FACULDADE DE TECNOLOGIA, CIÊNCIAS E EDUCAÇÃO

Graduação

GRADUAÇÃO EM BACHARELADO EM CIÊNCIAS DA COMPUTAÇÃO

As Leis de Escalabilidade de Modelos Neurais de Grande Escala: Implicações para Modelos Futuros

Johnathan Spoljaric de Oliveira
Professora Dra. Julyette Priscila Redling

Resumo

Diante do cenário atual de inteligência artificial, com o crescimento do investimento global em modelos de inteligência artificial de grande escala, promessas de que os mesmos vão continuar melhorando indefinidamente e que estamos próximos de desenvolver AGI (Artificial General Intelligence), surge a necessidade de verificar quão realista são as expectativas de crescimento da área. O objetivo geral deste trabalho é estabelecer quais são os fatores limitantes para o crescimento dos modelos de inteligência artificial utilizados hoje, analisar esses fatores e concluir se é viável esperar que os modelos atuais continuem progredindo ou se os fatores limitantes indicam um teto para o crescimento desses modelos. São utilizados artigos científicos que tratam dos assuntos abordados publicados majoritariamente pelas equipes de pesquisa que desenvolveram os “modelos de fronteira” atuais, constam também notícias de eventos de interesse da área. Foram encontradas os fatores mais comumente utilizados para otimizar o desenvolvimento de modelos de grande escala: o número de parâmetros do modelo, a capacidade computacional dedicada ao treinamento desse modelo e as o tamanho das bases de dados utilizadas. Dentre os três, o tamanho das bases de dados e a capacidade computacional se mostraram ser os fatores em que existem maiores problemas para superar, esses problemas foram estudados, demonstrando como eles podem vir a limitar o crescimento dos modelos de grande escala atuais. Enquanto que ainda há bastante potencial de crescimento para os modelos atuais, visto que o limite teórico do método de escala atual (a entropia semântica) não foi atingido, existem fatores que limitam esse crescimento muito mais relevantes do que os limites teóricos e que tornam a obtenção dos recursos necessários para a evolução dos modelos atuais extremamente mais custosos, de modo que é mais provável que o limite do crescimento dos modelos será a restrição dos recursos necessários e o crescimento dos custos dos mesmos, principalmente se não for encontrada uma justificativa econômica de efeito imediato para os investimentos imensos necessários.

Palavras-chave: inteligência artificial, modelos de grande escala, leis de escala, limitações de escala.

Abstract

Given the current landscape of artificial intelligence, with the growing global investment in large-scale AI models, promises that they will continue improving indefinitely and that we are close to developing AGI (Artificial General Intelligence), the need arises to assess how realistic the growth expectations for the field are. The overall objective of this work is to identify the limiting factors for the growth of the artificial intelligence models used today, analyze these factors and determine whether it is viable to expect current models to continue improving or if the limiting factors suggest a ceiling for their growth. Scientific articles addressing the topics discussed, primarily published by the research teams that developed the current “frontier models,” are used, along with news about events of interest in the field. The most commonly used factors for optimizing the development of large-scale models were identified: the number of parameters in the model, the amount of compute dedicated to training the model, and the size of the datasets used. Of the three, dataset size and compute were found to be the factors with the greatest challenges to overcome. These issues were studied, demonstrating how they may limit the growth of current large-scale models. While there is still significant potential for growth in current models, given that the theoretical limit of the current scaling method (semantic entropy) has not been reached, there are factors that limit this growth far more significantly than the theoretical limits, making the acquisition of the resources required for the evolution of current models considerably more expensive, so that the ceiling for these models is more likely to be the restriction of the necessary resources and the rising costs, especially if an immediate economic justification for the massive investments required is not found.

Keywords: artificial intelligence, large-scale models, scaling laws, limitations of scaling.

Introdução

Inteligência artificial é um campo que é objeto de fascinação de muitos, desde ficção científica até pesquisas acadêmicas; com o crescimento do investimento global em modelos de inteligência artificial de grande escala, promessas de que os mesmos vão continuar melhorando indefinidamente e que estamos próximos de desenvolver AGIs (Artificial General Intelligence) – sistemas altamente autônomos que superam o desempenho humano na maioria das áreas de trabalho economicamente valiosas – surge a necessidade de verificar quão realista é essa possibilidade.

Neste contexto apresenta-se o problema de pesquisa: quais são os parâmetros principais utilizados para otimizar os modelos atuais e como esses parâmetros influenciam a possibilidade de crescimento desses modelos?

O objetivo geral deste trabalho é estabelecer quais são os fatores limitantes para o crescimento dos modelos de inteligência artificial utilizados hoje, analisar esses fatores e concluir se é viável esperar que os modelos atuais continuem progredindo ou se os fatores limitantes indicam um teto para o crescimento desses modelos.

Os objetivos específicos incluem fazer um paralelo histórico com o cenário de IA em torno da década de 60; analisar eventos recentes que levaram ao cenário de IA atual; estabelecer quais são os parâmetros principais utilizados para a otimização dos modelos de grande escala atuais e analisar como esses parâmetros podem afetar o desenvolvimento de novos modelos.

1 Material de pesquisa

A maioria das fontes da pesquisa são artigos científicos que tratam dos assuntos abordados publicados majoritariamente pelas equipes de pesquisa que desenvolveram os modelos de fronteira atuais, constam também notícias de eventos de interesse da área.

2 A Primeira Primavera e o Primeiro Inverno de IA

O uso do termo Inteligência Artificial como campo de pesquisa é atribuído ao *Dartmouth Summer Research Project on Artificial Intelligence*, um evento organizado por John McCarthy, realizado em 1956; das pessoas presentes no evento, muitas se tornaram líderes na área e foram pioneiros na área de pesquisa. A área em si atraiu investidores que estavam interessados nas aplicações práticas que o campo possibilitaria. Esse período é considerado começo do primeiro “AI Boom” ou a primeira “Primavera de IA”.

Durante a Guerra Fria, o governo americano estava particularmente interessado na tradução de textos e artigos científicos de línguas estrangeiras para inglês, investindo agressivamente na área (PIERCE, John R., 1966, prefácio).

