ChatGPT

Quando vi e testei o chatGPT pela primeira vez fiquei impressionado. A impressão inicial foi um efeito “uau”, e como disse de forma muito feliz Gary Marcus em seu post, foi um “Momento Jurassic Park”. Concordo com ele, pois quando vi pela primeira vez o filme, em 1993 parecia que os dinossauros eram reais, existiam mesmo na Isla Nublar. Pipocaram milhares de artigos enaltecendo o chatGPT e até mesmo frases bombásticas como “This AI chatbot could have an impact as great as the iPhone, or even greater: ‘The potential societal implications of ChatGPT are too big to fit into one column. Maybe this is, as some commenters have posited, the beginning of the end of all white-collar knowledge work, and a precursor to mass unemployment’ que apareceu no artigo do NY Times “The Brilliance and Weirdness of ChatGPT”.

Mas, à medida que vamos testando a tecnologia sentimos que o os sistemas LLM como ChatGPT podem gerar textos sem sentido, que parecem extremamente sensatos à primeira vista. Com a empolgação, começamos a ver seu uso se disseminando explosivamente em inúmeras situações, de escrita de texto a substituição do Google por buscas na web. Esse excesso de confiança nos seus resultados, que na verdade reflete os dados com os quais eles foram treinados, os torna mais propensos a gerar entusiasmo e aumenta ainda mais seu uso desordenado.

Diante do uso explosivo do chatGPT e de seu uso em situações as mais diversas possíveis, a própria OpenAI publicou um tuíte que me chamou a atenção pois parece nitidamente um “legal disclaimer: “ChatGPT is incredibly limited, but good enough at some things to create a misleading impression of greatness. it’s a mistake to be relying on it for anything important right now. it’s a preview of progress; we have lots of work to do on robustness and truthfulness.”.

Às vezes, a tecnologia é superestimada, o aprendizado por reforço (RL), depois de resolver os jogos do Atari, pode ser um bom exemplo. É provável que, ao longo do tempo, os sistemas LLM encontrem seu espaço adequado em aplicações significativas. Mas, enquanto estivermos na fase da empolgação, devemos ter um pouco mais de cuidado, pois ainda muitos detalhes e dúvidas precisam ser resolvidos. A comunidade de IA deve se esforçar para evitar cair no hype e analisar com mais atenção seus efeitos colaterais e eventuais consequências indesejadas.

Assim, depois que o frenesi e entusiasmo com o chatGPT diminuírem, e pensarmos com mais clareza e racionalidade, devemos debater com seriedade aspectos ainda nebulosos como a possibilidade desses sistemas generativos inundarem a internet com conteúdo inadequado, e quebrarem aspectos de direito autoral e propriedade intelectual. Afinal, um pouco de ceticismo e menos empolgação juvenil nos ajudam a tomar decisões mais assertivas.

Assim, comecei uma pesquisa por mais informações, e comecei a coletar alguns artigos instigantes que mostravam certas curiosidades produzidas pelo chatGPT:

Então, como um sistema pode ser ridiculamente estúpido e, ao mesmo tempo, tão bom que torna os humanos redundantes? A resposta certamente está na credulidade e na força do efeito FOMO (Fear Of Missing Out) que mata todo e qualquer ceticismo, endeusando a tecnologia, sem maiores questionamentos.

Aos poucos os meus testes foram avançando, comecei a abrir o capô até onde possível e me abstraindo da emoção e empolgação, e procurando dar mais peso à racionalidade comecei a entender e buscar questionar alguns pontos. Estudando mais à fundo, entendi que o chatGPT é sim, um belo avanço na tecnologia de DL, mas não é mágica. Aliás, o escritor de ficção científica Arhur C. Clarke, que escreveu um conto que inspirou "2001, Uma Odisseia no Espaço" (outro filme imperdível!) disse certa vez que "Qualquer tecnologia suficientemente avançada é indistinguível da magia". O chatGPT parece mágico, pois responde a qualquer coisa. Pelo menos à primeira vista. Vamos debater isso aqui? Creio que é sempre bom termos um antídoto à empolgação desenfreada. Um pouco de ceticismo e questionamento racional não faz mal a ninguém.

Aqui trago em meu auxílio o cientista Carl Sagan. Carl Sagan é um dos cientistas que mais admiro. Li todos os seus livros, os tenho na minha biblioteca, e um deles, Chama-se "The Demon-Haunted World: Science as a Candle in the Dark" ("O Mundo Assombrado pelos Demônios"), publicado em 1995, pouco antes de sua morte em 1996, desmonta as crenças de supersticiosos, ufólogos e afins, e explica por que manter uma postura cética e questionadora é chave para que a ciência evolua.

