Como a China criou o modelo de IA DeepSeek e chocou o mundo

Políticas governamentais, generosos investimentos e graduados em engenharia de prompt, tem ajudado empresas chinesas a desenvolverem avançadas LLMs.

Por Gemma Conroy & Smriti Mallapaty

A tecnologia chinesa da startup DeepSeek tem deixado o mundo tech em maus lençóis com o lançamento de dois grandes modelos de linguagem (LLMs) que rivalizam com dominantes ferramentes desenvolvidas por grandes empresas de tecnologias americanas, mas criadas com uma fração de custo e poder computacional.

Em 20 de Janeiro a companhia sediada na cidade de Hangzhou lançou o DeepSeek-R1, um modelo parcialmente "racional" que consegue resolver problemas científicos de maneira similar ao o1, OpenAI's mais avançado LLM, qual a companhia, sediada em São francisco, Califórnia, revelou no último ano. E no inicio desta semana, a DeepSeek lançou outro modelo, chamado Janus-Pro-7B, qual consegue gerar imagens a partir de textos de prompt muito semelhante ao OpenAI's DALL-E 3 e Stable Diffusion (outro modelo LLM), feito pela Stabiility AI em Londres.

Se a performance surpreendeu muitas pessoas fora da China, pesquisadores do país dissertam que o sucesso da startup já era esperado e combina com as ambições governamentais ao se tornarem lideres em Inteligência Artificial (IA) no mundo tudo.

Foi inevitável que uma companhia tal qual a DeepSeek emergiria da China, dado o grande investimento de capital de risco em empresas de desenvolvimento de LLMs e muitas pessoas cujas quais possuem doutorados em ciência, tecnologia, engenharia ou outros campos relacionados a exatas, incluindo IA, disse Yunji Chen, um Cientista da Computação que tem trabalhado em chips de IA no instituto de Tecnologia Computacional da acadêmia chinesa de ciências, Beijing. "Se não fosse a DeepSeek, alguma outra LLM chinesa poderia fazer um excelente trabalho".

De fato, existe. Em 29 de Janeiro, a tech Behemonth Alibaba lançou seu mais avançado LLM até o momento, Qwen2.5-max, qual a companhia disse performar melhor que o próprio DeepSeek's V3, outro LLM com firme lançamento em Dezembro. E na última semana, a Moonshot AI e ByteDance lançaram novos modelos racionais, Kimi 1.5 e 1.5-pro, qual companhias afirmaram conseguir performar melhor que o1 (OpenAI's) em alguns testes de referência.

Prioridade Governamental

Em 2017, o governo chines anunciou suas intenções para o país, ao se tornarem lideres mundiais em IA em 2030. Seu Encerramento industrial com o completar da avanço em IA "tais tecnologias e aplicações alcançaram um nível de liderança mundial " até 2025. Desenvolvimento de prompt de "talento IA" se tornaram uma prioridade. Até 2022, o ministério da educação chinesa tinha aprovado 440 universidades a oferecer graduações em especialização em IA, de acordo com uma reportagem vinda do Centro de Segurança e tecnologias emergentes (CSET) na universidade de Geórgia em Washington DC. Naquele ano, a China proporcionou mais da metade da liderança de pesquisas em IA, enquanto os EUA relataram apenas 18% de acordo ao Think Tank Macropolo em Chicago, Illinois

A DeepSeek provavelmente beneficiada por investimentos governamentais em educação e desenvolvimento de talentos em IA, qual incluiu inúmeras bolsas de estudos, subsídios de pesquisas e parcerias entre acadêmias e industrias, disse Marina Zhang, uma cientista política pesquisadora na universidade de Tecnologia de Sydney na Austrália que está focada em inovações na China. Por exemplo, ela acrescenta, iniciativas de apoio estatal tais quais os de laboratório de engenharia nacional para tecnologias de Deep Learning (aprendizado de maquina profundo) e aplicações, quais são lideradas pela companhia de tecnologia Baidu em Beijing, que tem treinado milhares de especialistas em IA.

Números exatos da força de trabalho da DeepSeek são difíceis de definir, mas seu fundador Liang Wenfeng disse a mídia chinesa que a companhia tem recrutado graduados e estudantes de doutorados do mais alto ranking de universidades chinesas. Alguns membros de liderança da empresa são jovens de menos de 35 anos e tem presenciado o crescimento da China como uma potência tecnológica, disse Zhang. "Eles estão profundamente motivados por uma autoconfiança em inovação".