Em 1964, após 8 anos de investimentos em diversas agências, o National Research Concil (NRC), preocupado com a falta de progresso na área, cria o Automatic Language Processing Advisory Committee (ALPAC) para investigar o progresso feito pelos pesquisadores. ALPAC libera um relatório em 1966 mostrando que tradução via máquinas era mais caro, menos preciso e mais lento que tradução humana (PIERCE, John R., 1966, p.19, p.64); após 10 anos de pesquisa e 20 milhões de dólares gastos (PIERCE, John R., 1966, p.29), o relatório resultou no encerramento do suporte a pesquisa em tradução de máquina (HUTCHINS, John, 2003, p.131).

Em 1969, o então Senador Mike Mansfield autorou uma emenda que proibia financiamento militar de pesquisas que não tivessem um relacionamento direto ou aparente com uso militar (U.S., 1969, p.206). A Emenda Mansfield foi modificada posteriormente, direcionando o Departamento de Defesa ao suporte de pesquisas aplicadas de curto prazo nas universidades (U.S. Congress, 1991, p.61). Como a pesquisa da ARPA indicava que a maioria da pesquisa em inteligência artificial não teria nenhum resultado prático de curto prazo, o financiamento de projetos de inteligência artificial diminuiu drasticamente.

Em 1972, James Lighthill foi comissionado pelo Science Research Concil (SRC) para produzir um relatório sobre o estado da pesquisa em inteligência artificial, principalmente no Reino Unido (LIGHTHILL, 2012, overview). Em seu relatório, Lighthill criticou a falha da inteligência artificial de atingir seus objetivos, concluindo que nada feito com inteligência artificial não poderia ser feito de maneira melhor por outras áreas de pesquisa, também implicando que mesmo os algoritmos mais bem-sucedidos de inteligência artificial eram incapazes de lidar com a complexidade do mundo real, servindo apenas para resolver “problemas de brinquedo” (LIGHTHILL, 1973, p.16-17).

O relatório foi criticado por pesquisadores de inteligência artificial da época (LIGHTHILL, 2012, part II-III), resultando em um debate na série “Controversy” da BCC em 1973; o debate “The general purpose robot is a mirage” apresentava Lighthill versus Donald Michie, John McCarthy e Richard Gregory (BBC, 1973).

Apesar do debate, o relatório levou ao corte da pesquisa em inteligência artificial no Reino Unido, marcando o que é considerado o primeiro “Inverno de IA”, um período em que o financiamento de pesquisa acadêmica relacionado a IA diminuiu drasticamente.

3 Eventos Recentes que Possibilitaram o Cenário Atual de IA

Na década de 80 houve uma segunda “Primavera de IA”, porém dessa vez focada em programas especialistas, com funções práticas imediatas em vez de objetivos efêmeros de longo prazo; os sistemas especialistas se provaram efetivos porém difíceis de atualizar, sendo eventualmente abandonados. Na década de 90 começou o segundo “Inverno de IA”, este durando até o fim dos anos 2000.

O cenário atual de inteligência artificial é similar ao da década de 60, os avanços na área geraram uma onda de investimentos e expectativas grandiosas que alguns prometem estar logo no horizonte, porém o que possibilitou esse cenário?

Em 2012, AlexNet (KRIZHEVSKY, 2012) foi a primeira rede neural a conquistar o primeiro lugar no ILSVRC (ImageNet Large Scale Visual Recognition Challenge) - um desafio organizado pelo ImageNet, uma base de dados feita para o

uso em softwares de reconhecimento visual de objetos. Redes neurais são uma tecnologia que tem suas raízes na década de 40 (MCCULLOCH, 1943), porém o custo computacional envolvido limitava os resultados que eram possíveis de se obter através das mesmas.

Com o passar das décadas a capacidade computacional cresceu exponencialmente e AlexNet provou ao público que, providos os recursos computacionais necessários, redes neurais poderiam ter um desempenho competitivo com os outros modelos da época, retomando o interesse do público em redes neurais. De 2012 até a descontinuação do ILSVRC em 2017 todos os primeiros colocados no desafio foram redes neurais. Em 2015, Ilya Sutskever, um dos coautores do artigo do AlexNet, cofundou a OpenAI.

Em 2017, uma equipe de pesquisadores do Google publicou seus resultados de sua pesquisa acadêmica, a pesquisa tratava de um novo modelo de rede neural que tinha como foco tradução de texto (VASWANI, 2017, p.10).

Dentre as qualidades notáveis do modelo estavam: complexidade computacional por camada similar ou abaixo dos outros tipos de camada comumente utilizados na época e um número baixo de operações sequenciais necessárias, o que aumentava a quantidade de operações computacionais que poderiam ser paralelizadas; a distância entre dependências na rede, o fato dos caminhos que os sinais têm que atravessar serem mais curtos nesse tipo de camada tornava mais fácil para o modelo aprender dependências de longa distância e o modelo tinha a capacidade de ser mais interpretável, o que permitia fazer um balanceamento mais compreensivo dos parâmetros (VASWANI, 2017, p.6-7).

Este modelo, nomeado Transformer, é considerado um dos principais preponentes dos modelos atuais, estes sendo, de certa forma, versões deste modelo escaladas em ordens de magnitudes. Esse modelo é o que leva aos modelos GPT (Generative Pre-trained Transformer), que por sua vez foram o que atraiu o interesse do público em massa em 2022 com o lançamento do ChatGPT3 preview.

4 Fatores Limitadores de Modelos Neurais

Em 2020, os pesquisadores da OpenAI publicaram um estudo que demonstrou empiricamente como a performance (medida em função de cross-entropy loss) de modelos neurais de linguagem seguem leis de escala. Cross-entropy loss, ou perda de entropia cruzada, escala como lei de potência com o número de parâmetros do modelo, tamanho da base de dados e a quantidade de computação usada no treinamento (KAPLAN, 2020).

No estudo, os valores inferidos através dessas leis se mostraram verdadeiros por mais de seis ordens de magnitude e extrapolando o valor do começo da curva de treinamento era possível prever razoavelmente a perda que seria obtida caso o modelo fosse treinado por um tempo mais longo (KAPLAN, 2020, p.3).

As leis obtidas nesse estudo, apesar de serem ajustadas com o passar dos anos pelos próprios autores e através de outros estudos (HENIGHAN, 2020; HOFFMANN, 2022), são usadas até hoje para desenvolver modelos de grande escala e demonstram de uma maneira previsível como valores necessários para diminuir a perda entrópica cruzada crescem.

