Desigualdades Geográficas e Impacto Ambiental no Desenvolvimento de Modelos de Inteligência Artificial

Matheus Felipe Bosa[1], **Luiz Gomes-Jr**[1]

[1]PPGCA – Universidade Tecnológica Federal do Paraná (UTFPR)
Curitiba – PR – Brazil

bosa@alunos.utfpr.edu.br, lcjunior@utfpr.edu.br

Abstract. *The advancement of Artificial Intelligence (AI) raises questions about resource concentration, geographical inequalities, and environmental impact. This study analyzes data from the EpochAI research institute to investigate patterns in the development of frontier models. Through a cluster analysis, an academic group was distinguished from an industrial one among the main model developers. A linear regression analysis confirmed that model size and the hardware used are dominant factors for financial cost (*$R^2 = 0.871$*) and environmental impact (*$R^2 = 0.762$*). It is concluded that AI development is dominated by large North American companies, but with significant contributions from universities, evidencing a strong concentration of resources with pockets of academic innovation.*

Resumo. *O avanço da Inteligência Artificial (IA) levanta questões sobre concentração de recursos, desigualdades geográficas e impacto ambiental. Este estudo analisa dados do instituto de pesquisa EpochAI para investigar padrões no desenvolvimento de modelos de fronteira. Por meio de uma análise de clusters, distinguiu-se um grupo acadêmico de um industrial entre os principais desenvolvedores de modelos. Uma análise de regressão linear confirmou que o tamanho do modelo e o hardware utilizado são fatores dominantes para o custo financeiro (*$R^2 = 0,871$*) e para o impacto ambiental (*$R^2 = 0,762$*). Conclui-se que o desenvolvimento de IA é dominado por grandes empresas norte-americanas, mas com contribuições significativas de universidades, evidenciando uma forte concentração de recursos com focos de inovação acadêmica.*

1. Introdução

Os avanços recentes das capacidades da Inteligência Artificial (IA) ocorrem à medida em que os custos de treinamento crescem aproximadamente 2,4 vezes ao ano desde 2016 [Cottier et al. 2024]. De maneira semelhante, a energia demandada para treinamento de modelos apresenta crescimento aproximado de 2,1 vezes ao ano desde 2012 [Frymire and Rahman 2024]. Todavia, esse avanço não ocorre de maneira uniforme, sendo marcado por desigualdades geográficas devido ao fato de que são poucas as nações e corporações que possuem os recursos financeiros para operar no desenvolvimento de modelos de fronteira tecnológica [Cottier et al. 2024]. Isso pode ser visualizado na Figura 1. Dessa forma, ocorre uma concentração de organizações que, somada ao custo computacional e energético crescente para o treinamento de modelos, levanta preocupações sobre a sustentabilidade e a acessibilidade desse progresso técnico.

Este trabalho analisa as disparidades no desenvolvimento de modelos de IA, analisando a distribuição geográfica, os custos computacionais, o tipo de *hardware* utilizado e o contexto organizacional relacionado com modelos proprietários e de código-aberto. Para guiar esta análise, foram formuladas três perguntas de pesquisa: (a) É possível identificar grupos distintos de modelos de IA com base em suas características técnicas, acessibilidade e contexto organizacional? (b) Como as características técnicas do *hardware*, o tamanho do modelo e a escala do treinamento influenciam o custo financeiro e o impacto ambiental? (c) Existem organizações que alcançam um desempenho notável e inesperado considerando o seu contexto operacional? Os resultados revelam uma divisão no ecossistema de IA, segmentado em um *cluster* acadêmico, com modelos de menor escala focados em visão computacional, e um *cluster* industrial, dominado por modelos de linguagem de grande escala que foram desenvolvidos por grandes empresas. Foi também demonstrado através de um modelo de regressão que o custo financeiro e o impacto ambiental, estimado através do consumo energético, podem ser explicados pelas características técnicas e pela escala do treinamento. Adicionalmente, a análise de anomalias identificou universidades e organizações que apresentaram uma capacidade de produzir modelos de impacto na sociedade.

