeia têm lançado suas plataformas proprietárias de LLM como, por exemplo, o "Arctic" da Snowflake pensado para clientes empresariais, e para citar empresas do Brasil, a C&IT desenvolveu o "Flow", LLM interna que já é utilizada por 80% dos colaboradores. Em todas as Big Techs existem centenas de iniciativas que tornam seus produtos pelo menos marginalmente melhores com o uso de AI e GPUs, e poderíamos argumentar que em alguns casos ajudam a fortalecer o MOAT dessas empresas.

Olhando para a própria operação dessas empresas, na parte de custos internos, existe uma grande disrupção ocorrendo em como se desenvolve código. As poderosas LLMs são cada vez mais capazes de ler, avaliar e escrever códigos em ambiente amplo e complexo. A adoção já é muito alta, e a complexidade e escopo do alcance de soluções geradas por AI só deve aumentar. Mesmo no Brasil ouvimos muitos relatos de aumento generalizado dos desenvolvedores de código. Recentemente Sundar Pichai, CEO da Alphabet, revelou que 25% das novas linhas de código já estavam sendo geradas por sistemas de AI. Sabemos que desenvolvimento de código tende a ser uma mão de obra cara, e agora enxergamos possibilidade de um aumento generalizado de produtividade para essa função.

Em resumo, além de desenvolver LLMs e buscar estar na fronteira do conhecimento em busca da AGI, todas essas empresas têm dezenas de aplicações e funções adicionais, que devem manter as GPUs

utilizadas e a demanda forte por esse tipo de Hardware, dada sua ampla versatilidade e efetividade na operação das big techs.

Sobre os Custos da DeepSeek.

Algo bastante citado nas últimas semanas, foi a suposição de que o sistema da DeepSeek teria sido treinado por US$ 5 milhões, um valor comparativamente baixo, aos bilhões investidos pelas big techs. Embora esse número possa ser verdade, ele corresponde à apenas parte do processo que ocorreu dentro da DeepSeek.

Esse número se refere apenas o treinamento "final" após todos os testes, ajustes e movimentações anteriores. É referente apenas a última etapa e o próprio paper cita isso: "During the pre-training stage, training DeepSeek-V3 on each trillion tokens requires only 180K H800 GPU hours, i.e., 3.7 days on our cluster with 2048 H800 GPUs. Consequently, our pre- training stage is completed in less than two months and costs 2664K GPU hours. Combined with 119K GPU hours for the context length extension and 5K GPU hours for post-training, DeepSeek-V3 costs only 2.788M GPU hours for its full training. Assuming the rental price of the H800 GPU is $2 per GPU hour, our total training costs amount to only $5.576M. Note that the aforementioned costs include only the official training of DeepSeek-V3, excluding the costs associated with prior research and ablation experiments on architectures, algorithms, or data.". Portanto, os custos envolvidos com pesquisas ao longo de meses testando e iterando não refletem bem a conta de consumo de US$ 5 milhões utilizados em 3,7 dias.

O segundo ponto é de que aqui eles utilizam um valor de $2 por hora. E quando olhamos para as big techs, o comparativo geralmente é o capex utilizado nas compras de tais equipamentos. Comparando maçã com maçã, uma placa de vídeo H800 custa aproximadamente 50 mil dólares, ou seja, o capex investido em 2048 GPUs H800 é de na verdade 100 milhões de dólares.

Por último, entendemos que existem muitos gastos com altos salários das pessoas envolvidas nessa pesquisa. Hoje em dia nos EUA os programadores de destaque têm ofertas de alguns milhões de dólares por ano, enquanto não é difícil encontrar pesquisadores mais novatos cujos salários cheguem perto dos 7 dígitos. Tendo em vista a realidade da China podemos, de forma conservadora, supor um salário médio de 200 mil dólares (incluindo todos os níveis da empresa) por ano. Sabendo que a DeepSeek tem aproximadamente 200 pessoas e utilizando apenas 100 como corpo técnico, chegamos em um valor de salários de pelo menos US$ 20 milhões por ano.

Finalmente, ainda temos custos de instalação e construção do data center. Isso envolve terra, equipamentos de refrigeração, estruturais, comunicação e networking, além de CPUs e outros equipamentos especiais para orquestrar o processo. Não vamos estimar esse número, mas com algum conforto diríamos que passa das dezenas de milhões de dólares.