Wenfeng aos 39 anos, é para ele mesmo um jovem empresário, graduado em Ciência da computação pela Universidade de Zhenjiang, um líder no instituto em Hangzhou. Ele é cofundador da empresa Hedge onde alçou grandes voos por mais de uma década atrás e se estabeleceu na DeepSeek em 2023.

Jacob Feldgoise, que estuda talentos de IA na China no CSET, disse que as políticas nacionais que promovem um eco sistema de desenvolvimento de modelos para IA's, ajudou de certa forma empresas como a DeepSeek, em termos de atração de financiamentos e talentos.

Mas apesar do crescimento em cursos de IA's em universidades, Feldgoise disse que não está claro como muitos estudantes estão se graduando com ênfase em bacharelados em IA e se eles estão sendo instruídos de quais habilidades as empresas realmente precisam. Companhias de IA chinesas tem queixado nos recentes anos que "Os graduados vindos desses programas não tinham qualidade que eles esperado por", ele disse, algumas importantes empresas tem firmado parcerias com universidades.

"Eficiência sobre restrições"

Talvez o mais impressionante elemento do sucesso da DeepSeek, disseram cientistas, é que ela desenvolveu a DeepSeek-R1 e Janus-Pro-7B entre o controle de exportações do governo americano, qual tem bloqueado o acesso a China à chips avançados de computação para AI desde 2022. Zhang disse que a liderança da DeepSeek encorpora uma distinta abordagem chinesa para inovação, enfatizando a eficiência sobre restrições. Entretanto, a companhia não declarou detalhes específicos sobre quanto hardware eles usaram, ela adiciona.

A DeepSeek disse que usou aproximadamente 2 mil chips H800, construídos pela Nvidia para o treinamento DeepSeek-V3, um modelo lançado em Dezembro que superou a OpenAI's LLM GPT-4o, anunciado em Maio do último ano, em testes de referência. Por contraste, o modelo Llama 3.1 405B, um sofisticado LLM lançado em Julho desenvolvido pela Meta no Menlo Park, Califórnia, conta com mais de 16 mil dos mais avançados chips H100 da Nvidia. Em um post em 2020 no rede social WeChat, Hige-Flyer disse que ele tinha 10 mil chips antigos modelo A100 da Nvidia, qual provavelmente a DeepSeek teve acesso. A DeepSeek usou menos poder dos chips o que provavelmente fez seus modelos terem baixo custo. "O problema que nos tivemos que lidar nunca foi dinheiro, mas o banimento ao acesso de chips de alta performance" Wenfeng falou para a mídia Chinesa em Junho de 2024.

A DeepSeek se baseia em uma variedades de abordagens de aumento de eficiência de seus modelos. Por exemplo, suas implementações a uma "mistura de expertises" em sua arquitetura, um método de aprendizado de maquina (Machine Learning) que treina modelos mais rápidos do que técnicas convencionais, e poucos parâmetros. Isso possibilita a companhia a treinar seus modelos com menos chips, disse Chang Xu, uma Cientista da Computação da universidade de Sydney. Ela também usa uma versão inovada de um outra técnica, chamada de "Atenção latente de múltiplas cabeças" qual permite o modelo armazenar mais dados com uso mais baixo de memória.

Nesta semana, as mídias reportaram as sugestões que a OpenAI estava revendo afirmações de que a DeepSeek treinou seus modelos usando outputs vindos dos modelos da OpenAI. (A OpenAI está sendo processa por infringir propriedades intelectuais por organizações de notícias). A DeepSeek ainda não respondeu as afirmações feitas pela OpenAI. Até mesmo se isso for verdade, isso não diminuiria os feitos da DeepSeek em criar o R1, disse Lewis Tunstall, um pesquisador da plataforma Hugging face open-science, sediada em Berna, Suíça. Seus avanços estão na abordagem de aprendizado a inserção de habilidades "racionais" dentro de uma LLM, qual experimentos já tem gerado resultados, disse ele. Hugging Face é um projeto principal que tenta recriar o R1 do zero. "Eu espero que possamos aprender rapidamente se de alguma forma os dados sintéticos da OpenAI são ou não realmente necessários.

Os feitos da DeepSeek poderiam oferecer um plano para países que possuem ambição em IA porém sobre de carência de recursos e hardware para o treino massivo de LLMs usando padrões de abordagem do Vale do Silício, disse Yanbo, um cientista político pesquisador que tem foco em inovações na Universidade de Hong Kong. "Isso poderia ser um convite para a criação de largos exércitos de novos modelos", Disse ele.

doi: https://doi.org/10.1038/d41586-025-00259-0