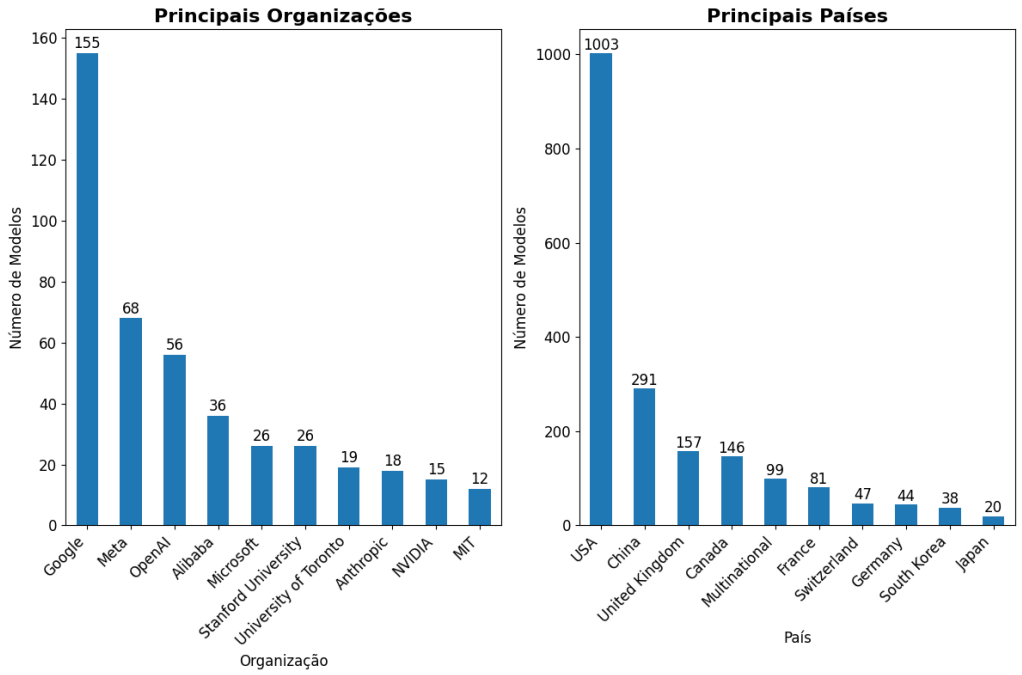

Figura 1. Principais organizações e países no desenvolvimento de modelos de IA de 1950 até abril de 2025. Fonte: Elaborado pelos autores com dados públicos do instituto EpochAI [Epoch AI 2025a, Epoch AI 2025c].

2. Trabalhos Relacionados

O trabalho [Cottier et al. 2024] identifica que o crescimento dos custos de treinamento levará a gastos superiores a um bilhão de dólares por treinamento até 2027. Isso significa que apenas as organizações mais bem financiadas poderão desenvolver modelos de IA de fronteira. Essa constatação está diretamente relacionada com a investigação sobre as desigualdades geográficas e organizacionais, especialmente a observada concentração

do desenvolvimento nos Estados Unidos da América e em poucas organizações conhecidas como *Big Techs*. O objeto científico destaca ainda que essa concentração pode limitar perspectivas e abordagens, levantando preocupações sobre supervisão pública. Já o artigo [Cottier et al. 2023] descreve o crescimento exponencial do treinamento ao longo dos anos, destacando aumentos de até dez milhões de vezes em organizações líderes como o Google. No contexto de sustentabilidade, o trabalho [Wu et al. 2022] caracteriza a pegada de carbono da computação de IA examinando seu ciclo de desenvolvimento e a vida útil do *hardware*. Com relação ao contexto de sustentabilidade ambiental, o trabalho [Wu et al. 2022] se dedica a explorar o impacto no meio ambiente das tendências de crescimento de modelos de IA a partir de uma perspectiva holística. Caracteriza a pegada de carbono da computação de IA ao examinar o ciclo de desenvolvimento de modelos, composto pelas etapas de *Dados*, *Experimentação*, *Treinamento* e *Inferência*, assim como analisa o ciclo de vida do *hardware* utilizado nesse contexto. Em relação à acessibilidade, [Vake et al. 2025] utiliza dados da plataforma *Hugging Face* para analisar modelos de código-aberto, indicando que os altos custos favorecem o modelo proprietário para garantir retorno sobre investimento (ROI), o que gera dependência da comunidade de código-aberto em relação às *Big Techs*. Por fim, o trabalho [Samborska 2025] destaca que os sistemas mais recentes de IA avançaram suas capacidades principalmente através do aumento de escala desses sistemas do que através de inovação científica. Argumenta-se que para escalar um sistema de IA, três fatores principais precisam ser coordenados: quantidade de dados para treinamento, número de parâmetros do modelo e recursos computacionais.