Também é importante notar que para muitas dessas equações foi encontrada uma constante irredutível, que se denomina a entropia semântica (HENIGHAN, 2020, p.7), indicando um limite teórico para a escala desses modelos. Em um campo específico como linguagem, essa entropia semântica pode ser demonstrada pelo fato que duas frases, apesar de terem formas diferentes, podem ter o mesmo significado.

“Por exemplo, um modelo que está incerto sobre se deve gerar ‘Da França, a capital é Paris’ ou ‘Paris é a capital da França’ não está incerto em nenhum sentido importante. Ainda assim, no nível de tokens, o modelo está incerto entre duas formas do mesmo significado” (KUHN, 2023, p.1, tradução nossa).

4.1 Bases de Dados

Os modelos atuais são treinados em quantidades exorbitantes de dados públicos obtidos através da internet (SHUMAILOV, 2023, p.1; DOHMATOB, 2024, p.1-2; VILLALOBOS, 2024, p.1) e há a necessidade de uma quantidade cada vez maior de dados, porém, estudos especulam que as bases de dados atuais já utilizam uma grande porcentagem da informação pública existente, com a possibilidade de modelos que utilizam todas as fontes públicas de informação surgirão em poucos anos.

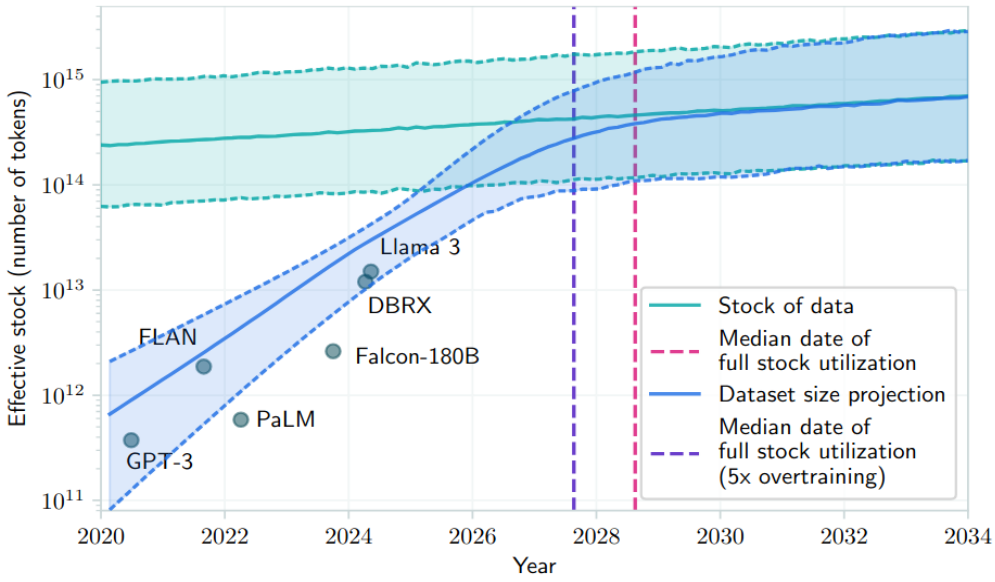


Figura 1: Projeção de tamanho total dos textos públicos gerados por serem humanos e das bases de dados utilizadas para treinar LLMs notáveis, com a previsão de quando esse textos serão totalmente utilizados. (VILLALOBOS, 2024, p.1).

Em um relatório ao *House of Lords Communications and Digital Committee* do parlamento britânico, a OpenAI declarou que “seria impossível treinar os modelos de IA líderes de hoje sem usar materiais protegidos por copyright” (OPENAI, 2024, tradução nossa), mostrando que materiais protegidos por copyright são um fator limitante para o crescimento desses modelos, que demandam uma quantidade cada vez maior de dados.

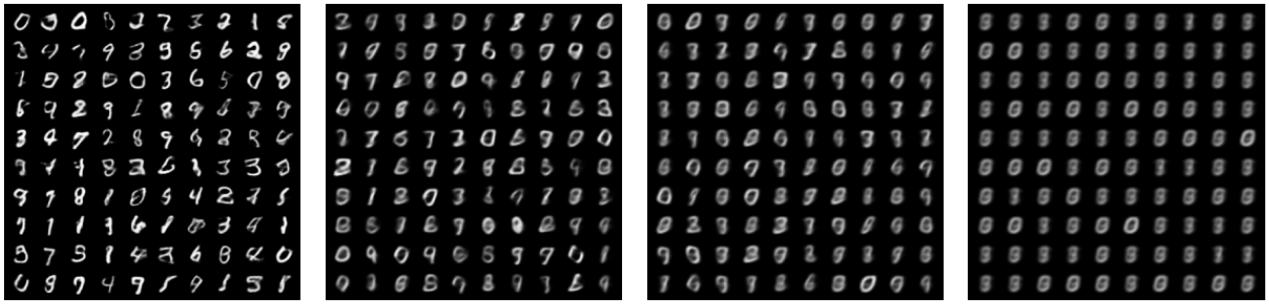
Longpre demostra que houve um aumento nas restrições nas bases de dados públicas que são comumente utilizadas para treinar esses modelos (LONGPRE, 2024, p.4-5), tanto no “full corpus” – o corpus completo das bases de dados públicas C4 (Colossal Clean Crawled Corpus), RefinedWeb e Dolma, quanto no “head distributions” – os 2 mil sites de cada base de dados ranqueados pelo número de tokens, com uma união de 3,95 mil sites; essa distribuição representa os maiores sites, mais ativamente mantidos e domínios críticos para o treinamento de IA (LONGPRE, 2024, p.3).

“De uma perspectiva relativa, de abril de 2023 à abril de 2024 essas restrições subiram 500%+ para o ‘full corpus’ de ambos C4 e o RefinedWeb, e 1000%+ para ‘head distributions’ de ambos o C4 e da RefinedWeb. Nota-se que essas medidas só capturam os domínios com restrição total, e os números são mais altos para domínios parcialmente restritos.” (LONGPRE, 2024, p.5-6, tradução nossa)

Enquanto algumas empresas como Reddit e StackOverflow têm licenciado o acesso às suas bases de dados para empresas como Google e OpenAI (Reddit […], 2024; Stack […], 2024; How […], 2024), a crescente demanda por dados só torna mais aparente o quanto as restrições ao acesso do conteúdo dessas bases de dados vai afetar a possibilidade de crescimento desses modelos.