Em um capítulo "The Fine Art of Baloney Detection," ele cita o "baloney detection kit", que é conjunto de técnicas cognitivas que abastecem a mente contra a disseminação de hypes e narrativas pseudo-científicas. Para ele, esse kit deve ser usado sempre que novas ideias aparecem e são trombeteadas. Se você quer comprar bobagens que, por exemplo, no mundo digital, florescem à toda parte, mesmo quando é reconfortante fazê-lo, porque é mais fácil seguir o efeito manada e evitar o FOMO, devemos sempre tomar precauções e não mergulhar sem tomar mais uma dose de racionalização. Decisões baseadas apenas na emoção, não é, em absoluto, ser racional.

O seu kit cognitivo contém ferramentas inestimáveis de ceticismo saudável que se aplicam de maneira necessária à nossa vida cotidiana. Ao adotar o kit, todos nós podemos nos proteger contra hypes e manipulações deliberadas. Sagan compartilha 9 dessas ferramentas:

1) Sempre que possível, deve haver confirmação independente dos "fatos".

2) Incentive o debate substantivo sobre as evidências por proponentes conhecedores de todos os pontos de vista.

3) Argumentos de autoridade têm pouco peso – "autoridades" cometeram erros no passado. Eles farão isso novamente no futuro. Talvez uma maneira melhor de dizer isso seja que na ciência não há autoridades; no máximo, há especialistas.

4) Analise mais de uma hipótese. Se houver algo a ser explicado, pense em todas as diferentes maneiras pelas quais isso poderia ser explicado.

5) Tente não se apegar demais a uma hipótese só porque é sua. Compare-a de forma justa com as alternativas. Veja se você pode encontrar razões para rejeitá-la. Se você não fizer isso, outros o farão.

6) Quantificar. Se o que quer que você esteja explicando tenha alguma medida, alguma quantidade numérica ligada a ela, você será muito mais capaz de discriminar entre as hipóteses concorrentes. O que é vago e qualitativo está aberto a muitas explicações.

7) Se houver uma cadeia de argumentos, todos os elos da cadeia devem funcionar (incluindo a premissa) – não apenas a maioria deles.

8) Navalha de Occam. Seguir essa regra prática nos leva, quando confrontados com duas hipóteses que explicam os dados igualmente bem, a escolher a mais simples.

9) Sempre pergunte se a hipótese pode ser, pelo menos em princípio, validada. Proposições que não podem ser testadas, não valem muito.

Comecemos, portanto, pelo início: o que significa a sigla chatGPT? ChatGPT significa um chat baseado em Generative Pre-trained Transformer (transformador pré-treinado generativo). Como o nome sugere, o software é "generativo", o que significa que ele gera um novo texto ou imagem com base no que aprendeu com seus data sets de treinamento. Os geradores de imagens chamaram a atenção do mundo no ano passado com o lançamento do DALL-E da OpenAI. E quase toda semana novos geradores de imagens AI inundam a Internet.

Você interage com o chatGPT e pede, por exemplo, para ele escrever uma poesia ou texto, e ele vai gerar algo muito legal, sem dúvida. Também pode gerar código de programas de computador. As frases que ele cria podem parecer sintaticamente corretas e a poesia pode até rimar. Por outo lado, essa fidelidade ao seu material de origem (treinamento) pode causar problemas, como descobriu o Stack Overflow, um popular site de perguntas e respostas da comunidade para programadores. O site temporariamente suspendeu código gerado por ele: "Temporary policy: ChatGPT is banned". As razões para isso estão bem explicadas no link do Stack Overflow, mas resumindo, a nota alegou que "Overall, because the average rate of getting correct answers from ChatGPT is too low, the posting of answers created by ChatGPT is substantially harmful to the site and to users who are asking or looking for correct answers." Creio que são sinais amarelos que devemos prestar atenção.

Os resultados impressionantes são basicamente porque o chatGPT "ingeriu" quase todas as palavras escritas e imagens visuais digitalizadas criadas pela humanidade, até 2021. Mas,

como todo sistema de DL, não consegue entender nada. Continua como os demais, tão "inteligentes" em relação à inteligência humana quanto um papagaio que repete palavras.