Diante disso, o trabalho apresentado neste artigo busca explorar os impactos do desenvolvimento dos modelos de IA, focando nos aspectos de concentração de recursos e sustentabilidade ambiental. Enquanto a literatura relacionada, como [Cottier et al. 2024, Wu et al. 2022], aponta qualitativamente para os altos custos e a pegada ambiental, e [Samborska 2025] identifica a escala como o principal impulsionador para o avanço, este estudo busca medir essas relações. As principais contribuições deste artigo são um mapeamento dos polos de inovação acadêmicos e industriais, que oferece uma perspectiva sobre o desenvolvimento em IA, e a análise quantitativa do impacto financeiro e ambiental a partir de fatores como tamanho e características do *hardware*.

3. Metodologia

Foram formuladas três perguntas de pesquisa para guiar o desenvolvimento deste trabalho, a saber: (a) É possível identificar grupos distintos de modelos de IA com base em suas características técnicas, acessibilidade e contexto organizacional? (b) Como as características técnicas do *hardware*, o tamanho do modelo e a escala do treinamento influenciam o custo financeiro e o impacto ambiental? (c) Existem organizações que alcançam um desempenho notável e inesperado considerando o seu contexto operacional? Para responder a essas perguntas, adotou-se uma abordagem quantitativa que foi estruturada nas seguintes etapas.

3.1. Coleta e Processamento de Dados

A base para este estudo foram os dados abertos que o instituto EpochAI disponibiliza de maneira pública [Epoch AI 2025d]. Trata-se de um instituto de pesquisa multidisciplinar sem fins lucrativos dedicado a investigar o futuro da IA, analisando seus fatores

impulsionadores e prevendo seu impacto econômico e social. Utilizaram-se três conjuntos de dados: ***Large-Scale AI Models*** com dados sobre mais de 200 modelos de IA treinados com mais de 10^{23} operações de ponto flutuante, na vanguarda da escala e das capacidades [Epoch AI 2025a]; ***Notable AI Models*** que acompanha o progresso da IA de 1950 até os dias atuais, com mais de 900 modelos notáveis e mais de 400 estimativas de computação de treinamento [Epoch AI 2025c]; ***Machine Learning Hardware*** com informações acerca de mais de 160 aceleradores de IA, como GPUs e TPUs, utilizados no desenvolvimento e implantação de modelos de aprendizado profundo [Epoch AI 2025b].

Os dados foram integrados e processados utilizando a biblioteca Pandas em ambiente Python. O processo incluiu a limpeza de dados, tratamento de valores ausentes e a seleção de características relevantes para a análise, resultando em um conjunto de dados final composto por $1\,216$ modelos de IA. As visualizações gráficas para todas as análises foram geradas com as bibliotecas Matplotlib e Seaborn. Embora a base de dados do EpochAI seja referência para modelos de fronteira, reconhece-se que seu uso exclusivo pode introduzir um viés de seleção, sub-representando iniciativas de empresas *startups*, projetos independentes e modelos oriundos de países em desenvolvimento.

3.2. Análise de *Clusters* para Segmentação de Modelos

Para responder à primeira pergunta de pesquisa (a), foi empregada uma análise de *clusters* a fim de identificar agrupamentos de modelos. O pré-processamento das variáveis para a análise foi realizado da seguinte forma: as variáveis numéricas `Parameters`, `Training compute (FLOP)` e `Training dataset size (datapoints)` passaram por uma transformação logarítmica para normalizar suas escalas e, subsequentemente, foram padronizadas junto com `Publication year`. A transformação logarítmica foi necessária no pré-processamento devido à assimetria e à distribuição de cauda longa inerente aos dados de escala computacional. Para as variáveis categóricas, adotou-se a criação de variáveis indicadoras (*dummy variables*), uma prática usual para incorporar dados não numéricos em modelos quantitativos [Hair et al. 2018]. Especificamente, as variáveis de rótulo único (`Model accessibility`, `Training code accessibility`, `Frontier model`) foram convertidas usando a abordagem *One-Hot*. Já as variáveis de múltiplos rótulos (`Domain`, `Organization categorization`, `Country (from Organization)`, `Notability criteria`) foram tratadas por meio de uma binarização *multilabel* [Tsoumakas and Katakis 2007], um processo que cria uma coluna binária para cada rótulo possível. Com o objetivo de reduzir a dimensionalidade, a binarização foi restrita às categorias mais prevalentes, neste caso, aquelas com frequência superior a 5% nas amostras, limitando-se a um máximo de 8 categorias por variável.