Quantidades massivas de dados não são o único fator limitante de uma base de dados, modelos necessitam de dados de alta qualidade, capazes de indicar como um profissional qualificado executaria uma tarefa, porém, com a popularização dos modelos de IA generativa, mais e mais informações sintéticas (informações geradas através de IA generativa) têm sido inseridas nas bases de dados públicas (SHUMAILOV, 2023, p.1; DOHMATOB, 2024, p.1).

Enquanto que a ideia de usar dados sintéticos para complementar os dados de alta qualidade soa promissora, tornando possível aumentar o tamanho de uma base de dados de maneira artificial, a realidade é que modelos treinados em dados sintéticos tendem a regredir em vez de melhorar.



(a) Original model (b) Generation 5 (c) Generation 10 (d) Generation 20

Figura 2: Demonstração de colapso de modelo (SHUMAILOV, 2023, p.10).

Quanto maior a porcentagem de dados sintéticos numa base de dados, maior a possibilidade de ocorrer um “colapso de modelo”, um processo de aprendizado degenerativo em que modelos esquecem eventos improváveis conforme o modelo assume mais e mais da informação sintética como realidade. Modelos treinados em dados sintéticos tendem a super-representar os dados mais estatisticamente representados e esquecer os dados menos representados, resultando em representações distorcidas da realidade.

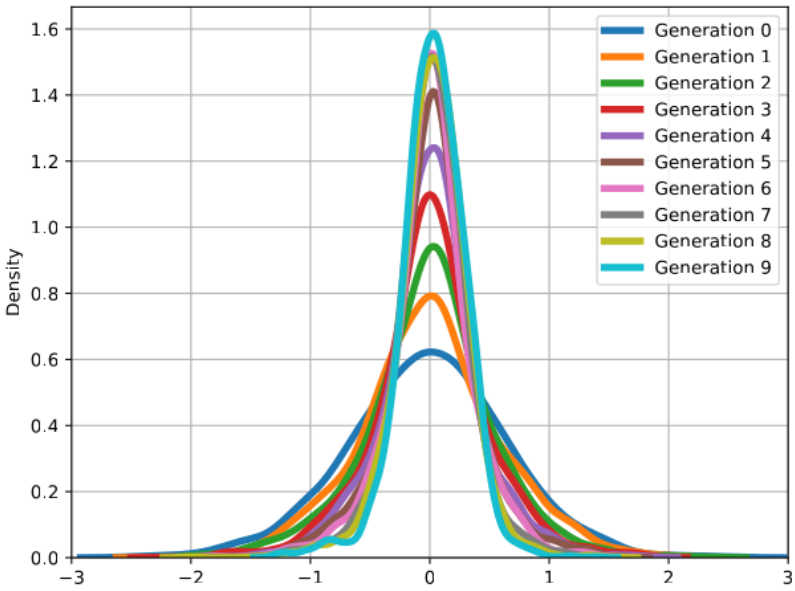


Figura 3: Distribuição gaussiana onde cada geração é treinada somente na geração anterior (SHUMAILOV, 2023, p.10).

Quanto mais avançados os modelos ficam, mais eles são utilizados para geração de dados em bases de dados públicas, e por sua vez, mais difícil fica de distinguir dados reais, criados por seres humanos, de dados sintéticos, criados por IA generativa. Isso gera um ciclo de poluição das bases de dados públicas.

4.2 Custos Computacionais

Stanford University publicou em abril o *AI Index 2024 Annual Report* (MASLEJ, 2024), a maior base de dados sobre inteligência artificial até o hoje, a partir dessa base de dados surgiram estudos que abordam como o custo de modelos de fronteira cresceram desde 2016 até hoje.

Foi demonstrado uma tendência de crescimento do custo da etapa de treinamento final de 2.6x ao ano para modelos treinados usando computação em nuvem e de 2.4x ao ano para modelos usando hardware próprio (COTTIER, 2024, p.5-6).

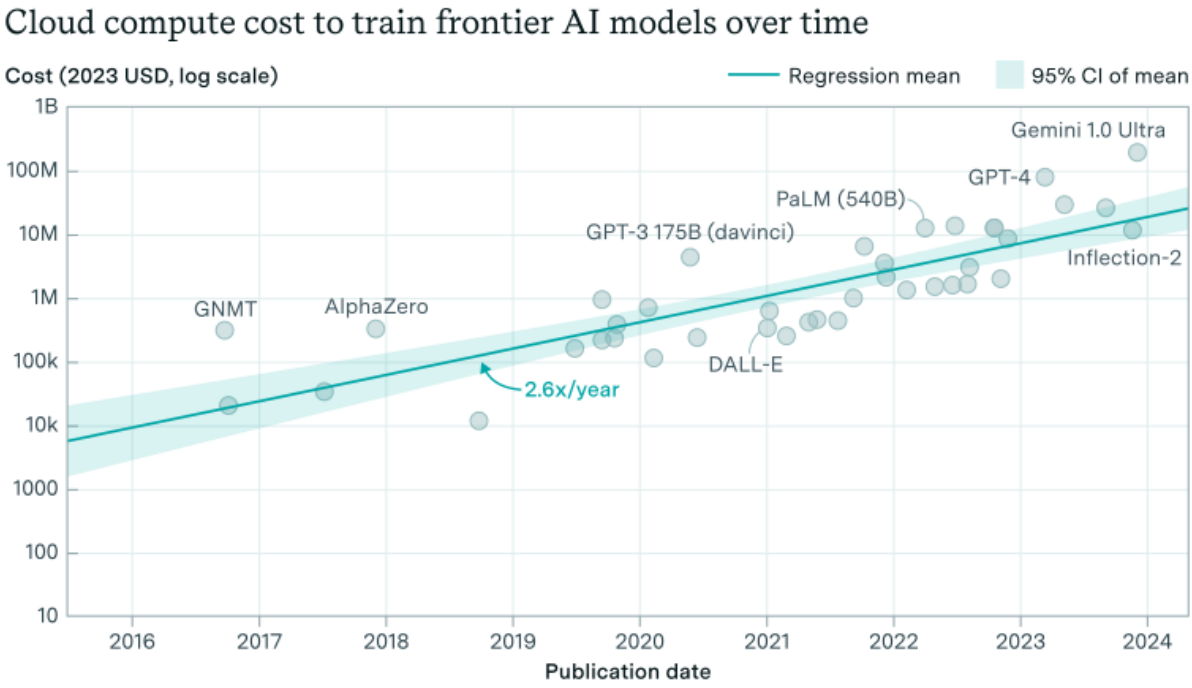


Figura 4: Custo de Computação em Nuvem da última etapa de treinamento (COTTIER, 2024, p.6).