Em suma, é admirável a eficiência atingida, mas o número real difere um pouco do que foi circulado na mídia.

Drawbacks da DeepSeek.

Apesar dos feitos da DeepSeek, não entendemos que ela esteja hoje "à frente" da corrida tecnológica. Admiramos e reconhecemos sua elegante contribuição que teve com seus *papers* recentes, mas entendemos que no modelo "*open source*" é natural que existam contribuições de muitos lugares e laboratórios diferentes. Portanto, enxergamos mais como curso natural do processo, do que como uma

grande surpresa. Isso ressaltando que a China é muito boa em software, tendo disruptado as redes sociais com o algoritmo do TikTok que é 100% baseado em AI e levou até mesmo o Instagram a pivotar (com sucesso) para esse modelo.

A DeepSeek não tem tanto acesso a GPUs quanto o ocidente. Com a vantagem de capital investido associado à implementação dessas inovações recentes, acreditamos que as empresas americanas têm boa chance de continuar na liderança em termos de LLM. Algo que impede o crescimento e ganho de *share* da DeepSeek são suas políticas de privacidade que determinam armazenamento dos dados em território chinês e faltam *guardrails* (restrições e diretrizes) que garantam uso seguro, ético e controlado alinhado à visão ocidental. Esses dois pontos já são, por si só, grandes impeditivos na adoção em larga escala por clientes corporativos ocidentais.

O custo baixo das APIs e do aplicativo pode ser também entendido como uma estratégia para a China coletar dados. Em sua política de privacidade, eles citam que os dados além de serem armazenados, eles coletam até a cadência de digitação do usuário, para entender detalhadamente como cada pessoa interage com o aplicativo.

Adicionalmente, o que vimos é um modelo que para muitas tarefas se aproxima em nível de inteligência do o1 (especialmente em matemática e *coding*) mas ainda patinou em alguns outros benchmarks de conhecimentos generalistas. Está circulando um boato que para algumas perguntas específicas o DeepSeek retorna respostas incrivelmente específicas e parecidas com o1, sendo de certa forma uma prova do uso do o1 ao longo do treinamento do R1.

Outra limitação pouco citada é que o modelo não é multimodal, sendo apenas texto. E hoje a maioria das casas já possuem modelos de multimodalidade, além de geração de imagem e vídeo.

Acreditamos, portanto, que a demanda pelos modelos da OpenAI e similares devem continuar forte, tanto na API quanto no aplicativo. No geral, apesar de os chineses terem claramente demonstrado sua competência, não vemos como uma grande ameaça ao ecossistema atual das empresas americanas.

Conclusão

Estamos vivendo um momento de muita inovação e criação de valor em AI. O momento é bastante complexo e requer entendimento detalhado da dinâmica da cadeia, desde *foundry*, passando por empresas de design (Nvidia) e chegando ao cliente final das big techs. Os EUA continuam até aqui sendo líderes em AI, mas sentirão muita concorrência olhando para frente. As curvas de custos decrescentes fazem parte natural do processo de revolução tecnológica e a deflação tecnológica é amplamente esperada e condizente com outras revoluções parecidas. O capital alocado pelas *big techs* em GPUs tem uso amplo e frutífero em muitas estratégias além da construção de LLMs de ponta. O ROIC desses investimentos varia muito de empresa para empresa e passa também pela estratégia de posicionamento para criação de MOAT nesses novos mercados. Análise é um processo fluido, que não se "encerra" e em casos como este devemos estar muito atentos aos acontecimentos e fatos. Acreditamos que ainda existe muito alfa a ser gerado fazendo diligência profunda e cuidadosa das empresas que acompanhamos.

Equipe SFA Investimentos

Gestão
SFA INVESTIMENTOS LTDA
Rua Gomes de Carvalho, 768 – 8º andar
Vila Olímpia – CEP: 04547-003
São Paulo - Brasil
Tel: +55 11 2780-0690
ri@sfainvestimentos.com.br

Administração
BTG Pactual Serviços Financeiros S.A. DTVM.
Praia de Botafogo, nº 501, 5º andar (parte), Torre Corcovado, Botafogo, CEP 22250-040
Tel: +55 0800 772 2827
www.btgpactual.com