Com os dados pré-processados, o algoritmo *k-means* foi utilizado [MacQueen 1967] para aprender sobre $1\,216$ modelos com um total de 16 características. Para determinar o número ideal de *clusters* (k), foi empregado o Método da Silhueta [Rousseeuw 1987], com k variando de 2 a 6. A análise dos coeficientes médios indicou um valor máximo em $k = 2$ de $0,399$. Este coeficiente afere a qualidade da clusterização, avaliando quão similar um objeto é ao seu próprio grupo em comparação com outros grupos. Para visualizar e interpretar a distribuição dos *clusters*, foi aplicada a Análise de Componentes Principais (PCA) [Hotelling 1933].

3.3. Análise de Regressão para Custo e Impacto Ambiental

Visando responder à segunda pergunta de pesquisa (b), foram desenvolvidos dois modelos de regressão linear para investigar os fatores que influenciam o custo financeiro total, segundo cotação de dólares estadunidenses em 2023, e o impacto ambiental, estimado a partir do consumo de energia. As variáveis preditoras incluíram características do *hardware* (`TDP (W)`, `Hardware quantity`), a escala do modelo (`Parameters`, `Training compute (FLOP)`, `Training dataset size`) e métricas de treinamento (`Training time (hours)`). Antes da modelagem, as variáveis foram submetidas a uma transformação logarítmica. A multicolinearidade foi avaliada pelo Fator de Inflação de Variância (VIF) [Fox 2016] e mitigada com o uso da Análise de Componentes Principais (PCA) [Hotelling 1933], que gerou componentes ortogonais a serem utilizados como preditores. Como o tratamento de dados ausentes resultou em diferentes tamanhos amostrais para cada variável dependente, sendo 61 amostras para custo financeiro e 78 para impacto ambiental, a PCA foi realizada separadamente para cada caso. A interpretação dos componentes principais foi conduzida com base nos coeficientes de carga ou *loadings*, os quais indicam a contribuição relativa de cada variável original para a formação de cada componente. Cada análise resultou em um conjunto distinto de seis componentes principais (PC1 a PC6), representando combinações linearmente independentes das variáveis originais. A adequação dos modelos foi verificada pela análise de resíduos e gráficos Quantil-Quantil (Q-Q).

3.4. Detecção de Anomalias para Identificação de Destaques

Para abordar a terceira pergunta de pesquisa (c), foi realizada uma análise de detecção de anomalias ou *outliers* focada no impacto científico dos modelos, medido pelo número de citações (`Citations`). As variáveis de contexto incluíram o número de parâmetros (`Parameters`), custo de treinamento (`Training compute cost (2023 USD)`), organizações e países envolvidos no desenvolvimento (`Organization` e `Country (from Organization)`), categorização da organização em indústria ou academia (`Organization categorization`), assim como o enriquecimento dos dados com a identificação se a organização é uma *Big Tech* ou não. As afiliações organizacionais dos modelos foram padronizadas para agrupar variantes, como "Google Brain" e "Google DeepMind" mapeadas para "Google". Para detectar anomalias, foram combinados dois métodos: ***Z-Score*** para identificar *outliers* de forma relativa e calculado a partir de $Z = \frac{x-\mu}{\sigma}$, onde x é o número de citações do modelo, μ é a média e σ o desvio padrão da organização. Um modelo foi classificado como anômalo por este método se o valor absoluto de seu *Z-Score* fosse superior a 2, indicando que suas citações estavam a mais de dois desvios padrão da média de seu grupo [Montgomery and Runger 2018]; ***Isolation Forest*** foi utilizado para detectar modelos *outliers* no contexto global envolvendo o conjunto de dados. O algoritmo foi configurado com um hiperparâmetro de $contamination = 0.1$, estimando que aproximadamente 10% dos modelos poderiam ser considerados anomalias globais [Liu et al. 2008]. Finalmente, os modelos identificados como *outliers* por ambos os métodos simultaneamente foram considerados como anomalias confirmadas. Convencionou-se o termo "anomalia positiva" para denotar modelos de impacto excepcional. Assim, uma anomalia foi classificada como positiva se o seu número de citações excedesse a mediana de sua organização, indicando um sucesso acadêmico, e negativa caso representasse um modelo com características extremas, como alto custo, mas com baixo número de citações.