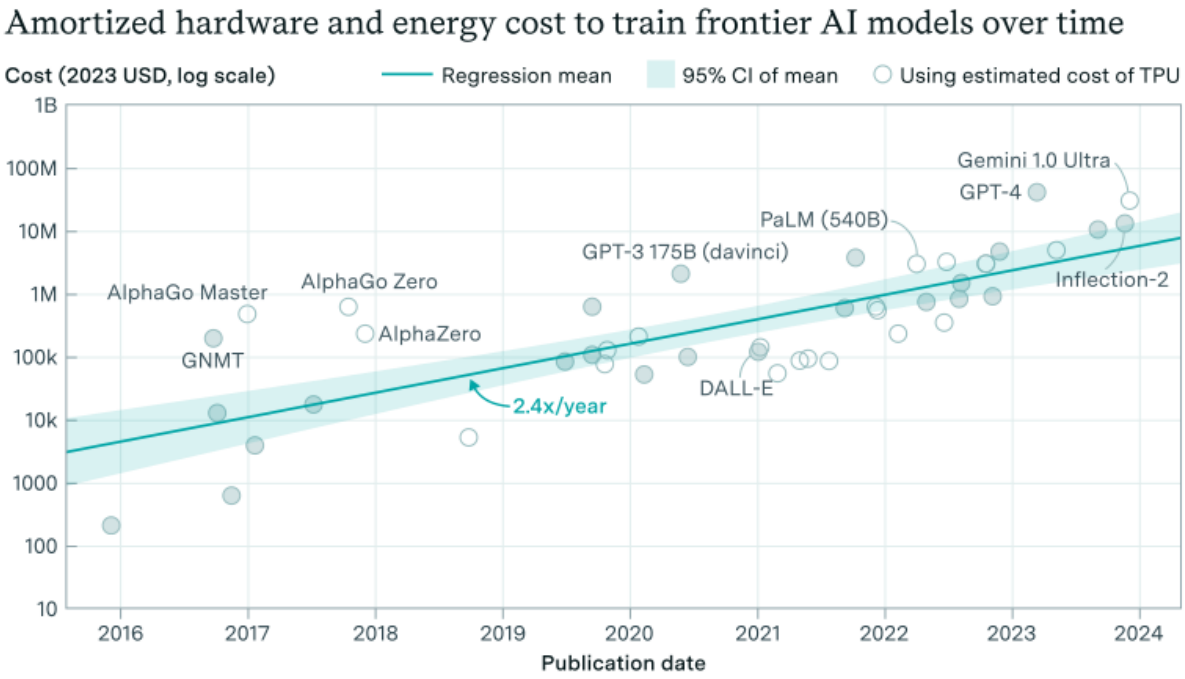


Figura 5: Custo amortizado de hardware e energia da última etapa de treinamento (COTTIER, 2024, p.5).

Já existem modelos que custam dezenas de milhões de dólares – a última etapa de treinamento do modelo Gemini 1.0 Ultra teve um custo estimado de 30 milhões de dólares; a última etapa de treinamento do modelo GPT4 teve um custo estimado de 40 milhões de dólares (COTTIER, 2024, p.1) com o CEO da OpenAI confirmando que o preço total do modelo GPT4 supera 100 milhões de dólares (OpenAI's […], 2023). Esses valores indicam que, com um crescimento de 2.4x ao ano, até o fim de 2027 os custos de treinamento vão ultrapassar 1 bilhão de dólares (COTTIER, 2024, p.6).

A pesquisa também indica que houve uma tendência de crescimento da potência elétrica necessária para esses modelos de 2.0x ao ano.

Um modelo como Gemini Ultra requer uma potência elétrica de aproximadamente 35MW, projetando esses números, em 2029 poderemos ter uma etapa de treinamento consumindo 1GW de potência elétrica (COTTIER, 2024, p.20).

Em comparação, a usina hidroelétrica de Itaipu, a segunda maior hidroelétrica do mundo, gera aproximadamente 14GW de potência de elétrica.

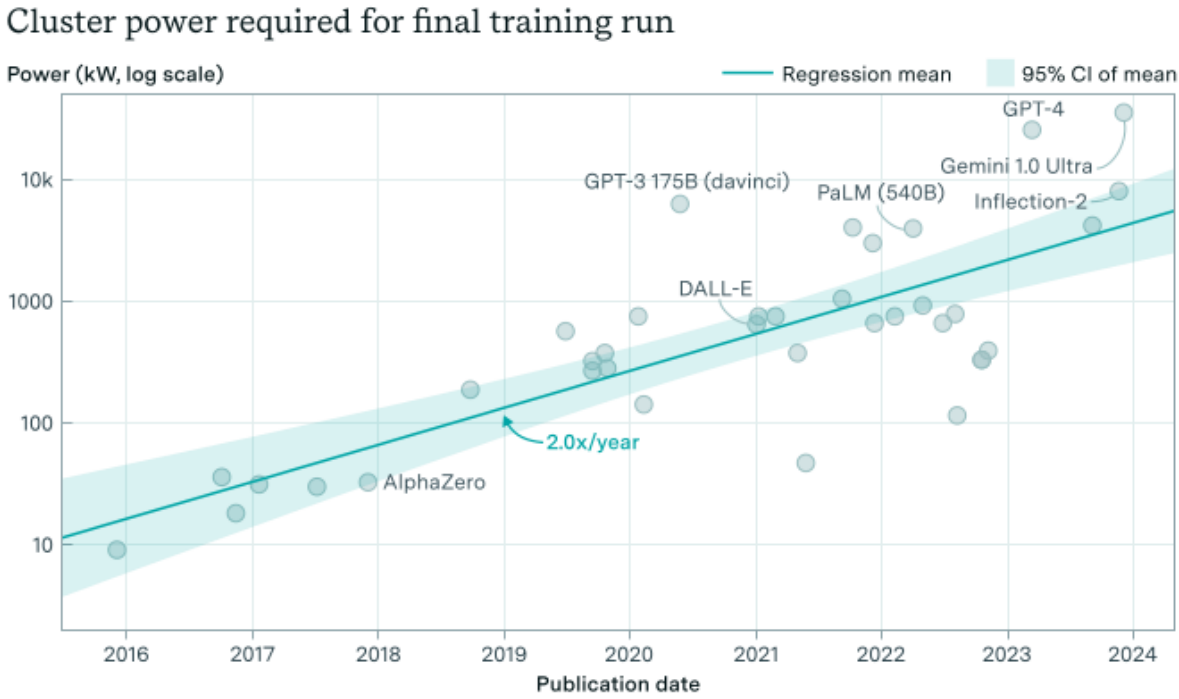


Figura 6: Custos energéticos necessários para a última etapa de treinamento (COTTIER, 2024, p.20).