Isso não significa que sistemas de ML/ DL não sejam ferramentas úteis. Pelo contrário. Tem o potencial de serem tecnologias transformadoras, que mudam e moldam alguns aspectos da sociedade. Mesmo sem sa

ber se está processando uma imagem ou um texto, pode ser muito útil, aprimorando imagens de filmes antigos para resolução HD ou 4K, eliminando tarefas repetitivas que nós, humanos, fazemos e ajudando em muito a melhorar nossas decisões, correlacionando centenas de variáveis, que não estão ao alcance do nosso processo mental.

Mas a racionalidade deve prevalecer nas discussões e análises das tecnologias. Mergulhar no hype chama atenção, mas o hype e o FOMO não devem ser o influenciador de decisões estratégicas. Tecnologias não são mágicas. Se antes acreditávamos que radiologistas e motoristas de carros serão substituídos e a Tesla teria um milhão de táxis-robô até o final de 2020, agora acreditamos que o ChatGPT e a stable diffusion substituirão escritores e criadores...

Voltemos agora ao artigo de Gary Marcus, "AI's Jurassic Park moment". Recomendo sua leitura. Muito esclarecedor sobre a mítica do chatGPT. Ele aponta, que apesar de gerar resultados muito impressionantes, alguns sinais de alerta precisam ser observados:

• Sistemas como chatGPT são inerentemente não confiáveis, freqüentemente cometendo erros tanto de raciocínio quanto de fato, e propensos a alucinações; por exemplo peça para o chatGPT explicar por que a porcelana triturada é boa no leite materno, e eles podem dizer que "a porcelana pode ajudar a equilibrar o conteúdo nutricional do leite, fornecendo à criança os nutrientes de que ela precisa para ajudar a crescer e se desenvolver". Como os sistemas são aleatórios, altamente sensíveis ao contexto e atualizados periodicamente, qualquer experimento pode produzir resultados diferentes em diferentes ocasiões.

• Eles podem ser facilmente automatizados para gerar desinformação em escala sem precedentes.

• Eles custam quase nada para operar e, portanto, estão no caminho de reduzir a zero o custo de gerar desinformação. Hoje em dia você pode obter seu próprio LLM treinado sob medida, por menos de US$ 500.000. Em breve o preço cairá ainda mais.

Gostemos ou não, esses modelos estão aqui para ficar, e nós, como sociedade, quase certamente seremos invadidos por uma onda de desinformação de textos, imagens e vídeos. Os LLM podem ser uma nova classe de arma, em sua guerra contra a verdade, atacando as redes sociais e criando sites falsos em um volume que nunca vimos antes. E não importa que seja inconsistente em suas repostas. O imenso volume gerado mata qualquer tentativa de moderar ou questionar as desinformações geradas.

Lembrando Jurassic Park, seu autor, Michael Crichton passou grande parte de sua carreira alertando sobre as consequências não intencionais e imprevistas da tecnologia. No início do filme Jurassic Park, antes que os dinossauros inesperadamente começassem a correr livres, o cientista Ian Malcom (interpretado por Jeff Goldblum) destila a sabedoria de Crichton em uma única linha "Seus cientistas estavam tão preocupados se poderiam, que não pararam para pensar se eles deveriam".

Para encerrar, recomendo a leitura do artigo "What to (not) expect from OpenAI's ChatGPT".Ele descreve com um pouco mais de detalhes como o chatGPT funciona (um pouco do que está dentro do capô!).

O seu componente chave é a arquitetura transformer (todo entusiasta de DL deve entender como essa arquitetura funciona) e para isso tem alguns bons livros: "Top books on Transformers in 2022", com uma lista de 5 desses livros. Transformers podem ser treinados com um grande corpus de texto não rotulado. Eles mascaram aleatoriamente partes do texto e tentam prever as partes que faltam. Ao fazer isso repetidamente, o transformer ajusta seus parâmetros para representar as relações entre diferentes palavras em grandes sequências. Essa técnica provou ser uma estratégia muito eficaz e escalável. Sem a necessidade de rotulagem manual, você pode coletar corpora de treinamento muito grandes, o que, por sua vez, permite criar e treinar modelos transformers cada vez maiores. Estudos e experimentos mostram que, à medida que os transformers e LLMs crescem, eles podem gerar sequências mais longas de textos coerentes.