4. Resultados

Nesta seção, apresentam-se os resultados obtidos para cada uma das perguntas de pesquisa levantadas para o estudo.

4.1. Análise de *Clusters* para Segmentação de Modelos

O algoritmo *k-means* [MacQueen 1967] segmentou os modelos em dois grupos, resultando em ***Cluster* 0** com 221 modelos $(18,2\%)$ e ***Cluster* 1** com 995 modelos $(81,8\%)$.

Para visualização da separação dos *clusters*, foi utilizada PCA [Hotelling 1933] para redução da dimensionalidade. Os dois primeiros componentes principais explicaram $64,2\%$ da variância total nos dados. A separação geométrica dos *clusters* foi confirmada visualmente através da projeção nos dois componentes principais.

Na Figura 2 constata-se que o ***Cluster* 0** é composto por modelos com ordens de magnitude superiores na quantidade de parâmetros, tamanho do *dataset* e custo computacional de treinamento, medido em operações de ponto flutuante por segundo (FLOPs). Em contraste, o ***Cluster* 1** agrupa modelos com uma escala técnica relativamente menor. Também observa-se que o ***Cluster* 0** concentra os modelos mais recentes. Quanto à acessibilidade dos parâmetros utilizados nos modelos, o ***Cluster* 0** é marcado pela falta de transparência, com $76,9\%$ dos modelos possuindo acessibilidade desconhecida e $18,6\%$ não publicada. Em contraste, o ***Cluster* 1** apresenta maior abertura, com 19% dos modelos oferecendo parâmetros abertos para uso irrestrito e $11,2\%$ via acesso por API. Em relação ao contexto organizacional, o ***Cluster* 0** apresenta um predomínio de instituições acadêmicas $(89,6\%)$, e o ***Cluster* 1** mostra uma maioria de organizações da indústria $(58,7\%)$. Ainda, com relação a domínios de aplicação, observa-se que 39,6% dos modelos do ***Cluster* 0** são utilizados em visão computacional, e 50,0% dos modelos do ***Cluster* 1** em Processamento de Linguagem Natural (PLN). Assim sendo, os *clusters* identificados podem ser descritos da seguinte forma: o ***Cluster* 0** agrupa modelos de menor escala, majoritariamente de origem acadêmica, com foco em visão computacional e menos recentes; enquanto o ***Cluster* 1** é caracterizado por modelos de alto custo computacional, predominantemente da indústria, focados em PLN e publicados mais recentemente. Essa segmentação reflete barreiras econômicas e objetivos estratégicos distintos que moldam o ecossistema. Enquanto o setor acadêmico (***Cluster* 0**) enfrenta restrições de financiamento e prioriza a pesquisa fundamental através de modelos mais acessíveis, o setor industrial (***Cluster* 1**) é tracionado pela necessidade de retorno sobre o investimento (ROI). Consequentemente, as grandes empresas concentram recursos na infraestrutura massiva exigida pelos modelos de linguagem de fronteira, que tendem a ser mantidos como proprietários para preservar vantagens comerciais.

4.2. Análise de Regressão para Custo e Impacto Ambiental

A análise do Fator de Inflação de Variância (VIF) [Fox 2016] revelou multicolinearidade alta entre os preditores, com valores elevados para as variáveis `Training compute (FLOP) log` ($VIF = 996$) e `Hardware - Memory bandwidth (byte/s) log` ($VIF = 730$).