Esses gastos energéticos são tão altos que empresas como Amazon, Google e Microsoft já anunciaram estarem investindo em usinas de energia nuclear (Hungry […], 2024), já que outras fontes de energia verde como energia eólica e energia solar não produzem energia da maneira constante que os datacenters demandam.

O treinamento e a operação desses modelos se mostram vez mais custosos, de modo que os custos superam a renda gerada pelos mesmos (OpenAI […], 2024), sendo necessário investimentos externos para possibilitar a continuidade desses projetos.

Apesar do investimento imenso feitos nos modelos atuais, estudos mostram que o comportamento desses modelos não necessariamente indica que eles sejam capazes de raciocínio lógico, mas sim que fazem reconhecimento de padrões probabilísticos. A capacidade de raciocínio dos modelos atuais é julgada baseada

em benchmarks (métricas padronizadas) adotadas por toda a indústria, essas benchmarks são constituídas de milhares de perguntas estáticas que o modelo deve responder, e a porcentagem de perguntas que o modelo responde corretamente é a sua “nota”.

Srivastava e Mirzadeh segurem versões funcionais dessas perguntas para os benchmarks MATH e GSM8K – versões em que os parâmetros dessas questões são alterados – e a partir disso demonstram que há uma queda nos resultados (SRIVASTAVA, 2024; MIRZADEH, 2024), essa queda nos resultados é denominada lacuna de raciocínio.

Srivastava (2024, p.9, tradução nossa) mostra que “avaliado sobre o Q1’24 (consistindo de outubro, novembro e dezembro de 2023), nós descobrimos que os modelos têm uma lacuna de raciocínio variando entre 58.35% e 80.31%”. Essa lacuna entre a métrica original e a apresentada por Srivastava indica que as métricas utilizadas para medir esses modelos originalmente estão defasadas.

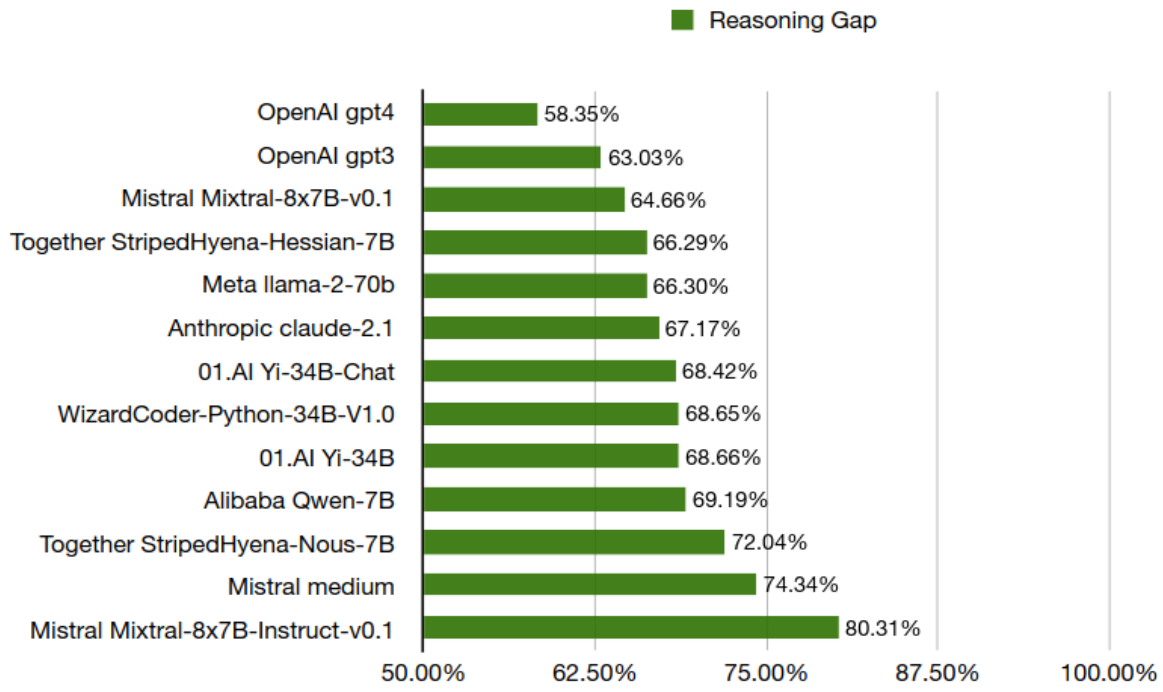


Figura 7: Lacuna de raciocínio apresentada pelos modelos quando testados na benchmark estática MATH versus a benchmark funcional MATH() (SRIVASTAVA, 2024, p.11).

“Modelos de linguagem de última geração exibem algumas capacidades de raciocínio, porém a avaliação precisa disso está defasada. Em particular, as capacidades de LLMs podem ter sido superestimadas porque elas são testadas usando métricas de desempenho que são suficientes para a compreensão linguística (questão textual, resposta) porém podem não medir a capacidade de raciocínio com precisão” (SRIVASTAVA, 2024, p.1, tradução nossa).