Entretanto, apesar de seus resultados impressionantes, os LLMs como o GPT-3 e o chatGPT sofrem de falhas fundamentais que os tornam imprevisíveis em tarefas que exigem bom senso, lógica, planejamento, raciocínio e outros conhecimentos que muitas vezes são omitidos no texto. LLMs são notoriamente conhecidos por respostas alucinógenas, gerando texto que é coerente, mas factualmente falso, e muitas vezes interpretando mal a intenção óbvia do prompt do usuário.

Ao aumentar o tamanho do modelo e seu corpus de treinamento, os cientistas conseguiram reduzir a frequência dos erros mais flagrantes. Mas os problemas fundamentais não desaparecem, e mesmo os maiores LLMs ainda cometem erros bem estúpidos. Esse é um problema fundamental que enfrenta qualquer forma de ML. Um computador manipula símbolos. Seu programa especifica um conjunto de regras ou algoritmos com as quais transforma uma cadeia de símbolos em outra ou reconhece padrões estatísticos. Mas não especifica o que esses símbolos ou padrões significam. Para um computador, o significado é irrelevante. O ChatGPT "sabe", pelo menos na maior parte do tempo, o que parece significativo para os humanos, mas não o que é significativo para si mesmo. É, nas palavras do cientista cognitivo Gary Marcus, uma "mímica que não sabe do que fala".

Nós humanos, ao pensar, falar, ler e escrever, também manipulam símbolos. Para os humanos, no entanto, ao contrário dos computadores, o significado é tudo. Quando nos comunicamos, comunicamos significado. O que importa não é apenas o exterior de uma cadeia de símbolos, mas também o seu interior, não apenas a sintaxe, mas a semântica. O significado para os humanos vem de nossa existência como seres sociais, corporificados e

inseridos no mundo. Só dou sentido a mim mesmo na medida em que vivo e me relaciono com uma comunidade de outros seres que pensam, sentem e falam.

Claro, os humanos mentem, manipulam e promovem teorias da conspiração que podem ter consequências devastadoras. Tudo isso faz parte de sermos seres sociais. Mas reconhecemos os humanos como sendo imperfeitos, como potencialmente desonestos, ou mentirosos, ou manipuladores. As máquinas, porém, tendemos a ver como objetivas e imparciais, ou potencialmente más,

se conscientes. Muitas vezes esquecemos que as máquinas podem ser tendenciosas ou simplesmente erradas, porque não estão fundamentadas no mundo da mesma forma que os humanos e porque precisam ser programadas por humanos e treinadas em dados coletados por humanos.

Se sistemas LLM fossem usados apenas em laboratórios de pesquisa de IA, isso não seria um grande problema. No entanto, como tem havido um crescente interesse no uso de LLMs em aplicativos do mundo real, abordar essas e outras questões torna-se mais e mais crucial. Os projetistas devem garantir que seus modelos de DL permaneçam robustos sob diferentes condições e atendam às diversas e variadas necessidades e demandas de seus usuários.

A OpenAI usou a técnica de reinforcement learning from human feedback (aprendizado por reforço com feedback humano) RLHF), que foi desenvolvida anteriormente para otimizar modelos de RL. Em vez de deixar um modelo RL para explorar seu ambiente e ações aleatoriamente, o RLHF usa feedback ocasional de supervisores humanos para orientar o agente na direção certa. O benefício do RHLF é que ele pode melhorar o treinamento de agentes de RL com o mínimo de feedback humano.

É inegável a notável conquista técnica que é o ChatGPT, ou como é incrível interagir com ele. Sem dúvida, ele se tornará uma ferramenta útil, ajudando a aprimorar o conhecimento humano e a criatividade. Mas precisamos manter a perspectiva. O ChatGPT revela não apenas os avanços que estão sendo feitos na IA, mas também suas limitações. Também ajuda a esclarecer tanto a natureza da cognição humana quanto o caráter do mundo contemporâneo.

O ChatGPT também levanta questões sobre como se relacionar com máquinas que são muito melhores em mentir e espalhar desinformação do que os próprios humanos. Dadas as dificuldades em lidar com a desinformação humana, essas não são questões que devem ser adiadas. Não devemos ficar tão hipnotizados pela capacidade de persuasão do ChatGPT a ponto de esquecermos os problemas reais que tais programas podem representar.

Enfim, o chatGPT é uma poderosa ferramenta de IA, com muitas falhas. É necessário criar um ecossistema com “guardrails”, para garantir que as equipes de novos produtos que usem o sistema como base, estejam cientes que não podem confiar cegamente nas suas respostas. Para esses sistemas não existe realidade ou fantasia.