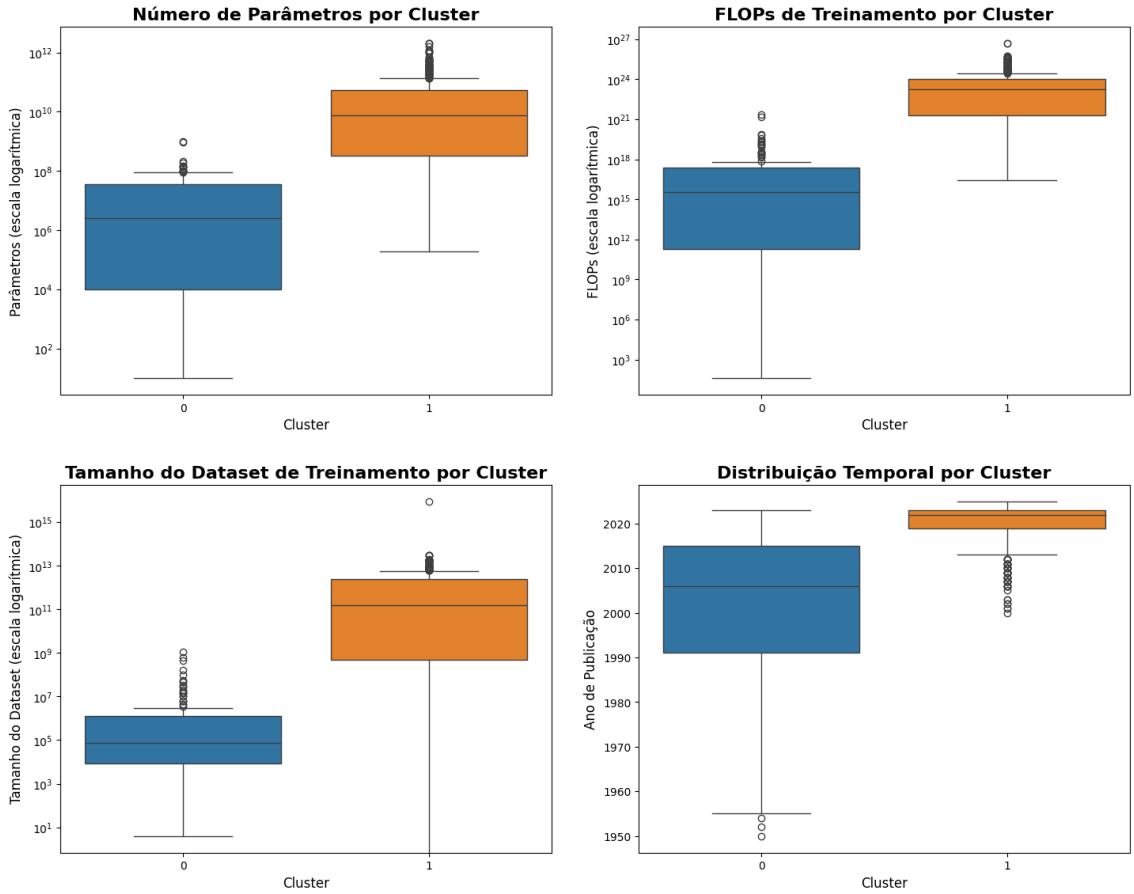

Figura 2. Distribuição das características técnicas por *cluster*.

4.2.1. Análise de Custo Financeiro

Os resultados da regressão para o custo financeiro são apresentados na Tabela 1 e explicam 85,7% da variância dos dados. Os componentes PC1, PC2, PC4, PC5 e PC6 foram preditores estatisticamente significantes. Conforme esperado, a "Escala do Modelo" (PC1) teve um impacto positivo. O componente PC2 apresentou um coeficiente negativo ($\beta = -0,143$), indicando que a estratégia de utilizar menor quantidade de *hardware*, porém mais potente, tende a ser mais econômica. Pode-se interpretar da seguinte forma os componentes principais com maior poder explicativo de acordo com seus coeficientes *loadings*: o **PC1 – Escala do Modelo** é caracterizado por coeficientes de carga positivos em todas as variáveis, com maior contribuição em `Training compute (FLOP) log` (coeficiente $+0,470$) e `Parameters log` ($+0,446$), representando a magnitude geral do treinamento; já o **PC2 – Potência vs. Quantidade de Hardware** contrasta a potência do *hardware*, como `Hardware - TDP (W)` (coeficiente $+0,680$), com a quantidade de *hardware* utilizada ($-0,589$), de modo que um valor alto neste componente representa o uso de poucas unidades de *hardware* mais potentes.