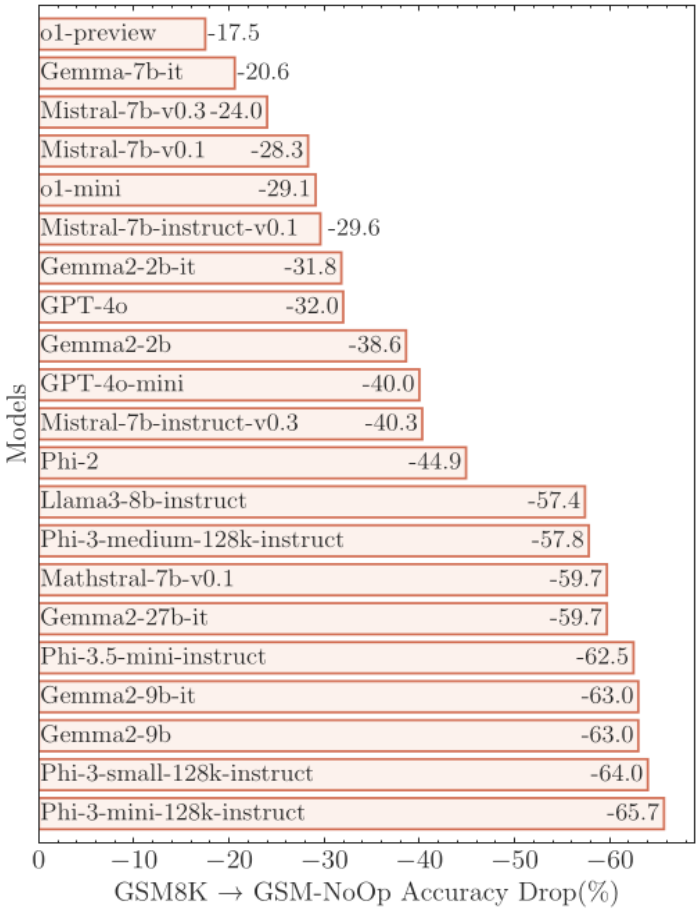


Figura 8: Lacuna de raciocínio apresentada pelos modelos quando testados na benchmark estática GSM8K versus a benchmark funcional GSM-NoOp (MIRZADEH, 2024, p.11).

Mirzadeh demonstra que ao adicionar cláusulas que *parecem relevantes* ao prompt, a performance dos modelos cai até 65% em vários modelos de fronteira, mesmo que a cláusula não contribua em nada para a linha de raciocínio necessária para chegar na resposta (MIRZADEH, 2024, p.10).

Essa variação entre métricas comumente usadas atualmente e esses estudos apontam para a possibilidade que a capacidade de raciocínio dos modelos de fronteira atuais é superestimada, o que vai de contramão às crescentes expectativas de que esses modelos possam ser usados confiavelmente para fins comerciais.

5 Resultados

Através desse estudo foi possível verificar que, para os modelos de larga escala que são o foco global de desenvolvimento da atualidade, são utilizados três fatores como principais vetores de crescimento. São estes: o número de parâmetros do modelo, a capacidade computacional dedicada ao treinamento desse modelo e o tamanho das bases de dados utilizadas para o treinamento.

Dos três fatores não foi possível encontrar nenhuma limitação relevante para o número de parâmetros, o mesmo sendo limitado mais por lógica programacional do que por fatores físicos. Esta limitação se mostra pouco relevante quando comparada às limitações dos outros fatores.

Quanto ao tamanho das bases de dados, foi demonstrado que existem indicações reais que a disponibilidade das bases de dado, principalmente bases de dado de alta qualidade, tem recebido restrições parciais e totais de acesso às mesmas através de copyright. Outra limitação das bases de dado foi o aumento dos dados sintéticos criados por inteligência artificial generativa que tende a poluí-las, aumentando a possibilidade de colapso de modelo durante o treinamento, sendo necessário filtrar os dados para a remoção de dados sintéticos num cenário em que conforme os modelos ficam mais capazes de se comportar de maneira humana, mais difícil é filtrar informação sintética das bases de dados.

Quanto à capacidade computacional dedicada durante ao treinamento, foi demonstrado como os modelos de fronteira têm crescido em custo computacional e energético de maneira exponencial, com alguns já custando dezenas de milhões de dólares durante a sua etapa final de treinamento e como, se as previsões de consumo energético se manterem, é possível projetar que haverão modelos que, só

durante a etapa final de treinamento, consumirão 1GW de potência elétrica até o fim da década.

Esses custos extramente altos precisam ser justificados economicamente, porém os retornos ainda não estão presentes. Há também evidência de que os testes utilizados para avaliar a capacidade de executar tarefas desses modelos são frágeis quando expostos à informações não relevantes para a resolução das tarefas propostas, que são comumente encontradas num cenário real.

A partir disso vê-se que enquanto há um potencial teórico de continuar escalando os modelos de grande escala atuais – com base de que ainda não foi alcançada a entropia semântica demonstrada a partir das leis de escala – os recursos necessários para alcançar esse limite são de uma magnitude extraordinária; a disponibilidade dos dados é um problema inevitável e os altos custos computacionais requerem investimentos externos recorrentes, porém ainda não há um retorno econômico que justifique esses investimentos.

Considerações Finais

Esse estudo demonstrou de maneira bem-sucedida os principais fatores que são utilizados atualmente para otimizar os maiores modelos em desenvolvimento no mercado atual (número de parâmetros do modelo, quantidade computacional dedicada ao modelo e o tamanho da base de dados), fatores que, apesar de não serem os únicos a determinar o resultado do treinamento desses modelos, são os que demonstram ter a maior influência na diminuição da perda entrópica cruzada, o valor para qual todos os modelos estudados tentam otimizar.

Nenhum desses modelos chegou no limite teórico dessa otimização, a entropia semântica da especialização de modelo, e em alguns casos, como modelos de linguagem, ainda não foi determinado um valor definido para a entropia semântica. Isso demonstra que existe a possibilidade de melhorar ainda mais esses modelos em relação a essa métrica de otimização.

Apesar disso, esse estudo demonstrou que existem limitações físicas referentes aos fatores de otimização. Enquanto que a quantidade de recursos necessários para o treinamento de modelos de fronteira cresce exponencialmente, a disponibilidade desses recursos fica cada dia mais restrita e mais custosa, dificultando não só a obtenção dos recursos necessários, mas também torna mais difícil justificar economicamente os bilhões de dólares que serão necessários para o treinamento e a manutenção desses modelos de fronteira.

Até mesmo a empresa com maior parcela do mercado de inteligência artificial generativa do mundo a OpenAI tem deficit previsto de 5 bilhões de dólares para 2024 (OpenAI […], 2024). Os recursos necessários para treinar modelos de fronteira são tão grandes que empresas como Amazon, Google e Microsoft anunciaram planos para investir em usinas de energia nuclear para suprir os gastos energéticos constantes dos seus datacenters (Hungry […], 2024).