4.2.2. Análise de Impacto Ambiental

Na Tabela 1, observa-se que o modelo de regressão para consumo de energia explicou 74,2% da variância. Os preditores significativos foram PC1, PC3 e PC5. A "Escala do Modelo" (PC1) novamente teve um impacto positivo. Notavelmente, o PC3 apresentou um coeficiente negativo ($\beta = -0,224$), sugerindo que a estratégia de realizar treinos mais longos em modelos menores, representada por um PC3 alto, está associada a um menor consumo de energia. Os componentes principais mais explicativos podem ter a seguinte interpretação: o **PC1 – Escala do Modelo**, de forma similar ao modelo de custo, possui coeficientes de carga positivos em todas as variáveis, representando a magnitude geral do trabalho computacional; enquanto o **PC3 – Duração de Treinamento *vs.* Tamanho do Modelo** captura uma oposição entre o tempo de treinamento (`Training time (hours)`), com coeficiente de $+0,708$, e o número de parâmetros (`Parameters log`), com coeficiente de $-0,404$, indicando que um valor alto representa treinos mais longos em modelos com menos parâmetros. A adequação dos modelos de regressão para custo financeiro e consumo de energia foi confirmada por meio da análise de resíduos, que se apresentaram dispersos e sem um padrão claro, e pela verificação de normalidade nos gráficos Quantil-Quantil (Q-Q).

Tabela 1. Modelos de regressão linear para custo financeiro e consumo de energia

	Custo Financeiro				Consumo de Energia			
Preditores	β	**EP**	**t**	**p**	β	**EP**	**t**	**p**
Constante	4,530	0,061	74,61	<0,001***	5,124	0,062	82,88	<0,001***
PC1	0,571	0,032	18,09	<0,001***	0,448	0,031	14,22	<0,001***
PC2	-0,143	0,054	-2,65	0,010**	0,013	0,063	0,20	0,843
PC3	-0,109	0,073	-1,48	0,144	-0,224	0,070	-3,20	0,002**
PC4	0,219	0,081	2,70	0,009**	0,054	0,076	0,72	0,476
PC5	-0,351	0,093	-3,76	<0,001***	-0,327	0,101	-3,23	0,002**
PC6	0,324	0,123	2,63	0,011*	-0,230	0,123	-1,86	0,066
$R^2 = 0,871$ (Custo Financeiro); $R^2 = 0,762$ (Consumo de Energia)								

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$. EP = Erro Padrão.

4.3. Detecção de Anomalias para Identificação de Destaques

Empregou-se *Z-Score* [Montgomery and Runger 2018] para a identificação de modelos em que o número de citações é um *outlier* relativo aos outros modelos da mesma organização ($Z - Score > 2$), revelando 58 anomalias. Outro método utilizado foi *Isolation Forest* [Liu et al. 2008] com uma contaminação de 10% para revelar modelos *outliers* no contexto global do conjunto de dados analisado, sinalizando a presença de 142 anomalias. Da intersecção dos dois métodos e da distinção de anomalias de alto e baixo impacto, obteve-se um total de 44 modelos que foram classificados como anomalias confirmadas positivas.

A análise quantificou a produção de anomalias positivas por instituição. Destacaram-se as *Big Techs* norte-americanas, lideradas por Google (9 anomalias em 213 modelos), Microsoft (2 em 49) e Meta (4 em 91). No cenário acadêmico, evidenciou-se

a relevância de instituições como a Universidade de Stanford (3 anomalias em 48 modelos), Carnegie Mellon (2 em 31 modelos), Berkeley (2 em 26), além da Universidade de Toronto (2 em 29) no Canadá e Oxford (2 em 18) no Reino Unido. A Tabela 2 apresenta modelos que foram identificados como anomalias positivas. Esses trabalhos originam-se tanto da indústria quanto da academia e representam avanços no estado da arte com significativo número de citações.

Tabela 2. Modelos anômalos positivos mais citados

Modelo	Data de Publicação	Citações
ResNet-152 / ResNet-110	10/12/2015	175 697
ADAM (CIFAR-10)	22/12/2014	141 441
AlexNet	30/09/2012	113 243
Transformer	12/06/2017	108 604
VGG19 / VGG16	04/09/2014	94 013
BERT-Large	11/10/2018	83 567
LSTM	15/11/1997	83 231

5. Conclusão

O estudo examinou características do desenvolvimento de modelos de IA, identificando por análise de *cluster* um grupo predominantemente industrial, com modelos de linguagem recentes e de grande escala, e outro majoritariamente acadêmico, focado em visão computacional. Uma regressão mostrou que a escala do modelo e as especificações de *hardware* são preditores significativos do custo financeiro e do impacto ambiental. Os resultados contribuem para compreender melhor o que pode ser descrito como oligopolização no desenvolvimento de IA de fronteira. A dependência de *hardware* massivo e custos exponenciais não apenas impõe severas barreiras de entrada para instituições menores e países em desenvolvimento, mas também levanta pontos de atenção sobre a trajetória de sustentabilidade ambiental dessa corrida tecnológica.

Para trabalhos futuros, mostra-se relevante explorar estruturas mais complexas de clusterização como base para uma análise comparativa entre diferentes domínios de aplicação, como visão computacional e linguagem. Além disso, o uso de um modelo linear para regressão pode não ter capturado relações não-lineares complexas inerentes ao contexto de IA e hardware. Como desdobramento, sugere-se a aplicação de modelos não-lineares mais robustos, como *Random Forest*, *Gradient Boosting* ou redes neurais, para comparar o desempenho das estimativas. Quanto à detecção de anomalias, o uso exclusivo de citações como estimativa de impacto apresenta limitações, pois carrega vieses temporais e de domínio, uma vez que modelos mais antigos tendem a acumular mais citações. Recomenda-se que estudos futuros explorem a combinação de citações acadêmicas com indicadores complementares de impacto tecnológico e comercial, como o registro de patentes, a adoção por empresas, a influência em *benchmarks* ou menções em políticas públicas. Por fim, para mitigar possíveis vieses de seleção da base de dados, estudos posteriores devem integrar repositórios comunitários, como *Hugging Face* e *PapersWithCode*, para incluir modelos de regiões com menor visibilidade científica global, como América Latina, África e sudeste asiático.

Referências

Cottier, B., Besiroglu, T., and Owen, D. (2023). Who is leading in ai? an analysis of industry ai research. *arXiv preprint arXiv:2312.00043*.

Cottier, B., Rahman, R., Fattorini, L., Maslej, N., and Owen, D. (2024). The rising costs of training frontier ai models. *arXiv preprint arXiv:2405.21015*.

Epoch AI (2025a). Data on large-scale ai models. `https://epoch.ai/data/large-scale-ai-models`. Acessado em: 2025-06-21.

Epoch AI (2025b). Data on machine learning hardware. `https://epoch.ai/data/machine-learning-hardware`. Acessado em: 2025-06-21.

Epoch AI (2025c). Data on notable ai models. `https://epoch.ai/data/notable-ai-models`. Acessado em: 2025-06-21.

Epoch AI (2025d). Epoch AI. `https://epoch.ai/`. Acessado em: 2026-03-05.

Fox, J. (2016). *Applied Regression Analysis and Generalized Linear Models*. Sage Publications, Thousand Oaks, CA, 3rd edition.

Frymire, L. and Rahman, R. (2024). The power required to train frontier ai models is doubling annually. `https://epoch.ai/data-insights/power-usage-trend`. Acessado em: 2026-03-05.

Hair, J. F., Black, W. C., Babin, B. J., and Anderson, R. E. (2018). *Multivariate Data Analysis*. Cengage, Boston, MA, 8th edition.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6):417–441.

Liu, F. T., Ting, K. M., and Zhou, Z. H. (2008). Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pages 413–422, Pisa, Italy.

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297.

Montgomery, D. C. and Runger, G. C. (2018). *Applied Statistics and Probability for Engineers*. Wiley, Hoboken, NJ, 7th edition.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

Samborska, V. (2025). Scaling up: How increasing inputs has made artificial intelligence more capable. `https://ourworldindata.org/scaling-up-ai`. Acessado em: 2026-03-05.

Tsoumakas, G. and Katakis, I. (2007). Multi-label classification: An overview. *International Journal of Data Warehousing and Mining*, 3(3):1–13.

Vake, D., Šinik, B., Vičič, J., and Tošić, A. (2025). Is open source the future of ai? a data-driven approach. *Applied Sciences*, 15(5):2790.

Wu, C.-J. et al. (2022). Sustainable ai: Environmental implications, challenges and opportunities. In *Proceedings of Machine Learning and Systems*, volume 4, pages 795–813.