É possível que nos próximos anos nós cheguemos no limite do que é viável de escalar o modelo Transformer, pelo menos economicamente. Para haver progresso significativo na área de IA a partir dos modelos que existem hoje, será necessário encontrar um avanço lógico – como o modelo Transformer possibilitou para redes neurais – e não simplesmente um aumento de escala, que é o foco principal hoje.

Referências

BBC. Série Controversy. O debate **“The general purpose robot is a mirage”.** Video. 81min. 1973. Disponível em: https://media.aiai.ed.ac.uk/Video/Lighthill1973/. Acesso em: 01 nov. 2024.

COTTIER, Ben; RAHMAN, Robi; FATTORINI, Loredana et al. The rising costs of training frontier AI models. **arXiv preprint arXiv:2405.21015**, 2024.

DOHMATOB, Elvis; FENG, Yunzhen; YANG, Pu et al. A tale of tails: Model collapse as a change of scaling laws. **arXiv preprint arXiv:2402.07043**, 2024.

HENIGHAN, Tom; KAPLAN, Jared; KATZ, Mor et al. Scaling laws for autoregressive generative modeling. **arXiv preprint arXiv:2010.14701**, 2020.

HOFFMANN, Jordan; BORGEAUD, Sebastian; MENSCH, Arthur et al. Training compute-optimal large language models. **arXiv preprint arXiv:2203.15556**, 2022.

HOW Stack Overflow is partnering with Google to encourage socially responsible AI. **Stack Overflow**, 2024. Disponível em: https://stackoverflow.blog/2024/03/12/how-stack-overflow-is-partnering-with-google-to-encourage-socially-responsible-ai/. Acesso em: 01 nov. 2024.

HUNGRY for Energy, Amazon, Google and Microsoft Turn to Nuclear Power. **NY Times**, 2024. Disponível em: https://www.nytimes.com/2024/10/16/business/energy-environment/amazon-google-microsoft-nuclear-energy.html?unlocked_article_code=1.Sk4.D1TP.D-hnKAIGWj0m. Acesso em: 01 nov. 2024.

HUTCHINS, John. ALPAC: the (in) famous report. **Readings in machine translation**, v. 14, p. 131-135, 2003.

KAPLAN, Jared; MCCANDLISH, Sam; HENIGHAN, Tom et al. Scaling laws for neural language models. **arXiv preprint arXiv:2001.08361**, 2020.

KRIZHEVSKY, Alex; SUTSKEVER, Ilya; HINTON, Geoffrey E. Imagenet classification with deep convolutional neural networks. **Advances in neural information processing systems**, v. 25, 2012.

KUHN, Lorenz; GAL, Yarin; FARQUHAR, Sebastian. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. **arXiv preprint arXiv:2302.09664**, 2023.

LIGHTHILL, James. Artificial intelligence: A general survey. In: **Artificial Intelligence: a paper symposium**. London: Science Research Council, 1973. p. 1-21.

LIGHTHILL Report – Documents and Commentary. **Chilton Computing**, 2012. Disponível em: http://www.chilton-computing.org.uk/inf/literature/reports/lighthill_report/contents.htm. Acesso em: 01 nov. 2024.

LONGPRE, Shayne; MAHARI, Robert; LEE, Ariel et al. Consent in crisis: The rapid decline of the ai data commons. **arXiv preprint arXiv:2407.14933**, 2024.

MASLEJ, Nestor; FATTORINI, Loredana; PERRAULT, Raymond, “The AI Index 2024 Annual Report,” AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA, April 2024.

MCCULLOCH, Warren S.; PITTS, Walter. A logical calculus of the ideas immanent in nervous activity. **The bulletin of mathematical biophysics**, v. 5, p. 115-133, 1943.

MIRZADEH, Iman; ALIZADEH, Keivan; SHAHROKHI, Hooman et al. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models. **arXiv preprint arXiv:2410.05229**, 2024.

OPENAI. Written evidence (LLM0113) to the House of Lords Communications and Digital Select Committee inquiry: Large Language Models. 2024.

OPENAI sees roughly $5 billion loss this year on $3.7 billion in revenue. **CNBC**, 2024. Disponível em: https://www.cnbc.com/2024/09/27/openai-sees-5-billion-loss-this-year-on-3point7-billion-in-revenue.html. Acesso em: 01 nov. 2024.

OPENAI's CEO Says the Age of Giant AI Models Is Already Over. **Wired**, 2023. Disponível em: https://www.wired.com/story/openai-ceo-sam-altman-the-age-of-giant-ai-models-is-already-over/. Acesso em: 01 nov. 2024.

PIERCE, John R.; CARROLL John B.; HAMP, Eric P. et al. Languages and Machines – Computers in Translation and Linguistics. **ALPAC Report, National Academy of Sciences, National Research Concil, Washington, DC**, 1966.

REDDIT cashes in on AI gold rush with $203M in LLM training license fees. **Ars Technica**, 2024. Disponível em: https://arstechnica.com/ai/2024/02/reddit-has-already-booked-203m-in-revenue-licensing-data-for-ai-training/. Acesso em: 01 nov. 2024.

SHUMAILOV, Ilia; SHUMAILOV, Zakhar; ZHAO, Yiren et al. The curse of recursion: Training on generated data makes models forget. **arXiv preprint arXiv:2305.17493**, 2023.

SRIVASTAVA, Saurabh et al. Functional benchmarks for robust evaluation of reasoning performance, and the reasoning gap. **arXiv preprint arXiv:2402.19450**, 2024.

STACK Overflow and OpenAI Partner to Strengthen the World's Most Popular Large Language Models. **Stack Overflow**, 2024. Disponível em: https://stackoverflow.co/company/press/archive/openai-partnership. Acesso em: 01 nov. 2024.

U.S., **Public Law 91-121**, Fiscal year 1970 Military Authorization Act, Sec.203, p.206, 1969.

U.S. Congress, Office of Technology Assessment, Federally Funded Research: Decisions for a Decade, **OTA-SET-490 (Washington, DC: U.S. Government Printing Office**, 1991.

VASWANI, Ashish; SHAZEER, Noam; PARMAR, Niki et al. Attention is all you need. **Advances in Neural Information Processing Systems**, 2017.

VILLALOBOS, Pablo; HO, Anson; SEVILLA, Jaime et al. Position: Will we run out of data? Limits of LLM scaling based on human-generated data. In: **Forty-first International Conference on Machine Learning**, 2024.